Chapter 7

Probability and Statistics

Probability Theory is the mathematics for dealing with randomness. Formal probability theory first developed from a desire to analyze games of chance, then quickly spread to applications in insurance, and from there to all disciplines needing to deal with randomness, uncertainty, and even subjectivity. One application of probability theory is in **statistics**, which is concerned with methods for learning about a population using information contained in a sample from that population.

7.1 Probability Theory

7.1.1 *Probability Functions*

A random phenomenon is some activity (human or otherwise) with a range of possible outcomes, where which outcome occurs can be thought of, at some level, as due to chance. An analysis of random phenomena using probabilities begins by identifying the set of all possible outcomes (the **sample space** or **outcome space**), and then assigning probabilities to all events of interest (an **event** is any subset of the sample space). Probabilities assigned should make sense from a subject-matter point of view, and must satisfy the **probability axioms**: if Ω is the sample space, then we must have

(a) $\Pr(A) \geq 0$ for all events $A \subset \Omega$,

(b) $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if A and B are disjoint events, i.e., if $A \cap B = \emptyset$, and

(c) $\Pr(\Omega) = 1$

where "$\Pr(A)$" stands for "the probability of event A occurring".

Example 7.1 Suppose a box contains 20 red balls, 5 green balls and 75 blue balls. The activity is to randomly draw one ball from the box, and the outcome of interest is the color of the ball. The sample space is $\Omega = \{red, green, blue\}$. To model this activity, we assign probabilities to events such as "a blue ball is drawn" $A = \{blue\}$, or "either a red ball or a green ball is drawn" $B = \{red, green\}$, or "either a green ball or a blue ball is drawn" $C = \{green, blue\}$, and so on. The probabilities must satisfy the probability axioms.

For simple examples like this where there are a finite number of possible outcomes, the easiest way to make probability assignments that satisfy the probability axioms is to assign probabilities between zero and one to each of the elementary outcomes $\{red\}$, $\{green\}$ and $\{blue\}$ such that all the probabilities sum to one, and then define the probability of events

to be the sum of the probabilities of the outcomes that make up the event. For example, after assigning $\Pr(\{green\})$ and $\Pr(\{blue\})$, define $\Pr(\{green, blue\}) = \Pr(\{green\}) + \Pr(\{blue\})$.

Since the draw is random (you mix the balls up well, and draw a ball without looking into the box), it seems reasonable to assume that $\Pr(\{red\}) = 0.20$, $\Pr(\{green\}) = 0.05$ and $\Pr(\{blue\}) = 0.75$. This results in the probability assignments in Table 7.1.

Table 7.1. Probability assignments for a draw from a box with 20 red, 5 green and 75 blue balls.

Event	Probability of Event
$\emptyset$	0
$\{red\}$	0.20
$\{green\}$	0.05
$\{blue\}$	0.75
$\{red, green\}$	0.25
$\{red, blue\}$	0.95
$\{green, blue\}$	0.80
$\{red, green, blue\}$	1

The mapping from events to probabilities in Table 7.1 is called a **probability function**. It is an example of a *set function*, i.e., a function whose domain is a set of sets. For completeness, we include the empty set in our list of events, though this "empty event" will be given probability zero.

The following theorem regarding probability functions reflects behaviour that we expect of probabilities:

Theorem 7.1 *For any probability function, and for any events A and B, we have*

(a) $\Pr(A) \leq 1$,

(b) $\Pr(B \cap A^c) = \Pr(B) - \Pr(A \cap B)$,

(c) $\Pr(A^c) = 1 - \Pr(A)$ *where* $A^c = \Omega - A$ *and* Ω *is the sample space,*

(d) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$,

(e) $A \subset B \Rightarrow \Pr(A) \leq \Pr(B)$.

These are straightforward to prove, and easy to see by appealing to Venn diagrams such as Fig. 7.1 where the entire rectangle represents the sample space Ω and the sets represent events. In such a diagram, the area of a set represents the probability with which the corresponding event occurs. The area of the whole rectangle is $\Pr(\Omega) = 1$. Two events with the same probability of occurring are represented by sets of the same area, such as the sets representing events A and B in Fig. 7.1.

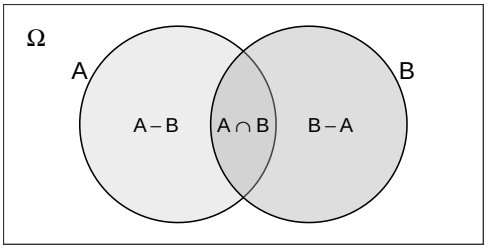

Fig. 7.1. Representing events and event probabilities with a Venn diagram.

7.1.2 *Conditional Probabilities*

Sometimes we want to update the probability of an event given new information. Suppose a ball is drawn from the box in Example 7.1. Without any information regarding the color of the ball, you would say that there is a $1/5$ probability that a ball is red. If you are told that the drawn ball is either red or green, you would update the probability of a red ball to $4/5$, since out of the 25 red and green balls in the box, 20 are red. In terms of the Venn diagram in Fig. 7.1, if we know that event B has occurred, then the sample space is "reset" to the set B, and the probability of A occurring is the ratio of the area of the set $A \cap B$ to the area of the set B. That is, the **conditional probability of A given B**, is

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{7.1}$$

for all sets A and B such that $\Pr(B) \neq 0$. Likewise, we have

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)}\,. \tag{7.2}$$

Example 7.2 A deck of playing cards has 52 cards divided equally into four suites: clubs, diamonds, hearts and spades. Each suite of 13 cards comprises an ace, nine cards labeled 2 to 10, and three picture cards Jack, Queen and King. If the deck is well-shuffled, we can assume that each card has a $1/52$ chance of getting drawn. What is the probability that two cards randomly drawn from the deck are both aces?

