# STAT 4203

# MATHEMATICAL STATISTICS I

Fall, 2002

**Lecture Notes**

**Joshua M. Tebbs**

**Department of Statistics**

**Oklahoma State University**

# Contents

# 1 Probability

Complimentary 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."

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.

How do we **assign** probabilities to events? Here, we present three 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 an **experiment** can be repeated under identical conditions.

**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 "two." Let $A$ denote this event. The relative frequency approach says that

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

where $f$ denotes the **frequency** of the event, and $n$ denotes the number of times the experiment was conducted. The ratio $f/n$ is sometimes called the **relative frequency**.

*NOTATION*: The symbol $P(A)$ is shorthand for "the probability that $A$ occurs."

Continuing with our example, suppose that $f = 158$. Then, we would **estimate** $P(A)$ with $158/1000 = 0.158$. If we performed this experiment repeatedly, the relative frequency approach says that $f/n \to P(A)$ as $n \to \infty$. Of course, if the die is **unbiased**, $f/n \to P(A) = 1/6$.

3. *Axiomatic approach.*

   - based on set theory; this is where we will start.

## 1.2 Sample spaces and basic set theory

In probability applications, it is common to **perform** an experiment and then **observe** an outcome.

*TERMINOLOGY*: The set of all possible outcomes for an experiment is called the **sample space**, hereafter denoted $S$.

**Example 1.2.** In each of the following situations, we provide the sample space, $S$.

(a) A rat is selected and we observe the sex of the rat:

$$S = \{\text{male, female}\}.$$

Had we observed sex ($x$) and weight ($y$), then the sample space might be

$$S = \{(x, y) : x \in \{\text{male, female}\},\ y > 0\}.$$

(b) The Michigan state lottery calls for a three-digit integer to be selected:

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

(c) 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

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

*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.

*TERMINOLOGY*: A **countable set** is one whose elements can be put into a one-to-one correspondence with $\mathcal{N} = \{1, 2, ...\}$, the set of natural numbers. 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 the context of probability applications, $S$ will usually denote a **sample space**, $A$ will often represent an **event** to which we wish to assign a probability, and $\omega$ usually denotes a possible **experimental outcome**. If $\omega \in A$, we would say that "the event $A$ has occurred."

**Example 1.3.** Suppose that $S = \{1, 2, 3, 4, 5, 6\}$, $A = \{2, 4, 6\}$, and $B = \{1, 2\}$. Clearly, $A \subset S$ and $B \subset S$. Note that $A \not\subseteq B$ and $B \not\subseteq A$.

*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 $\omega$ notation,

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

In Example 1.3, $A \cup B = \{1, 2, 4, 6\}$.

*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 $\omega$ notation,

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

In Example 1.3, $A \cap B = \{2\}$.

Of course, 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

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

**Example 1.4.** Suppose that $S = (0, 1)$, $A_1 = (0, 1/3]$, $A_2 = (1/3, 2/3]$, and $A_3 = (2/3, 1)$. Here, $A_1 \cup A_2 \cup A_3 = (0, 1) = S$ and $A_1 \cap A_2 \cap A_3 = \emptyset$.

Now, suppose that $A_1, A_2, ...$ is a **countable** sequence of sets. The union of this infinite collection of sets is

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

and the intersection of this infinite collection is

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

**Example 1.5.** Suppose that $S = (0, \infty)$ and $A_n = (1 - 1/n, 1 + 1/n)$ for $n = 1, 2, ....$ Here,

$$\bigcup_{n=1}^{\infty} A_n = (0, 2) \quad \text{and} \quad \bigcap_{n=1}^{\infty} A_n = \{1\}.$$

*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 $\omega$ notation,

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

In Example 1.3, $\overline{A} = \{1, 3, 5\}$ and $\overline{B} = \{3, 4, 5, 6\}$.

*VENN DIAGRAMS*: see pages 24-26 WMS.

*TERMINOLOGY*: We say that $A$ is a **subset** of $B$, and write $A \subset B$ (or $A \subseteq B$) if $\omega \in A \Rightarrow \omega \in B$.

Suppose that $A$ and $B$ are events in an experiment and $A \subset B$, then, if $A$ occurs, $B$ must occur.

*TWO USEFUL SET THEORY FACTS*:

**Distributive Laws:**

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

**DeMorgans Laws:**

$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$
$$\overline{A \cup B} = \overline{A} \cap \overline{B}$$

*TERMINOLOGY*: We call two events $A$ and $B$ **mutually exclusive**, or **disjoint**, if $A \cap B = \emptyset$.

## 1.3 Discrete probability models and events

WMS describe two types of events:

(a) simple events

(b) compound events

*TERMINOLOGY*: A **simple event** is one that can not be decomposed. That is, a simple event corresponds to exactly one sample point $\omega$. **Compound events** are those events that contain more than one sample point.

**Example 1.6.** In a field experiment, we observe the number of insects that test positive for a disease. If $n$ insects are under investigation, $S = \{0, 1, 2, ..., n\}$. Define

$$A_1 = \{\text{exactly one insect is infected}\} = \{1\}$$

$$A_2 = \{\text{no insects are infected}\} = \{0\}$$

$$A_3 = \{\text{more than one insect is infected}\} = \{2, 3, ..., n\}$$

$$A_4 = \{\text{at most one insect is infected}\} = \{0, 1\}$$

Because $A_1$ and $A_2$ only contain one sample point each, they are simple events. The events $A_3$ and $A_4$ 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.

**Example 1.7.** Suppose we roll a die. Here, $S = \{1, 2, 3, 4, 5, 6\}$, and suppose that $A_1 = \{1\}$, and $A_2 = \{3, 4\}$. Clearly, $A_1$ is a simple event, and $P(A_1) = 1/6$ (assuming that the die is unbiased). On the other hand, $A_2$ is a compound event.

$$P(A_2) = \underbrace{P(\{3\}) + P(\{4\}) = 1/6 + 1/6}_{\text{adding individual probabilities}} = 1/3,$$

assuming, again, that the die is unbiased. In this example, counting the number of sample points associated with $A_2$ is trivial. In some problems, it may not be so straightforward. We will need to discuss methods (tools) for counting sample points (see Section 2.6 WMS).

*A NOTE ON WMS's NOTATION*: We have used the symbol $\omega$ to denote an element in a set (i.e., a sample point in an event). In a more probabilistic spirit, the authors use

the symbol $E_i$ to denote the $i$th sample point (i.e., simple event). Thus, if $A$ denotes any compound event,

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

That is, we are simply summing up the simple event probabilities $P(E_i)$ for all $i$ such that $E_i \in A$.

## 1.4 Discrete sample spaces

*TERMINOLOGY*: If a sample space for an experiment contains a finite or countable number of sample points, we call it a **discrete sample space**. We denote the sample space as $S$. Then, events $A$ are just subsets of $S$. Recall,

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

*THE THREE AXIOMS OF PROBABILITY*: Given a nonempty sample space $S$,

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

(2) $P(S) = 1$,

(3) Given a countable sequence of mutually exclusive (disjoint) events $A_1, A_2, ...$,

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

**Example 1.8.** Four equally qualified applicants $(a, b, c, d)$ are up for two positions. Applicant $a$ is a minority. Positions are chosen at random. What is the probability that the minority is hired?

SOLUTION: Here, $S = \{ab, ac, ad, bc, bd, cd\}$ (we are assuming that the order of the

positions is not important). If the positions are assigned at random, each of the six sample points has a 1/6 probability (follows from Axiom 2). Thus,

$$P(\{\text{“minority hired”}\}) = \underbrace{P(\{ab\}) + P(\{ac\}) + P(\{ad\})}_{\text{Axiom 3}} = 1/6 + 1/6 + 1/6 = 1/2.$$

## 1.5 Tools for counting sample points

**Theorem 1.1.** 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 $i$th 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,

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

### 1.5.1 The multiplication rule

*THE MULTIPLICATION RULE FOR COUNTING*: Consider an experiment consisting of $k$ stages. Let

$$n_1 = \text{number of ways stage 1 can occur}$$
$$n_2 = \text{number of ways stage 2 can occur}$$
$$\vdots$$

$$n_k = \text{number of ways stage } k \text{ can occur.}$$

Then, there are

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

different ways that the experiment can occur.

**Example 1.9.** 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.

**Example 1.10.** 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 1.11.** 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) $26 \times 25 \times 24 \times 10 \times 9 \times 8 \times 7 = 78,624,000$

### 1.5.2 Permutations

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

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.

Thus, 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 1.12.** 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 1.13.** Now, suppose that in Example 1.12 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.

**Theorem 1.2.** 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

$$P_r^n \equiv \frac{n!}{(n-r)!}.$$

The symbol $P_r^n$ 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 $r$th 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 1.14.** 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

$$P_3^5 = \frac{5!}{(5-3)!} = \frac{120}{2} = 60$$

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

**Example 1.15.** 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 fertilizers. How

many ways can I choose these four plots and then assign fertilizers?

SOLUTION: There are

$$P_4^{10} = \frac{10!}{(10-4)!} = 5040$$

different permutations. Here, we are, again, assuming order is important, since each fertilizer is presumably different.

**Example 1.16.** In Example 1.15,

(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 of interest (to permute) are *not* distinct?

**Example 1.17.** 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

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

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.,

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

This example should motivate the following theorem.

**Theorem 1.3.** The number of ways to partition $n$ distinct objects into $k$ groups containing $n_1$, $n_2$, ..., $n_k$ objects, respectively, is given by the **multinomial coefficient**

$$\left( \begin{array}{cccc} & & n & \\ n_1 & n_2 & \cdots & n_k \end{array} \right) \equiv \frac{n!}{n_1!\ n_2!\ \cdots\ n_k!}.$$

**Example 1.18.** 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:

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

**Example 1.19.** In Example 1.18, 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. Let $A = \{\text{all four white flags are grouped together}\}$.

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 1.18 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 | $n_2$ |

From Theorem 1.3, we know that

$$n_2 = \frac{5!}{2!\ 3!} = 10.$$

Thus, $n_a = 6 \times 10 = 60$. Finally, since we have equally likely outcomes, Theorem 1.1 says $P(A) = n_a/N = 60/1260 \approx 0.0476$.

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 $n_a$ by, again, 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, Theorem 1.1 says $P(A) = n_a/N = 17280/9! \approx 0.0476$.

*MORAL*: How we structure our sample space completely determines the counting tools we use.

### 1.5.3 Combinations

*TERMINOLOGY*: A **combination** is a selection of objects without regard to order. *Order is not important.*

**Theorem 1.4.** Given $n$ distinct objects, the number of ways to choose $r$ of them ($r \leq n$), without regard to order, is given by

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

The symbol $C_r^n$ 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 groups:

Group 1 $r$ chosen

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

By Theorem 1.3, there are $\frac{n!}{r!(n-r)!}$ ways to do this. □

*REMARK*: We will adopt the notation $\binom{n}{r}$, read "$n$ choose $r$," as the symbol for $C_r^n$. 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 1.20.** Return to Example 1.14. 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.

From Examples 1.14 and 1.20 (or Theorems 1.2 and 1.4), one should note that

$$P_r^n = r! \times C_r^n.$$

Recall that combinations do not regard order as important. Thus, once we have chosen our $r$ objects (there are $C_r^n$ 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 1.21.** 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), Theorem 1.1 says $P(A) = n_a/N = 6188/15504 \approx 0.399$.

## 1.6 Conditional probability and independence

In some applications, 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, then, 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 1.22.** A couple has two children. What is the probability that both are girls?

SOLUTION: 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

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

Then, clearly, $A_1 \cap A_2 = \{(F, F)\}$ and $P(A_1 \cap A_2) = 1/4$, assuming equally likely outcomes.

What is the probability that both are girls, $\underbrace{\text{if the eldest is a girl?}}_{\text{"extra information"}}$

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

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

In a sense the "new information" (here, that the eldest is a girl) **induces** a new sample space (or **restricted** sample space) given by

$$S_{\text{new}} = \{(F, M), (F, F)\}.$$

On this space, $P(A_2) = 1/2$ (computed with respect to $S_{\text{new}}$). So, whether you compute $P(A_2|A_1)$ with the original sample space $S$ or compute $P(A_2)$ with the restricted space $S_{\text{new}}$, you will get the same answer.

**Example 1.23.** 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 conditional 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:

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

*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

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

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 1.24.** 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

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

the events $A$ and $B$ are independent.

**Example 1.25.** 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 $P(A_1 \cap A_2) = (0.95)^2 = 0.9025$.

The concept of independence 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

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

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 $i$th trial, and the trials are

**independent**,

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

**Example 1.26.** An unbiased die is rolled six times. Let

$$A_i = \{i \text{ appears on roll } i\},$$

$i = 1, 2, ..., 6$. Then, $P(A_i) = 1/6$. Assuming independence,

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

**Theorem 1.5.** If $A$ is an event in a non-empty sample space $S$, then $P(\overline{A}) = 1 - P(A)$.

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

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

and, hence, $P(\overline{A}) = 1 - P(A)$ as desired. $\square$

It is oftentimes easier to compute $P(\overline{A})$ than to compute $P(A)$ directly.

**Theorem 1.6.** If $A$ and $B$ are independent events, so are

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

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

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

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

$$\begin{aligned} P(\overline{A} \cap B) &= P(\overline{A}|B)P(B) \\ &= [1 - P(A|B)]P(B) \\ &= [1 - P(A)]P(B) \\ &= P(\overline{A})P(B). \ \square \end{aligned}$$

As you may suspect, independence of events and their complements holds for any finite sequence $A_1, A_2, ..., A_n$.

**Example 1.27.** Referring to Example 1.26, 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, $P(B) = 1 - P(\overline{B}) = 1 - 0.335 = 0.665$.

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

**Example 1.28.** Suppose that in a certain population, individuals have a certain disease with probability $p$. Let $A_i = \{\text{person } i \text{ has the disease}\}$. Then, $P(A_i) = p$ and $P(\overline{A}_i) = 1 - p$. Suppose that 10 individuals are observed. Assuming independence among individuals, what is the probability that

(a) no one has the disease?

(b) at least one has the disease?

(c) exactly one has the disease?

(d) exactly $k$ have the disease ($k = 0, 1, 2, ..., 10$)?

## 1.7 Some probability laws

*THE 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. If either $P(A)$ or $P(B)$ is zero, this still holds, but the proof is more difficult. $\square$

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

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

This suggests that we can compute probabilities like $P(A \cap B \cap C)$ "sequentially" by first computing $P(A)$, then $P(B|A)$, then $P[C|(A \cap B)]$.

Of course, the probability of a $k$-fold intersection can be computed similarly.

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

*THE ADDITIVE LAW OF PROBABILITY*: Suppose $A$ and $B$ are events in a nonempty sample space $S$. 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,

$$P(A \cup B) = P(A) + P(\overline{A} \cap B) \tag{1.1}$$

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, this last statement with (1.1) gives the result. $\square$

**Example 1.29.** 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. Let $A_i$ denote

the event that train $i$ is on time. What is the probability that at least one train is on time?

SOLUTION:

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

What is the probability that neither train is on time?

SOLUTION:

$$\begin{aligned} P(\overline{A}_1 \cap \overline{A}_2) &= P(\overline{A_1 \cup A_2}) \\ &= 1 - P(A_1 \cup A_2) \\ &= 1 - 0.98 = 0.02 \end{aligned}$$

The additive law of probability can be extended to any finite sequence of sets $A_1, A_2, ..., A_n$. For example, if $n = 3$,

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

Section 2.9 WMS presents the **Decomposition Method**, which simply uses the multiplication laws, the addition laws, and combinations thereof.

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** (for two events $A$ and $B$). We will often abbreviate this as LOTP.

*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) $B \cup \overline{B} = S$.

The LOTP is often helpful. Sometimes computing $P(A|B)$, $P(A|\overline{B})$, and $P(B)$ may be easily computed with the available information whereas computing $P(A)$ directly may be difficult.

**Example 1.30.** 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. What is the probability that a new policy-holder will have an accident?

SOLUTION:

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

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

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

SOLUTION: We want $P(B|A)$. But,

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

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 1.31.** 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,

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

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,

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

Table 1.1 summarizes our recent calculations.

*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

Table 1.1: *The general Bayesian scheme.*

| Guess before test | | Result | | Updated guess |
|---|---|---|---|---|
| $P(D)$ | | $F$ | | $P(D\|F)$ |
| 0.005 | $\longrightarrow$ | ✠ | $\longrightarrow$ | 0.323 |
| 0.005 | $\longrightarrow$ | $\overline{✠}$ | $\longrightarrow$ | 0.00025 |

(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)*: Let $B_1, B_2, ..., B_k$ form 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

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

Thus,

$$\begin{aligned} P(A) &= P\left[\bigcup_{i=1}^{k} (A \cap B_i)\right] \\ &= \sum_{i=1}^{k} P(A \cap B_i) \\ &= \sum_{i=1}^{k} P(A|B_i)P(B_i), \end{aligned}$$

where the last step follows from applying the multiplication law of probability for each $i$. $\square$

*BAYES RULE (restated)*: Let $B_1, B_2, ..., B_k$ form 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. □

**Example 1.32.** A manufacturer buys 60 percent of a raw material from Supplier A, 30 percent from Supplier B, and 10 percent from Supplier C. For each supplier, defect rates are as follows: A: 0.01, B: 0.02, and C: 0.03. Suppose that the manufacturer observes a defective box of raw material. What is the probability that it came from Supplier B?

SOLUTION: Let $D = \{\text{observe defective}\}$, and $A$, $B$, and $C$, respectively, denote the events that the defective comes from Supplier A, B, and C. By Bayes Rule (noting that $\{A, B, C\}$ partitions the space of possible suppliers),

$$\begin{aligned} P(B|D) &= \frac{P(D|B)P(B)}{P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C)} \\ &= \frac{(0.02)(0.3)}{(0.01)(0.6) + (0.02)(0.3) + (0.03)(0.1)} \\ &= 0.40 \end{aligned}$$

What is the probability that the defective did not come from Supplier C?

SOLUTION: First, compute $P(C|D)$. By Bayes Rule,

$$\begin{aligned} P(C|D) &= \frac{P(D|C)P(C)}{P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C)} \\ &= \frac{(0.03)(0.1)}{(0.01)(0.6) + (0.02)(0.3) + (0.03)(0.1)} \\ &= 0.20 \end{aligned}$$

Thus, $P(\overline{C}|D) = 1 - P(C|D) = 1 - 0.2 = 0.8$.

*NOTE*: Read Sections 2.11 and 2.12 from WMS.

# 2 Discrete Random Variables and Their Probability Distributions

Complimentary reading: Chapter 3 (WMS).

## 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 $\mathbf{R}$. That is, $Y : S \to \mathbf{R}$.

*WORKING DEFINITION*: A **random variable** is a variable whose observed value is determined by chance (as you may suspect, **probability** is involved.)

**Example 2.1.** Suppose that we flip two fair coins. The sample space consists of four sample points:

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

Now, let $Y$ = the number of **heads** observed. Of course, before we perform the experiment, we do not know, with certainty, the value of $Y$. What are the **possible values** of $Y$? Consider the following table:

| Sample point | $Y$ |
|---|---|
| $(H,H)$ | 2 |
| $(T,H)$ | 1 |
| $(H,T)$ | 1 |
| $(T,T)$ | 0 |

Thus, 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.,

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

and hence,

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

From these probability calculations; note that we can

- work on the original sample space $S$ and compute probabilities using the sample points, or
- work on $\mathbf{R}$ and compute probabilities for events of the form $\{Y = y\}$ (in fact, this is what we will do from here on).

*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 Stillwater. Here, $Y$ is random variable (it is unknown and random). After the baby is born, we observe $y = 128$. How might we compute a probability like $P(Y > 150)$? To do this, we have to discuss the concept of a **probability distribution**.

## 2.2 Probability distributions for discrete random variables

*TERMINOLOGY AND NOTATION*: 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 $R$.

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

*GOAL*: 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, theoretically, one would have to determine all sample points $E_i \in S$ such that $Y(E_i) = y$ and then compute

$$P(Y = y) \equiv p_Y(y) = \sum_{i:Y(E_i)=y} P(E_i)$$

for all $y \in R$.

*TERMINOLOGY*: The function $p_Y(y)$ is called the **probability distribution** or **probability distribution function** (**pdf**) for the discrete random variable $Y$.

*FACTS*: The probability distribution for a discrete random variable $Y$ consists of **two** parts:

(a) $R$, the support set of $Y$, and

(b) $P(Y = y) = p_Y(y)$ for all $y \in R$.

Furthermore, the following must be true (for the probability distribution to be **valid**):

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

(2)

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

**Example 2.4.** Suppose that we roll two unbiased dice. 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. Then $R_Y = \{2, 3, ..., 12\}$.

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

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

The calculation $P(Y = y)$ is performed similarly for other values of $y = 4, 5, ..., 12$. The pdf for $Y$ can be given as a **formula**, **table**, or **graph**. A formula for $p_Y(y)$ is given by

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

We could also display the distribution in tabular form:

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

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

$$\sum_{y=2}^{12} p_Y(y) = 1.$$

**Example 2.5.** An experiment consists of rolling an unbiased die until the first six is observed. Let $X$ denote the number of rolls needed. Here, $R_X = \{1, 2, ..., \}$. Assuming independent trials, we have

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

In general, we have

$$P(X=x) = \left(\frac{5}{6}\right)^{x-1} \times \frac{1}{6},$$

for all $x = 1, 2, ....$ This can be expressed succinctly as

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

This formula represents the **pdf** for $X$. Is this a **valid** pdf? Clearly, $p_X(x) > 0$ for all $x \in R_X$. Also,

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

Above, we have used the following fact concerning **infinite geometric sums.**

*FACT*: Let $a$ be any real number, and suppose that $|r| < 1$. Then,

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

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

## 2.3 Expected values

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

$$E(Y) = \sum_{\text{all } y} y p_Y(y).$$

In words, the expected value for discrete random variable is a **weighted average** of possible values the variable can assume, each value being weighted with the probability that the random variable assumes the corresponding value.

*MATHEMATICAL ASIDE*: In computing the expected value of a discrete random variable, it must be the case that the sum above is **absolutely convergent**; i.e.,

$$\sum_{\text{all } y} |y| p_Y(y) < \infty.$$

Otherwise, $E(Y) = \infty$.

**Example 2.6.** Let the random variable $Y$ have pdf

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

The expected value of $Y$ is given by

$$\begin{aligned} \sum_{\text{all } y} y p_Y(y) &= \sum_{y=1}^{4} y \frac{1}{10}(5-y) \\ &= 1(4/10) + 2(3/10) + 3(2/10) + 4(1/10) = 2. \end{aligned}$$

How is $E(Y)$ interpreted? We can think of in various ways:

(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 $\mu$ or $\mu_Y$ when discussing it; that is

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

In statistical settings, $\mu_Y$ often denotes a population **parameter** that we wish to **estimate**.

*EXPECTATIONS OF FUNCTIONS OF $Y$*. Let $g$ be a real-valued function and let $Y$ be a discrete random variable. Then, $g(Y)$ is a random variable and

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

*MATHEMATICAL ASIDE*: Again, in computing $E[g(Y)]$, it must be the case that

$$\sum_{\text{all } y} |g(y)|p_Y(y) < \infty$$

i.e., that this sum is **absolutely convergent**. Otherwise, $E[g(Y)] = \infty$. The proof of this widely-used fact appears on page 90 (WMS).

**Example 2.7.** Consider the random variable $Y$ in Example 2.6. Find $E(Y^2)$ and $E(e^Y)$.
SOLUTION: Clearly, $g_1(Y) = Y^2$ and $g_2(Y) = e^Y$ are just real functions of the random variable $Y$.

$$\begin{aligned} E(Y^2) &= \sum_{\text{all } y} y^2 p_Y(y) \\ &= \sum_{y=1}^{4} y^2 \frac{1}{10}(5-y) \\ &= 1^2(4/10) + 2^2(3/10) + 3^2(2/10) + 4^2(1/10) = 5.0. \end{aligned}$$

Also,

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

*PROPERTIES OF EXPECTATIONS*: Let $Y$ be a discrete random variable with pdf $p_Y(y)$, let $g, g_1, g_2, ..., g_k$ denote 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 $E(\cdot)$ a **linear operator**. Proofs to these facts are given on pages 92-93 of WMS.

**Example 2.8.** In a one-hour period, the number of gallons of antifreeze that can be produced, say $Y$, has the pdf

| $y$ | 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 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

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

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

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

Now, we use the aforementioned linearity properties to compute

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

Thus, the expected hourly cost is \$27.60.

## 2.4 Variance

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** (or **spread**) in the distribution.

*TERMINOLOGY*: Suppose that the discrete random variable $Y$ has pdf $p_Y(y)$ and let $\mu_Y = E(Y)$. The **variance** of $Y$, $V(Y)$, is given by

$$V(Y) \equiv E[(Y - \mu_Y)^2] = \sum_{\text{all } y} (y - \mu_Y)^2 p_Y(y).$$

The **standard deviation** of $Y$ is given by the positive square root of the variance.

Table 2.2: *A review of probability and statistics notation for means, variances, and standard deviations.*

| Quantity | Probability | Statistics |
|---|---|---|
| Mean | $E(Y)$ | $\mu_Y$ |
| Variance | $V(Y)$ | $\sigma_Y^2$ |
| Standard deviation | $SD(Y)$ | $\sigma_Y$ |

*FACTS ABOUT THE VARIANCE AND STANDARD DEVIATION*:

(a) $V(Y) \geq 0$.

(b) $V(Y) = 0$ 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 $V(Y)$ is, the more spread in the possible values of $Y$ about the mean.

(d) $V(Y)$ is measured in (units)$^2$ and $SD(Y)$ is measured in the original units.

*NOTE*: Facts (a), (b), and (c) above are true if we replace $V(Y)$ with $SD(Y)$.

**Example 2.9.** Define two random variables, $X$ and $Y$, with pdfs

$$p_X(x) = \begin{cases} \frac{1}{3}, & x = -1, 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

and

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

respectively. Compute $V(X)$ and $V(Y)$.

SOLUTION: First,

$$E(X) = \sum_{\text{all } x} x p_X(x) = -1(1/3) + 0(1/3) + 1(1/3) = 0$$

$$E(Y) = \sum_{\text{all } y} y p_Y(y) = -2(1/3) + 0(1/3) + 2(1/3) = 0.$$

Thus, both random variables have zero mean. Now,

$$\begin{aligned} V(X) &= \sum_{\text{all } x} (x - \mu_X)^2 p_X(x) \\ &= (-1-0)^2(1/3) + (0-0)^2(1/3) + (1-0)^2(1/3) = 2/3 \end{aligned}$$

$$\begin{aligned} V(Y) &= \sum_{\text{all } y} (y - \mu_Y)^2 p_Y(y) \\ &= (-2-0)^2(1/3) + (0-0)^2(1/3) + (2-0)^2(1/3) = 8/3. \end{aligned}$$

Thus, $Y$ has a larger variance (and a larger standard deviation) that $X$.

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

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

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

*Proof.*

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

Thus, the result follows. □

*RESULT*: Let $Y$ denote a random variable (not necessarily a discrete random variable) and suppose $a$ and $b$ are real constants. Then

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

*Proof.* Exercise.

From the above result, one should immediately deduce that $V(b) = 0$ for any **constant** $b$. This makes sense intuitively; the variance is a **measure of variability** for a random variable. A constant (such as $b$) does not vary!

## 2.5 Moment generating functions

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

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

provided $E(e^{tY})$ is finite for $t$ in an open neighborhood about 0.

*TERMINOLOGY*: In general, we call $E(Y^k)$ the $k$**th moment** of the random variable $Y$. The statistical symbol for the $k$th moment is $\mu_k'$. That is, $E(Y^k) = \mu_k'$.

| | |
|---|---|
| $E(Y)$ | 1st moment (mean!) |
| $E(Y^2)$ | 2nd moment |
| $E(Y^3)$ | 3rd moment |
| $\vdots$ | $\vdots$ |

*REMARK*: The moment generating function, hereafter denoted **mgf**, is used to generate moments. From mathematical analysis, it follows that if the mgf exists, it characterizes an infinite set of moments. So, how do we **generate** moments?

**Theorem 2.1.** Let $Y$ denote a random variable (not necessarily a discrete random variable) with mgf $m_Y(t)$. Then,

$$\mu_k' \equiv E(Y^k) = \left. \frac{d^k m_Y(t)}{dt^k} \right|_{t=0}.$$

*NOTE*: In the above expression, note that derivatives are taken with respect to $t$.

*Proof of Theorem* 2.1. With $k = 1$, we have

$$\begin{aligned} \frac{d}{dt} m_Y(t) &= \frac{d}{dt} \sum_{\text{all } y} e^{ty} p_Y(y) \\ &= \sum_{\text{all } y} \frac{d}{dt} e^{ty} p_Y(y) \\ &= \sum_{\text{all } y} y e^{ty} p_Y(y) = E(Y e^{tY}). \end{aligned}$$

Thus, it follows that

$$\left. \frac{dm_Y(t)}{dt} \right|_{t=0} = E(Y e^{tY}) \Big|_{t=0} = E(Y) = \mu_Y = \mu_1'$$

Proceeding in a similar manner; i.e., continuing to take higher-order derivatives, we can prove that

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

Thus, the result follows. □

*MATHEMATICAL ASIDE*: In the second line of the proof of Theorem 2.1, we interchanged the derivative and (possibly infinite) sum. This is permitted as long as $m_Y(t) = E(e^{tY})$ exists.

*REMARK*: The text proves Theorem 2.1 by making use of the McLaurin series expansion (i.e., a Taylor series expansion about 0) of the function $e^z$. Recall from calculus that this is given by

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots$$

Then, making the substitution $z = tY$, the result follows. See page 133 WMS.

*COMPUTING MEANS AND VARIANCES WITH THE MGF*: 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

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

Thus,

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

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

$$E(Y) = \sum_{\text{all } y} y p_Y(y)$$

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

**Example 2.10.** Let the random variable $Y$ have pdf $p_Y(y)$ given by

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

Simple calculations show that $E(Y) = 2/3$ and $E(Y^2) = 1$. Thus, $V(Y) = E(Y^2) - [E(Y)]^2 = 1 - (2/3)^2 = 5/9$. Now, we compute $E(Y)$ and $V(Y)$ using the mgf. First, the mgf is given by

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

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,

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

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

and, hence, $V(Y) = E(Y^2) - [E(Y)]^2 = 1 - (2/3)^2 = 5/9$. So, we can use the mgf to get $E(Y)$ and $E(Y^2)$, or we can compute $E(Y)$ and $E(Y^2)$ directly−we get the same answer (as we should).

**Example 2.11.** Suppose that $X$ is a random variable with pdf

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

The mgf of $X$ is given by

$$\begin{aligned} m_X(t) = E(e^{tX}) &= \sum_{\text{all } x} e^{tx} p_X(x) \\ &= \sum_{x=1}^{\infty} e^{tx} \left(\frac{1}{2}\right)^x \\ &= \sum_{x=1}^{\infty} \left(\frac{e^t}{2}\right)^x \\ &= \left[\sum_{x=0}^{\infty} \left(\frac{e^t}{2}\right)^x\right] - 1 \\ &= \frac{1}{1-\frac{e^t}{2}} - 1 = \frac{e^t}{2-e^t}, \end{aligned}$$

for values of $t < \ln 2$. Thus,

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

*FINAL REMARK ABOUT MGFS*: The importance of the mgf can not be overemphasized in statistical theory. 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 is **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 theory of Laplace transforms). This turns out to be an amazingly helpful result! For now, however, it suffices to envision the mgf as a "special expectation" that, provided it exists, generates moments. This, in turn, helps us to compute means and variances of random variables.

## 2.6 Binomial distribution

Many experiments consist of a **sequence of trials**, where each trial can result in a "success" or "failure" (i.e., there are only 2 possible outcomes per trial). If

(i) there are $n$ trials (where $n$ is **fixed** in advance)

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

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

then we call such trials **Bernoulli trials**.

*TERMINOLOGY*: With a sequence of $n$ Bernoulli trials, let $Y$ count the **number of successes** (out of $n$). We call $Y$ a **binomial random variable**, and say "$Y$ has a **binomial distribution** with parameters $n$ (the number of trials performed) and success probability $p$." Shorthand notation for this last sentence is $Y \sim \text{bin}(n, p)$.

**Example 2.12.** 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, $Y \sim \text{bin}(n = 10, p = 0.5)$.

(b) In a field 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 $Y \sim \text{bin}(n = 4, p = 0.4)$.

(c) In a large African city, the prevalence rate for HIV is about 12 percent. Let $Y$ denote the number of HIV infecteds in a sample of 500 individuals. Here, $Y \sim \text{bin}(n = 500, p = 0.12)$.

(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, $Y \sim \text{bin}(n = 40, p = 0.001)$.

**Example 2.13.** Explain why the following are **not** binomial experiments.

(a) I draw 3 cards from an ordinary deck and count $Y$, the number of aces. Drawing is done without replacement.

(b) A couple decides to have children until a girl is born. Let $Y$ denote the number of children the couple will have.

(c) In a sample of 5000 individuals, I record the age of each person, denoted as $Y$.

(d) A chemist repeats a solubility test ten times on the same substance. Each test is conducted at a temperature 10 degrees higher than the previous test. Let $Y$ denote the number of times the substance dissolves completely.

We now derive the pdf of a binomial random variable. That is, we need $p_Y(y)$ for each possible value of $Y$. Recall, that $Y$ is the **number of successes** in $n$ Bernoulli trials, and $p$ is the probability of success on any one trial. How can we get **exactly** $y$ successes? Denoting

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

a possible sample point may be

$$SSFSFSFFS \cdots FSF$$

By **independence** of trials, 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? By Theorem 1.3 (notes), we know that there are $\binom{n}{y}$ ways to do this. Thus, by independence of trials and the multiplication rule, we may deduce that the pdf for $Y$ is, for $0 < p < 1$,

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

**Example 2.14.** In Example 2.12(b), assume $Y \sim \text{bin}(n = 4, p = 0.4)$. Thus,

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

Observe that $\sum_{y=0}^{4} p_Y(y) = 1$.

**Example 2.15.** 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 independence among the patients. Here, $Y \sim \text{bin}(n = 20, p)$, where $p$ denotes the probability of response to the treatment. In a **statistics** problem, $p$ might be an unknown **parameter** that we might want to **estimate**. For this problem, we'll assume that $p = 0.7$. We want to compute (a) $P(Y = 15)$, (b) $P(Y \geq 15)$, and (c) $P(Y < 10)$.

(a) $P(Y = 15) = \binom{20}{15}(0.7)^{15}(0.3)^{20-15} = 0.1789$.

(b) $P(Y \geq 15) = \sum_{y=15}^{20} \binom{20}{y}(0.7)^y(0.3)^{20-y}$ (have to use the binomial pdf 6 times and add up the results!!)

- Instead of computing the sum $\sum_{y=15}^{20} \binom{20}{y}(0.7)^y(0.3)^{20-y}$ directly, we could write $P(Y \geq 15) = 1 - P(Y \leq 14)$ (complement rule). We do this because WMS's Appendix III (pages 783-785) contains binomial probability calculations of the form $P(Y \leq y)$ for various values of $n$ and $p$. In fact, with $n = 20$ and $p = 0.7$, we see from Appendix III that $P(Y \leq 14) = 0.584$. Thus, $P(Y \geq 15) = 1 - 0.584 = 0.416$.

(c) $P(Y < 10) = P(Y \leq 9) = 0.017$, from Appendix III.

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

$$[p + (1-p)]^n = \sum_{y=0}^{n} \binom{n}{y} p^y (1-p)^{n-y}. \tag{2.2}$$

However, the LHS (left-hand side) of (2.2) clearly equals 1, and the RHS (right-hand side) of (2.2) represents the bin$(n, p)$ pdf. Thus, $p_Y(y)$ is valid.

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

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

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 $Y \sim \text{bin}(n, p)$. 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,

$$V(Y) = E(Y^2) - [E(Y)]^2 = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

*REMARK*: You'll see that on pages 103-105 (WMS), the authors get the same results for $E(Y)$ and $V(Y)$. Their arguments, however, are made by computing $E(Y)$ and $V(Y)$ directly (i.e., **not** using the mgf).

**Example 2.16.** Suppose that 15 seeds are planted in identical soils and temperatures, and let $Y$ denote the number of seeds that germinate. If 60% of all seeds germinate, on average, and we assume a bin$(15, 0.6)$ **probability model** for $Y$, the mean number of seeds that germinate is $E(Y) = np = 15(.6) = 9$. Also, the variance is given by $V(Y) = np(1-p) = 15(0.6)(0.4) = 3.6$, and the standard deviation is $SD(Y) = \sqrt{3.6} \approx 1.9$.

*THE BERNOULLI DISTRIBUTION*: When $n = 1$, the binomial pdf 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**. As we will show later on, a binomial distribution can be motivated by forming the **sum** of independent Bernoulli random variables.

## 2.7 Geometric distribution

*TERMINOLOGY*: Imagine an experiment where Bernoulli trials are continually observed. If $Y$ denotes the trial on which the **first success** occurs, then $Y$ has 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 pdf for $Y$ is given by

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

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

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

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

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

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

**Example 2.17.** A coin is repeatedly tossed until the first tail is observed. Let $Y$ denote the trial at which the 1st tail is observed. Then $Y \sim \text{geom}(p = 0.5)$.

$$\begin{aligned}P(Y=1) &= (1-0.5)^{1-1}(0.5) = 0.5 \\ P(Y=2) &= (1-0.5)^{2-1}(0.5) = 0.25 \\ P(Y=3) &= (1-0.5)^{3-1}(0.5) = 0.125 \\ &\vdots\end{aligned}$$

EXERCISE: Let $A = \{\text{first tail is observed on an odd-numbered toss}\}$. Find $P(A)$.

**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 in the first five flies checked?

SOLUTION: Let $Y$ denote the number of flies needed to observe the first white-eyed fly. Assuming independence among flies, $Y \sim \text{geom}(p = 0.25)$, and

$$\begin{aligned}P(Y=1) = (1-0.25)^{1-1}(0.25) &= 0.25 \\ P(Y=2) = (1-0.25)^{2-1}(0.25) &= 0.1875 \\ P(Y=3) = (1-0.25)^{3-1}(0.25) &= 0.140625 \\ P(Y=4) = (1-0.25)^{4-1}(0.25) &= 0.10546875 \\ P(Y=5) = (1-0.25)^{5-1}(0.25) &= 0.0791015625.\end{aligned}$$

Thus, $P(Y \leq 5) = \sum_{y=1}^{5} P(Y=y) \approx 0.76$.

*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

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

Finally,

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

*REMARK*: You'll see that on page 112 (WMS), the authors get the same results for $E(Y)$ and $V(Y)$. Their arguments, however, are made by computing $E(Y)$ and $V(Y)$ directly (i.e., **not** using the mgf).

**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 only four percent of bags are underfilled. Then, if $Y$ denotes the the number of bags observed, $Y \sim \text{geom}(p = 0.04)$, $E(Y) = (1/0.04) = 25$, and $V(Y) = \frac{0.96}{(0.04)^2} = 600$.

## 2.8 Negative binomial distribution

- a generalization of the geometric

- 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 continually 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 $Y \sim \text{nib}(r, p)$. The pdf for $Y$ 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 $\text{nib}(r, p)$ pdf reduces to the $\text{geom}(p)$.

The logic behind the form of $p_Y(y)$ is as follows. If the $r$th success occurs on the $y$th 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 $y$th trial, we observe the $r$th success (this occurs with probability $p$). Thus, putting it all together, we get

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

**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 ($p = 0.30$). Assuming independence among trees, 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 $P(Y = 6 \text{ or } Y = 7)$.

$$P(Y = 6) = \binom{6-1}{3-1}(0.3)^3(1-0.3)^{6-3} = 0.09261$$

$$P(Y = 7) = \binom{7-1}{3-1}(0.3)^3(1-0.3)^{7-3} = 0.09724$$

Thus, $P(Y = 6 \text{ or } Y = 7) = P(Y = 6) + P(Y = 7) = 0.09261 + 0.09724 = 0.1899$.

*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.

*MGF FOR THE NEGATIVE BINOMIAL DISTRIBUTION*: Suppose that $Y \sim \text{nib}(r, p)$. Then 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 $t$. 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

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

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

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

Now, consider the McLaurin Series expansion (i.e., a Taylor Series expansion about $w = 0$). This 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 $0 < q < 1$.

Now, to compute the mgf of the nib$(r, p)$ random variable, we will use the result in the above lemma. With $q = 1 - p$,

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

where the penultimate step follows from the above lemma. □

*REMARK*: Showing that the nib$(r, p)$ 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

Consider a collection of $N$ similar objects (e.g., people, poker chips, plots of land, etc.) and suppose that we have **two** dichotomous classes. For example,

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

We would like to compute the probability that $r$ of the objects belong to a specific class. If $r$ belong to Class 1, say, then $N - r$ belong to Class 2.

*REMARK*: This is sort of like a binomial setup! However, the difference with this situation is that $N$, the population size, is **not** finite (it is assumed infinite in the binomial model). In this case, if we sample without replacement, then the "success" probability will definitely change from trial to trial! This, of course, violates the binomial model assumptions.

*TERMINOLOGY*: Envision a collection of $n$ objects sampled (at random and without replacement) from a population of size $N$ ($n \leq N$), and let $Y$ denote the number of objects that belong to a specific class. Then, we say $Y$ has a hypergeometric distribution, and write $Y \sim \text{hyper}(N, n, r)$, where

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

The pdf for $Y$ is given by

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

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

In the hyper$(N, n, r)$ pdf, we have three parts:

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

*REMARK*: To show that $p_Y(y)$ sums to one, see Exercise 3.176 (WMS).

**Example 2.21.** In a small pond, 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), $n = 7$ (number in the sample), $r = 10$ (tagged fish; class 1), $N - r = 50 - 10 = 40$ (untagged fish; class 2), and $y = 2$ (number of tagged fish caught). Thus,

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

What about the probability that my catch contains **at most two** tagged fish? Here, we want

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

**Example 2.21.** 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. Define

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

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, of course, 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 **OC curve**. Of course, as $r$ (or equivalently, $p$) increases, the probability of accepting the lot decreases. Acceptance sampling is a big part of **statistical process control**. In practice, lot sizes may be very large (e.g., $N = 1000$, etc.), and developing mathematically-sound sampling plans is crucial in order to avoid using defective parts in finished products.

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

$$E(Y) = \frac{nr}{N}$$

and

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

Direct proofs of these are possible (they involve using some obscure combinatoric identities). However, there is a nicer way to prove these results in Chapter 5 (WMS). Thus, we defer the proofs until then.

*RELATIONSHIP WITH THE BINOMIAL*: As noted earlier, the binomial distribution and the hypergeometric are similar. The key difference is that for the binomial experiment, $p$ does not change from trial to trial, but it does in the hypergeometric setting, especially if $N$ is small. However, one can show (see Exercise 3.172 WMS) that

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

if $r/N \to p$. The upshot of the result is this: if $N$ is large, a binomial probability calculation, with $p = r/N$, closely **approximates** the corresponding hypergeometric probability calculation. Of course, since the binomial and hypergeometric distributions are similar in situations where $N$ is large, their means and variances are similar too. See the discussion on page 122 WMS.

## 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.

Suppose that an experiment satisfies the above three conditions, and let $Y$ denote the number of occurrences in an interval of length one.

*GOAL*: Find an expression for $p_Y(y) = P(Y = y)$, the pdf of $Y$.

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 $\lambda/n$, where $\lambda$ 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,

$$\begin{aligned}\lim_{n\to\infty} P(Y = y) &= \lim_{n\to\infty} \frac{n!}{y!\,(n-y)!}\lambda^y\Big(\frac{1}{n}\Big)^y\Big(1 - \frac{\lambda}{n}\Big)^n\left(\frac{1}{1-\frac{\lambda}{n}}\right)^y \\ &= \lim_{n\to\infty} \underbrace{\frac{n(n-1)\cdots(n-y+1)}{n^y}}_{A}\underbrace{\frac{\lambda^y}{y!}}_{B}\underbrace{\Big(1 - \frac{\lambda}{n}\Big)^n}_{C}\underbrace{\left(\frac{1}{1-\frac{\lambda}{n}}\right)^y}_{D} \qquad (2.3)\end{aligned}$$

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

$$\begin{aligned}\lim_{n\to\infty} A &= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-y+1)}{n^y} = 1 \\ \lim_{n\to\infty} B &= \lim_{n\to\infty} \frac{\lambda^y}{y!} = \frac{\lambda^y}{y!} \\ \lim_{n\to\infty} C &= \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda} \\ \lim_{n\to\infty} D &= \lim_{n\to\infty} \left(\frac{1}{1-\frac{\lambda}{n}}\right)^y = 1.\end{aligned}$$

Thus, (2.3) becomes

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

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

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

since $e^{\lambda} = \sum_{y=0}^{\infty} \frac{\lambda^y}{y!}$ (the McLaurin series expansion of $e^{\lambda}$).

*SOME EXAMPLES OF POISSON PROCESSES* (i.e., situations where a Poisson probability model may be reasonable):

(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 $\alpha$-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 earthquakes in California in a given year.

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

**Example 2.22.** The number of cars abandoned weekly on a certain highway has 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?

SOLUTION.

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

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

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

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

*REMARK*: WMS's Appendix III, (Table 3) includes an impressive table for Poisson probabilities of the form
$$P(Y \leq y) = \sum_{j=0}^{y} \frac{\lambda^j e^{-\lambda}}{j!}.$$
This makes computing compound event probabilities much easier.

*MGF FOR THE POISSON DISTRIBUTION*: Suppose that $Y \sim \text{Poisson}(\lambda)$. Then the mgf of $Y$, for all $t$, is given by
$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \sum_{y=0}^{\infty} e^{ty} \frac{\lambda^y e^{-\lambda}}{y!} \\ &= e^{-\lambda} \underbrace{\sum_{y=0}^{\infty} \frac{(\lambda e^t)^y}{y!}}_{e^{\lambda e^t}} \\ &= e^{-\lambda} e^{\lambda e^t} = \exp[\lambda(e^t - 1)]. \end{aligned}$$

*NOTATION*: Here, we have used the conventional notation $\exp[x] \equiv e^x$.

*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,

$$E(Y^2) = \lambda + \lambda^2.$$

Finally,

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

Thus, for the Poisson distribution, the mean and variance are **always** equal.

**Example 2.23.** Suppose that $Y$ denotes the number of defects, per month, observed at an automotive plant. From past experience, it is believed that $Y \sim$ Poisson(7). What is the probability that, in any given month, we observe 11 or more defectives?

SOLUTION. We want to compute

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

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

SOLUTION. First, we assume independence among the 12 months (is this reasonable?), and call the event {11 or more defects in a month} a "success." Thus, under our independence assumption and viewing each month as a "trial," we have a sequence of 12 Bernoulli trials with "success" probability $p = 0.099$.

Let $X$ denote the number of months where we observe 11 or more defects. Then, $X \sim \text{bin}(12, 0.099)$, and thus,

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

*EXTENDING THE POISSON PROCESS TO INTERVALS OF ARBITRARY LENGTH*: If events or occurrences in a Poisson process occur at a rate of $\lambda$ per unit time or space, then the expected number of occurrences in an interval of length $t$ is $\lambda t$.

**Example 2.24.** Phone calls arrive at a switchboard according to a Poisson process, at a rate of $\lambda = 3$ per minute. Thus, if $X$ represents the number of calls received in 5 minutes, we have that $X \sim \text{Poisson}(3 \cdot 5 = 15)$, and the probability that 8 or fewer calls come in during a 5-minute span is given by

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

from Appendix III.

Skip Sections 3.10 and 3.11. You will not be responsible for them.

# 3 Continuous Random Variables and Their Probability Distributions

Complimentary reading: Chapter 4 (WMS).

## 3.1 Cumulative distribution functions

*RECALL*: In the last chapter, we spent a lot of time looking at **discrete random variables**. Recall that a discrete random variable is one that can only assume a finite or countable number of values. We also learned about **probability distribution functions** (**pdfs**). Loosely speaking, these were functions that told us how to assign probabilities and to which points we assign probabilities to. The concept of a pdf will also be used this chapter with **continuous random variables**.

*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.

*TERMINOLOGY*: The term **probability density function** is synonymous with **probability distribution function**. The term "probability density function" is still abbreviated **pdf**.

We now introduce a new function associated with any random variable.

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

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

*REMARK*: The cumulative distribution function is defined for **all real** $y$ (not just for those values of $y$ in $R_Y$, the support set of $Y$).

*NOTATION*: Sometimes we abbreviate "cumulative distribution function" as **cdf**.

*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 pdf

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

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 $y \in \mathbf{R}$. Some points are worth mentioning concerning the graphs of the **pdf** and **cdf** (for this example, but these main points hold more generally when discussing **discrete** random variables).

- **PDF**
  - The height of the bar above $y$ denotes 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 ASSOCIATED WITH ANY CDF*: 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 $y_1 \leq y_2$, $F_Y(y_1) \leq F_Y(y_2)$.

## 3.2 Continuous random variables

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

*RECALL*: Cdfs associated with discrete random variables are **step-functions** (these 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 $f_Y(y)$, is given by

$$f_Y(y) = \frac{d}{dy}F_Y(y), \tag{3.4}$$

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

$$F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt. \tag{3.5}$$

*REMARK*: Equations (3.4) and (3.5) are key relationships linking pdfs and cdfs for **continuous** random variables!!

**Theorem 3.1.** Suppose that $Y$ is a continuous random variable with pdf $f_Y(y)$. Then,

(1) $f_Y(y) > 0$ for all $y \in R_Y$, and

(2) $\int_{-\infty}^{\infty} f_Y(y)dy = 1$.

*REMARK*: Compare these to the analog results for the **discrete** case (see page 29 notes). The only difference is that in the continuous case, integrals replace sums.

**Example 3.2.** Suppose that $Y$ has the pdf

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

We want to find the cdf $F_Y(y)$. To do this, we need to compute $P(Y \leq y)$ for all possible cases (remember, $F_Y(y)$ is defined for all $y$):

- 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

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

- when $y \geq 2$, we have

$$\begin{aligned} F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt &= \int_{-\infty}^{0} 0dt + \int_{0}^{2} \frac{1}{2}dt + \int_{2}^{y} 0dt \\ &= 0 + \frac{t}{2}\Big|_0^2 + 0 = 1. \end{aligned}$$

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.** In Example 3.2, what is $P(Y \leq 1)$?

SOLUTION. We know that $P(Y \leq 1) = F_Y(1) = 1/2$.

**Example 3.4.** The lifetime of a small electric motor ($Y$, measured in hours) has **cdf**

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

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 $y \geq 0$,

$$\begin{aligned} f_Y(y) = \frac{d}{dy}F_Y(y) &= \frac{d}{dy}\Big(1 - e^{-y/3}\Big) \\ &= 0 - \Big(-\frac{1}{3}\Big)e^{-y/3} = \frac{1}{3}e^{-y/3}. \end{aligned}$$

Thus, putting it all together we get

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

**Example 3.5.** For the cdf $F_Y(y)$ in Example 3.4, show that the four theoretical cdf properties hold.

**Example 3.6.** In Example 3.4, what is $P(Y \leq 2)$?

SOLUTION. We know that $P(Y \leq 2) = F_Y(2) = 1 - e^{-2/3} \approx 0.487$.

**Theorem 3.2.** If $Y$ is a **continuous** random variable with pdf $f_Y(y)$, then

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

**Corollary 3.3.** If $Y$ is a **continuous** random variable with cdf $F_Y(y)$, then

$$P(a \leq Y \leq b) = F_Y(b) - F_Y(a).$$

**Example 3.7.** In Example 3.4, what is the probability that a randomly selected motor lasts between 2 and 5 hours? That is, what is $P(2 \leq Y \leq 5)$?

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

- **PDF**

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

- **CDF**

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

*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 from Theorem 3.2 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

- intervals with **nonnegative** probability, and
- specific 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 3.8.** Suppose that $X$ represents the time (in seconds) until a certain chemical reaction takes place (in a manufacturing process, say), and varies according to the pdf

$$f_X(x) = \begin{cases} cxe^{-x/2}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

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

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

SOLUTION. To find the $c$, recall that $\int_{-\infty}^{\infty} f_Y(y)dy = 1$. Thus,

$$c\int_0^{\infty} xe^{-x/2}dx = 1. \tag{3.6}$$

Using an **integration by parts** argument with $u = x$ and $dv = e^{-x/2}dx$, we have that the LHS of (3.6) is

$$\begin{aligned} &= c\left[-2xe^{-x/2}\Big|_{x=0}^{\infty} + \int_{x=0}^{\infty} 2e^{-x/2}dx\right] \\ &= c\left[(0+0) + 2(-2)e^{-x/2}\Big|_{x=0}^{\infty}\right] \\ &= c[(-4)(0-1)] = 4c. \end{aligned}$$

Solving for $c$, we get $c = 1/4$.