Let A be the event that an ace is drawn on the first draw and B be the event that an ace is drawn on the second draw. What is the probability of drawing two aces? The probability of drawing an ace on the first draw is $\Pr(A) = 4/52$. If an ace was drawn on the first draw, then there are 3 aces out of the 51 cards remaining, so the conditional probability of drawing an ace on the second draw given that an ace was drawn in the first draw is

$\Pr(B \mid A) = 3/51$. From (7.1), we can calculate the probability of A and B both occurring (the probability that we get aces on both draws) to be

$$\Pr(A \cap B) = \Pr(B \mid A) \Pr(A) = \frac{3}{51}\frac{4}{52} = \frac{1}{221}\,.$$

The events A and B in Example 7.2 are dependent. The probability of drawing an ace in the second draw depends on whether or not an ace was drawn in the first draw. Two events A and B are **independent events** if

$$\Pr(A \cap B) = \Pr(B) \Pr(A)\,. \tag{7.3}$$

If (7.3) holds, then

$$\Pr(B \mid A) = \Pr(B) \quad \text{and} \quad \Pr(A \mid B) = \Pr(A)\,.$$

Whether or not one event occurs has no bearing on the probability of the other event occurring.

Example 7.3 Consider two random tosses of a coin. Let A be the event that the first toss is heads (the event that the first toss is tails is A^c), and let B be the event that the second toss is heads. The fact that heads is obtained on the first toss should not affect the probability of getting heads on the second toss, i.e., A and B should be independent events, so $\Pr(B \mid A) = \Pr(A)$, and $\Pr(A \cap B) = \Pr(A)\Pr(B)$. If the probability of getting heads is p, then the probability of getting two heads from two tosses of the coin is p^2.

It is straightforward to show that if A and B are independent events, then A and B^c, A^c and B, and A^c and B^c are also pairwise independent, meaning that

$$\Pr(A \cap B^c) = \Pr(A)\Pr(B^c)\,, \quad \Pr(A^c \cap B) = \Pr(A^c)\Pr(B)$$
$$\text{and} \quad \Pr(A^c \cap B^c) = \Pr(A^c)\Pr(B^c)$$

all hold. For instance, we have

$$\begin{aligned}\Pr(A \cap B^c) &= \Pr(A) - \Pr(A \cap B)\\ &= \Pr(A) - \Pr(A)\Pr(B)\\ &= \Pr(A)(1 - \Pr(B)) = \Pr(A)\Pr(B^c)\,.\end{aligned}$$

Referring back to Example 7.3, the probability of heads followed by tails, and tails followed by heads, are both $p(1-p)$, and the probability of two tails is $(1-p)^2$.

Do not confuse independent events with mutually exclusive events. If A and B are mutually exclusive events, i.e., if A and B are disjoint, then the

occurrence of one of them means that the other did not occur. Mutually exclusive events are very much dependent events!

A set of three or more events are independent if the probability of the intersection of any selection of the events is equal to the product of the unconditional probabilities of the selected events. Pairwise independence is not enough. For example, three events A_1, A_2 and A_3 are independent if

$$\Pr(A_1 \cap A_2) = \Pr(A_1)\Pr(A_2)\,,\ \Pr(A_1 \cap A_3) = \Pr(A_1)\Pr(A_3)\,,$$
$$\Pr(A_2 \cap A_3) = \Pr(A_2)\Pr(A_3)$$
$$\text{and}\quad \Pr(A_1 \cap A_2 \cap A_3) = \Pr(A_1)\Pr(A_2)\Pr(A_3)\,.$$

Example 7.4 Suppose a coin is tossed three times. The sample space is

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}\,.$$

Suppose each of the eight outcomes in Ω occur with equal probability. Let

$A = \{HHH, HHT, HTH, HTT\}$ (heads on the first toss)
$B = \{HHH, HHT, THH, THT\}$ (heads on the second toss)
$C = \{HTH, HTT, THH, THT\}$ (different outcomes on first two tosses)

Each of these has probability

$$\Pr(A) = \Pr(B) = \Pr(C) = 1/2\,.$$

It is easy to verify that these three events are pairwise independent:

$$\Pr(A \cap B) = 1/4 = \Pr(A)\Pr(B)\,,$$
$$\Pr(A \cap C) = 1/4 = \Pr(A)\Pr(C) \quad \text{and} \quad \Pr(B \cap C) = 1/4 = \Pr(B)\Pr(C)\,.$$

However, we have

$$\Pr(A \cap B \cap C) = 0 \neq \Pr(A)\Pr(B)\Pr(C)\,.$$

Although A, B and C are pairwise independent, the events C and $A \cap B$ are mutually exclusive.

On the other hand, if we replace the event C with

$D = \{HHH, HTH, THH, TTH\}$ (heads on the third toss)

then the events A, B and D are independent. We have already shown $\Pr(A \cap B) = \Pr(A)\Pr(B)$. We have

$$\Pr(A \cap D) = \{HHH, HTH\} = 1/4 = \Pr(A)\Pr(D)\,,$$
$$\Pr(B \cap D) = \{HHH, THH\} = 1/4 = \Pr(B)\Pr(D)$$
$$\text{and}\quad \Pr(A \cap B \cap D) = \{HHH\} = 1/8 = \Pr(A)\Pr(B)\Pr(D)\,.$$

7.1.3 *Bayes' Theorm*

The identities

$$\Pr(A \mid B)\Pr(B) = \Pr(A \cap B) = \Pr(B \mid A)\Pr(A)$$

imply **Bayes' Theorem**:[1]

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)} \quad \text{and} \quad \Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A)}\,. \quad (7.4)$$

Bayes' Theorem is useful because it shows us how to "turn conditional probabilities around".