(b) Using Theorem 3.2, we get

$$P(3.5 \leq X < 4.5) = \int_{3.5}^{4.5} \frac{1}{4}xe^{-x/2}dx \approx 0.135.$$

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

## 3.3 Expected values

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

$$E(Y) = \int_{-\infty}^{\infty} yf_Y(y)dy.$$

*MATHEMATICAL ASIDE*: Above, we assume that $\int_{-\infty}^{\infty} |y|f_Y(y)dy < \infty$; if this is not true, then $E(Y) = \infty$.

*RECALL*: When $Y$ is a **discrete** random variable with pdf $p_Y(y)$, the **expected value** (or **mean**) of $Y$ is given by

$$E(Y) = \sum_{\text{all } y} yp_Y(y).$$

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

**Example 3.9.** Suppose that $Y$ has a pdf given by

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

Here, the expected value of $Y$ is given by

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} y f_Y(y) dy \\ &= \int_0^1 y \times 2y \; dy \\ &= \int_0^1 2y^2 dy \\ &= 2\frac{y^3}{3}\Big|_0^1 = 2\left(\frac{1}{3} - 0\right) = 2/3. \end{aligned}$$

*EXPECTATIONS OF FUNCTIONS OF $Y$.* Let $g$ be a real-valued function and let $Y$ be a **continuous** random variable. Then, $g(Y)$ is a random variable and

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y) f_Y(y) dy.$$

*MATHEMATICAL ASIDE*: Again, in computing $E[g(Y)]$, it must be the case that $\int_{-\infty}^{\infty} |g(y)| f_Y(y) dy < \infty$. Otherwise, $E[g(Y)] = \infty$.

**Example 3.10.** With the pdf in Example 3.9, compute $E(Y^2)$ and $E(\ln Y)$.

SOLUTION. (a) We note that $g_1(Y) = Y^2$ is just a function of $Y$. Thus,

$$\begin{aligned} E(Y^2) = \int_0^1 y^2 \times 2y \; dy &= \int_0^1 2y^3 dy \\ &= 2\,\frac{y^4}{4}\Big|_0^1 = 1/2. \end{aligned}$$

Also, since $g_2(Y) = \ln Y$ is just a function of $Y$, we have

$$\begin{aligned} E(\ln Y) = \int_0^1 \ln y \times 2y \; dy &= 2\int_0^1 y \ln y \; dy \\ &= \frac{1}{2}y^2 \ln y \Big|_0^1 - \int_0^1 \frac{1}{2}y^2 \times \frac{1}{y} \; dy \qquad (3.7) \\ &= (0-0) - \frac{1}{2}\left(\frac{y^2}{2}\Big|_0^1\right) \\ &= -\frac{1}{2} \times \frac{1}{2} = -\frac{1}{4}. \end{aligned}$$

Note that (3.7) follows from an **integration by parts** argument (with $u = \ln y$ and $dv = ydy$).

*A SPECIAL EXPECTATION*: For a **continuous** random variable $Y$, the **variance** of $Y$, $V(Y)$, is given by

$$V(Y) \equiv E[(Y - \mu_Y)^2] = \int_{-\infty}^{\infty} (y - \mu_Y)^2 f_Y(y) dy,$$

where $\mu_Y = E(Y)$.

**Example 3.11.** With the pdf in Example 3.9, compute $V(Y)$.

SOLUTION. Recall that $\mu_Y = E(Y) = 2/3$, from Example 3.9. Now, we could compute

$$V(Y) = \int_0^1 (y - \mu_Y)^2 2y \; dy$$

directly (try it!). However, recall that $V(Y) = E(Y^2) - [E(Y)]^2$ (**the computing formula**); thus, $V(Y) = 1/2 - (2/3)^2 = 1/18$ (recall that in Example 3.10, we computed the second moment, $E(Y^2)$).

*ANOTHER SPECIAL EXPECTATION*: For a **continuous** random variable $Y$, the **moment generating function** for $Y$, denoted $m_Y(t)$, is given by

$$m_Y(t) = E(e^{tY}) = \int_{-\infty}^{\infty} e^{ty} f_Y(y) dy,$$

provided $E(e^{tY})$ is finite for $t$ in an open neighborhood about 0.

*MOMENTS AND CENTRAL MOMENTS*: If $Y$ is a random variable, the $k$**th moment about the origin** (or just $k$**th moment**) is given by

$$\mu_k \equiv E(Y^k), \; k = 1, 2, ...$$

and the $k$**th moment about the mean** (or $k$**th central moment**) is given by

$$\mu_k' \equiv E[(Y-\mu)^k], \ k = 1, 2, ...$$

QUESTION: What is another name for the 2nd central moment?

**Example 3.12.** 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.

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

for values of $t < 1$.

With the mgf, we can derive the mean and variance. Differentiating the mgf, we get

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

Thus,

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

Now, we need to 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,

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

Finally,

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

EXERCISE. Find $E(Y)$ and $V(Y)$ directly (from the definitions of mean and variance).

*PROPERTIES OF EXPECTATIONS*: Let $Y$ be a **continuous** random variable with pdf $f_Y(y)$, let $g, g_1, g_2, ..., g_k$ denote real-valued functions, and let $c$ be any real constant. Then,

(a) $E(c) = c$

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

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

You'll note that these properties are **identical** to those when $Y$ was discrete.

## 3.4 Uniform distribution

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

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

*REMARKS AND NOTATION*: Shorthand notation is $Y \sim U(\theta_1, \theta_2)$. Sometimes, we call $\theta_1$ and $\theta_2$ the **model parameters**. A popular case is the $U(0, 1)$ distribution; i.e., a uniform distribution with $\theta_1 = 0$ and $\theta_2 = 1$.

EXERCISE. Show that the $U(\theta_1, \theta_2)$ pdf integrates to one.

EXERCISE. Derive the cdf $F_Y(y)$ for a $U(\theta_1, \theta_2)$ distribution.

**Example 3.12.** 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, $Y \sim U(0.1, 0.5)$, and

$$\begin{aligned} P(Y < 0.4) &= \int_{0.1}^{0.4} \frac{1}{0.5 - 0.1} dy \\ &= \frac{y}{0.4}\Big|_{0.1}^{0.4} \\ &= \frac{0.3}{0.4} = 0.75. \end{aligned}$$

*MEAN AND VARIANCE OF THE UNIFORM DISTRIBUTION*: If $Y \sim U(\theta_1, \theta_2)$, then

$$E(Y) = \frac{\theta_1 + \theta_2}{2}$$

and

$$V(Y) = \frac{(\theta_2 - \theta_1)^2}{12}.$$

*Proof.* Exercise.

*MGF FOR THE UNIFORM DISTRIBUTION*: Suppose that $Y \sim U(\theta_1, \theta_2)$. Then the mgf of $Y$ is given by

$$m_Y(t) = \frac{1}{\theta_2 - \theta_1}\left[\frac{e^{\theta_2 t} - e^{\theta_1 t}}{t}\right].$$

*Proof.* Exercise.

*AN INTERESTING FACT*: Suppose that in a Poisson process we examine the interval $(0, t)$. If it is known that **exactly one** occurrence is between 0 and $t$, then the **time** at which it occurred is uniformly distributed between 0 and $t$. See page 167 WMS.

*REMARK*: Uniform random variables play a big role in **simulation**.

## 3.5 Normal distribution

*TERMINOLOGY*: The random variable $Y$ is said to have a **normal distribution** if its pdf is given by

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

*NOTATION*: Shorthand notation for this is $Y \sim N(\mu, \sigma^2)$. There are two parameters in the normal distribution:

$$\begin{aligned} \mu &= \text{mean} \\ \sigma^2 &= \text{variance} \end{aligned}$$

The **parameter space** (i.e., the collection of all possible values of the parameters) is given by $\{(\mu, \sigma^2) : -\infty < \mu < \infty,\ \sigma^2 > 0\}$.

*FACTS ABOUT ANY NORMAL DISTRIBUTION*:

(a) The pdf is **symmetric** about $\mu$; that is, for any real constant $a$, $f_Y(\mu - a) = f_Y(\mu + a)$.

(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 notation to let $Z$ denote the **standard normal random variable**; we often write $Z \sim N(0, 1)$.

*IMPORTANT*: Tabled values of the cdf of $Z$ are given in Appendix III (Table 4) of WMS. This table enables us to find probabilities associated with the standard normal distribution. This turns out to be helpful since the integral

$$\int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy$$

does not exist in closed form!

First, let's prove that the normal pdf **integrates to one**. Let $z = \frac{y-\mu}{\sigma}$. Then, $dz = \frac{1}{\sigma}dy$ and $dy = \sigma dz$. Now, define

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz. \end{aligned}$$

Since $I > 0$, it suffices to show that $I^2 = 1$.

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \times \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-\left(\frac{x^2+y^2}{2}\right)\right]. \end{aligned}$$

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 $dxdy = rdrd\theta$ (the **Jacobian** of the transformation from $(x, y)$ space to $(r, \theta)$ space). Thus, we write

$$\begin{aligned} I^2 &= \int_{\theta=0}^{2\pi}\int_{r=0}^{\infty} \frac{1}{2\pi} e^{-\frac{r^2}{2}} r dr d\theta \\ &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} \left[ \int_{r=0}^{\infty} r e^{-\frac{r^2}{2}} dr \right] d\theta \\ &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} \left[ -e^{-\frac{r^2}{2}} \Big|_{r=0}^{\infty} \right] d\theta \\ &= \frac{1}{2\pi} \int_{\theta=0}^{2\pi} d\theta \\ &= \frac{\theta}{2\pi} \Big|_{\theta=0}^{2\pi} = 1. \end{aligned}$$

Hence, the normal density integrates to one. □

*MGF FOR THE NORMAL DISTRIBUTION*: Suppose that $Y \sim N(\mu, \sigma^2)$. Then 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.*

$$\begin{aligned} m_Y(t) &= E(e^{tY}) = \int_{-\infty}^{\infty} e^{ty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{ty - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy. \end{aligned} \tag{3.8}$$

Define $b = ty - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2$, the exponent in the last integral. Then,

$$\begin{aligned} b &= ty - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 \\ &= ty - \frac{1}{2\sigma^2}(y^2 - 2\mu y + \mu^2) \\ &= -\frac{1}{2\sigma^2}(y^2 - 2\mu y - 2\sigma^2 ty + \mu^2) \\ &= -\frac{1}{2\sigma^2}\big[\underbrace{y^2 - 2(\mu + \sigma^2 t)y}_{\text{complete the square}} + \mu^2\big] \\ &= -\frac{1}{2\sigma^2}\Big[y^2 - 2(\mu + \sigma^2 t)y + \underbrace{(\mu + \sigma^2 t)^2 - (\mu + \sigma^2 t)^2}_{\text{add and subtract}} + \mu^2\Big] \\ &= -\frac{1}{2\sigma^2}\Big[[y - (\mu + \sigma^2 t)]^2\Big] + \frac{1}{2\sigma^2}\Big[(\mu + \sigma^2 t)^2 - \mu^2\Big] \\ &= -\frac{1}{2\sigma^2}(y - a)^2 + \underbrace{\frac{1}{2\sigma^2}(\mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2 - \mu^2)}_{=c, \text{ say}}, \end{aligned}$$

where $a = (\mu + \sigma^2 t)$. Thus, the integral in (3.8) is equal to

$$\left( \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-a)^2}}_{N(a,\sigma^2) \text{ density}} dy \right) \times e^c$$

Now, finally note $e^c \equiv \exp(c) = \exp(\mu t + \sigma^2 t^2/2)$. Thus, the result follows. □

**Theorem 3.4.** Suppose that $Y \sim N(\mu, \sigma^2)$. Then, the random variable

$$Z = \frac{Y - \mu}{\sigma}$$

has a normal distribution with mean 0 and variance 1. That is, $Z \sim N(0, 1)$.

*Proof.* Let $Z = \frac{1}{\sigma}(Y - \mu)$. The mgf of $Z$ is given by

$$\begin{aligned} m_Z(t) = E(e^{tZ}) &= E[\exp(tZ)] \\ &= E\left[\exp\left[t\left(\frac{Y-\mu}{\sigma}\right)\right]\right] \\ &= E\left[\exp(-\mu t/\sigma)\exp\left(\frac{t}{\sigma}Y\right)\right] \\ &= \exp(-\mu t/\sigma)\underbrace{E\left[\exp\left(\frac{t}{\sigma}Y\right)\right]}_{m_Y(t/\sigma)} \\ &= \exp(-\mu t/\sigma) \times \exp\left[\mu(t/\sigma) + \frac{\sigma^2(t/\sigma)^2}{2}\right] \\ &= \exp[t^2/2], \end{aligned}$$

which is the mgf of a $N(0,1)$ random variable. Thus, by the **uniqueness** of moment generating functions, we know that $Z \sim N(0,1)$.

*REMARK*: Sometimes, we call $Z$ the **standardized value** of $Y$.

*FACTS ABOUT STANDARDIZED VALUES*:

- **unitless** quantities
- indicates how many standard deviations an observation falls above (below) the mean $\mu$.

Applying the above standardization idea to each component of an event of interest involving a random variable $Y$, we see that

$$(y_1 < Y < y_2) \iff \left(\frac{y_1-\mu}{\sigma} < \frac{Y-\mu}{\sigma} < \frac{y_2-\mu}{\sigma}\right) \iff \left(\frac{y_1-\mu}{\sigma} < Z < \frac{y_2-\mu}{\sigma}\right).$$

Similarly, we have

$$(Y < y) \iff \left(\frac{Y-\mu}{\sigma} < \frac{y-\mu}{\sigma}\right) \iff \left(Z < \frac{y-\mu}{\sigma}\right),$$

and

$$(Y > y) \iff \left(\frac{Y-\mu}{\sigma} > \frac{y-\mu}{\sigma}\right) \iff \left(Z > \frac{y-\mu}{\sigma}\right),$$

Thus, if we want to find probabilities associated with **any** normal random variable, and we know $\mu$ and $\sigma^2$, all we need is a table of probabilities for the **standard normal random variable** $Z$. Recall that WMS provides such a table (Table 4, page 792).

**Example 3.12.** Mercury contamination of edible freshwater fish may be a health risk. In a certain lake in Florida, young large-mouth bass were studied to examine the level of mercury contamination, $Y$, which for the fish in the lake, varies according to a normal distribution with mean 18 and variance 16.

(a) What proportion of contamination levels are between 11 and 21 parts per million? Here, we want $P(11 < Y < 21)$. By standardizing, we see that

$$\begin{aligned}
P(11 < Y < 21) &= P\left(\frac{11-18}{4} < \frac{Y-18}{4} < \frac{21-18}{4}\right) \\
&= P\left(\frac{11-18}{4} < Z < \frac{21-18}{4}\right) \\
&= P(-1.75 < Z < 0.75) \\
&= P(Z < 0.75) - P(Z < -1.75) \\
&= [1 - .2266] - 0.0401 \\
&= 0.7333.
\end{aligned}$$

Hence, about 73 percent of the fish have contamination levels between 11 and 21.

(b) For this model, ninety percent of all contamination levels will be above what mercury level? Here, we want to find the **10th percentile** of the $Y \sim \mathcal{N}(18, 16)$ distribution; i.e., we want $y$ such that $P(Y > y) = 0.90$. To find $y$, first find the $z$ so that $P(Z > z) = 0.90$, then "unstandardize" $y$. First, $z = -1.28$ (Table 4). Now, equate the standardized value of $y$ with $-1.28$,

$$\frac{y-18}{4} = -1.28,$$

and solve for $y$. Thus, $y = (-1.28)(4) + 18 = 12.88$, so 90 percent of all contamination levels are larger that 12.88 parts per million.

## 3.6 The gamma family of pdfs

In this section, we examine an important family of probability distributions; namely, those in the **gamma family**. There are three in particular:

- **exponential** distribution
- **gamma** distribution
- $\chi^2$ distribution

*NOTE*: The exponential and $\chi^2$ distributions are really just **special** gamma distributions!

*NOTE*: The exponential and gamma distributions are very popular models for **lifetime random variables**; i.e., random variables that record "time to event" measurements. Other lifetime distributions include the lognormal and Weibull probability models (and there are others as well).

### 3.6.1 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 $Y \sim$ exponential($\beta$). The value of $\beta$ determines the **shape** of the distribution. That the exponential density function integrates to one is easily shown (try it!).

*MGF FOR THE EXPONENTIAL DISTRIBUTION*: Suppose that $Y \sim$ exponential($\beta$). Then the mgf of $Y$ is given by

$$m_Y(t) = \left(\frac{1}{1-\beta t}\right),$$

for values of $t < 1/\beta$.

*Proof.* Define $\theta = \beta(1-\beta t)^{-1}$. Then,

$$\begin{aligned}
m_Y(t) = E(e^{tY}) &= \int_0^\infty e^{ty}\frac{1}{\beta}e^{-y/\beta}dy \\
&= \frac{1}{\beta}\int_0^\infty e^{-y/\theta}dy \\
&= -\frac{\theta}{\beta}e^{-y/\theta}\Big|_{y=0}^{\infty} \\
&= \lim_{b\to\infty}\frac{-1}{(1-\beta t)}e^{-y/\theta}\Big|_{y=0}^{b} \\
&= \lim_{b\to\infty}\frac{1}{(1-\beta t)}e^{-y/\theta}\Big|_{y=b}^{0} \\
&= \lim_{b\to\infty}\left[\frac{1}{(1-\beta t)} - \frac{1}{(1-\beta t)}e^{-b/\theta}\right] \\
&= \left(\frac{1}{1-\beta t}\right),
\end{aligned}$$

for values of $t < \frac{1}{\beta}$. (Why?)

*MEAN AND VARIANCE OF THE EXPONENTIAL DISTRIBUTION*: Suppose that $Y \sim$ exponential($\beta$). Then

$$E(Y) = \beta$$

and

$$V(Y) = \beta^2.$$

*Proof.* Exercise.

**Example 3.13.** The lifetime of a certain electrical component has an exponential distribution with mean $\beta = 500$ hours. What is the probability that a randomly selected component fails before 100 hours? lasts between 250 and 750 hours?

SOLUTION. 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

$$P(< 250 < Y < 750) = \int_{250}^{750} \frac{1}{500} e^{-y/500} dy \approx 0.383.$$

*CDF FOR THE EXPONENTIAL DISTRIBUTION*: Suppose that $Y \sim$ exponential($\beta$). Then the cdf of $Y$ exists in closed form and is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - \exp(-y/\beta), & y > 0. \end{cases}$$

*Proof.* Exercise.

Thus, when dealing with the exponential distribution, we can use the cdf to compute probabilities as well.

*THE FORGETFULNESS PROPERTY*: Suppose that $Y \sim$ exponential($\beta$), and suppose that $a$ and $b$ are both positive constants. Then

$$P(Y > a + b | Y > a) = P(Y > b).$$

That is, given the lifetime $Y$ has exceeded $a$, the probability that $Y$ exceeds $a + b$ **is the same as** if we were to look at $Y$ unconditionally lasting until time $b$. Put another way, that $Y$ has actually made it to time $a$ has been "forgotten." The exponential random variable is the only continuous random variable that enjoys the forgetfulness property! See page 179 WMS for details.

### 3.6.2 Gamma distribution

*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

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

*NOTATION*: Shorthand notation is $Y \sim \text{gamma}(\alpha, \beta)$ or $Y \sim \Gamma(\alpha, \beta)$

*REMARK*: Sometimes we call $\alpha$ the **scale parameter** and $\beta$ the **shape parameter**. The gamma probability model is **extremely flexible!** By changing the values of $\alpha$ and $\beta$, the gamma pdf can assume many shapes. Thus, the gamma model is very popular for modelling lifetime variables.

*REMARK*: Of course, when $\alpha = 1$, the gamma distribution reduces to the exponential($\beta$) distribution!

*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

$$\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} dy = \frac{1}{\Gamma(\alpha)} \int_0^\infty u^{\alpha-1} e^{-u} du = \frac{\Gamma(\alpha)}{\Gamma(\alpha)} = 1.$$

*THE GAMMA FUNCTION*: The quantity $\Gamma(t)$ is called the **gamma function**; it 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 integration by parts argument shows that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ for all $\alpha > 1$.

(2) If $\alpha$ is an integer, $\Gamma(\alpha) = (\alpha - 1)!$. For example, $\Gamma(5) = 4! = 24$.

*MGF FOR THE GAMMA DISTRIBUTION*: Suppose that $Y \sim \text{gamma}(\alpha, \beta)$. Then the mgf of $Y$ is given by

$$m_Y(t) = \left( \frac{1}{1 - \beta t} \right)^\alpha,$$

for values of $t < 1/\beta$.

*Proof.*

$$\begin{aligned} m_Y(t) = E(e^{tY}) &= \int_0^\infty e^{ty} \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta} dy \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty y^{\alpha-1} \exp\bigg[-(1-\beta t)\frac{y}{\beta}\bigg] dy \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty y^{\alpha-1} \exp[-y/\theta] dy \qquad (3.9) \end{aligned}$$

where $\theta = \beta(1-\beta t)^{-1}$. Thus, we see that (3.9) is equal to

$$\bigg(\frac{\theta}{\beta}\bigg)^\alpha \int_0^\infty \underbrace{\frac{1}{\Gamma(\alpha)\theta^\alpha} y^{\alpha-1} e^{-y/\theta}}_{\Gamma(\alpha,\theta) \text{ density}} dy = \bigg(\frac{\theta}{\beta}\bigg)^\alpha = \bigg(\frac{1}{1-\beta t}\bigg)^\alpha$$

Thus, the result follows. □

*MEAN AND VARIANCE OF THE GAMMA DISTRIBUTION*: If $Y \sim \text{gamma}(\alpha, \beta)$, then

$$E(Y) = \alpha\beta$$

and

$$V(Y) = \alpha\beta^2.$$

*Proof.* Exercise.

*TERMINOLOGY*: When talking about the $\Gamma(\alpha, \beta)$ density function, it is customary to break the formula into two parts:

- The **kernel**: $y^{\alpha-1}e^{-y/\beta}$
- The **constant**: $[\Gamma(\alpha)\beta^\alpha]^{-1}$.

**Example 3.14.** Suppose that $Y$ has pdf given by

$$f_Y(y) = \begin{cases} cy^2 e^{-y/4}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(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 $Y$?

### 3.6.3 $\chi^2$ distribution

*TERMINOLOGY*: In the $\Gamma(\alpha, \beta)$ family, when $\alpha = \nu/2$ (for any integer $\nu$) and $\beta = 2$, we call the resulting distribution a $\chi^2$ **distribution with $\nu$ degrees of freedom**.

*NOTATION*: If $Y$ has a $\chi^2$ distribution with $\nu$ degrees of freedom, shorthand notation for this $Y \sim \chi^2_\nu$.

*PDF OF A $\chi^2$ RANDOM VARIABLE*: If $Y \sim \chi^2_\nu$, then the pdf of $Y$ is given by

$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}} y^{(\nu/2)-1} e^{-y/2}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

*MEAN AND VARIANCE OF THE $\chi^2$ DISTRIBUTION*: If $Y \sim \chi^2_\nu$, then

$$E(Y) = \nu$$

and

$$V(Y) = 2\nu.$$

*Proof.* Once you establish the mean and variance for the $\Gamma(\alpha, \beta)$ random variable, this result follows immediately. □

*FACT*: If $Y \sim$ gamma$(\alpha, \beta)$, then $2Y/\beta$ has a $\chi^2_{2\alpha}$ distribution. We will use this fact over and over again.

*Proof.* Exercise.

*REMARK*: The $\chi^2$ distribution is used extensively in **applied statistics**.

*REMARK*: WMS catalogue the $\chi^2$ areas on pages 794-795.

## 3.7 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

$$f_Y(y) = \begin{cases} \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

*NOTATION*: 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$(\alpha, \beta)$ density function, it is customary to break the formula into two parts:

- The **kernel**: $y^{\alpha-1}(1-y)^{\beta-1}$
- The **constant**: $\frac{1}{B(\alpha,\beta)}$.

*REMARK*: Since the support of $Y$ is $0 < y < 1$, the beta distribution is a popular probability model for **proportions**.

*NOTES ON THE SHAPE OF THE BETA PDF*: The beta pdf is very **flexible**. That is, by changing the values of $\alpha$ and $\beta$, we can come up with many different pdf shapes.

- 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 $U(0, 1)$ pdf.

*REMARK*: The mgf of a beta$(\alpha, \beta)$ random variable exists, but, unfortunately, not in closed form. Hence, we'll have to compute moments directly.

*MEAN AND VARIANCE OF THE BETA DISTRIBUTION*: If $Y \sim$ beta$(\alpha, \beta)$, then

$$E(Y) = \frac{\alpha}{\alpha + \beta}$$

and

$$V(Y) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

*Proof.* Exercise.

**Example 3.15.** A filling station is supplied with gasoline once per week. Its weekly volume of sales (in 1000s of gallons) is a random variable, say $Y$, and 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? That is, what are the values of $\alpha$ and $\beta$?

(b) What is the **average** capacity?

(b) What need the capacity of the tank be so that the probability of the supply being exhausted in any week is 0.01?

SOLUTION. (a) $\alpha = 1$ and $\beta = 5$.

(b) $E(Y) = \frac{1}{1+5} = 1/6$. Thus, the average capacity is about 166.66 gallons.

(c) We want to find the capacity, say $c$, such that $P(Y > c) = 0.01$. This means that

$$P(Y > c) = \int_c^1 5(1-y)^4 dy = 0.01.$$

Thus, we need to solve this equation for $c$. Using a change of variable $y = 1 - x$, we see that

$$\begin{aligned} P(Y > c) &= \int_c^1 5(1-y)^4 dy \\ &= -\int_{1-c}^0 5x^4 dx \\ &= \int_0^{1-c} 5x^4 dx \\ &= x^5\Big|_0^{1-c} = (1-c)^5. \end{aligned}$$

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.

## 3.8 Some final comments on the various probability models

So far, we have talked about quite a few named **probability models**:

- **Discrete**: binomial, Bernoulli, geometric, negative binomial, hypergeometric, Poisson
- **Continuous**: uniform, normal, gamma, exponential, $\chi^2$, beta

There are many more!! You may hear about other "named distributions" such as the **Weibull**, **Gumbel**, **lognormal**, **Maxwell**, **Burr**, **LaPlace**, **Cauchy**, etc. All of these are used as probability models for the processes that generate those observations (i.e., data!).

*CURIOSITY*: How do we know how to pick the "right" probability model? Is there ever a "right" one?