Example 7.5 Imagine that an infectious disease enters a population of 101000 people but that a large percentage — 100000 out of 101000 members of the population have vaccinated themselves against this disease. Of the 1000 not vaccinated, 50 percent (500 people) caught the disease. Only one percent of these 100000 vaccinated people (1000 people) eventually caught the disease, so the vaccine is effective.

Of those that caught the disease, many more were vaccinated than unvaccinated (1000 out of 1500, or about 67 percent). But this is simply because a large proportion of the population received the vaccine, and 1 percent of 100000 is more than 50 percent of 1000. The proportion of those that caught the disease who had previously received the vaccine is analogous to the *probability of having been vaccinated conditional on having caught the disease*, Pr(vaccinated | infected). But this proportion doesn't say much about the effectiveness of the vaccine, if that is what we are interested in. What we want instead is the *probability of getting infected conditional on having been vaccinated*, i.e., Pr(infected | vaccinated), which is 1000/100000.

How do we get from

$$\Pr(\text{vaccinated} \mid \text{infected}) = \frac{1000}{1500} \quad \text{to} \quad \Pr(\text{infected} \mid \text{vaccinated}) = \frac{1000}{100000}\ ?$$

Bayes' Theorem tells us that

$$\begin{aligned}\Pr(\text{infected} \mid \text{vaccinated}) &= \frac{\Pr(\text{vaccinated} \mid \text{infected})\Pr(\text{infected})}{\Pr(\text{vaccinated})} \\ &= \frac{\frac{1000}{1500} \cdot \frac{1500}{101000}}{\frac{100000}{101000}} = \frac{1000}{100000} = 0.01\,.\end{aligned}$$

[1] Thomas Bayes (1701-1761), an English Presbytarian minister, philosopher and probability theorist. The theorem appears in a work published only after his death. The theorem is the starting point of an approach to statistics called *Bayesian Statistics* which treats parameters as random variables, and probability as a measure of uncertainty regarding these variables. The approach to statistics that we present later in this chapter is known as *classical statistics*, or *frequentist statistics*, which views probability as the frequency of events over large number of trials.

Another way to state Bayes' Theorem is to use the fact that if a set of events $\{A_1, A_2, ..., A_m\}$ partitions the sample space, then

$$\Pr(B) = \Pr(B \mid A_1)\Pr(A_1) + \Pr(B \mid A_2)\Pr(A_2) + \cdots + \Pr(B \mid A_m)\Pr(A_m)\,.$$

This is the **Law of Total Probabilities**, illustrated in Fig. 7.2 for a partitioning set of events $\{A_1, A_2, \ldots, A_6\}$. Since this set of events partitions the entire sample space, it also partitions B. The total probability of event B is then the sum $\sum_{j=1}^{6} \Pr(B \cap A_j)$, which is the same as $\sum_{j=1}^{6} \Pr(B \mid A_j)\Pr(A_j)$.

Bayes' Theorem can therefore be written, for any A_i in a partitioning set of events $\{A_j\}_{i=1}^{m}$, as

$$\Pr(A_i \mid B) = \frac{\Pr(B \mid A_i)\Pr(A_i)}{\Pr(B)} = \frac{\Pr(B \mid A_i)\Pr(A_i)}{\sum_{j=1}^{m} \Pr(B \mid A_j)\Pr(A_j)}\,. \tag{7.5}$$

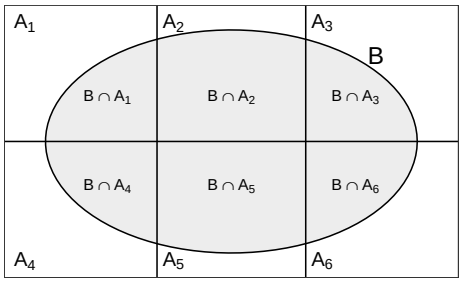

Fig. 7.2. Law of Total Probabilities.

Example 7.6 Suppose 0.05 of males (m) in a certain population are colorblind (cb), whereas only 0.005 of females (f) are colorblind. Moreover, suppose there are equal numbers of males and females in the population. Suppose a randomly drawn person from this population is colorblind. What are the chances that this person is male?

We seek $\Pr(m \mid cb)$ given $\Pr(cb \mid m) = 0.05$, $\Pr(cb \mid f) = 0.005$ and $\Pr(m) = \Pr(f) = 0.5$. From Bayes' Theorem we have

$$\Pr(m \mid cb) = \frac{\Pr(cb \mid m)\Pr(m)}{\Pr(cb \mid m)\Pr(m) + \Pr(cb \mid f)\Pr(f)}$$

$$= \frac{0.05 \cdot 0.5}{0.05 \cdot 0.5 + 0.005 \cdot 0.5} \approx 0.909\,.$$

There is an approximately 91% chance that a randomly drawn person from this population is male, given that the person drawn is colorblind.

7.1.4 *Random Variables*

If we map the possible outcomes $\{red, green, blue\}$ in Example 7.1 to numerical values, we get a **random variable**.

Example 7.7 You play a game where you ante up \$100 to draw a ball from the box of red, green and blue balls in Example 7.1. If a blue ball is drawn, the \$100 is returned to you. If a red ball is drawn, you lose the \$100. If a green ball is drawn, you win \$1000, a sum which includes the initial \$100 that you put up. Let X be your winnings (or losses) from playing this game. Then X is a random variable with possible outcomes $\{-100, 0, 1000\}$ dollars, with **probability distribution function (pdf)** $f_X(x)$ given by

$$f_X(x) = \Pr(X = x) = \begin{cases} 0.20 & \text{for } x = -100 \\ 0.75 & \text{for } x = 0 \\ 0.05 & \text{for } x = 1000\,. \end{cases}$$

The subscript X in $f_X(x)$ indicates the name of the random variable, although it is sometimes omitted.[2]

The probability of events can be obtained from the pdf, e.g.,

$$\begin{aligned} \Pr(X \geq 0) &= \Pr(X = 0 \text{ or } X = 1000) \\ &= \Pr(X = 0) + \Pr(X = 1000) = 0.80. \end{aligned}$$

The set of all possible values of a random variable is called its **range**. The following are some examples of random variables and their probability distribution functions.