> "All models are wrong; some are useful." −George Box

Don't worry about Sections 4.10 and 4.11. You will not be responsible for them.

# 4 Multivariate Distributions

Complimentary reading: Chapter 5 (WMS).

So far, we have only discussed single random variables. However, investigators are often interested in probability statements concerning **two or more** random variables.

**Example 4.1.** Suppose that in a field trial, we record $Y_1$ = yield (bushels/acre) and $Y_2$ = nitrogen content. We want to understand the **relationship** between $Y_1$ and $Y_2$.

**Example 4.2.** In a psychology study, the researcher records $(Y_1, Y_2)$, where $Y_1$ = pretest score and $Y_2$ = posttest score. The goal is **predict** $Y_2$ from the value of $Y_1$.

**Example 4.3.** A marketing study is undertaken to **forecast** next month's sales $Y_{n+1}$, based on the past sales $Y_1, Y_2, ..., Y_n$.

*TERMINOLOGY*: If $Y_1$ and $Y_2$ are random variables, then $(Y_1, Y_2)$ is called a **bivariate random vector**. In general, if $Y_1, Y_2, ..., Y_n$ are all 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. Of course, all ideas discussed herein extend naturally to higher dimensional settings.

## 4.1 Discrete random vectors

*TERMINOLOGY*: Let $Y_1$ and $Y_2$ be **discrete** random variables. The joint distribution 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 two dimensional **support** set of $(Y_1, Y_2)$. The function $p_{Y_1,Y_2}(y_1, y_2)$ is sometimes called a **joint** probability distribution function (pdf).

The requirements for $p_{Y_1,Y_2}(y_1, y_2)$ to be **valid** are that

(1) $p_{Y_1,Y_2}(y_1, y_2) > 0$ for all $(y_1, y_2) \in R_{Y_1,Y_2}$, and

(2) $\sum p_{Y_1,Y_2}(y_1, y_2) = 1$, where the sum is taken over all points $(y_1, y_2)$ in $R_{Y_1,Y_2}$.

**Example 4.4.** An urn contains 3 red balls, 4 white balls, and 5 blue balls. Let $(Y_1, Y_2)$ denote the bivariate random vector where, out of 3 randomly selected balls,

$$\begin{aligned} Y_1 &= \text{number of red balls} \\ Y_2 &= \text{number of white balls.} \end{aligned}$$

The pdf of $(Y_1, Y_2)$, $p_{Y_1,Y_2}(y_1, y_2)$, is given by

$$\begin{aligned} p_{Y_1,Y_2}(0,0) &= \frac{\binom{3}{0}\binom{4}{0}\binom{5}{3}}{\binom{12}{3}} = \frac{10}{220} \\ p_{Y_1,Y_2}(0,1) &= \frac{\binom{3}{0}\binom{4}{1}\binom{5}{2}}{\binom{12}{3}} = \frac{40}{220} \\ p_{Y_1,Y_2}(0,2) &= \frac{\binom{3}{0}\binom{4}{2}\binom{5}{1}}{\binom{12}{3}} = \frac{30}{220} \\ p_{Y_1,Y_2}(0,3) &= \frac{\binom{3}{0}\binom{4}{3}\binom{5}{0}}{\binom{12}{3}} = \frac{4}{220} \\ p_{Y_1,Y_2}(1,0) &= \frac{\binom{3}{1}\binom{4}{0}\binom{5}{2}}{\binom{12}{3}} = \frac{30}{220} \\ p_{Y_1,Y_2}(1,1) &= \frac{\binom{3}{1}\binom{4}{1}\binom{5}{1}}{\binom{12}{3}} = \frac{60}{220}, \end{aligned}$$

and similarly, $p_{Y_1,Y_2}(1,2) = \frac{18}{220}$, $p_{Y_1,Y_2}(2,0) = \frac{15}{220}$, $p_{Y_1,Y_2}(2,1) = \frac{12}{220}$, and $p_{Y_1,Y_2}(3,0) = \frac{1}{220}$. Here, the support set of $(Y_1, Y_2)$ is

$$R_{Y_1,Y_2} = \{(0,0), (0,1), (0,2), (0,3), (1,0), (1,1), (1,2), (2,0), (2,1), (3,0)\}.$$

We can also display the pdf in a tabular form. See Table 4.3.

*COMPUTING PROBABILITIES WITH THE DISCRETE JOINT PDF*: Suppose that $(Y_1, Y_2)$ is a discrete random vector with pdf $p_{Y_1,Y_2}(y_1, y_2)$. For any set $A \subset \mathcal{R}^2$,

$$P((Y_1, Y_2) \in A) = \sum_{\text{all } (y_1,y_2) \in A} p_{Y_1,Y_2}(y_1, y_2).$$

Table 4.3: *Joint pdf of* $Y_1$ *and* $Y_2$.

| | $y_2 = 0$ | $y_2 = 1$ | $y_2 = 2$ | $y_2 = 3$ | Row Sum |
|---|---|---|---|---|---|
| $y_1 = 0$ | $\frac{10}{220}$ | $\frac{40}{220}$ | $\frac{30}{220}$ | $\frac{4}{220}$ | $\frac{84}{220}$ |
| $y_1 = 1$ | $\frac{30}{220}$ | $\frac{60}{220}$ | $\frac{18}{220}$ | 0 | $\frac{108}{220}$ |
| $y_1 = 2$ | $\frac{15}{220}$ | $\frac{12}{220}$ | 0 | 0 | $\frac{27}{220}$ |
| $y_1 = 3$ | $\frac{1}{220}$ | 0 | 0 | 0 | $\frac{1}{220}$ |
| Column sum | $\frac{56}{220}$ | $\frac{112}{220}$ | $\frac{48}{220}$ | $\frac{4}{220}$ | 1 |

(It is straightforward to see that $\sum_{\text{all } (y_1,y_2)} p_{Y_1,Y_2}(y_1, y_2) = 1$).

**Example 4.5.** In Example 4.4, what is $P(Y_1 \leq 1, Y_2 \leq 1)$? Here, the event $A = \{(0,0), (0,1), (1,0), (1,1)\}$, and

$$\begin{aligned} P(Y_1 \leq 1, Y_2 \leq 1) &= p_{Y_1,Y_2}(0,0) + p_{Y_1,Y_2}(0,1) + p_{Y_1,Y_2}(1,0) + p_{Y_1,Y_2}(1,1) \\ &= \frac{10}{220} + \frac{40}{220} + \frac{30}{220} + \frac{60}{220} \\ &= \frac{140}{220} \end{aligned}$$

is the desired probability.

*TERMINOLOGY*: The **joint cumulative distribution function** of $(Y_1, Y_2)$ is given by

$$F_{Y_1,Y_2}(y_1, y_2) \equiv P(Y_1 \leq y_1, Y_2 \leq y_2)$$

for all $(y_1, y_2) \in \mathcal{R}^2$. This definition holds whether or not $(Y_1, Y_2)$ is **discrete**, **continuous**, or a **mixture** random vector.

**Example 4.6.** In Example 4.5, we computed $F_{Y_1,Y_2}(1,1) = P(Y_1 \leq 1, Y_2 \leq 1)$.

## 4.2 Continuous random vectors

*TERMINOLOGY*: If $Y_1$ and $Y_2$ are both **continuous** random variables, then $(Y_1, Y_2)$ is called a **continuous random vector**, and the joint distribution (or density) function

(pdf) of $(Y_1, Y_2)$ is denoted as $f_{Y_1,Y_2}(y_1, y_2)$ for $(y_1, y_2) \in R_{Y_1,Y_2}$.

For the pdf to be valid, it must be true that

(1)
$$f_{Y_1,Y_2}(y_1, y_2) > 0, \quad \text{for all } (y_1, y_2) \in R_{Y_1,Y_2}, \text{ and}$$

(2)
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2 = 1.$$

*RELATIONSHIP BETWEEN THE CDF AND PDF*: 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 for $(Y_1, Y_2)$ is given by

$$F_{Y_1,Y_2}(y_1, y_2) = \int_{-\infty}^{y_1}\int_{-\infty}^{y_2} f_{Y_1,Y_2}(t_1, t_2)dt_1dt_2,$$

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.

*COMPUTING PROBABILITIES WITH THE CONTINUOUS JOINT PDF*: Suppose that $(Y_1, Y_2)$ is a **continuous** random vector with pdf $f_{Y_1,Y_2}(y_1, y_2)$. For any set $A \subset \mathcal{R}^2$,

$$P((Y_1, Y_2) \in A) = \int\int_{(y_1,y_2)\in A} f_{Y_1,Y_2}(y_1, y_2)dy_1dy_2.$$

Geometrically, $P((Y_1, Y_2) \in A)$ represents the **volume** under $f_{Y_1,Y_2}(y_1, y_2)$ above $A$.

**Example 4.7.** Suppose that in a controlled experiment, we observe $(Y_1, Y_2)$, where $Y_1 =$ temperature (in Celcius) and $Y_2 =$ precipitation level (in inches). 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 $P(Y_2 > Y_1/5)$.

(a) We know that

$$\int_{y_1=10}^{20}\int_{y_2=0}^{3} cy_1y_2dy_1dy_2 = 1.$$

Thus,

$$\begin{aligned} 1 = \int_{y_1=10}^{20}\int_{y_2=0}^{3} cy_1y_2dy_1dy_2 &= c\int_{y_1=10}^{20} y_1\left(\frac{y_2^2}{2}\bigg|_0^3\right)dy_1 \\ &= \frac{9c}{2}\left(\frac{y_1^2}{2}\bigg|_{10}^{20}\right) \\ &= \frac{9c}{2}(200-50) = 675c. \end{aligned}$$

Thus, $c = 1/675$.

(b) Let $A = \{(y_1, y_2) : y_1 > 15, y_2 < 1\}$. Then,

$$\begin{aligned} P((Y_1, Y_2) \in A) &= \int_{y_1=15}^{20}\int_{y_2=0}^{1} \frac{1}{675}y_1y_2dy_1dy_2 \\ &= \frac{1}{675}\int_{y_1=15}^{20} y_1\left(\frac{y_2^2}{2}\bigg|_0^1\right)dy_1 \\ &= \frac{1}{1350}\left(\frac{y_1^2}{2}\bigg|_{15}^{20}\right) \\ &= \frac{1}{1350}\left(200 - \frac{225}{2}\right) \approx 0.065. \end{aligned}$$

(c) Let $A = \{(y_1, y_2) : y_2 > y_1/5\}$. Then,

$$\begin{aligned}
P((Y_1, Y_2) \in A) &= \int_{y_2=2}^{3} \int_{y_1=10}^{5y_2} \frac{1}{675} y_1 y_2 dy_1 dy_2 \\
&= \frac{1}{675} \int_{y_2=2}^{3} y_2 \left( \frac{y_1^2}{2} \Big|_{10}^{5y_2} \right) dy_2 \\
&= \frac{1}{675} \int_{y_2=2}^{3} y_2 \frac{1}{2} (25y_2^2 - 100) dy_2 \\
&= \frac{1}{1350} \int_{y_2=2}^{3} (25y_2^3 - 100y_2) dy_2 \\
&= \frac{1}{1350} \left( \frac{25y_2^4}{4} - 50y_2^2 \Big|_2^3 \right) \\
&= \frac{1}{1350} \left[ \frac{25(3)^4}{4} - 50(3)^2 - \frac{25(2)^4}{4} + 50(2)^2 \right] \\
&\approx 0.116.
\end{aligned}$$

## 4.3 Marginal distributions

Recall the joint pdf of $(Y_1, Y_2)$ in Example 4.4. 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, from Table 4.3 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** $Y_2$.

*TERMINOLOGY*: Let $(Y_1, Y_2)$ be a **discrete** random vector with pdf $p_{Y_1,Y_2}(y_1, y_2)$. Then the **marginal pdf** 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 pdf** of $Y_2$ is

$$p_{Y_2}(y_2) = \sum_{\text{all } y_1} p_{Y_1,Y_2}(y_1, y_2).$$

Thus, in the two-dimensional discrete case, marginal pdfs 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

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)dy_1.$$

Thus, in the two-dimensional continuous case, marginal pdfs are obtained by integrating out over the other variable.

**Example 4.8.** Suppose that $(Y_1, Y_2)$ has joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 6y_1, & 0 < y_1 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Now, for values of $y_2 \in [y_1, 1]$, 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(2,2) distribution! That is, marginally, $Y_1 \sim$ beta$(2, 2)$.

Similarly, for values of $y_1 \in [0, y_2]$, 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(3,1) distribution! That is, marginally, $Y_2 \sim$ beta$(3, 1)$.

**Example 4.9.** In Example 4.8, compute $P(Y_2 > 0.5)$.

SOLUTION. This can be done two different ways.

(1) Use the **joint** distribution $f_{Y_1,Y_2}(y_1, y_2)$ and compute

$$\int_{y_2=0.5}^{1} \int_{y_1=0}^{y_2} 6y_1 dy_1 dy_2.$$

(2) Use the **marginal** distribution $f_{Y_2}(y_2)$ and compute

$$\int_{y_2=0.5}^{1} 3y_2^2 dy_2.$$

*DISCLAIMER*: When you derive a marginal distribution, it is good to check that it integrates to one. Of course, if it does not, you have committed an error in deriving it!

## 4.4 Conditional distributions

Recall that if $A$ and $B$ are events in a non-empty sample space $S$, we defined

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

as long as $P(B) > 0$. Letting $B = \{Y_2 = y_2\}$ and $A = \{Y_1 = y_1\}$, we obtain

$$P(A|B) = \frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_2 = y_2)}. \tag{4.10}$$

Now, when $(Y_1, Y_2)$ is a discrete random vector, we see that the RHS of (4.10) is given by

$$\frac{P(Y_1 = y_1, Y_2 = y_2)}{P(Y_2 = y_2)} = \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_2}(y_2)}.$$

*TERMINOLOGY*: Suppose that $(Y_1, Y_2)$ is a discrete random vector with joint pdf $p_{Y_1,Y_2}(y_1, y_2)$. We define the **conditional distribution** 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 $p_{Y_2}(y_2) > 0$. The conditional distribution of $Y_2$, given $Y_1 = y_1$, is defined analogously.

**Example 4.10.** Using the joint pdf of $(Y_1, Y_2)$, in Example 4.4, find the conditional distribution of $Y_1$, when $Y_2 = y_2 = 1$.

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, given that $Y_2 = y_2 = 1$, $p_{Y_1|Y_2}(y_1|y_2 = 1)$ tells us how $Y_1$ is distributed.

EXERCISE. Compute $p_{Y_1|Y_2}(y_1|y_2 = 2)$ and $p_{Y_2|Y_1}(y_2|y_1 = 0)$.

When $(Y_1, Y_2)$ is a **continuous** random vector, then we have to alter the way we motivate conditional distributions; this is because the quantity

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_2}(y_2)} \tag{4.11}$$

always has a zero denominator!

*RESULT*: It turns out that (4.11) is the correct formula for conditional distributions involving continuous random vectors; however, we have to motivate its construction in a slightly different way.

Consider multiplying the LHS and RHS of (4.11) by $dy_1$ and $(dy_1dy_2)/dy_2$, respectively, to obtain

$$f_{Y_1|Y_2}(y_1|y_2)dy_1 = \frac{f_{Y_1,Y_2}(y_1,y_2)dy_1dy_2}{f_{Y_2}(y_2)dy_2}. \tag{4.12}$$

Now, the RHS of (4.12) is

$$\begin{aligned} &\approx \frac{P(y_1 \leq Y_1 \leq y_1 + dy_1, y_2 \leq Y_2 \leq y_2 + dy_2)}{P(y_2 \leq Y_2 \leq y_2 + dy_2)} \\ &= P(y_1 \leq Y_1 \leq y_1 + dy_1 | y_2 \leq Y_2 \leq y_2 + dy_2). \end{aligned}$$

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 $y_2 + dy_2$.

*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 $A$,

$$P(Y_1 \in A | Y_2 = y_2) = \sum_A f_{Y_1|Y_2}(y_1|y_2).$$

If $Y_1$ and $Y_2$ are jointly **continuous**, then for any set $A$,

$$P(Y_1 \in A | Y_2 = y_2) = \int_A f_{Y_1|Y_2}(y_1|y_2)dy_1.$$

In particular, if $A = (-\infty, y_1]$, then we get (for the continuous case)

$$P(Y_1 \leq y_1 | Y_2 = y_2) = \int_{-\infty}^{y_1} f_{Y_1|Y_2}(t_1|y_2)dt_1,$$

which is $F_{Y_1|Y_2}(y_1|y_2)$, the **conditional cumulative distribution function** (cdf) of $Y_1$, given $Y_2 = y_2$.

**Example 4.11.** Consider the bivariate pdf in Example 4.8:

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} 6y_1, & 0 < y_1 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Derive the conditional distributions $f_{Y_1|Y_2}(y_1|y_2)$ and $f_{Y_2|Y_1}(y_2|y_1)$.

SOLUTION. We showed in Example 4.8 that

$$f_{Y_1}(y_1) = \begin{cases} 6y_1(1-y_1), & 0 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

and

$$f_{Y_2}(y_2) = \begin{cases} 3y_2^2, & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Now, suppose that we fix $Y_2 = y_2$. Then, for values of $y_1 \in (0, y_2)$, it follows that

$$\frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_2}(y_2)} = \frac{6y_1}{3y_2^2} = \frac{2y_1}{y_2^2},$$

and, thus, this is the value of $f_{Y_1|Y_2}(y_1|y_2)$ when $y_1 \in (0, y_2)$. Remember, once we condition on $Y_2 = y_2$, then we simply regard $y_2$ as some constant! **This is an important point to understand.** Thus, the conditional distribution 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 distribution of $Y_2$ given $Y_1$, we fix $Y_1 = y_1$; then for all values of $y_2 \in (y_1, 1)$, we have

$$\frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_1}(y_1)} = \frac{6y_1}{6y_1(1-y_1)} = \frac{1}{1-y_1},$$

and, thus, this is the value of $f_{Y_2|Y_1}(y_2|y_1)$ when $y_2 \in (y_1, 1)$. Remember, once we condition on $Y_1 = y_1$, then we simply regard $y_1$ as some constant! Thus, the conditional distribution 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 $Y_1 = y_1$, $Y_2 \sim U(y_1, 1)$!

## 4.5 Independent random variables

*TERMINOLOGY*: Let $(Y_1, Y_2)$ be 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 $Y_2$ as $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$, respectively. We say that the random variables $Y_1$ and $Y_2$ are **independent** if and only if

$$F_{Y_1,Y_2}(y_1, y_2) = F_{Y_1}(y_1) \times F_{Y_2}(y_2)$$

for all values of $y_1$ and $y_2$.

*IN PRACTICE*: Instead of working with the cdfs, it is often easier to work with the cdfs.

*RESULT*: Let $(Y_1, Y_2)$ be a random vector (discrete or continuous) with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$, and denote the marginal pdfs of $Y_1$ and $Y_2$ as $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$, respectively. Then, the random variables $Y_1$ and $Y_2$ are **independent** if and only if

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1) \times f_{Y_2}(y_2)$$

for all values of $y_1$ and $y_2$. Thus, if $f_{Y_1,Y_2}(y_1, y_2) \neq f_{Y_1}(y_1) \times f_{Y_2}(y_2)$, then it follows that $Y_1$ and $Y_2$ are dependent random variables.

**Example 4.12.** 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 for any $y_1 \in (0,1)$ and $y_2 \in (0,1)$, we have that $f_{Y_1,Y_2}(y_1,y_2) = y_1 + y_2 \neq (y_1 + \frac{1}{2})(y_2 + \frac{1}{2}) = f_{Y_1}(y_1) \times f_{Y_2}(y_2)$, and, thus, $Y_1$ and $Y_2$ are dependent.

*A CONVENIENT RESULT*: Let $(Y_1, Y_2)$ be a random vector (discrete or continuous) with pdf $f_{Y_1,Y_2}(y_1,y_2)$, Then, 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 $f_{Y_1,Y_2}(y_1,y_2)$ into two nonnegative expressions

$$f_{Y_1,Y_2}(y_1,y_2) = g(y_1) \times h(y_2),$$

then $Y_1$ and $Y_2$ are **independent** random variables. Note that $g(y_1)$ and $h(y_2)$ are simply functions; **they need not be pdfs**. 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.

**Example 4.13.** Suppose that the random vector $(Y_1, Y_2)$ has joint pdf

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} \lambda e^{-\lambda y_1}(\lambda y_1)^{\alpha+\beta-1} \frac{y_2^{\alpha-1}(1-y_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} & y_1 > 0, 0 < y_2 < 1 \\ 0, & \text{otherwise} \end{cases}$$

for $\lambda > 0$, $\alpha > 0$, and $\beta > 0$. Then, it follows immediately that $Y_1$ and $Y_2$ are independent, since we can write

$$f_{Y_1,Y_2}(y_1,y_2) = \underbrace{\lambda e^{-\lambda y_1}(\lambda y_1)^{\alpha+\beta-1}}_{g(y_1)} \times \underbrace{\frac{y_2^{\alpha-1}(1-y_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}}_{h(y_2)}.$$

Note that we are **not** saying that $g(y_1)$ and $h(y_2)$ are marginal distributions of $Y_1$ and $Y_2$ (in fact, they are **not** the marginal distributions).

*NOTATION*: We want to generalize results for **independence** in terms of $n$-variate random vectors. We will 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 distribution of $\boldsymbol{Y}$ as $f_{\boldsymbol{Y}}(\boldsymbol{y})$.

*TERMINOLOGY*: Suppose that the random vector $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ has joint pdf $f_{\boldsymbol{Y}}(\boldsymbol{y})$, and that the random variable $Y_i$ has pdf $f_{Y_i}(y_i)$; $i = 1, 2, ..., n$. 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 distribution can be factored into the product of the marginal distributions.

## 4.6 Expectations of functions of random variables

Rules for expectations are similar to situations wherein only one variable is of interest.

*RESULT*: Suppose that the random vector $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ has joint pdf $f_{\boldsymbol{Y}}(\boldsymbol{y})$ (continuous) or $p_{\boldsymbol{Y}}(\boldsymbol{y})$ (discrete), and suppose that $g(\boldsymbol{Y}) = g(Y_1, Y_2, ..., Y_n)$ is any real vector valued function of $Y_1, Y_2, ..., Y_n$. Then,

- if $\boldsymbol{Y}$ is discrete,
$$E[g(\boldsymbol{Y})] = \sum_{\text{all } y_k} \cdots \sum_{\text{all } y_2} \sum_{\text{all } y_1} g(\boldsymbol{y}) p_{\boldsymbol{Y}}(\boldsymbol{y}),$$
- and if $\boldsymbol{Y}$ is continuous,
$$E[g(\boldsymbol{Y})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\boldsymbol{y}) f_{\boldsymbol{Y}}(\boldsymbol{y}) d\boldsymbol{y}.$$

If these quantities are not finite, then we say that $E[g(\boldsymbol{Y})]$ does not exist.

*PROPERTIES OF EXPECTATIONS*: Let $\boldsymbol{Y}$ be a random vector with pdf $f_{\boldsymbol{Y}}(\boldsymbol{y})$, let $g, g_1, g_2, ..., g_k$ denote real vector valued functions, and let $c$ be a real constant. Then,

(a) $E(c) = c$

(b) $E[cg(\boldsymbol{Y})] = cE[g(\boldsymbol{Y})]$

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

*REMARK*: Examples 5.15-5.21 in WMS (pages 242-246) provide excellent examples of these aforementioned facts concerning expectations.

**Example 4.14.** Suppose that $Y_1 \sim \mathcal{N}(\mu, \sigma_1^2)$, and $Y_2 \sim \mathcal{N}(\mu, \sigma_2^2)$. Define the random variables $T_1 = Y_1 + Y_2$, $T_2 = Y_1 - Y_2$, and $T_3 = Y_1^2 + Y_2^2$. Find $E(T_1)$, $E(T_2)$, and $E(T_3)$.
SOLUTION. We know that $E(T_1) = E(Y_1) + E(Y_2) = \mu + \mu = 2\mu$, and that $E(T_2) = E(Y_1) - E(Y_2) = \mu - \mu = 0$. Now, recall from the **variance computing formula**

$$V(Y_1) = E(Y_1^2) - [E(Y_1)]^2.$$

Thus, solving for $E(Y_1^2)$, we get $E(Y_1^2) = V(Y_1) + [E(Y_1)]^2 = \sigma_1^2 + \mu^2$. Similarly, $E(Y_2^2) = V(Y_2) + [E(Y_2)]^2 = \sigma_2^2 + \mu^2$. Thus,

$$\begin{aligned} E(T_3) &= E(Y_1^2 + Y_2^2) \\ &= E(Y_1^2) + E(Y_2^2) \\ &= (\sigma_1^2 + \mu^2) + (\sigma_2^2 + \mu^2) \\ &= \sigma_1^2 + \sigma_2^2 + 2\mu^2. \end{aligned}$$

*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, of course, that all expectations exist.
*Proof.* See WMS, pages 245-6.

**Example 4.15.** In Example 4.14, compute $E(Y_1Y_2^2)$ under the **assumption** that $Y_1$ and $Y_2$ are independent.
SOLUTION. Since $Y_1$ and $Y_2$ are independent, it follows that

$$E[Y_1Y_2^2] = E[Y_1]E[Y_2^2] = \mu(\sigma_2^2 + \mu^2) = \mu\sigma_2^2 + \mu^3.$$

## 4.7 Covariance and correlation

*NOTATION*: In discussing a bivariate pair of random variables, for this section, I will use the notation $(X, Y)$. There are good reasons for doing this, and I may allude to these reasons along the way.

*TERMINOLOGY*: Suppose that $X$ and $Y$ are random variables with means $\mu_X$ and $\mu_Y$, respectively. The **covariance** between $X$ and $Y$ is given by

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

The covariance gives us information about how $X$ and $Y$ are **linearly** related.

*THE COVARIANCE COMPUTING FORMULA*: It is easy to show that

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

This latter expression is sometimes easier to work with, and is sometimes called the **covariance computing formula**.

*NOTES ON THE COVARIANCE*:

- If $\text{Cov}(X, Y) > 0$, then $X$ and $Y$ are **positively** linearly related.
- If $\text{Cov}(X, Y) < 0$, then $X$ and $Y$ are **negatively** linearly related.
- If $\text{Cov}(X, Y) = 0$, then $X$ and $Y$ are **not** linearly related.

*REMARK*: If $\text{Cov}(X, Y) = 0$, this does **not** necessarily mean that $X$ and $Y$ are independent!

*RESULT*: If $X$ and $Y$ are independent, then $\text{Cov}(X, Y) = 0$.

*Proof.* Using the covariance computing formula, we have

$$\begin{aligned} \text{Cov}(X, Y) &= E(XY) - \mu_X\mu_Y \\ &= E(X)E(Y) - \mu_X\mu_Y = 0. \end{aligned}$$

Thus, the result follows. □

*CONTRAPOSITVELY*: If $\text{Cov}(X, Y) \neq 0$, then $X$ and $Y$ are dependent.

*MAIN POINT*: If two random variables are independent, then they have zero covariance; however, zero covariance does not necessarily imply independence.

**Example 4.16.** *An example of two dependent variables with zero covariance.* Suppose that $X \sim U(-1, 1)$, and let $Y = X^2$. It is straightforward to show that $E(X) = 0$, $E(XY) = E(X^3) = 0$, and $E(Y) = E(X^2) = V(X) = 1/3$ (you should verify all of these calculations). Thus,

$$\text{Cov}(X, Y) = E(XY) - \mu_X \mu_Y = 0 - 0(1/3) = 0.$$

However, not only are $X$ and $Y$ related, they are **perfectly** related! But, the relationship is not linear (it is quadratic).

*IMPORTANT RESULT*: Suppose that $X$ and $Y$ are random variables. Then

$$V(X + Y) = V(X) + V(Y) + 2\text{Cov}(X, Y)$$

and

$$V(X - Y) = V(X) + V(Y) - 2\text{Cov}(X, Y)$$

*Proof.*

$$\begin{aligned}
V(X+Y) &= E\{[(X+Y) - E(X+Y)]^2\} \\
&= E\{[X+Y-E(X)-E(Y)]^2\} \\
&= E\{[(X-E(X)) + (Y-E(Y))]^2\} \\
&= E[(X-E(X))^2 + (Y-E(Y))^2 + \underbrace{2(X-E(X))(Y-E(Y))}_{\text{cross product}}] \\
&= E[(X-E(X))^2] + E[(Y-E(Y))^2] + 2E[(X-E(X))(Y-E(Y))] \\
&= V(X) + V(Y) + 2\text{Cov}(X, Y).
\end{aligned}$$

That $V(X - Y) = V(X) + V(Y) - 2\text{Cov}(X, Y)$ is shown similarly. □

*RESULT*: Suppose that $X$ and $Y$ are **independent** random variables. Then

$$V(X+Y) = V(X) + V(Y).$$

*Proof.* We know that, in general, $V(X+Y) = V(X) + V(Y) + 2\text{Cov}(X,Y)$. However, since $X$ and $Y$ are independent, we have that $\text{Cov}(X,Y) = 0$. Thus, the result follows immediately. □

*LEMMA*: Suppose that $X$ and $Y$ are random variables with means $\mu_X$ and $\mu_Y$, respectively, and suppose that $a$, $b$, $c$, and $d$ are all constants. Then,

$$\text{Cov}(a+bX, c+dY) = bd\text{Cov}(X,Y).$$

*Proof.* Exercise.

*LEMMA*: Suppose that $X$ and $Y$ are random variables, then it follows that

(a) $\text{Cov}(X,Y) = \text{Cov}(Y,X)$

(b) $\text{Cov}(X,X) = V(X)$.

*Proof.* Exercise.

*PROBLEM*: Suppose that $X$ and $Y$ are random variables, and 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 $\beta_0$ and $\beta_1$. In this situation, the **error in prediction** is given by

$$Y - (\beta_0 + \beta_1 X).$$

Errors 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

$$E\{[Y - (\beta_0 + \beta_1 X)]^2\}.$$

It turns out that the mean squared error of prediction is minimized when

$$\beta_1 = \frac{\text{Cov}(X,Y)}{V(X)}$$

and

$$\beta_0 = E(Y) - \frac{\text{Cov}(X,Y)}{V(X)}E(X).$$

However, note that the value of $\beta_1$, algebraically, is equal to

$$\begin{aligned}\beta_1 &= \frac{\text{Cov}(X,Y)}{V(X)} \\ &= \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} \times \frac{\sigma_Y}{\sigma_X} \\ &= \rho_{X,Y} \times \frac{\sigma_Y}{\sigma_X},\end{aligned}$$

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.

*WARNING*: 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.

*REMARK*: Bivariate relationships form the basis for a statistical technique called **regression**.

*REMARK*: The correlation measure is oftentimes preferred over the covariance since $\rho_{X,Y}$ is on a bounded, unitless scale. It follows that $\text{Cov}(X, Y)$ can be any real number!

## 4.8 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. Then,

$$U_1 = \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

is called a **linear combination** of the random variables $Y_1, Y_2, ..., Y_n$.

*EXPECTED VALUE OF A LINEAR COMBINATION*:

$$E(U_1) = E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E(Y_i)$$

*VARIANCE OF A LINEAR COMBINATION*:

$$\begin{aligned} V(U_1) = V\left(\sum_{i=1}^{n} a_i Y_i\right) &= \sum_{i=1}^{n} a_i^2 V(Y_i) + 2\sum_{i<j} a_i a_j \text{Cov}(Y_i, Y_j) \\ &= \sum_{i=1}^{n} a_i^2 V(Y_i) + \sum_{i \neq j} a_i a_j \text{Cov}(Y_i, Y_j) \end{aligned}$$

*COVARIANCE BETWEEN TWO LINEAR COMBINATIONS*: Suppose that

$$U_2 = \sum_{i=1}^{m} b_j X_j = b_1 X_1 + b_2 X_2 + \cdots + b_m X_m.$$

Then, it follows that

$$\text{Cov}(U_1, U_2) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \text{Cov}(Y_i, X_j).$$

*REMARK*: Proofs of the last two results are given on pages 257-8 in WMS. The result involving the expectation of a linear combination follows directly from the linearity properties of the expected value operator.

*CASES WHEN* $n = 2$: Interest will often focus on situations wherein we have a linear combination of two random variables. In this setting,

$$E(a_1Y_1 + a_2Y_2) = a_1E(Y_1) + a_2E(Y_2)$$

and

$$V(a_1Y_1 + a_2Y_2) = a_1^2V(Y_1) + a_2^2V(Y_2) + 2a_1a_2\text{Cov}(Y_1, Y_2).$$

Similarly,

$$\begin{aligned}\text{Cov}(a_1Y_1 + a_2Y_2, b_1X_1 + b_2X_2) &= a_1b_1\text{Cov}(Y_1, X_1) + a_1b_2\text{Cov}(Y_1, X_2) \\ &\quad + a_2b_1\text{Cov}(Y_2, X_1) + a_2b_2\text{Cov}(Y_2, X_2)\end{aligned}$$

**Example 4.17.** Suppose that $Y_1 \sim \mathcal{N}(12, 4)$, $Y_2 \sim \mathcal{N}(16, 9)$, and $Y_3 \sim \mathcal{N}(20, 16)$. Also, suppose that $Y_1$ and $Y_2$ are independent, $\text{Cov}(Y_1, Y_3) = 0.8$, and $\text{Cov}(Y_2, Y_3) = -6.7$. Define the linear combinations

$$U_1 = 0.5Y_1 - 2Y_2 + Y_3 \quad \text{and} \quad U_2 = 3Y_1 - 2Y_2 - Y_3.$$

Find $E(U_1)$, $V(U_1)$, and $\text{Cov}(U_1, U_2)$.

SOLUTION.

$$\begin{aligned}E(U_1) &= E(0.5Y_1 - 2Y_2 + Y_3) \\ &= 0.5E(Y_1) - 2E(Y_2) + E(Y_3) \\ &= 0.5(12) - 2(16) + 20 = -6.\end{aligned}$$

$$\begin{aligned} V(U_1) &= V(0.5Y_1 - 2Y_2 + Y_3) \\ &= 0.5^2V(Y_1) + 2^2V(Y_2) + V(Y_3) \\ &\quad + 2(0.5)(-2)\underbrace{\text{Cov}(Y_1, Y_2)}_{=0} + 2(0.5)(1)\text{Cov}(Y_1, Y_3) + 2(-2)(1)\text{Cov}(Y_2, Y_3) \\ &= (0.25)(4) + 4(9) + 16 + 2(0.5)(-2)(0) + 2(0.5)(0.8) + 2(-2)(-6.7) = 80.6. \end{aligned}$$

$$\begin{aligned} \text{Cov}(U_1, U_2) &= \text{Cov}(0.5Y_1 - 2Y_2 + Y_3, 3Y_1 - 2Y_2 - Y_3) \\ &= (0.5)(3)\underbrace{\text{Cov}(Y_1, Y_1)}_{V(Y_1)} + (0.5)(-2)\underbrace{\text{Cov}(Y_1, Y_2)}_{=0} + (0.5)(-1)\text{Cov}(Y_1, Y_3) \\ &\quad + (-2)(3)\underbrace{\text{Cov}(Y_2, Y_1)}_{=0} + (-2)(-2)\underbrace{\text{Cov}(Y_2, Y_2)}_{V(Y_2)} + (-2)(-1)\text{Cov}(Y_2, Y_3) \\ &\quad + (1)(3)\text{Cov}(Y_3, Y_1) + (1)(-2)\text{Cov}(Y_3, Y_2) + (1)(-1)\underbrace{\text{Cov}(Y_3, Y_3)}_{V(Y_3)} \\ &= 28. \end{aligned}$$

EXERCISE. Compute $\rho_{U_1,U_2}$, the correlation between $U_1$ and $U_2$.

*IMPORTANT EXAMPLES FROM WMS*: Examples 5.27 and 5.28 (pages 258-9).

## 4.9 The multinomial model

*RECALL*: When we discussed the binomial model, we saw that each Bernoulli trial resulted in a "success" or a "failure;" that is, on each trial, only two outcomes were possible.

*TERMINOLOGY*: A **multinomial experiment** is simply a generalization of the binomial experiment.

(1) The experiment consists of $n$ trials (fixed).

(2) The outcome for any trial belongs to exactly one of $k \geq 2$ classes.

(3) The probability that an outcome for a single trial falls into class $i$ is given by $p_i$, for $i = 1, 2, ..., k$. Furthermore, $p_i$ remains constant from trial to trial.

(4) The trials are independent.

(5) $Y_i$ denotes the number of outcomes in class $i$, and $Y_1 + Y_2 + \cdots + Y_k = n$.

If each of these assumptions is true, and $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_k)$, then we call $\boldsymbol{Y}$ a **multinomial random vector**. We often write $\boldsymbol{Y} \sim \text{mult}(n, p_1, p_2, ..., p_k; \sum_i p_i = 1)$

*NOTE*: When $k = 2$, this reduces to our well-known binomial variable; when $k = 3$, we sometimes call $\boldsymbol{Y}$ a **trinomial** random vector, etc.

*PROBABILITY DENSITY FUNCTION*: If $\boldsymbol{Y} \sim \text{mult}(n, p_1, p_2, ..., p_k; \sum_i p_i = 1)$, then the pdf 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 4.18.** 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 independence among parts, then one may posit that $\boldsymbol{Y} = (Y_1, Y_2, Y_3) \sim \text{mult}(10, p_1, p_2, p_3; \sum_i p_i = 1)$.

**Example 4.19.** Suppose that in Example 4.18, $p_1 = 0.90$, $p_2 = 0.03$, and $p_3 = 0.07$. What is the probability that my sample (of 10) contains 8 non-defective parts, 1 defective part, and 1 reworkable part?

SOLUTION. We want to compute $p_{Y_1,Y_2,Y_3}(y_1 = 8, y_2 = 1, y_3 = 1)$.

$$\begin{aligned} p_{Y_1,Y_2,Y_3}(y_1 = 8, y_2 = 1, y_3 = 1) &= \frac{10!}{8!1!1!}(0.90)^8(0.03)^1(0.07)^1 \\ &\approx 0.041. \end{aligned}$$

*MISCELLANEOUS FACTS*: If $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_k) \sim \text{mult}(n, p_1, p_2, ..., p_k; \sum_i p_i = 1)$,

- $E(Y_i) = np_i$, for $i = 1, 2, ..., k$.
- $V(Y_i) = np_i(1 - p_i)$, for $i = 1, 2, ..., k$.
- $\text{Cov}(Y_i, Y_j) = -np_ip_j$, for $i \neq j$.
- The marginal distribution of $Y_i$ is binomial$(n, p_i)$, for $i = 1, 2, ..., k$.
- The distribution of $(Y_i, Y_j)$ is trinomial$(n, p_i, p_j, 1 - p_i - p_j)$.

## 4.10 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

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{y_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{y_1 - \mu_1}{\sigma_1}\right)\left(\frac{y_2 - \mu_2}{\sigma_2}\right) + \left(\frac{y_2 - \mu_2}{\sigma_2}\right)^2\right]$$

Shorthand notation for this is $(Y_1, Y_2) \sim$ bivariate normal$(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

*FACTS ABOUT THE BIVARIATE NORMAL DISTRIBUTION*:

- Marginally, $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

- If $\rho = 0$, then $Y_1$ and $Y_2$ are independent. This is only true in the bivariate normal setting!
- We can extend the bivariate normal distribution to higher dimensional settings.

## 4.11 Conditional expectations

*TERMINOLOGY*: Suppose that $Y_1$ and $Y_2$ are continuous random variables, that $g(Y_1)$ and $h(Y_2)$ are functions of $Y_1$ and $Y_2$, respectively, and that the conditional distributions $f_{Y_1|Y_2}(y_1|y_2)$ and $f_{Y_2|Y_1}(y_2|y_1)$ are defined. Then,

$$\underbrace{E[g(Y_1)|Y_2 = y_2]}_{\text{function of } y_2} = \int_{-\infty}^{\infty} g(y_1) f_{Y_1|Y_2}(y_1|y_2) dy_1$$

and

$$\underbrace{E[h(Y_2)|Y_1 = y_1]}_{\text{function of } y_1} = \int_{-\infty}^{\infty} h(y_2) f_{Y_2|Y_1}(y_2|y_1) dy_2.$$

If $Y_1$ and $Y_2$ are discrete, then sums replace the integrals.

*TERMINOLOGY*: If $g(Y_1) = Y_1$, we call $E[Y_1|Y_2 = y_2]$ is called the **conditional mean** of $Y_1$, given $Y_2 = y_2$. A similar statement holds for $E[Y_2|Y_1 = y_1]$.

**Example 4.20.** In Example 4.8, we considered the joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 6y_1, & 0 < y_1 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

and in Example 4.11, we derived the conditional distributions

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} 2y_1/y_2^2, & 0 < y_1 < y_2 \\ 0, & \text{otherwise} \end{cases}$$

and

$$f_{Y_2|Y_1}(y_2|y_1) = \begin{cases} \frac{1}{1-y_1}, & y_1 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the conditional mean of $Y_1$, given $Y_2 = y_2$ is given by

$$\begin{aligned} E(Y_1|Y_2 = y_2) &= \int_0^{y_2} y_1 f_{Y_1|Y_2}(y_1|y_2)dy_1 \\ &= \int_0^{y_2} y_1 \times 2y_1/y_2^2 dy_1 \\ &= \frac{2}{y_2^2}\int_0^{y_2} 2y_1^2 dy_1 \\ &= \frac{2}{y_2^2}\frac{y_1^3}{3}\Big|_0^{y_2} = \frac{2y_2}{3}. \end{aligned}$$

Similarly, we can show that $E(Y_2|Y_1 = y_1) = \frac{1}{2}(y_1 + 1)$.

*REMARK*: Since, in general, $E(Y_1|Y_2)$ is a function of $Y_2$, it must be a random variable itself! Thus, it has a mean and variance associated with it!!

*CONDITIONAL MEANS AND VARIANCES*: Suppose that $Y_1$ and $Y_2$ are random variables. Then the **laws of iterated expectation** and **variance**, respectively, are given by

$$E(Y_1) = E[E(Y_1|Y_2)] \tag{4.13}$$

and

$$V(Y_1) = E[V(Y_1|Y_2)] + V[E(Y_1|Y_2)].$$

*WARNING*: Be careful with these expectations. For example, on the RHS of (4.13), the inner expectation is taken with respect to the conditional distribution of $Y_1$ given $Y_2$. However, since $E(Y_1|Y_2)$ is a function of $Y_2$, the outer expectation on the RHS of (4.13) is taken with respect to the distribution of $Y_2$. Both of these results are proven on pages 271-272 in WMS.

**Example 4.21.** 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** itself; specifically, we suppose that the random variable $P$ (using capital letter notation like we usually do for random variables) varies according to a beta$(\alpha, \beta)$ distribution. Thus, we assume a **hierarchical structure**:

$$\begin{aligned} Y|P=p &\sim \text{binomial}(n,p) \\ P &\sim \text{beta}(\alpha,\beta). \end{aligned}$$

In this setting, the (unconditional) mean of $Y$ can be computed using an iterated expectation argument.

$$E(Y) = E[E(Y|P)] = E[nP] = nE(P) = n\left(\frac{\alpha}{\alpha+\beta}\right).$$

Also, the (unconditional) variance of $Y$ is given by

$$\begin{aligned} V(Y) &= E[V(Y|P)] + V[E(Y|P)] \\ &= E[nP(1-P)] + V[nP] \\ &= nE(P-P^2) + n^2V(P) \\ &= nE(P) - nE(P^2) + n^2V(P) \\ &= nE(P) - n[V(P) + E^2(P)] + n^2V(P) \\ &= n\left(\frac{\alpha}{\alpha+\beta}\right) - n\left[\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \left(\frac{\alpha}{\alpha+\beta}\right)^2\right] + \frac{n^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= n\left(\frac{\alpha}{\alpha+\beta}\right)\left[1-\left(\frac{\alpha}{\alpha+\beta}\right)\right] + \underbrace{\frac{n(n-1)\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}_{\text{extra variation}} \end{aligned}$$

Unconditionally, we say that $Y$ has a **beta-binomial** distribution. This is a popular probability model for situations wherein one observes binomial type responses with an inflated variance (this may be due to extra variation arising from "other factors").

# 5 Functions of Random Variables

Complimentary reading: Chapter 6 (WMS).

In many applications, interest lies in obtaining information about a function of a random variable, say, $g(Y)$, instead of the random variable $Y$ itself.

**Example 5.1.** In a medical experiment, $Y$ denotes the systolic blood pressure for a group of patients. However, oftentimes measurements are observed on the log scale. How is $g(Y) = \log Y$ distributed?

**Example 5.2.** In a manufacturing setting, $Y$ denotes machine down-time. How is the associated cost function $g(Y) = 2Y^2 + Y + 50$ distributed?

**Example 5.3.** An agricultural experiment is undertaken to study $Y$, the diameter of a certain species of eggs. How is $g(Y) = \sqrt{Y}$ distributed?

This chapter deals with finding distributions of **functions** of random variables. We will investigate three main techniques for doing this:

(1) Method of distribution functions

(2) Method of transformations

(3) Method of moment generating functions.

## 5.1 The method of distribution functions

*SETTING*: Suppose that $Y$ is a continuous random variable with cdf $F_Y(y)$. The general approach is to compute $F_U(u) = P(U \leq u)$, the cdf of $U = g(Y)$, and then to differentiate it (with respect to $u$) to find the pdf of $U$.

**Example 5.4.** Suppose that $Y \sim U(0, 1)$. Find the distribution of $U = g(Y) = -\log Y$.

SOLUTION. Using the method of distribution functions, we have

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ &= P(-\log Y \leq u) \\ &= P(\log Y \geq -u) \\ &= P(Y \geq e^{-u}) \\ &= 1 - P(Y \leq e^{-u}) \\ &= 1 - F_Y(e^{-u}) \end{aligned}$$

We know that $F_Y(y) = y$ for $0 < y < 1$. Thus, for $0 < e^{-u} < 1$, we have $F_U(u) = 1 - F_Y(e^{-u}) = 1 - e^{-u}$. Hence, taking derivatives, we get

$$f_U(u) = \frac{d}{du} F_U(u) = \frac{d}{du}(1 - e^{-u}) = e^{-u}$$

Now, when $0 < y < 1$, it must be that $u = -\log y > 0$. Thus,

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

That is, $U \sim$ exponential(1).

**Example 5.5.** Suppose that $Y \sim$ exponential(1). Find the distribution of $U = g(Y) = Y + \theta$, for $\theta > 0$.

SOLUTION. Using the method of distribution functions, we have

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ &= P(Y + \theta \leq u) \\ &= P(Y \leq u - \theta) \\ &= 1 - F_Y(u - \theta) \end{aligned}$$

We know that $F_Y(y) = 1 - e^{-y}$ for $y > 0$. Thus, for $u - \theta > 0$; i.e., $u > \theta$, we have $F_U(u) = 1 - F_Y(u - \theta) = 1 - e^{-(u-\theta)}$. Hence, taking derivatives, we get

$$f_U(u) = \frac{d}{du} F_U(u) = \frac{d}{du}\left[1 - e^{-(u-\theta)}\right] = e^{-(u-\theta)}.$$

Thus,

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

The random variable $U$ is said to have a **shifted exponential distribution**, where $\theta$ denotes the shift parameter.

**Example 5.6.** In Example 5.5, what is $E(U)$? $V(U)$?

SOLUTION. We could compute appeal to **The Law of the Unconscious Statistician** (LUS) and compute

$$E(U) = \int_\theta^\infty u f_U(u) du$$

or, equivalently,

$$E(Y+\theta) = \int_0^\infty (y+\theta) f_Y(y) dy.$$

The LUS says that both approaches will yield the same answer. However, in this situation, it would be easier to use the linearity properties of $E(\cdot)$ and compute $E(U) = E(Y+\theta) = E(Y) + \theta = 1 + \theta$. Also, $V(U) = V(Y+\theta) = V(Y) = 1$.

## 5.2 The method of transformations

*SETTING*: Suppose that $Y$ is a continuous random variable with cdf $F_Y(y)$, and that $U = g(Y)$, where $g$ is a **continuous one-to-one** function of $Y$.

*NOTE*: Examples of continuous one-to-one functions $g : \mathcal{R} \to \mathcal{R}$ include continuous (strictly) **increasing** and continuous (strictly) **decreasing** functions.

*NOTE*: If a function $g$ is one-to-one, it has an unique inverse $g^{-1}$. Furthermore, if $g$ is increasing (decreasing), so is $g^{-1}$.

Without loss, suppose that $g(y)$ is a strictly increasing function of $y$. Then, it follows that

$$u = g(y) \Leftrightarrow g^{-1}(u) = y,$$

and

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ &= P(g(Y) \leq u) \\ &= P(Y \leq g^{-1}(u)) \\ &= F_Y(g^{-1}(u)). \end{aligned}$$

Hence, differentiating $F_U(u)$, we get

$$f_U(u) = \frac{d}{du}F_U(u) = \frac{d}{du}F_Y(g^{-1}(u)) = f_Y(g^{-1}(u)) \underbrace{\frac{d}{du}g^{-1}(u)}_{\text{chain rule}}.$$

Now if $g$ is increasing, so is $g^{-1}$; thus, $\frac{d}{du}g^{-1}(u) > 0$. If $g$ is decreasing, $g^{-1}$ is as well and $\frac{d}{du}g^{-1}(u) < 0$. To handle both cases (as to avoid negative density functions), we use an absolute value measure. Summarizing, the density function of $U$, where positive, is given by

$$f_U(u) = f_Y(g^{-1}(u))\left|\frac{d}{du}g^{-1}(u)\right|.$$

*WARNING*: It is important to keep track of the support for $U$. If $R_Y$ denotes the support set of $Y$, then $R_U$, the support set for $U$, is given by $R_U = \{u : u = g(y); y \in R_Y\}$.

**Example 5.7.** Suppose that $Y \sim$ exponential($\beta$). Let $U = g(Y) = \sqrt{Y}$. What is the distribution of $U$?

SOLUTION. First, we should verify that the transformation is one-to-one. Clearly, $g(y) = \sqrt{y}$ is a continuous increasing function of $y$ over $(0, \infty)$, and, thus, $g(y)$ is one-to-one. Now,

$$g(y) = u = \sqrt{y} \Leftrightarrow \underbrace{y = g^{-1}(u) = u^2}_{\text{inverse transformation}}$$

and

$$\frac{d}{du}g^{-1}(u) = \frac{d}{du}u^2 = 2u.$$

Next, we need to find the support set of $U$. This is easy since $y > 0$ implies $u = \sqrt{y} > 0$.

Thus, $R_U = \{u : u > 0\}$. So for $u > 0$,

$$\begin{aligned} f_U(u) &= f_Y(g^{-1}(u))\left|\frac{d}{du}g^{-1}(u)\right| \\ &= \frac{1}{\beta}e^{-u^2/\beta} \times |2u| \\ &= \frac{2u}{\beta}e^{-u^2/\beta}, \end{aligned}$$

and $f_U(u) = 0$, otherwise. This is a **Weibull distribution** with parameters $\beta$ and $m = 2$ (see Exercise 4.152 on page 206 WMS).

**Example 5.8.** Suppose that $Y \sim \text{beta}(\alpha = 6, \beta = 2)$; i.e.,

$$f_Y(y) = \begin{cases} 42y^5(1-y), & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

What is the distribution of $U = g(Y) = 1 - Y$?

SOLUTION. First, we should verify that the transformation is one-to-one. Clearly, $g(y) = 1 - y$ is a continuous decreasing function of $y$ over $(0, 1)$, and, thus, $g(y)$ is one-to-one. Now,

$$g(y) = u = 1 - y \Leftrightarrow \underbrace{y = g^{-1}(u) = 1 - u}_{\text{inverse transformation}}$$

and

$$\frac{d}{du}g^{-1}(u) = \frac{d}{du}(1-u) = -1.$$

Next, we need to find the support set of $U$. This is easy since $0 < y < 1$ implies $0 < u < 1$. Thus, $R_U = \{u : 0 < u < 1\}$. So for $0 < u < 1$,

$$\begin{aligned} f_U(u) &= f_Y(g^{-1}(u))\left|\frac{d}{du}g^{-1}(u)\right| \\ &= 42(1-u)^5[1-(1-u)] \times |-1| \\ &= 42u(1-u)^5, \end{aligned}$$

and $f_U(u) = 0$, otherwise. Of course, we recognize this is a **beta distribution** with parameters $\alpha = 2$ and $\beta = 6$.

*IMPORTANT QUESTION*: What happens if $g$ is not one-to-one? In this situation, we can still use the method of transformations, but we have "break up" the transformation $g : R_Y \to R_U$ into disjoint regions where $g$ *is* one-to-one.

**Theorem 5.1.** Suppose that $Y$ is a continuous random variable with pdf $f_Y(y)$ and that $U = g(Y)$, not necessarily a one-to-one function (but continuous) of $y$. Suppose that we can partition the support $R_Y$ into a finite collection, say, $A_1, A_2, ..., A_k$, where

- $P(Y_i \in A_i) > 0$ for all $i$, and
- $f_Y(y)$ is continuous on each $A_i$.