Example 7.8 Suppose you have a coin which may or may not be properly balanced. If it is not balanced, then it may be more likely to obtain "heads" than "tails", or vice versa. Let p be the probability of obtaining heads. If we let X be the outcome of a single toss of the coin, where we code heads as 1 and tails as 0, then X is a random variable with range $\{0, 1\}$ and its pdf is

$$f_X(x) = \begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1\,. \end{cases}$$

[2]In formal mathematical terms, the random variable and its pdf are two separate things. A random variable is a function that assigns numerical values to the possible outcomes of an activity, which might not be numerical. In our example, the possible outcomes are red, green and blue, and the random variable is the function that assigns -100 to red, 1000 to green, and 0 to blue. The pdf is the function that gives the probabilities of numerical outcomes (and sets of numerical outcomes). Since there is nothing random about the function that maps outcomes to numerical values, you will sometimes hear the exclamation "Random variables are not random!" Really, nothing in mathematics is random, but we can nonetheless use mathematics to model randomness.

Such a random variable is called a **Bernoulli random variable**[3] with *parameter* p, and its pdf is called a **Bernoulli distribution**. We write $X \sim \text{Bernoulli}(p)$. The Bernoulli pdf is usually written more concisely as

$$f_X(x) = p^x(1-p)^{1-x}\,,\; x = 0, 1\,. \tag{7.6}$$

The Bernoulli pdf is used to model any activity with a random binary outcome, such as tossing a coin, or sampling an individual from a population and asking if they are a smoker.

Example 7.9 A die is a small cube with dots on each side corresponding to integers 1 to 6. If X is the number of dots on the upper face after a random roll of the die (and if the die is evenly weighted) then X is a random variable with pdf

$$f_X(x) = \frac{1}{6}\,,\; x = 1, 2, 3, 4, 5, 6\,.$$

It is an example of a uniformly distributed random variable.

Random variables with a finite or countably infinite number of possible outcomes are called **discrete random variables**. The random variables in the previous examples are discrete, with finite range. The following is a discrete random variable with countably infinite range.

Example 7.10 A production line produces stem bolts one at a time. Faulty stem bolts occur with probability p, independently of whether previously produced stem bolts were good or faulty. Let X be the number of good stem bolts produced before a faulty one appears. The probability that $X = 0$ (a faulty stem bolt occurs immediately) is p. The probability that $X = 1$ (one good stem bolt is obtained before a faulty one appears) is $p(1-p)$. The probability that $X = 2$ (two good stem bolts before a faulty one) is $p(1-p)^2$, and so on. Its pdf is

$$f_X(x) = \Pr(X = x) = p(1-p)^x\,,\; x = 0, 1, 2, 3, ... \tag{7.7}$$

This is the **Geometric distribution**. We say that X is a **Geometric random variable** and write $X \sim \text{Geometric}(p)$. The Geometric pdf with $p = 0.25$ is shown in Fig. 7.3(a), for x from 0 to 10.

Sometimes we work with the **cumulative distribution function (cdf)** of a random variable, defined as

$$F_X(x) = \Pr(X \leq x)\,,\; x \in \mathbb{R}\,,$$

[3]Jacob (Jacques) Bernoulli (1655-1705) made numerous fundamental contributions to probability theory and calculus. He is one of several members of the Bernoulli family to make important contributions to science, including Johann (a.k.a. Jean, John, 667-1748, brother), Nicolaus I (1687-1759, son of Nicolaus, another brother), Nicolaus II (1695-1726, son of Johann), Daniel (1700-1782, son of Johann, to whom we owe the concept of *expected utility*), Johann II (1720-1790, son of Johann), Johann III (1744-1807, son of Johann II), and Jacob II (1759-1789, son of Johann II).

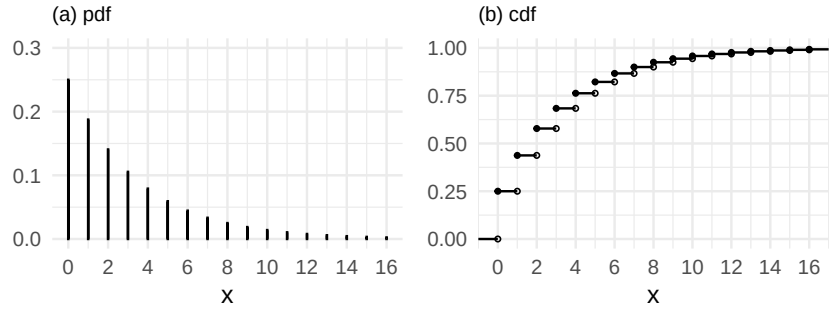

Fig. 7.3. Geometric pdf and cdf.

instead of its pdf. The cdf of a Geometric random variable is

$$F_X(x) = \begin{cases} 0 & \text{when } x < 0 \\ 1-(1-p)^{k+1} & \text{when } k \leq x < k+1\,,\ k = 0,1,2,... \end{cases} \tag{7.8}$$

which follows from the formula for the sum of a finite number of terms in a geometric progression:

$$\begin{aligned} \Pr(X \leq k) &= \sum_{i=0}^{k} p(1-p)^i \\ &= \frac{p - p(1-p)^{k+1}}{1-(1-p)} \\ &= 1-(1-p)^{k+1}\,,\ k = 0,1,2,.... \end{aligned}$$

The Geometric cdf with $p = 0.25$ is shown in Fig. 7.3(b). Notice that the pdf is defined over the range of the random variable whereas the cdf is defined over the entire real line. Notice also that the cdf of a discrete random variable has jumps.

A **continuous random variable** is one whose range is a *continuum*, such as an interval $[a, b]$, or the entire real line.[4]

Example 7.11 A random variable X has a **Uniform(a,b) distribution** if its pdf is

$$f_X(x) = \frac{1}{b-a}\,,\ x \in [a,b]$$

[4]We have omitted a few "measure theoretic" details regarding the assignment of probabilities to events contained in sample spaces that are uncountable. We will not need to worry about these details in this book.

and zero for all other values of x. We write $X \sim \text{Uniform}(a, b)$ or $X \sim U(a, b)$. The Uniform$(0, 1)$ distribution is called the **Standard Uniform distribution**, and is particularly important.

Whereas the pdf of a discrete random variable has the interpretation as $f_X(x) = \Pr(X = x)$, this interpretation must be modified for continuous random variables. For continuous random variables, the probability of obtaining an outcome between c and d is the area between the pdf and the x-axis from $x = c$ to $x = d$. That is,

$$\Pr(c \leq X \leq d) = \int_c^d f_X(u)\, du.$$

It doesn't matter whether the inequalities are strict or non-strict. For any particular value of x, we have $\Pr(X = x) = 0$. Just as the probabilities in a discrete pdf must sum to one, the pdf of a continuous random variable must integrate over the entire range to one. By its definition, the cdf of a random variable is never decreasing (it is monotone increasing). Furthermore, wherever $f_X(x)$ is continuous, we have

$$f_X(x) = \frac{d}{dx} F_X(x)\,.$$

$\Pr(c \leq X \leq d)$ is trivial to compute for Standard Uniform random variables. If $X \sim \text{Uniform}(0, 1)$, then for $0 \leq c \leq d \leq 1$, we have

$$\Pr(c < X < d) = \int_c^d 1\, du = d - c\,.$$

The Uniform$(0, 1)$ cdf is

$$F_X(x) = \int_{-\infty}^{x} f(u)\, du = \int_0^x 1\, du = x \quad \text{for all } x \in [0, 1]\,,$$

with $F_X(x) = 0$ for $x < 0$ and $F_X(x) = 1$ for $x > 1$. The pdf and cdf of X are shown in Fig. 7.4.

Digression: Random Number Generators. People have been flipping coins and rolling dice for thousands of years, sometimes for fun and sometimes to make life-altering decisions. We now use random number generators in a wide range of applications. Physical random number generators use naturally occurring randomness such as radioactive decay and even quantum systems to generate true random numbers.[5] These are used in cryptography to generate encryption/decryption

[5] See Stipčević and Koç (2014) and Herrero-Collantes and Garcia-Escartin (2017).

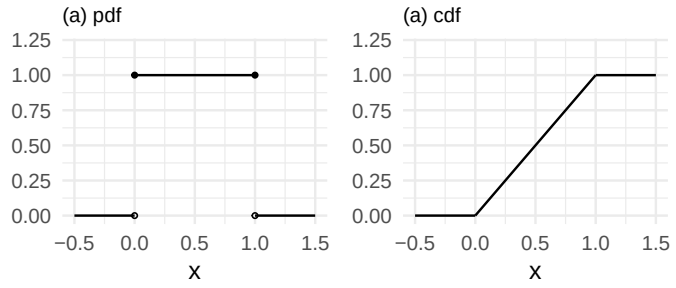

Fig. 7.4. Uniform(0,1) pdf and cdf.

keys, and also in lotteries. Algorithms called **Pseudo-Random Number Generators (PRNGs)** produce a *deterministic* sequence of values that can be replicated if the *seed* is known, but appears random otherwise. PRNGs have the advantage of being replicable, and are used in simulation experiments of complex dynamic systems, in obtaining a random sample from a population (see Section 7.2.1) and in computer-based statistical inference (see Section 7.6), among other applications. Programming languages like R and Python include PRNGs, and we will use a few of them in this chapter. For more information on random number generation, see, e.g., Johnson (2018). Most physical and algorithmic random number generators are designed to produce uniformly distributed random numbers, typically over a range of integers, or a continuum such as the interval $[0, 1]$. If required, it is possible to convert uniform random numbers into random numbers from other distributions (see Ex. 7.9).

The pdfs of random variables with countable outcomes are often called *probability mass functions*, while the pdfs of random variables with a continuum of outcomes are often called *probability density functions*. Sometimes probability mass functions are called *discrete probability density functions*. In this book, we will just use the term *probability distribution function* for both.

There is a long list of important probability distribution functions, each with its own characteristics and applications. We will study several of them later in Section 7.1.6 and Section 7.4.

7.1.5 *Expectations*

The expected value of a random variable measures the "location" of its probability distribution. Realizations of the random variable will tend to be "around" this value. The variance of a random variable measures the spread of the probability distribution around its expected value, and is an indicator of how close (or far) realizations of the random variable are likely

to be from its expected value.

Example 7.12 Suppose there are two boxes, Box A and Box B, each containing 100 balls. Each ball is labelled with a number from 1 to 8. Call a ball labelled i an "i-ball" (so we have 1-balls, 2-balls, and so on). Suppose that the numbers of each i-ball in the boxes are as shown in Table 7.2.

Table 7.2. Number of i-balls in Box A and Box B, $i = 1, 2, ..., 8$.