Further, suppose that there exist functions $g_1(y), g_2(y), ..., g_k(y)$ such that $g_i(y)$ is defined on $A_i$, $i = 1, 2, ..., k$, and $g_i(y)$ satisfy

(a) $g(y) = g_i(y)$ for all $y \in A_i$

(b) $g_i(y)$ is monotone on $A_i$.

Then,

$$f_U(u) = \begin{cases} \sum_{i=1}^{k} f_Y(g_i^{-1}(u)) \left| \frac{d}{du} g_i^{-1}(u) \right|, & u \in R_U \\ 0, & \text{otherwise.} \end{cases}$$

**Example 5.9.** Suppose that $Y \sim \mathcal{N}(0, 1)$; that is, $Y$ has a standard normal distribution; i.e.,

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Consider the transformation $U = g(Y) = Y^2$. This transformation is not one-to-one on $R_Y = \mathcal{R} = \{y : -\infty < y < \infty\}$, but it *is* one-to-one on $A_1 = (-\infty, 0)$ and $A_2 = [0, \infty)$ (separately) since $g(y) = y^2$ is decreasing on $A_1$ and increasing on $A_2$. Furthermore, note that $A_1 \cup A_2$ partitions $R_Y$. Summarizing,

| Partition | Transformation | Inverse transformation |
|---|---|---|
| $A_1 = (-\infty, 0)$ | $g_1(y) = y^2 = u$ | $g_1^{-1}(u) = -\sqrt{u} = y$ |
| $A_2 = [0, \infty)$ | $g_2(y) = y^2 = u$ | $g_2^{-1}(u) = \sqrt{u} = y$ |

And, on both sets $A_1$ and $A_2$,

$$\left|\frac{d}{du} g_i^{-1}(u)\right| = \frac{1}{2\sqrt{u}}$$

Clearly, $u = y^2 > 0$; thus, $R_U = \{u : u \geq 0\}$. By Theorem 5.1, we know that the pdf of $U$ is given by

$$f_U(u) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{u})^2/2} \left|\frac{1}{2\sqrt{u}}\right| + \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{u})^2/2} \left|\frac{1}{2\sqrt{u}}\right|, & u \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, for $u \geq 0$, and realizing that $\Gamma(1/2) = \sqrt{\pi}$, $f_U(u)$ collapses to

$$\begin{aligned} f_U(u) &= \frac{2}{\sqrt{2\pi}} e^{-u/2} \left|\frac{1}{2\sqrt{u}}\right| \\ &= \frac{1}{\sqrt{2\pi}} u^{\frac{1}{2}-1} e^{-u/2} \\ &= \frac{1}{\sqrt{\pi} 2^{1/2}} u^{\frac{1}{2}-1} e^{-u/2} \\ &= \frac{1}{\Gamma(1/2) 2^{1/2}} u^{\frac{1}{2}-1} e^{-u/2} \end{aligned}$$

That is, $U \sim \Gamma(1/2, 2)$, but recall that the $\Gamma(1/2, 2)$ distribution is the same as a $\chi^2$ distribution with 1 degree of freedom; that is, $U \sim \chi^2(1)$.

*NOTE*: That $\Gamma(1/2) = \sqrt{\pi}$ can be shown by writing $\Gamma(1/2) = \int_0^\infty e^{-w} w^{-1/2} dw$, and then making a change of variable $u = \sqrt{2w}$. See Exercise 4.162 in WMS.

## 5.3 The method of moment generating functions

In our experience with moment generating functions so far, we have briefly talked about the following result.

**Theorem 5.2.** Let $X$ and $Y$ have mgfs $m_X(t)$ and $m_Y(t)$, respectively. If $m_X(t) = m_Y(t)$ for all $t$, then $X$ and $Y$ have the same probability distribution. This is sometimes referred to as the **uniqueness property** of moment generating functions.

*PUNCHLINE*: The mgf completely determines the distribution! How can we use this result? Suppose that we have a transformation $U = g(Y)$. If we can compute $m_U(t)$, the mgf of $U$, and can recognize it as one we already know (e.g., Poisson, normal, gamma, binomial, etc.), then we can use this uniqueness property to conclude that $U$ has that distribution!

**Example 5.10.** Suppose that $Y \sim \Gamma(\alpha, \beta)$. Use the method of moment generating functions to show that $U = g(Y) = 2Y/\beta$ has a $\chi^2$ distribution with $2\alpha$ degrees of freedom.

Solution. We know that the mgf of a $\chi^2_{2\alpha}$ distribution is given by $(1-2t)^{-\alpha}$, for values of $t < 1/2$. Thus, our strategy is to derive the mgf of $U$ and show that it equals this mgf.

$$\begin{aligned} m_U(t) = E(e^{tU}) &= E[e^{t(2Y/\beta)}] \\ &= E[e^{(2t/\beta)Y}] \\ &= m_Y(2t/\beta) \\ &= \left[\frac{1}{\beta(2t/\beta)}\right]^{\alpha} \\ &= \left(\frac{1}{1-2t}\right)^{\alpha}, \end{aligned}$$

for values of $t < 1/2$. Thus, by the uniqueness property of mgfs, we can conclude that $U = 2Y/\beta \sim \chi^2_{2\alpha}$.

*IMPORTANT REMINDER*: Remember that the $\Gamma(\alpha, 2)$ and the $\chi^2_{2\alpha}$ models represent the same distribution.

*NOTE*: The method of moment generating functions is very useful (and commonly applied) when we have independent random variables $Y_1, Y_2, ..., Y_n$ and interest lies in deriving the distribution of the **sum** $U = Y_1 + Y_2 + \cdots + Y_n$.

**Theorem 5.3.** Suppose that $Y_1, Y_2, ..., Y_n$ are independent random variables where $Y_i$ has mgf $m_{Y_i}(t)$ for $i = 1, 2, ..., n$. Let $U = Y_1 + Y_2 + \cdots + Y_n$. Then,

$$m_U(t) = \prod_{i=1}^{n} m_{Y_i}(t) = m_{Y_1}(t) \times m_{Y_2}(t) \times \cdots \times m_{Y_n}(t).$$

*Proof.*

$$\begin{aligned} m_U(t) = E(e^{tU}) &= E[e^{t(Y_1+Y_2+\cdots+Y_n)}] \\ &= E(e^{tY_1}e^{tY_2}\cdots e^{tY_n}) \\ &= E(e^{tY_1})E(e^{tY_2})\cdots E(e^{tY_n}) \\ &= m_{Y_1}(t) \times m_{Y_2}(t) \times \cdots \times m_{Y_n}(t) \\ &= \prod_{i=1}^{n} m_{Y_i}(t) \end{aligned}$$

Thus, the result follows. □

**Example 5.11.** Suppose that $Y_1, Y_2, ..., Y_n$ are independent Bernoulli($p$) random variables. The pdf of $Y_i$, for each $i$, is given by

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

What is the distribution of $U = Y_1 + Y_2 + \cdots + Y_n$?

SOLUTION. Using Theorem 5.3, we know that the mgf of $U$ is given by

$$\begin{aligned} m_U(t) &= m_{Y_1}(t) \times m_{Y_2}(t) \times \cdots \times m_{Y_n}(t) \\ &= \underbrace{(q + pe^t) \times (q + pe^t) \times \cdots \times (q + pe^t)}_{n \text{ times}} \\ &= (q + pe^t)^n, \end{aligned}$$

which we recognize as the mgf of a binomial($n, p$) random variable! Thus, by the uniqueness property of mgfs, we have that $U = Y_1 + Y_2 + \cdots + Y_n \sim$ binomial($n, p$).

**Example 5.12.** Suppose that $Y_1, Y_2, ..., Y_n$ are independent $\Gamma(\alpha_i, \beta)$ random variables. What is the distribution of $U = Y_1 + Y_2 + \cdots + Y_n$?

SOLUTION. Recall that for $t < 1/\beta$,

$$m_{Y_i}(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha_i},$$

for each $i = 1, 2, ..., n$. Thus,

$$\begin{aligned} m_U(t) &= m_{Y_1}(t) \times m_{Y_2}(t) \times \cdots \times m_{Y_n}(t) \\ &= \left(\frac{1}{1-\beta t}\right)^{\alpha_1} \times \left(\frac{1}{1-\beta t}\right)^{\alpha_2} \times \cdots \times \left(\frac{1}{1-\beta t}\right)^{\alpha_n} \\ &= \left(\frac{1}{1-\beta t}\right)^{\sum_{i=1}^n \alpha_i} \end{aligned}$$

Thus, by the uniqueness property of mgfs, we have that $U = Y_1 + Y_2 + \cdots + Y_n \sim \Gamma(\sum_{i=1}^n \alpha_i, \beta)$.

**Example 5.13.** Suppose that $Y_1, Y_2, ..., Y_n$ are independent $\mathcal{N}(\mu_i, \sigma_i^2)$ random variables for $i = 1, 2, ..., n$, and let $a_1, a_2, ..., a_n$ be non-random real constants. What is the distribution of $U = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$?

SOLUTION. For each $i = 1, 2, ..., n$, let $X_i = a_iY_i$. Then, we can write

$$U = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n = X_1 + X_2 + \cdots + X_n.$$

Furthermore, $X_1, X_2, ..., X_n$ are independent (since $Y_1, Y_2, ..., Y_n$ are independent). Thus,

$$\begin{aligned} m_U(t) &= m_{X_1}(t) \times m_{X_2}(t) \times \cdots \times m_{X_n}(t) \\ &= E(e^{tX_1})E(e^{tX_2}) \cdots E(e^{tX_n}) \\ &= E[e^{t(a_1Y_1)}]E[e^{t(a_2Y_2)}] \cdots E[e^{t(a_nY_n)}] \\ &= E[e^{(a_1t)Y_1}]E[e^{(a_2t)Y_n}] \cdots E[e^{(a_nt)Y_n}] \\ &= m_{Y_1}(a_1t) \times m_{Y_2}(a_2t) \times \cdots \times m_{Y_n}(a_nt) \\ &= \exp[(a_1\mu_1)t + \sigma_1^2(a_1t)^2/2] \times \exp[(a_2\mu_2)t + \sigma_2^2(a_2t)^2/2] \times \\ &\quad \cdots \times \exp[(a_n\mu_n)t + \sigma_n^2(a_nt)^2/2] \\ &= \exp\left[\left(\sum_{i=1}^n a_i\mu_i\right)t + \left(\sum_{i=1}^n a_i^2\sigma_i^2\right)t^2/2\right]. \end{aligned}$$

Thus, by the uniqueness property of mgfs, we have that $U = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n \sim \mathcal{N}\big(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2\big)$.

*IMPORTANT PUNCHLINE*: The distribution of a linear combination of independent normal random variables is itself normally distributed! We'll use this result over and over again.

*IMPORTANT RESULT*: Even if the random variables are not independent, the distribution of a linear combination of normal random variables is itself normally distributed. In this setting, what are the mean and the variance of the linear combination? (Recall, our results from the last chapter).

**Example 5.14.** Suppose that $Y_1 \sim \mathcal{N}(10, 2)$, $Y_2 \sim \mathcal{N}(16, 4)$, and $Y_3 \sim \mathcal{N}(4, 1)$. Then, if $Y_1, Y_2$, and $Y_3$ are independent, $U = 2Y_1 - 3Y_2 + 6Y_3 \sim \mathcal{N}(\mu_U, \sigma_U^2)$, where

$$\mu_U = 2(10) - 3(16) + 6(4) = -4$$
$$\sigma_U^2 = 2^2(2) + (-3)^2(4) + 6^2(1) = 80.$$

EXERCISE. With the $U$ in Example 5.14, find the distribution of $U$ if $\text{Cov}(Y_1, Y_2) = -2$, $\text{Cov}(Y_1, Y_3) = 1$, and $\text{Cov}(Y_2, Y_3) = 4$.

**Example 5.15.** In Example 5.13, if $a_1 = a_2 = \cdots = a_n = 1$, then we see that $U = Y_1 + Y_2 + \cdots + Y_n \sim \mathcal{N}\big(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\big)$.

**Example 5.16.** Suppose that $Y_1, Y_2, ..., Y_n$ are independent $\mathcal{N}(\mu_i, \sigma_i^2)$ random variables for $i = 1, 2, ..., n$, and let

$$Z_i = \frac{Y_i - \mu_i}{\sigma_i},$$

for each $i$. Then, we know that $Z_1, Z_2, ..., Z_n$ are independent $\mathcal{N}(0, 1)$ random variables, and furthermore, from Example 5.9, we know that $Z_1^2, Z_2^2, ..., Z_n^2$ are independent $\chi^2$ random variables, each having 1 degree of freedom. Since $\chi_1^2$ and $\Gamma(\frac{1}{2}, 2)$ represent the same probability distribution, we know from Example 5.12 that $U = \sum_{i=1}^n Z_i^2 \sim \Gamma(n/2, 2)$; i.e., $U \sim \chi_n^2$.

EXERCISE. Suppose that $Y_1, Y_2, ..., Y_n$ are independent random variables. In each of the following situations, use the method of moment generating functions to find the distribution of $U = Y_1 + Y_2 + \cdots + Y_n$.

(a) $Y_i \sim \text{Poisson}(\lambda_i)$

(b) $Y_i \sim \text{exponental}(\beta)$

ANSWERS. (a) $U \sim \text{Poisson}(\sum_{i=1}^{n} \lambda_i)$ (b) $U \sim \Gamma(n, \beta)$.

## 5.4 Multivariate transformations using jacobians

So far in this chapter, we have talked about **univariate transformations**; i.e., transformations involving a single random variable $Y$. However, sometimes interest will lie in deriving the distribution of **two** random variables $U_1$ and $U_2$, where

$$\begin{aligned} U_1 &= g_1(Y_1, Y_2) \\ U_2 &= g_2(Y_1, Y_2). \end{aligned}$$

Such a transformation is called a **bivariate transformation**. For our entire discussion, we will assume that $Y_1$ and $Y_2$ are jointly continuous random variables. For the following methods to apply, the bivariate transformation needs to be **one-to-one**.

*THE BIVARIATE TRANSFORMATION METHOD*: Suppose that $(Y_1, Y_2)$ is a continuous random vector with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$. Without loss, assume that $f_{Y_1,Y_2}(y_1, y_2) > 0$ for all $(y_1, y_2) \in R_{Y_1,Y_2}$, the two-dimensional support set of $(Y_1, Y_2)$. Now, let $g : \mathcal{R}^2 \to \mathcal{R}^2$ be a one-to-one vector valued mapping from $R_{Y_1,Y_2}$ to $R_{U_1,U_2}$, where $U_1 = g_1(Y_1, Y_2)$, $U_2 = g_2(Y_1, Y_2)$, and $R_{U_1,U_2}$ denotes the two-dimensional support set of $(U_1, U_2)$. If $g_1^{-1}(u_1, u_2)$ and $g_2^{-1}(u_1, u_2)$ have continuous partial derivatives with respect to both $u_1$ and $u_2$, and the Jacobian, $J$, where, with "det" denoting "determinant",

$$J = \det \begin{vmatrix} \frac{\partial g_1^{-1}(u_1,u_2)}{\partial u_1} & \frac{\partial g_1^{-1}(u_1,u_2)}{\partial u_2} \\ \frac{\partial g_2^{-1}(u_1,u_2)}{\partial u_1} & \frac{\partial g_2^{-1}(u_1,u_2)}{\partial u_2} \end{vmatrix} \neq 0,$$

then

$$f_{U_1,U_2}(u_1,u_2) = \begin{cases} f_{Y_1,Y_2}[g_1^{-1}(u_1,u_2), g_2^{-1}(u_1,u_2)]|J|, & (u_1,u_2) \in R_{U_1,U_2} \\ 0, & \text{otherwise,} \end{cases}$$

where $|J|$ denotes the absolute value of $J$.

*RECALL*: The determinant of the $2 \times 2$ matrix

$$\det \begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

is given by $ad - bc$.

*IMPORTANT*: When performing a bivariate transformation, the function $g : \mathcal{R}^2 \to \mathcal{R}^2$ must be one-to-one. In addition, one needs to keep track of what the transformation $U_1 = g_1(Y_1, Y_2), U_2 = g_2(Y_1, Y_2)$ does to the support set $R_{Y_1,Y_2}$. Remember, $g$ is a vector valued function that maps points in the support set $R_{Y_1,Y_2}$ to the transformed support set $R_{U_1,U_2}$.

*MATHEMATICAL NOTE*: To verify that $g$ is one-to-one, one needs to show that

$$g(y_1,y_2) = g(y_1',y_2') \Rightarrow (y_1,y_2) = (y_1',y_2').$$

*AN EASIER APPROACH*: Expressing the transformation as

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = A \begin{pmatrix} y_1 \\ y_2 \end{pmatrix},$$

where $A$ is a $2 \times 2$ matrix of constants, it follows that $g$ is one-to-one if $\det(A) \neq 0$.

**Example 5.17.** Suppose that $Y_1 \sim \Gamma(\alpha, 1)$, $Y_2 \sim \Gamma(\beta, 1)$, and that $Y_1$ and $Y_2$ are independent. Define the transformation

$$\begin{aligned} U_1 &= g_1(Y_1,Y_2) = Y_1 + Y_2 \\ U_2 &= g_2(Y_1,Y_2) = \frac{Y_1}{Y_1 + Y_2}. \end{aligned}$$

Find each of the following distributions:

(a) $f_{U_1,U_2}(u_1,u_2)$, the joint distribution of $U_1$ and $U_2$,

(b) $f_{U_1}(u_1)$, the marginal distribution of $U_1$, and

(c) $f_{U_2}(u_2)$, the marginal distribution of $U_2$.

SOLUTIONS. (a) Since $Y_1$ and $Y_2$ are independent, it follows that the joint distribution of $Y_1$ and $Y_2$ is given by

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= f_{Y_1}(y_1) \times f_{Y_2}(y_2) \\ &= \frac{1}{\Gamma(\alpha)} y_1^{\alpha-1} e^{-y_1} \times \frac{1}{\Gamma(\beta)} y_2^{\beta-1} e^{-y_2} \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} y_1^{\alpha-1} y_2^{\beta-1} e^{-(y_1+y_2)}, \end{aligned}$$

for $y_1 > 0$, $y_2 > 0$, and 0, otherwise. Here, $R_{Y_1,Y_2} = \{(y_1,y_2) : y_1 > 0, y_2 > 0\}$. Now, what does the transformation $g \equiv (g_1, g_2)$ do to the support set $R_{Y_1,Y_2}$? By inspection, we see that $u_1 = y_1 + y_2 > 0$, and $u_2 = \frac{y_1}{y_1+y_2}$ must fall between 0 and 1. Thus, the support of $(U_1, U_2)$ is given by $R_{U_1,U_2} = \{(u_1,u_2) : u_1 > 0, 0 < u_2 < 1\}$. It also follows that this transformation is one-to-one.

The next step is to derive the inverse transformation. It follows that

$$\begin{array}{c} u_1 = g_1(y_1,y_2) = y_1 + y_2 \\ u_2 = g_2(y_1,y_2) = \frac{y_1}{y_1+y_2} \end{array} \Rightarrow \begin{array}{c} y_1 = g_1^{-1}(u_1,u_2) = u_1 u_2 \\ y_2 = g_2^{-1}(u_1,u_2) = u_1 - u_1 u_2 \end{array}.$$

Thus, the Jacobian is given by

$$J = \det \begin{vmatrix} \frac{\partial g_1^{-1}(u_1,u_2)}{\partial u_1} & \frac{\partial g_1^{-1}(u_1,u_2)}{\partial u_2} \\ \frac{\partial g_2^{-1}(u_1,u_2)}{\partial u_1} & \frac{\partial g_2^{-1}(u_1,u_2)}{\partial u_2} \end{vmatrix} = \det \begin{vmatrix} u_2 & u_1 \\ 1-u_2 & -u_1 \end{vmatrix} = -u_2 u_1 - u_1(1-u_2) = -u_1.$$

Thus, $|J| = |-u_1| = u_1$. So, for $u_1 > 0$ and $0 < u_2 < 1$, we have that

$$f_{U_1,U_2}(u_1,u_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} (u_1 u_2)^{\alpha-1} (u_1 - u_1 u_2)^{\beta-1} e^{-[u_1 u_2 + (u_1 - u_1 u_2)]} \times u_1.$$

Rewriting this expression, we get

$$f_{U_1,U_2}(u_1,u_2) = \begin{cases} \frac{u_2^{\alpha-1}(1-u_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} u_1^{\alpha+\beta-1} e^{-u_1}, & u_1 > 0, 0 < u_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

We can see that $U_1$ and $U_2$ are independent since the nonzero part of $f_{U_1,U_2}(u_1, u_2)$ can be factored into the two expressions $g(u_1)$ and $h(u_2)$, where

$$g(u_1) = u_1^{\alpha+\beta-1}e^{-u_1}$$

and

$$h(u_2) = \frac{u_2^{\alpha-1}(1-u_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}.$$

(b) To obtain the marginal distribution of $U_1$, we integrate the joint pdf $f_{U_1,U_2}(u_1, u_2)$ out over $u_2$. Thus, for $u_1 > 0$,

$$\begin{aligned} f_{U_1}(u_1) &= \int_{u_2=0}^{1} f_{U_1,U_2}(u_1, u_2)du_2 \\ &= \int_{u_2=0}^{1} \frac{u_2^{\alpha-1}(1-u_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} u_1^{\alpha+\beta-1}e^{-u_1}du_2 \\ &= u_1^{\alpha+\beta-1}e^{-u_1}\frac{1}{\Gamma(\alpha+\beta)} \underbrace{\int_{u_2=0}^{1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u_2^{\alpha-1}(1-u_2)^{\beta-1}du_2}_{=\,1}. \end{aligned}$$

Hence, we have that

$$f_{U_1}(u_1) = \begin{cases} \frac{1}{\Gamma(\alpha+\beta)}u_1^{\alpha+\beta-1}e^{-u_1}, & u_1 > 0, \\ 0, & \text{otherwise.} \end{cases}$$

But, we recognize this as a $\Gamma(\alpha+\beta, 1)$ density function. That is, marginally, $U_1 \sim \Gamma(\alpha+\beta, 1)$. Compare this with Example 5.12.

(c) To obtain the marginal distribution of $U_2$, we integrate the joint pdf $f_{U_1,U_2}(u_1, u_2)$ out over $u_1$. Thus, for $0 < u_2 < 1$,

$$\begin{aligned} f_{U_2}(u_2) &= \int_{u_1=0}^{\infty} f_{U_1,U_2}(u_1, u_2)du_1 \\ &= \int_{u_1=0}^{\infty} \frac{u_2^{\alpha-1}(1-u_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} u_1^{\alpha+\beta-1}e^{-u_1}du_1 \\ &= \frac{u_2^{\alpha-1}(1-u_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{\int_{u_1=0}^{\infty} u_1^{\alpha+\beta-1}e^{-u_1}du_1}_{=\,\Gamma(\alpha+\beta)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}u_2^{\alpha-1}(1-u_2)^{\beta-1}. \end{aligned}$$

Hence, we have that

$$f_{U_2}(u_2) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u_2^{\alpha-1}(1-u_2)^{\beta-1}, & 0 < u_2 < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, marginally, $U_2 \sim \text{beta}(\alpha, \beta)$.

*NOTE*: We can use this bivariate transformation technique to derive the $t$ and $F$ distributions (next chapter). These distributions are popular in **applied statistics**.

*NOTE*: We can extend this procedure to $n$ variable settings. See WMS (page 315).

## 5.5 Order statistics

Suppose that $Y_1, Y_2, ..., Y_n$ are **independent** observations with common cdf $F_Y(y)$ and pdf $f_Y(y)$. Define

$$\begin{aligned} Y_{(1)} &= \text{smallest of } Y_1, Y_2, ..., Y_n \\ Y_{(2)} &= \text{second smallest of } Y_1, Y_2, ..., Y_n \\ &\vdots \\ Y_{(n)} &= \text{largest of } Y_1, Y_2, ..., Y_n. \end{aligned}$$

Then, $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ are called **order statistics**.

*RECALL*: Since $Y_1, Y_2, ..., Y_n$ are independent, the joint distribution of $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_Y(y_1) f_Y(y_2) \cdots f_Y(y_n),$$

the product of the $n$ individual pdfs.

*FACT*: Since there are $n!$ different ways we can order the observations $Y_1, Y_2, ..., Y_n$, it follows that the joint distribution of $Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$ is given by

$$f_{Y_{(1)}, Y_{(2)}, ..., Y_{(n)}}(y_1, y_2, ..., y_n) = n! \times f_Y(y_1) f_Y(y_2) \cdots f_Y(y_n),$$

for $-\infty < y_1 < y_2 < \cdots < y_n < \infty$, and 0, otherwise.

*GOALS*: We often are interested in understanding how a single order statistic is distributed (e.g., the minimum, the maximum, or one in between). In addition, we might want to derive the distribution of a function of order statistics, say $Y_{(n)} - Y_{(1)}$. Thus, we'll want to know how to get joint distributions of multiple order statistics.

*MARGINAL DISTRIBUTION OF THE MINIMUM ORDER STATISTIC*: To derive $f_{Y_{(1)}}(y)$, we will use the **distribution function technique** of Section 5.1.

$$\begin{aligned} F_{Y_{(1)}}(y) &= P(Y_{(1)} \leq y) \\ &= 1 - P(Y_{(1)} > y) \\ &= 1 - P(Y_1 > y \text{ and } Y_2 > y \text{ and }, ..., \text{ and } Y_n > y) \\ &= 1 - P(Y_1 > y)P(Y_2 > y) \cdots P(Y_n > y) \\ &= 1 - [P(Y_1 > y)]^n \\ &= 1 - [1 - F_Y(y)]^n. \end{aligned}$$

Thus, for values of $y$ in the support of $Y_{(1)}$,

$$\begin{aligned} f_{Y_{(1)}}(y) &= \frac{d}{dy} F_{Y_{(1)}}(y) \\ &= \frac{d}{dy}\{1 - [1 - F_Y(y)]^n\} \\ &= -n[1 - F_Y(y)]^{n-1}[-f_Y(y)] \\ &= n f_Y(y)[1 - F_Y(y)]^{n-1}, \end{aligned}$$

and 0, otherwise.