	1-ball	2-ball	3-ball	4-ball	5-ball	6-ball	7-ball	8-ball
Box A	5	10	15	20	20	15	10	5
Box B	25	45	25	1	1	1	1	1

Let X be the value of a ball randomly drawn from Box A and Y be the value of a ball randomly selected from Box B. It seems reasonable to model X and Y as random variables with pdfs as in Table 7.3 and visualized in Fig. 7.5.

Table 7.3. Probability distribution functions of X and Y.

i	1	2	3	4	5	6	7	8
$f_X(i) = \Pr(X = i)$	0.05	0.10	0.15	0.20	0.20	0.15	0.10	0.05
$f_Y(i) = \Pr(Y = i)$	0.25	0.45	0.25	0.01	0.01	0.01	0.01	0.01

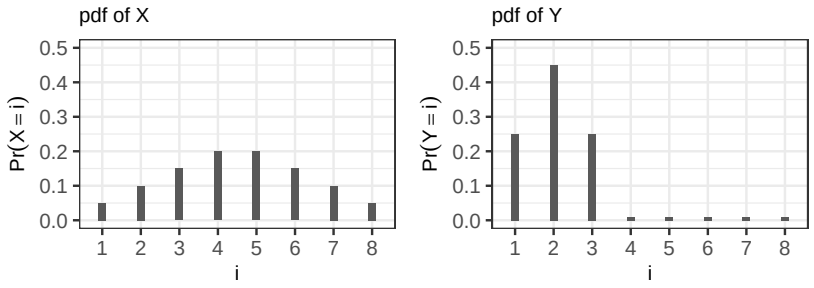

Fig. 7.5. Probability distribution functions of X and Y.

Drawing a ball from Box A will result in a value of X "around" 4 or 5, but there is a good chance of getting one of the extreme numbers. On the other hand, drawing from Box B will very likely result in $Y = 2 \pm 1$, since most of the probabilities are clustered around the values 1,2 and 3. The mean and variance of a random variable captures the idea of the "central location" and "spread" of the probabilities in a distribution.

The **mean** or **expected value** of a random variable X is defined as

$$E(X) = \begin{cases} \sum_x x f_X(x) = \sum_x x \Pr(X = x) & \text{if } X \text{ is discrete, and} \\ \int_{-\infty}^{\infty} x f_X(x)\, dx & \text{if } X \text{ is continuous.} \end{cases} \tag{7.9}$$

The symbol $\sum_x$ means "sum over the possible values of X".

The expected value of a random variable is essentially an average of its possible values, with the possible values weighted by their corresponding probabilities. For the random variable X and Y in Example 7.12, we have

$$\begin{aligned} E(X) &= 1 \cdot \Pr(X = 1) \ + \ 2 \cdot \Pr(X = 2) \ + \ \ldots \ + \ 7 \cdot \Pr(X = 7) \ + \ 8 \cdot \Pr(X = 8) \\ &= 1(0.05) + 2(0.10) + 3(0.15) + \cdots + 6(0.15) + 7(0.10) + 8(0.05) \\ &= 4.5 \end{aligned}$$

$$\begin{aligned} E(Y) &= 1 \cdot \Pr(Y = 1) \ + \ 2 \cdot \Pr(Y = 2) \ + \ \ldots \ + \ 7 \cdot \Pr(Y = 7) \ + \ 8 \cdot \Pr(Y = 8) \\ &= 1(0.25) + 2(0.45) + 3(0.25) + \cdots + 6(0.01) + 7(0.01) + 8(0.01) \\ &= 2.2 \end{aligned}$$

If you imagine that the probabilities are weights on a lever resting on a pivot, the mean is the location of the pivot such that the lever is in balance. The mean is sometimes called the **first moment** of a probability distribution.

If $X \sim \text{Bernoulli}(p)$, then

$$E(X) = 0 \cdot (1 - p) + 1 \cdot p = p\,.$$

If $X \sim \text{Uniform}(0, 1)$, then

$$E(X) = \int_0^1 x \cdot 1\, dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}\,.$$

If $X \sim \text{Geometric}(p)$, then using the fact that $\sum_{j=0}^{\infty} j r^j = \dfrac{r}{(1-r)^2}$ for $|r| < 1$, we have

$$E(X) = \sum_{x=0}^{\infty} x p (1-p)^x = \frac{p(1-p)}{(1-(1-p))^2} = \frac{1-p}{p}\,.$$

If $X \sim \text{Geometric}(p)$ is the number of non-defective products in a production line before a defective one occurs, and $p = 0.01$ is the probability of obtaining a faulty product, then the expected number of non-defective products before a faulty one is produced is $E(X) = 0.99/0.01 = 99$, which makes a lot of sense, since one in a hundred products made is defective.

If X is a random variable, then $g(X)$ is also a random variable, with expectation:

$$E(g(X)) = \begin{cases} \sum_x g(x) f_X(x) = \sum_x g(x) \Pr(X = x) & \text{if } X \text{ is discrete, and} \\ \int_{-\infty}^{\infty} g(x) f_X(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

For instance, if X is the random variable representing the value of a draw from Box A in Example 7.12, then $g(X) = X^2$ is a random variable with possible values $1, 4, 9, 16, 25, 36, 49, 64$ occurring with probabilities $0.05, 0.10, 0.15, 0.20, 0.20, 0.15, 0.10, 0.05$ respectively. The expectations of X^2 and Y^2 are:

$$\begin{aligned} E(X^2) = {} & 1^2(0.05) + 2^2(0.10) + 3^2(0.15) + 4^2(0.20) \\ & + 5^2(0.20) + 6^2(0.15) + 7^2(0.10) + 8^2(0.05) = 23.5 \end{aligned}$$

$$\begin{aligned} E(Y^2) = {} & 1^2(0.25) + 2^2(0.45) + 3^2(0.25) + 4^2(0.01) \\ & + 5^2(0.01) + 6^2(0.01) + 7^2(0.01) + 8^2(0.01) = 6.2\,. \end{aligned}$$

If $X \sim \text{Bernoulli}(p)$, we have

$$E(X^2) = 0^2(1-p) + 1^2 p = p\,.$$

If $X \sim \text{Uniform}(0, 1)$, then

$$E(X^2) = \int_0^1 x^2 \cdot 1\, dx = \left[\frac{x^3}{3}\right]_0^1 = \frac{1}{3}\,.$$

If $X \sim \text{Geometric}(p)$, then using the fact that $\sum_{j=0}^{\infty} j^2 r^j = \dfrac{r^2 + r}{(1-r)^3}$ for $|r| < 1$, we have

$$E(X^2) = \sum_{x=0}^{\infty} x^2 p(1-p)^x = \frac{p[(1-p)^2 + (1-p)}{(1-(1-p))^3} = \frac{(1-p)(2-p)}{p^2}\,.$$

It should also be clear from the definition of expectations and the properties of summation and integration that

$$E(ag(X) + bh(X)) = aE(g(X)) + bE(h(x))$$

where a and b are constants. Furthermore, the expectation of a constant is just the constant itself.

The **variance** of a random variable X is defined as its expected squared deviation from mean, i.e.,

$$\begin{aligned} Var(X) &= E\left((X - E(X))^2\right) \\ &= \begin{cases} \sum_x (x - E(X))^2 f_X(x) & \text{if } X \text{ is discrete, and} \\ \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x)\, dx & \text{if } X \text{ is continuous.} \end{cases} \end{aligned} \tag{7.10}$$

The variance is thus a measure of the spread of the probabilities about the mean. It is sometimes referred to as the **second central moment**. The square root of the variance of X is the **standard deviation** of X, and can be viewed as a measure of how far any given draw might be from the mean. Note that the unit of measurement of the standard deviation follows that of the variable itself. For instance, if X is measured in dollars, then the standard deviation is also measured in dollars, whereas the variance is measured in "squared dollars".

There is another expression for the variance that is often easier to use:

$$\begin{aligned} Var(X) &= E\left((X - E(X))^2\right) \\ &= E(X^2 - 2XE(X) + E(X)^2) = E(X^2) - E(X)^2 \,. \end{aligned} \tag{7.11}$$

For example, for the random variables X and Y in Example 7.12, we have

$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 = 23.5 - 4.5^2 = 3.25 \\ Var(Y) &= E(Y^2) - E(Y)^2 = 6.2 - 2.2^2 = 1.36 \,. \end{aligned}$$

The standard deviations are

$$sd(X) = \sqrt{3.25} = 1.803 \;\text{ and }\; sd(Y) = \sqrt{1.36} = 1.166 \,.$$

If $X \sim \text{Bernoulli}(p)$, we have

$$Var(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p) \,.$$

If $X \sim \text{Uniform}(0, 1)$, then

$$Var(X) = E(X^2) - E(X)^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12} \,.$$

If $X \sim \text{Geometric}(p)$, then

$$Var(X) = \frac{1-p}{p^2} \,.$$

Digression: Quantile-based measures of location and spread. An alternative to the mean and variance are quantile-based measures of location and spread. The **α-quantile** of the distribution of a random variable X is any value q_α such that

$$F_X(q_\alpha) = \Pr(X \leq q_\alpha) = \alpha \,, \;\; \alpha \in (0, 1) \,.$$

It is any value such that there is an α probability that a realization of X will fall on or below it. For instance, the 0.5-quantile (also known as the **median**) of the distribution of X is that value such that there is a 50-50 chance that a realization of X will fall above or below it. The 0.25-quantile (also known as the

first **quartile**) of the distribution of X is that value such that X is three times more likely to fall above it than below it. The α-quantile of a distribution is also known as the 100α-percentile of the distribution.

The median is often used as a measure of the location of a distribution. The **interquartile range**

$$IQR = q_{0.75} - q_{0.25}$$

is a popular quantile-measure of the spread of a distribution. If $X \sim \text{Uniform}(-1, 1)$, we have

$$\text{median}(X) = 0 \text{ and } IQR(X) = \frac{1}{2} - \left(-\frac{1}{2}\right) = 1\,.$$

If $X \sim \text{Uniform}(1, 3)$, then

$$\text{median}(X) = 2 \text{ and } IQR(X) = \frac{5}{2} - \frac{3}{2} = 1\,.$$

The quantile-based measures work well for continuous random variables with strictly increasing cdfs but can sometimes be awkward for discrete random variables, where the cdfs are not continuous nor strictly increasing. What are the medians and IQRs of the distributions displayed in Fig. 7.5? For the median of X in Fig. 7.5 we see that all values $m_x \in [4, 5)$ satisfy $\Pr(X \le m_x) = 0.5$. For Y in Fig. 7.5, the cdf jumps from $F_Y(m_y) = 0.25$ for all $m_y \in [1, 2)$ to $F_Y(m_y) = 0.7$ when $m_y = 2$, so there is no value m_Y such that $F_Y(m_y) = \Pr(Y \le m_y) = 0.5$.

In such cases, one convention is to take the median m_y to be the smallest value such that $F_Y(m_y) \ge 0.5$. Likewise for the other quantiles. Following this convention for the distributions of X and Y in Fig. 7.5, the median of X is 4 and the median of Y is 2. We also have $IQR(X) = 6 - 3 = 3$ and $IQR(Y) = 3 - 1 = 2$. One can show from (7.8) that if $X \sim \text{Geometric}(p)$, then median$(X)$ is the smallest integer larger than

$$\frac{\ln 0.5}{\ln(1-p)} - 1\,.$$

The next theorem regarding the mean and variance follows from definitions (7.9) and (7.10).