*MARGINAL DISTRIBUTION OF THE MAXIMUM ORDER STATISTIC*: To derive $f_{Y_{(n)}}(y)$, we will again use the **distribution function technique** of Section 5.1.

$$\begin{aligned} F_{Y_{(n)}}(y) &= P(Y_{(n)} \leq y) \\ &= P(Y_1 \leq y \text{ and } Y_2 \leq y \text{ and }, ..., \text{ and } Y_n \leq y) \\ &= P(Y_1 \leq y)P(Y_2 \leq y) \cdots P(Y_n \leq y) \\ &= [P(Y_1 \leq y)]^n \\ &= [F_Y(y)]^n. \end{aligned}$$

Thus, for values of $y$ in the support of $Y_{(n)}$,

$$\begin{aligned} f_{Y_{(n)}}(y) &= \frac{d}{dy}F_{Y_{(n)}}(y) \\ &= \frac{d}{dy}\{[F_Y(y)]^n\} \\ &= n[F_Y(y)]^{n-1}[f_Y(y)] \\ &= nf_Y(y)[F_Y(y)]^{n-1}, \end{aligned}$$

and 0, otherwise.

*MARGINAL DISTRIBUTION OF THE kth ORDER STATISTIC*: To derive $f_{Y_{(k)}}(y)$, we will appeal to the multinomial probability model. Define

| Class | Values of $Y$ | number |
|---|---|---|
| 1 | those values of $Y < y$ | $k-1$ |
| 2 | those values of $Y = y$ | 1 |
| 3 | those values of $Y > y$ | $n-k$ |

Thus, since $Y_1, Y_2, ..., Y_n$ are **independent**, we have, by appeal to the multinomial model,

$$f_{Y_{(k)}}(y) = \frac{n!}{(k-1)!1!(n-k)!}[F_Y(y)]^{k-1}[f_Y(y)]^1[1-F_Y(y)]^{n-k},$$

where we interpret

$$\begin{aligned} F_Y(y) &= P(Y_i < y) \\ f_Y(y) &= P(Y_i = y) \\ 1 - F_Y(y) &= P(Y_i > y). \end{aligned}$$

Thus, the pdf of the $k$th order statistic $Y_{(k)}$ is given by

$$f_{Y_{(k)}}(y) = \frac{n!}{(k-1)!(n-k)!}[F_Y(y)]^{k-1}f_Y(y)[1-F_Y(y)]^{n-k},$$

for values of $y$ in the support of $Y_{(k)}$, and 0, otherwise.

**Example 5.18.** Suppose that $Y_1, Y_2, ..., Y_{10}$ are independent random variables, each having a beta distribution with $\alpha = 2$ and $\beta = 1$. The pdf for each $Y_i$ is given by

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

(a) Find $f_{Y_{(1)}}(y)$ and compute $P(Y_{(1)} < 0.25)$.

(b) Find $f_{Y_{(10)}}(y)$ and compute $P(Y_{(10)} > 0.90)$.

(c) Find $f_{Y_{(6)}}(y)$ and compute $P(Y_{(6)} > 0.50)$.

SOLUTIONS. First, we need to derive the cdf of $Y$. For values of $0 < y < 1$,

$$F_Y(y) = \int_0^y 2t dt = y^2.$$

(a) The nonzero part of $f_{Y_{(1)}}(y)$ is given by

$$f_{Y_{(1)}}(y) = nf_Y(y)[1 - F_Y(y)]^{n-1} = 10(2y)[1 - y^2]^9.$$

After simplification, it follows that

$$f_{Y_{(1)}}(y) = \begin{cases} 20y[1-y^2]^9, & 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$P(Y_{(1)} < 0.25) = \int_0^{0.25} 20y[1-y^2]^9 dy \approx 0.476.$$

(b) The nonzero part of $f_{Y_{(10)}}(y)$ is given by

$$f_{Y_{(10)}}(y) = nf_Y(y)[F_Y(y)]^{n-1} = 10(2y)[y^2]^9.$$

After simplification, it follows that

$$f_{Y_{(10)}}(y) = \begin{cases} 20y^{19}, & 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$P(Y_{(10)} > 0.90) = \int_{0.90}^1 20y^{19} dy \approx 0.878.$$

(c) The nonzero part of $f_{Y_{(6)}}(y)$ is given by

$$\begin{aligned} f_{Y_{(6)}}(y) &= \frac{n!}{(k-1)!(n-k)!}[F_Y(y)]^{k-1} f_Y(y)[1 - F_Y(y)]^{n-k} \\ &= \frac{10!}{(6-1)!(10-6)!}[y^2]^{6-1} 2y[1-y^2]^{10-6}. \end{aligned}$$

After simplification, it follows that

$$f_{Y_{(6)}}(y) = \begin{cases} 2520y^{11}(1-y^2)^4, & 0 < y < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Thus,

$$P(Y_{(6)} > 0.50) = \int_{0.50}^{1} 2520y^{11}(1-y^2)^4 dy \approx 0.980.$$

**Example 5.19.** An engineering system consists of $n$ components placed in **series**; that is, the system fails when the first component fails. The $n$ component lifetimes $Y_1, Y_2, ..., Y_n$, are assumed to be independent and that each follows an exponential distribution with mean $\beta$. Since the system fails when the first component fails, system failures can be determined (probabilistically) by deriving the pdf of $Y_{(1)}$, the first order statistic.

$$\begin{aligned} f_{Y_{(1)}} &= nf_Y(y)[1-F_Y(y)]^{n-1} \\ &= n\left(\frac{1}{\beta}e^{-y/\beta}\right)\left[1-\left(1-e^{-y/\beta}\right)\right]^{n-1} \\ &= \frac{n}{\beta}e^{-y/\beta}\left(e^{-y/\beta}\right)^{n-1} \\ &= \frac{n}{\beta}e^{(-y/\beta)-[y(n-1)/\beta]} \\ &= \frac{1}{\beta/n}e^{-y/(\beta/n)}, \end{aligned}$$

for $y > 0$, and 0, otherwise. That is, $Y_{(1)} \sim$ exponential$(\beta/n)$.

**Example 5.20.** In Example 5.19, suppose that $n = 15$ and that $\beta = 1$ year. An engineer claims that a series system with these settings will likely last 6 months. Is there evidence to support his claim?

SOLUTION. We can compute the probability that the system lasts longer than 6 months (i.e., that $Y_{(1)} > 0.5$).

$$P(Y_{(1)} > 0.5) = \int_{1/2}^{\infty} \frac{1}{1/15}e^{-y/(1/15)}dy = \int_{1/2}^{\infty} 15e^{-15y}dy \approx 0.0006.$$

Thus, chances are the system would not last longer than six months. We don't have very much evidence to support his claim.

*REMARK*: It is also possible to think about joint distributions of two (or more) order statistics. For example, say that we wanted to get the joint distribution of $Y_{(1)}$ and $Y_{(n)}$, the minimum and maximum order statistics. Or, we might want to get the joint distribution of $Y_{(1)}$, $Y_{(2)}$, and $Y_{(3)}$. Such joint distributions can be derived by, again, appealing to the multinomial model. Details are on pages 321-322 of WMS, with an illustrative example appearing in Example 6.18 on pages 322-323.

# 6 Sampling Distributions and the Central Limit Theorem

Complimentary reading: Chapter 7 (WMS).

## 6.1 Independent and identically distributed random variables

In this chapter, and often in the subsequent course, we will often characterize $Y_1, Y_2, ..., Y_n$ as a **random sample**. We have to be clear what this means.

*IID OBSERVATIONS*: Suppose that $Y_1, Y_2, ..., Y_n$ are **independent** observations, where each $Y_i$ has the common pdf $f_Y(y;\theta)$. The model $f_Y(y;\theta)$ can be discrete or continuous. A succinct way to express this situation is "$Y_1, Y_2, ..., Y_n \sim$ iid $f_Y(y;\theta)$." The collection $Y_1, Y_2, ..., Y_n$ is called a **random sample**, and the model $f_Y(y;\theta)$ represents the distribution (population) from which the sample is drawn. The population is viewed as **infinite** in size.

*ACRONYM*: "iid" means "**i**ndependent and **i**dentically **d**istributed."

*NOTATION*: We emphasize the dependence of the pdf $f_Y(y;\theta)$ on the unknown parameter $\theta$. In a statistics problem, interest will often lie in **estimating** $\theta$ with the random sample $Y_1, Y_2, ..., Y_n$. The parameter $\theta$ could be single-valued or vector-valued.

**Example 6.1.** Suppose that $Y_1, Y_2, ..., Y_n \sim$ iid exponential($\theta$). Here $\theta$ denotes the **mean** of for the exponential distribution.

**Example 6.2.** Suppose that $Y_1, Y_2, ..., Y_n \sim$ iid binomial($m, \theta$), $m$ known. Here $\theta$ denotes the **success probability** for the binomial distribution.

**Example 6.3.** Suppose that $Y_1, Y_2, ..., Y_n \sim$ iid $\mathcal{N}(\mu, \sigma^2)$. Here, the parameter $\theta = (\mu, \sigma^2)$ is a vector of two parameters. As usual, $\mu$ denotes the **mean** and $\sigma^2$ denotes the **variance** of the underlying normal distribution.

*FRAME OF REFERENCE*: We envision $Y_1, Y_2, ..., Y_n$ as **data** that are observed in $n$ independent replicates of an experiment, and the distribution of each outcome is modelled by $f_Y(y; \theta)$.

**Example 6.4.** Suppose that $Y_1, Y_2, ..., Y_{20}$ is random sample from a $\mathcal{N}(75, 25)$ distribution. That is, $Y_1, Y_2, ..., Y_{20} \sim$ iid $\mathcal{N}(75, 25)$. Here, this normal distribution could be a model for salaries, test scores, hole diameters, wheat yields, etc.
QUESTION: What if the model parameters, here $\mu$ and $\sigma^2$, are **unknown**?

**Example 6.5.** Suppose that $Y_1, Y_2, ..., Y_9$ is random sample from a Poisson(4.4) distribution. That is, $Y_1, Y_2, ..., Y_9 \sim$ iid Poisson(4.4). Here, this Poisson distribution could be a model for the number of customers per hour, the number of insects per plot, the number of defects produced per day, etc.
QUESTION: What if the model parameter, here $\theta$, is **unknown**?

*RECALL*: Suppose that $Y_1, Y_2, ..., Y_n \sim$ iid $f_Y(y; \theta)$. The **joint distribution** of $Y_1, Y_2, ..., Y_n$ is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^{n} f_Y(y_i),$$

where $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ and $\boldsymbol{y} = (y_1, y_2, ..., y_n)$. If the random variables $Y_1, Y_2, ..., Y_n$ are **only independent**, then the joint distribution of $Y_1, Y_2, ..., Y_n$ is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i).$$

**Example 6.6.** Suppose that $Y_1, Y_2, ..., Y_n \sim$ iid exponential($\theta$). Then, for $y_i > 0$,

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \prod_{i=1}^{n} f_Y(y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} e^{-y_i/\theta} \\ &= \frac{1}{\theta^n} e^{-\sum_{i=1}^{n} y_i/\theta}, \end{aligned}$$

and 0, otherwise.

**Example 6.7.** Suppose that $Y_1, Y_2, ..., Y_n$ are independent random variables where $Y_i \sim$ exponential($\theta_i$); thus, the distribution changes as $i$ changes. Then, for $y_i > 0$,

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \prod_{i=1}^{n} f_{Y_i}(y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\theta_i} e^{-y_i/\theta_i} \\ &= \left( \prod_{i=1}^{n} \frac{1}{\theta_i} \right) e^{-\sum_{i=1}^{n}(y_i/\theta_i)}, \end{aligned}$$

and 0, otherwise.

## 6.2 Sampling distributions

*DEFINITION*: A **statistic**, say $T$, is a **function** of the random variables $Y_1, Y_2, ..., Y_n$. It could possibly depend on non-random constants, but it can **not** depend on unknown parameters. To emphasize the dependence of $T$ on $Y_1, Y_2, ..., Y_n$, we may sometimes write $T = T(Y_1, Y_2, ..., Y_n)$. When talking about a statistic $T$, it will often be the case that $Y_1, Y_2, ..., Y_n$ construe a **random sample** (i.e., that they are iid). In applied settings, we might view $Y_1, Y_2, ..., Y_n$ as **data** from an experiment or observational study.

**Example 6.8.** Suppose that $Y_1, Y_2, ..., Y_n \sim$ iid $f_Y(y; \theta)$. Here are examples of some statistics:

- $T(Y_1, Y_2, ..., Y_n) = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, the **sample mean**.
- $T(Y_1, Y_2, ..., Y_n) = \frac{1}{2}[Y_{(n/2)} + Y_{(n+1)/2}]$, the **sample median** (if $n$ is even).
- $T(Y_1, Y_2, ..., Y_n) = Y_{(1)}$, the **minimum order statistic**.
- $T(Y_1, Y_2, ..., Y_n) = Y_{(n)} - Y_{(1)}$, the **sample range**.
- $T(Y_1, Y_2, ..., Y_n) = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, the **sample variance**.

*AN IMPORTANT IDEA*: Since $Y_1, Y_2, ..., Y_n$ are random variables, any statistic $T = T(Y_1, Y_2, ..., Y_n)$ is also a random variable! Thus, $T$ has, among other characteristics,

- a mean,
- a variance,
- its **own probability distribution**!

*DEFINITION*: The **sampling distribution** associated with a statistic $T$ is the simply the probability distribution of the statistic. It characterizes how the statistic varies in **repeated sampling**.

**Example 6.9.** Suppose that $Y_1, Y_2, ..., Y_n \sim$ iid $\mathcal{N}(\mu, \sigma^2)$. What is the sampling distribution of the sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$?

SOLUTION. From Example 5.13 (notes), with $a_i = \frac{1}{n}$, $\mu_i = \mu$, and $\sigma_i^2 = \sigma^2$ for each $i$, it follows immediately that

$$\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n).$$

Furthermore,

$$Z = \frac{\overline{Y} - \mu}{\sqrt{\sigma^2/n}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1);$$

that is, $Z$ has a **standard normal distribution**. In applied settings, for known values of $\mu$ and $\sigma^2$, $Z$ is sometimes called the **one-sample $Z$ statistic**.

**Example 6.10.** Suppose that $Y_1, Y_2, ..., Y_9$ is a random sample from a $\mathcal{N}(100, 36)$ distribution. What is the sampling distribution of $\overline{Y} = \frac{1}{9}\sum_{i=1}^{9} Y_i$?

SOLUTION. We know from Example 6.9 that $\mu_{\overline{Y}} = \mu = 100$ and that $\sigma^2_{\overline{Y}} = \sigma^2/n = 36/9 = 4$. Hence, $\overline{Y} \sim \mathcal{N}(100, 4)$.

**Example 6.11.** In Example 6.10, what is $P(Y_1 > 106)$? $P(\overline{Y} > 106)$?

SOLUTION. $Y_1 \sim \mathcal{N}(100, 36)$. Thus, it follows that

$$P(Y_1 > 106) = P\left(\frac{Y_1 - 100}{6} > \frac{106 - 100}{6}\right) = P(Z > 1) = 0.1587.$$

Now, $\overline{Y} \sim \mathcal{N}(100, 4)$. Thus,

$$P(\overline{Y} > 106) = P\left(\frac{\overline{Y} - 100}{2} > \frac{106 - 100}{2}\right) = P(Z > 3) = 0.0013.$$

Note the very large difference between these two quantities.

**Theorem 6.1.** If $Y_1, Y_2, ..., Y_n$ are independent $\mathcal{N}(\mu_i, \sigma_i^2)$ random variables, then

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left( \frac{Y_i - \mu_i}{\sigma_i} \right)^2$$

has a $\chi_n^2$ sampling distribution.

PROOF. We already proved this in the last chapter. See Example 5.16.

*SPECIAL CASE*: If $Y_1, Y_2, ..., Y_n \sim$ iid $\mathcal{N}(\mu, \sigma^2)$ random variables, then

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left( \frac{Y_i - \mu}{\sigma} \right)^2$$

has a $\chi_n^2$ sampling distribution.

**Example 6.12.** Suppose that $Y_1, Y_2, ..., Y_9$ are iid $\mathcal{N}(100, 36)$. Then,

$$T = \sum_{i=1}^{9} \left( \frac{Y_i - 100}{6} \right)^2 \sim \chi_9^2,$$

and $P(T > 16.919) = 0.05$ (Appendix III, p. 794, WMS).

**Theorem 6.2.** If $Y_1, Y_2, ..., Y_n \sim$ iid $\mathcal{N}(\mu, \sigma^2)$ random variables, then

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left( \frac{Y_i - \overline{Y}}{\sigma} \right)^2$$

has a $\chi_{n-1}^2$ sampling distribution. In addition, $\overline{Y}$ and $S^2$ are **independent** statistics.

*REMARK*: We will not prove the independence result, in general. The text proves this for the $n = 2$ case. The statistics $\overline{Y}$ and $S^2$ are independent only if the observations $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{N}(\mu, \sigma^2)$. That is, if the probability model changes, then $\overline{Y}$ and $S^2$ are no longer independent.

PROOF OF THEOREM 6.2. We prove the first result; namely, that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Write

$$\begin{aligned}\underbrace{\sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2}_{W_1} &= \sum_{i=1}^{n}\left(\frac{Y_i-\overline{Y}+\overline{Y}-\mu}{\sigma}\right)^2 \\ &= \underbrace{\sum_{i=1}^{n}\left(\frac{Y_i-\overline{Y}}{\sigma}\right)^2}_{W_2}+\underbrace{\sum_{i=1}^{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right)^2}_{W_3},\end{aligned}$$

since the cross product

$$\sum_{i=1}^{n}\left(\frac{Y_i-\overline{Y}}{\sigma}\right)\left(\frac{\overline{Y}-\mu}{\sigma}\right)=0.$$

Now, from Theorem 6.1, we know that $W_1 \sim \chi^2_n$. Also, we can rewrite $W_3$ as

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right)^2 &= n\left(\frac{\overline{Y}-\mu}{\sigma}\right)^2 \\ &= \left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_1,\end{aligned}$$

since

$$\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

So, we have

$$\begin{aligned}W_1 &= W_2+W_3 \\ &= \frac{(n-1)S^2}{\sigma^2}+W_3,\end{aligned}$$

and, since $W_2$ is a function of $S^2$ and $W_3$ is a function of $\overline{Y}$, $W_2$ and $W_3$ are independent. Thus, the mgf of $W_1$ is given by

$$\begin{aligned}m_{W_1}(t)=E(e^{tW_1}) &= E\{e^{t[(n-1)S^2/\sigma^2+W_3]}\} \\ &= E\{e^{t[(n-1)S^2/\sigma^2]}e^{tW_3}\} \\ &= E\{e^{t[(n-1)S^2/\sigma^2]}\}E(e^{tW_3}).\end{aligned}$$

But, $m_{W_1}(t)=(1-2t)^{-n/2}$ since $W_1 \sim \chi^2_n$ and $m_{W_3}(t)=(1-2t)^{-1/2}$ since $W_3 \sim \chi^2_1$. Thus, it follows that

$$(1-2t)^{-n/2}=E\{e^{t[(n-1)S^2/\sigma^2]}\}(1-2t)^{-1/2}.$$

Hence, it must be the case that

$$E\{e^{t[(n-1)S^2/\sigma^2]}\} = E(e^{tW_2}) = m_{W_2}(t) = (1-2t)^{-(n-1)/2},$$

for values of $t < 1/2$. Thus, $W_2 \sim \chi^2_{n-1}$ by the uniqueness property of mgfs. □

**Example 6.13.** Suppose that $Y_1, Y_2, ..., Y_9 \sim$ iid $\mathcal{N}(100, 36)$. Then,

$$\frac{\sum_{i=1}^{9}(Y_i - \overline{Y})^2}{36} = \frac{8S^2}{36} \sim \chi^2_8.$$

### 6.2.1 The $t$ distribution

*THE T RANDOM VARIABLE*: Suppose that $Z \sim \mathcal{N}(0,1)$ and that $W \sim \chi^2_\nu$. If $Z$ and $W$ are **independent**, then the quantity

$$T = \frac{Z}{\sqrt{W/\nu}}$$

has a $t$ **distribution** with $\nu$ **degrees of freedom**.

*THE t PDF*: Suppose that the random variable $T$ has a $t$ distribution with $\nu$ degrees of freedom. Then, the pdf for $T$ is given by

$$f_T(t) = \begin{cases} \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\nu/2)}(1+t^2/\nu)^{-(\nu+1)/2}, & -\infty < t < \infty \\ 0, & \text{otherwise.} \end{cases}$$

*REMARK*: One could derive $t$ pdf using a bivariate transformation (from Section 5.4 in the notes). The good news is that, in practice, we will never use the formula for the $t$ pdf. Computing gives areas (probabilities) upon request; in addition, tabled values giving areas are readily available. See WMS, Table 5, page 793.

*FACTS ABOUT THE t DISTRIBUTION*:

- continuous and **symmetric** about 0.
- indexed by a parameter called the **degrees of freedom** (thus, there are infinitely many $t$ distributions!).

- in practice, $\nu$ will usually be an integer (and is often related to the sample size).
- As $\nu \to \infty$, $t_\nu \to \mathcal{N}(0,1)$; thus, when $\nu$ becomes larger, the $t_\nu$ and the $\mathcal{N}(0,1)$ distributions look more alike.
- $E(T) = 0$ and $V(T) = \frac{\nu}{\nu-2}$ for $\nu > 2$.
- When compared to the standard normal distribution, the $t$ distribution, in general, is less peaked, and has more mass in the tails. Note that $V(T) > 1$.

*THE CAUCHY DISTRIBUTION*: When $\nu = 1$, the $t$ pdf reduces to

$$f_T(t) = \begin{cases} \frac{1}{\pi(1+t^2)}, & -\infty < t < \infty \\ 0, & \text{otherwise.} \end{cases}$$

We call this a **Cauchy** pdf. An interesting fact about the Cauchy random variable is that its mean is not even finite! That is, if $Y \sim$ Cauchy, then $E(Y) = \infty$.

QUESTION: Would you ever consider using the Cauchy pdf as a model for data?

*ONE-SAMPLE $t$ STATISTIC*: Suppose $Y_1, Y_2, ..., Y_n \sim$ iid $\mathcal{N}(\mu, \sigma^2)$. Then, we know that

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

In addition, we know that $\overline{Y}$ and $S^2$ are **independent**. Thus, the quantity

$$t = \frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \sim \frac{\text{“}\mathcal{N}(0,1)\text{”}}{\sqrt{\text{“}\chi^2_{n-1}\text{”}/(n-1)}}. \tag{6.14}$$

has a $t_{n-1}$ distribution. But, simple algebra shows that the $t$ quantity in (6.14) reduces to

$$t = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

For a fixed (and known) value of $\mu$, $t$ is sometimes called the **one-sample $t$ statistic**. This quantity is used routinely in applied statistics.

**Example 6.14.** Suppose that $Y_1, Y_2, ..., Y_9 \sim$ iid $\mathcal{N}(100, 36)$. Then,

$$t = \frac{\overline{Y} - 100}{S/\sqrt{36}} \sim t_8,$$

where

$$\overline{Y} = \frac{1}{9}\sum_{i=1}^{9} Y_i \quad = \quad \text{sample mean}$$

$$S^2 = \frac{1}{9}\sum_{i=1}^{9}(Y_i - \overline{Y})^2 \quad = \quad \text{sample variance.}$$

### 6.2.2 The $F$ distribution

*THE F RANDOM VARIABLE*: Suppose that $W_1 \sim \chi^2_{\nu_1}$ and that $W_2 \sim \chi^2_{\nu_2}$. If $W_1$ and $W_2$ are **independent**, then the quantity

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has an $F$ **distribution** with $\nu_1$ (numerator) and $\nu_2$ (denominator) **degrees of freedom**.

*REMARK*: One could derive the pdf of the $F$ random variable using a bivariate transformation (from Section 5.4 in the notes); however, in practice, we will seldom need its actual form. Computing gives areas (probabilities) upon request; in addition, tabled values giving areas are readily available. See WMS, Table 7, pages 796-805.

*FACTS ABOUT THE F DISTRIBUTION*:

- continuous but **skewed right**.
- indexed by two **degree of freedom** parameters $\nu_1$ and $\nu_2$ (these are usually integers and are often related to sample sizes.)
- a popular probability distribution in applied statistics.

*AN IMPORTANT APPLICATION OF THE F DISTRIBUTION*: Suppose that we have **two independent samples**:

$$Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \text{iid } \mathcal{N}(\mu_1, \sigma_1^2)$$
$$Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \text{iid } \mathcal{N}(\mu_2, \sigma_2^2).$$

Define

$$\begin{aligned}
\overline{Y}_{1\cdot} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} &= \text{sample mean for sample 1} \\
\overline{Y}_{2\cdot} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j} &= \text{sample mean for sample 2} \\
S_1^2 = \frac{1}{n_1}\sum_{j=1}^{n_1}(Y_{1j} - \overline{Y}_{1\cdot})^2 &= \text{sample variance for sample 1} \\
S_2^2 = \frac{1}{n_2}\sum_{j=1}^{n_2}(Y_{2j} - \overline{Y}_{2\cdot})^2 &= \text{sample variance for sample 2.}
\end{aligned}$$

From Theorem 6.2, we know that

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1} \quad \text{and} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}.$$

Furthermore, as the samples are independent, $(n_1-1)S_1^2/\sigma_1^2$ and $(n_2-1)S_2^2/\sigma_2^2$ are as well. Thus, the quantity

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} \sim \frac{\text{“}\chi^2_{n_1-1}\text{”}/(n_1-1)}{\text{“}\chi^2_{n_2-1}\text{”}/(n_2-1)} \sim F_{n_1-1,n_2-1}. \tag{6.15}$$

But, the $F$ quantity in the LHS of (6.15) simplifies algebraically to

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

In addition, if the two variances $\sigma_1^2$ and $\sigma_2^2$ are equal; i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then the $F$ quantity in the LHS of (6.15) reduces to

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}.$$

*ANOTHER INTERESTING FACT*: Suppose that the random variable $T \sim t_\nu$. Then, it follows that

$$T^2 \sim F_{1,\nu}.$$

That is, the square of a random variable $T$ (where $T$ follows a $t_\nu$ distribution) has an $F_{1,\nu}$ distribution. This fact could be proven (for a general $\nu$) rigorously using a transformation argument. For the special case when $\nu = n-1$ (an integer), it is easy to show the result. We do this now.

Suppose that $Y_1, Y_2, ..., Y_n \sim$ iid $\mathcal{N}(\mu, \sigma^2)$. Recall, then, that

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Now, write

$$\begin{aligned} T^2 = \left(\frac{\overline{Y} - \mu}{S/\sqrt{n}}\right)^2 &= \left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}\right)^2 \frac{\sigma^2}{S^2} \\ &= \frac{\left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right)^2/1}{\frac{(n-1)S^2}{\sigma^2}/(n-1)} \\ &\sim \frac{\text{“}\chi^2_1\text{”}/1}{\text{“}\chi^2_{n-1}\text{”}/(n-1)} \sim F_{1,n-1}, \end{aligned}$$

since the numerator and denominator are **independent** (this follows since $\overline{Y}$ and $S^2$ are independent statistics when the underlying probability model is normal). Thus, the $t$ and $F$ distributions are connected by this relationship!

## 6.3 The Central Limit Theorem

Perhaps no result in statistics has more far-reaching implications than that of the Central Limit Theorem (CLT). Loosely speaking, the CLT states that “averages are approximately normally distributed” (there are some requirements and restrictions; we'll address these shortly).

*RECALL*: Before we state the CLT, let's recall an important result: Suppose that $Y_1, Y_2, ..., Y_n \sim$ iid $\mathcal{N}(\mu, \sigma^2)$. Then, from Example 6.9 (notes), we know that $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$. That is, the **sampling distribution** of $\overline{Y}$ is (exactly) normal with mean $\mu_{\overline{Y}} = \mu$ and variance $\sigma^2_{\overline{Y}} = \sigma^2/n$.

*REMARK*: In the previous remark, note that the underlying probability model is $\mathcal{N}(\mu, \sigma^2)$; that is, when the underlying model is normal, the sample mean $\overline{Y}$ has an exact normal distribution. The CLT states, loosely speaking, that $\overline{Y}$ has an **approximate** normal distribution, for $n$ sufficiently large, even if the underlying model is non-normal!

*MATHEMATICAL REQUIREMENTS*: For this last statement to hold, we need two requirements:

(i) $Y_1, Y_2, ..., Y_n$ construe an iid sample (i.e., a **random sample**).

(ii) $V(Y) = \sigma^2 < \infty$. Thus, the CLT does **not** hold if the underlying model is Cauchy, for example.

*THE CENTRAL LIMIT THEOREM*: Suppose that $Y_1, Y_2, ..., Y_n$ are iid random variables with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2 < \infty$. Let $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and define

$$U_n = \sqrt{n}\left(\frac{\overline{Y} - \mu}{\sigma}\right) = \left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}\right).$$

Then, as $n \to \infty$, the distribution function of $U_n$ converges to the $\mathcal{N}(0,1)$ distribution function.

*NOTATION*: The last sentence can be stated using the following notation:

$$U_n \xrightarrow{d} \mathcal{N}(0,1).$$

The symbol "$\xrightarrow{d}$" is read, "converges in distribution to."

*PUT ANOTHER WAY*: The mathematical statement that

$$U_n = \left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}\right) \xrightarrow{d} \mathcal{N}(0,1)$$

implies that $\overline{Y}$ has an **approximate** normal sampling distribution with mean $\mu$ and variance $\sigma^2/n$. Thus, it is common to write $\overline{Y} \sim \mathcal{AN}(\mu, \sigma^2/n)$.

*HOW GOOD IS THE APPROXIMATION?*: Since the CLT only offers an **approximate** sampling distribution for $\overline{Y}$, one might naturally wonder exactly how good the approximation is. In general, the goodness of the approximation **jointly** depends on

(a) the sample size, $n$, and

(b) the skewness in the underlying probability model $f_Y(y;\theta)$.

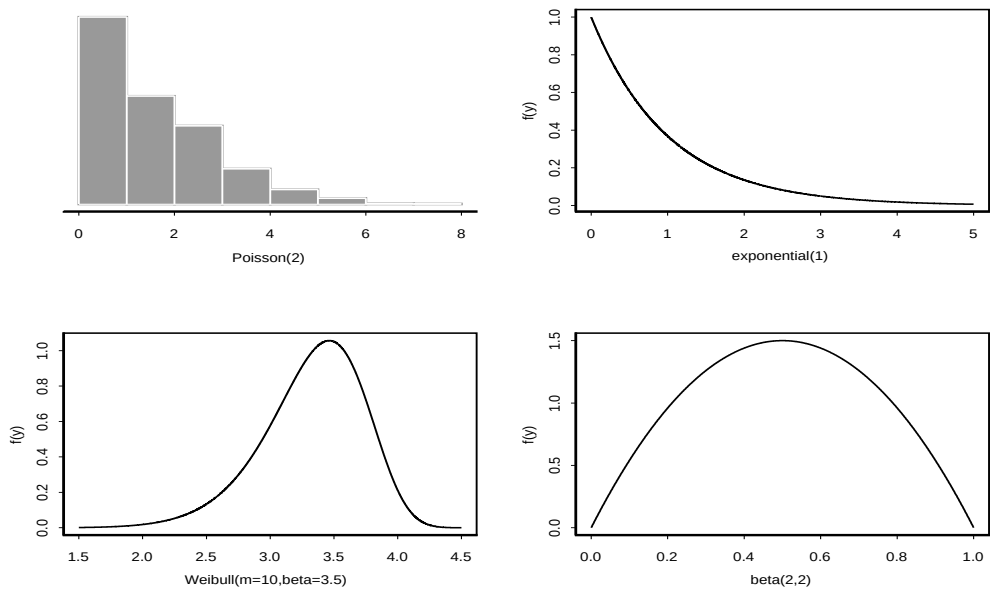


Figure 6.1: *Four different probability models.*

**Example 6.15.** In this example, we will investigate the CLT via **simulation**. For this demonstration, I have chosen four probability models (given below). In Figure 6.1, we see the probability density function for each model.

(1) $Y \sim$ **Poisson** with mean $\lambda = 2$; recall $E(Y) = \lambda = 2$ and $V(Y) = \lambda = 2$.

(2) $Y \sim$ **exponential** with mean $\beta = 1$; recall $E(Y) = \beta = 1$ and $V(Y) = \beta^2 = 1$.

(3) $Y \sim$ **Weibull** with parameters $m = 10$ and $\beta = 3.5$ (see Exercise 4.152 WMS page 206); here,

$$E(Y) = \beta\Gamma\left(1 + \tfrac{1}{m}\right) \approx 3.33 \quad \text{and} \quad V(Y) = \beta^2\left[\Gamma\left(1 + \tfrac{2}{m}\right) - \Gamma^2\left(1 + \tfrac{1}{m}\right)\right] \approx 0.16.$$

(4) $Y \sim$ **beta** with $\alpha = \beta = 2$; recall $E(Y) = \frac{\alpha}{\alpha+\beta} = 1/2$ and $V(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 1/20$.

*THINKING ABSTRACTLY*: Envision taking a random sample (i.e., an iid sample) of size $n = 2$ from each of the distributions considered and computing the sample mean of these two observations; that is,

$$\text{sample } Y_1, Y_2 \qquad \longrightarrow \qquad \text{compute } \overline{Y}.$$

**What is the sampling distribution of** $\overline{Y} = \frac{1}{2}(Y_1 + Y_2)$?

*REMARK*: In theory, one could derive this analytically by setting up a bivariate transformation, say with $U_1 = \frac{1}{2}(Y_1 + Y_2)$ and $U_2 = \frac{1}{2}(Y_1 - Y_2)$, deriving the joint distribution of the vector $(U_1, U_2)$, and then integrating over $u_2$ to obtain the marginal of $U_1$.

*FOR ILLUSTRATION*: To illustrate our results, we will rely on a **simulation study**. Here is the protocol:

- We are going to generate 10,000 iid samples, each of size $n = 2$ from each probability model of interest. I am going to use SPLUS for all simulations.
- For each of the 10,000 samples, we are going to compute $\overline{Y}$. Thus, for each of the four probability models, we will have 10,000 values of $\overline{Y}$.
- Graphing these 10,000 values, then, will provide an approximate representation of the true sampling distribution of $\overline{Y}$. Sometimes these are called **Monte Carlo distributions**.
- I will use a **kernel density estimator** to portray the approximate sampling distributions (this will just smooth out the resulting histograms with a **continuous** curve; this curve approximates the true sampling distribution of $\overline{Y}$).
- After we do this for the $n = 2$ case, we will repeat the whole process with $n = 10$, and again when $n = 30$.

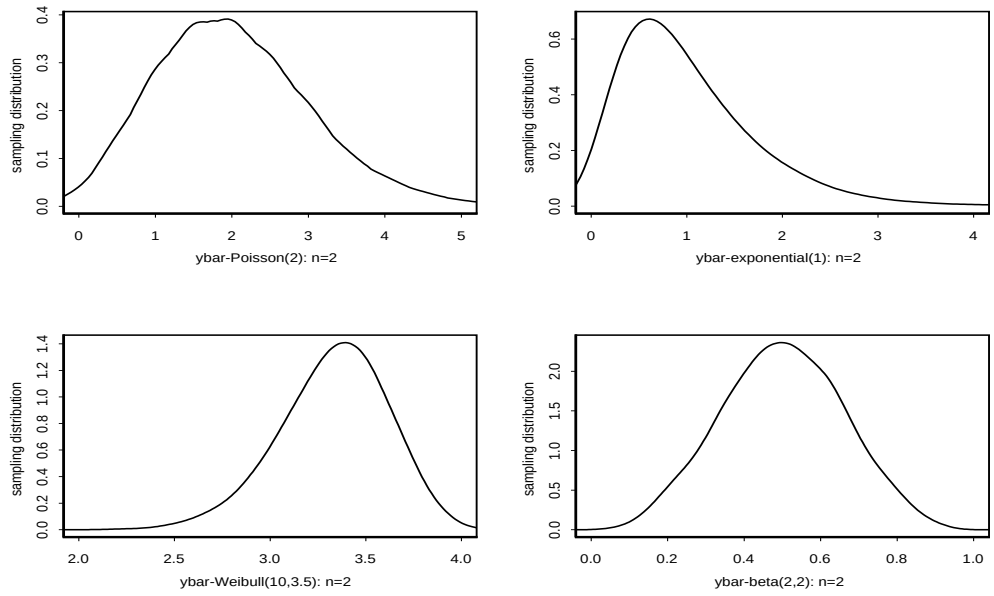


Figure 6.2: *The approximate sampling distributions for* $\overline{Y}$ *when* $n = 2$.

- Poisson: $E(\overline{Y}) = \mu = 2$, $V(\overline{Y}) = \sigma^2/n = 2/2 = 1$.
- exponential: $E(\overline{Y}) = \mu = 1$, $V(\overline{Y}) = \sigma^2/n = 1/2 = 0.5$.
- Weibull: $E(\overline{Y}) = \mu \approx 3.33$, $V(\overline{Y}) = \sigma^2/n \approx 0.16/2 = 0.08$.
- beta: $E(\overline{Y}) = \mu = 1/2$, $V(\overline{Y}) = \sigma^2/n = (1/20)/2 = 0.025$.

*OBSERVATIONS*: With only $n = 2$, we see some changes (from the graphs of the underlying probability models). Most notably, for the **exponential** case, the sampling distribution of $\overline{Y}$ looks to take a gamma-type shape (it is actually proportional to a gamma) with a high level of right skewness. In the **Poisson** case, the sampling distribution of $\overline{Y}$ looks jagged and skewed right (the jagged appearance results because of the discreteness of the Poisson distribution). For the **Weibull** case, the sampling distribution of $\overline{Y}$ looks skewed left. For the **beta** situation, the sampling distribution of $\overline{Y}$ already looks very symmetric, even when $n = 2$. In all cases, note how the variance (spread) has reduced slightly.

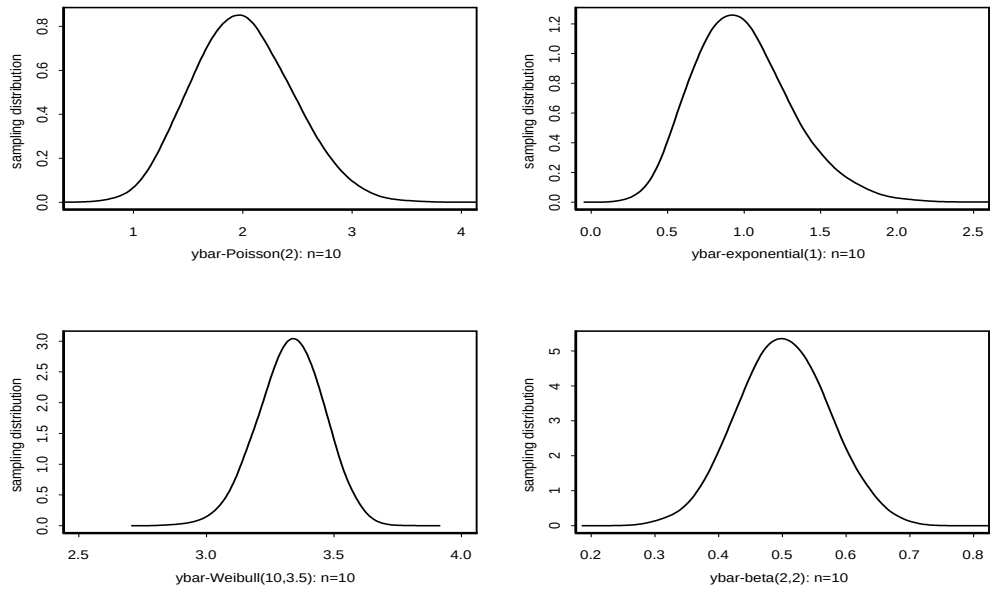


Figure 6.3: *The approximate sampling distributions for* $\overline{Y}$ *when* $n = 10$.

- Poisson: $E(\overline{Y}) = \mu = 2$, $V(\overline{Y}) = \sigma^2/n = 2/10 \approx 0.2$.
- exponential: $E(\overline{Y}) = \mu = 1$, $V(\overline{Y}) = \sigma^2/n = 1/10 = 0.1$.
- Weibull: $E(\overline{Y}) = \mu \approx 3.33$, $V(\overline{Y}) = \sigma^2/n \approx 0.16/10 = 0.016$.
- beta: $E(\overline{Y}) = \mu = 1/2$, $V(\overline{Y}) = \sigma^2/n = (1/20)/10 = 0.005$.

*OBSERVATIONS*: With $n = 10$ we start to see some real changes. In general, pretty much all four sampling distributions of $\overline{Y}$ look fairly symmetric; of course, some more than others. For example, in the **Poisson** and **exponential** cases, the sampling distribution of $\overline{Y}$ looks to be slightly skewed right (remember how the Poisson and exponential models are skewed right to begin with?). For the **Weibull** and **beta** cases, the sampling distribution of $\overline{Y}$ almost perfectly symmetric. Note that the variances of these distributions have decreased from when $n = 2$ (this makes sense since $V(\overline{Y}) = \sigma^2/n$ in general).

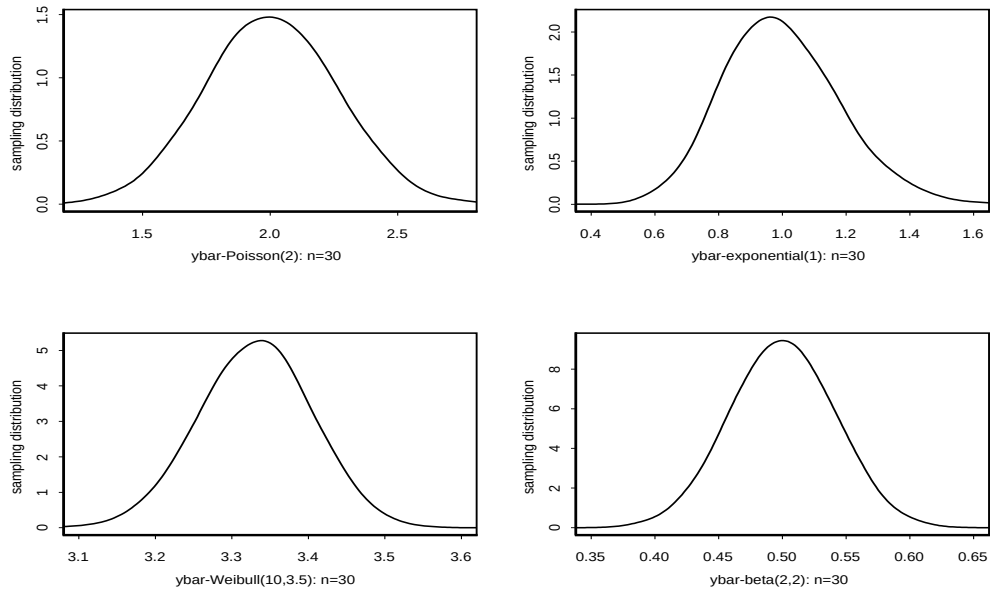


Figure 6.4: *The approximate sampling distributions for* $\overline{Y}$ *when* $n = 30$.

- Poisson: $E(\overline{Y}) = \mu = 2$, $V(\overline{Y}) = \sigma^2/n = 2/30 \approx 0.07$.
- exponential: $E(\overline{Y}) = \mu = 1$, $V(\overline{Y}) = \sigma^2/n = 1/30 \approx 0.03$.
- Weibull: $E(\overline{Y}) = \mu \approx 3.33$, $V(\overline{Y}) = \sigma^2/n \approx 0.16/30 \approx 0.005$.
- beta: $E(\overline{Y}) = \mu = 1/2$, $V(\overline{Y}) = \sigma^2/n = (1/20)/30 \approx 0.002$.

*OBSERVATIONS*: When $n = 30$, each of the four sampling distributions looks symmetric, each centered at $\mu$. One might detect a small amount of skewness in the **exponential** case. Indeed, in general, it looks like as $n$ increases, the sampling distribution of $\overline{Y}$ looks more like a normal distribution. Note again how the variance in these distributions has decreased compared to the case when $n = 10$. What would these sampling distributions look like when $n = 50$? $n = 100$? $n = 1000$? $n = 100,000$?

*PROOF OF THE CENTRAL LIMIT THEOREM*: See Section 7.4 of WMS pages 352-354. Not surprisingly, this important proof involves the moment generating function!

*EXAMPLES FROM WMS*: See Examples 7.7 and 7.8 on pages 348-349.

**Example 6.16.** In the interest of pollution control, an experimenter counts the number of bacteria per unit volume of water. Let $Y_1, Y_2, ..., Y_{36}$ denote the bacteria counts for $n = 36$ distinct water samples, and assume that the bacteria counts are iid, each following a Poisson distribution with mean $\lambda = 50$.

(1) What is the approximate sampling distribution of $\overline{Y}$? What is the mean of $\overline{Y}$? the variance of $\overline{Y}$?

(2) Approximate the probability that $\overline{Y}$ will exceed 52.

(3) Find the required value of $n$ such that $P(\overline{Y} < 49.5) \approx 0.01$ (this is a sample size-type calculation).

SOLUTIONS. (a) First, from the CLT, we have that $\overline{Y} \sim \mathcal{AN}(\mu, \sigma^2/n)$. Here, $\mu = \lambda = 50$ and $\sigma^2/n = 50/36 \approx 1.39$. Thus, $\overline{Y} \sim \mathcal{AN}(50, 1.39)$.

(b) This is a usual normal-type calculation.

$$P(\overline{Y} > 52) = P\left(\frac{\overline{Y} - 52}{2} > \frac{52 - 50}{1.39}\right) = P(Z > 1.44) = 0.0749.$$

(c) We want to find the $n$ such that

$$P(\overline{Y} < 49.5) = P\left(\frac{\overline{Y} - 50}{\sqrt{50/n}} < \frac{49.5 - 50}{\sqrt{50/n}}\right) = P\Bigg(Z < \underbrace{\frac{49.5 - 50}{\sqrt{50/n}}}_{= -2.33}\Bigg) \approx 0.01.$$

Thus, we need to solve

$$\frac{49.5 - 50}{\sqrt{50/n}} = -2.33$$

for $n$; it follows that $n \approx 1086$. Thus, an incredibly large sample size is needed so that $\overline{Y} > 49.5$ with high probability. This may not be a practical result.

EXERCISE. Redo part (c) where one only specifies $P(\overline{Y} < 49.5) \approx 0.10$. What is the required sample size? How does this compare to the case wherein $P(\overline{Y} < 49.5) \approx 0.01$?

## 6.4 The normal approximation the binomial

*RECALL*: Suppose that $Y_1, Y_2, ..., Y_n \sim$ iid Bernoulli($p$) random variables. That is $Y_i = 1$, if the $i$th trial is a "success," and $Y_i = 0$, otherwise. Recall that

$$\mu = E(Y_i) = p \quad \text{and} \quad \sigma^2 = V(Y_i) = p(1-p).$$

From Example 5.11 (notes), we know that

$$X = \sum_{i=1}^{n} Y_i$$

has a binomial distribution with parameters $n$ and $p$; that is, $X \sim$ binomial($n, p$). Define the **sample proportion** $\widehat{p}$ as

$$\widehat{p} = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

That is, $\widehat{p}$ is an **average** of iid values of 0 and 1; thus, the CLT must apply! It is easy to see that

$$\mu_{\widehat{p}} = p \quad \text{and} \quad \sigma^2_{\widehat{p}} = \frac{p(1-p)}{n}.$$

Thus, we can conclude from the CLT that when $n$ is sufficiently large,

$$\widehat{p} \sim \mathcal{AN}\left[p, \ \frac{p(1-p)}{n}\right].$$

*HOW GOOD IS THE APPROXIMATION?*: Since we are sampling from a "binary" population (almost as discrete as one can get), one might naturally wonder how well the normal distribution **approximates** the true sampling distribution of $\widehat{p}$.

*USEFUL FACT*: The approximation is **best** when

(a) $n$ is large (approximation improves as $n$ increases), and

(b) $p$ is close to 1/2.

*WARNING*: The CLT does not always provide a useful approximation for the sampling distribution of $\widehat{p}$.

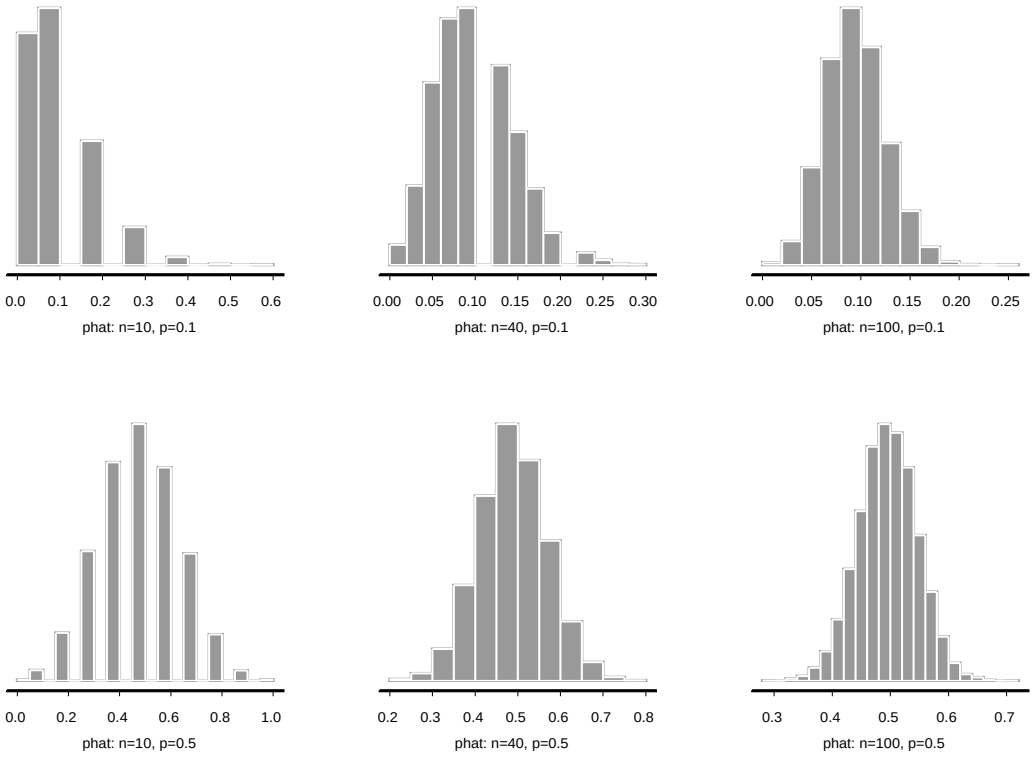


Figure 6.5: *The approximate sampling distributions for $\widehat{p}$ for different $n$ and $p$.*

*RULES OF THUMB*: One can feel comfortable using the normal approximation as long as $np$ and $n(1-p)$ are larger than 10. Other guidelines have been proposed in the literature.

**Example 6.17.** Figure 6.5 presents Monte Carlo distributions for 10,000 simulated values of $\widehat{p}$ for each of six cases: $(n, p) = (10, 0.1)$, $(40, 0.1)$, $(100, 0.1)$, $(10, 0.5)$, $(40, 0.5)$, and $(100, 0.5)$. One can clearly see that the normal approximation is not good when $p = 0.1$, except when $n$ is very large. On the other hand, when $p = 0.5$, the normal approximation is already pretty good when $n = 40$.

**Example 6.18.** 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. Currently, the worldwide seropreva-

lence rate of HCV is around 3%, and the World Health Organization has projected that HCV will be a major burden on the US health care system before the year 2020. A study was performed recently at the Blood Transfusion Service in Xuzhou City, China. The study involved a random sample of $n = 1875$ individuals and each was tested for the HCV antibody. If $p = 0.03$, what is the probability that 70 or more individuals will test positive?

SOLUTION. If $X$ denotes the number of infecteds, then $X \sim$ binomial$(1875, 0.03)$. The sample proportion $\widehat{p} = X/1875$ has mean and variance

$$\mu_{\widehat{p}} = 0.03 \quad \text{and} \quad \sigma^2_{\widehat{p}} = \frac{0.03(1-0.03)}{1875} \approx 0.00001552.$$

By the CLT, $\widehat{p} \sim \mathcal{AN}(0.03, 0.00001552)$; thus,

$$\begin{aligned} P(X \geq 70) &= P(\widehat{p} \geq 70/1875) \\ &= P\left(Z \geq \frac{70/1875 - 0.03}{\sqrt{0.00001552}}\right) \\ &= P(Z \geq 1.86) = 0.0314. \end{aligned}$$

This event $\{X \geq 70\}$ is not too likely under the assumption that $p = 0.03$. If we **did** observe an $X \geq 70$, what might this suggest about individuals living near Xuzhou City?

*RULE OF THUMB CHECK*: Here, $np = 1875 \times 0.03 = 56.25$ and $n(1-p) = 1875 \times 0.97 = 1818.75$. Thus, we can feel comfortable with the normal approximation.