Theorem 7.2 *If X is a random variable and a and b are constants, then*

(a) $E(aX + b) = aE(X) + b$,

(b) $Var(aX + b) = a^2\, Var(X)$.

Proof: We show (a) for continuous random variables. We have

$$\begin{aligned} E(aX+b) &= \int_{-\infty}^{\infty} (ax+b) f_X(x)\, dx \\ &= \int_{-\infty}^{\infty} ax f_X(x)\, dx + \int_{-\infty}^{\infty} b f_X(x)\, dx \\ &= a\int_{-\infty}^{\infty} x f_X(x)\, dx + b\int_{-\infty}^{\infty} f_X(x)\, dx = aE(X) + b\,. \end{aligned}$$

For (b) we note that $E((aX)^2) = a^2E(X^2)$. Therefore

$$\begin{aligned} Var(aX+b) &= E\left((aX+b-E(aX+b))^2\right) \\ &= E\left(a^2(X-E(X))^2\right) \\ &= a^2E\left(X-E(X))^2\right) \\ &= a^2\,Var(X)\,. \end{aligned}$$

Since $Var(X) \geq 0$, the identity $Var(X) = E(X^2) - E(X)^2$ implies that

$$E(X^2) \geq E(X)^2\,.$$

This is a special case of **Jensen's Inequality**, which we state below. The proof is left as an exercise.

Theorem 7.3 *For any random variable X, we have*

$$E(g(X)) \geq g(E(X)) \quad \textit{for any convex function } \; g\,. \tag{7.12}$$

The inequality is reversed if g is concave.

Proof: See Ex. 7.10.

Part (a) of Theorem 7.2 says that (7.12) holds with equality if the transformation is linear, i.e., if $g(X) = aX + b$.

Besides the mean and variance, the skewness S and kurtosis K are two other frequently used expectations-based measures of specific characteristics of a probability distribution. They are defined as

$$S = E((X-\mu)^3)/\sigma^3 \quad \text{and} \quad K = E((X-\mu)^4)/\sigma^4$$

where μ is the mean and σ is the standard deviation of X. The skewness is a measure of asymmetry: if a distribution is symmetric about its mean, then the skewness coefficient S is zero. Since deviations from the mean, $X - \mu$, retain their sign after taking cubes, the skewness coefficient is zero if corresponding negative and positive deviations from the mean have the same probability weight when taking expectations. This would be the case if the pdf is symmetric about the mean. The kurtosis emphasizes larger deviations from the mean over small deviations from the mean (deviations from the mean less than one become very small when raised to the fourth power). It is therefore a measure of the "fatness" of the tails of the distribution. The skewness and kurtosis measures are used heavily in finance.

Sometimes we want to find the distribution of a function of a random variable, e.g., if X has a certain distribution, what is the distribution of $Y = g(X)$? This is straightforward to do for discrete random variables, with $\Pr(Y = y)$ being the sum of all probabilities $\Pr(X = x)$ where $y = g(x)$.

For instance, if X is uniformly distributed over $x \in \{-1, 0, 1\}$, then the range of $Y = X^2$ is $\{0, 1\}$ with probabilities $\Pr(Y = 0) = \Pr(X = 0) = 1/3$ and $\Pr(Y = 1) = \Pr(X = -1 \text{ or } X = 1) = 2/3$. For continuous variables, we can often use the **cdf technique**, i.e., use the fact that

$$F_Y(y) = \Pr(Y \le y) = \Pr(g(X) \le y) = \int_{x:g(x)\le y} f_X(x)\, dx\,,$$

where "$\int_{x:g(x)\le y}$" means integrate over the region of x where $g(x) \le y$. The pdf of $Y = g(X)$ can then be obtained by differentiating $F_Y(y)$.

Example 7.13 If $X \sim \text{Uniform}(-1, 1)$, what is the distribution of $Y = X^2$? The pdf of X is $f_X(x) = 1/2$ for $-1 \le x \le 1$. We have

$$\begin{aligned} F_Y(y) &= \Pr(Y \le y) = \Pr(X^2 \le y) \\ &= \Pr(-\sqrt{y} \le X \le \sqrt{y}) \\ &= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2}\, dx = \sqrt{y}\,. \end{aligned}$$

Therefore $f_Y(y) = \dfrac{d}{dy}\sqrt{y} = \dfrac{1}{2\sqrt{y}}$, $0 < y \le 1$.

7.1.6 *The Normal and Log-Normal Distributions*

A random variable X has a **Normal distribution**, denoted $X \sim \text{Normal}(\mu, \sigma^2)$ or $X \sim N(\mu, \sigma^2)$, if its pdf is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},\ x \in \mathbb{R}\,, \tag{7.13}$$

where μ is the mean of X and σ^2 is its variance. The range of a Normal random variable is the entire real line. The graph of the pdf of a Normal random variable has the familiar symmetric bell-shape, centered at μ. The Normal distribution with mean 0 and variance 1 is called the **Standard Normal distribution**, which has no parameters. The Normal distribution has a special place in probability theory for reasons that will soon become clear. The Normal distribution is also called the **Gaussian** distribution.

Fig. 7.6 shows five Normal pdfs. The three centered at zero have mean zero. The thinner of these has variance 1/4, and the flatter, broader one has variance 4. The one in bold is the Standard Normal pdf. On either side are pdfs of Normal random variables with variance 1 and means -5 (left) and 5 (right).

Substituting $\mu = 0$ and $\sigma^2 = 1$ into (7.13) gives the pdf of the Standard Normal distribution, which is given the special notation $\phi(\cdot)$:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\},\ x \in \mathbb{R}. \tag{7.14}$$

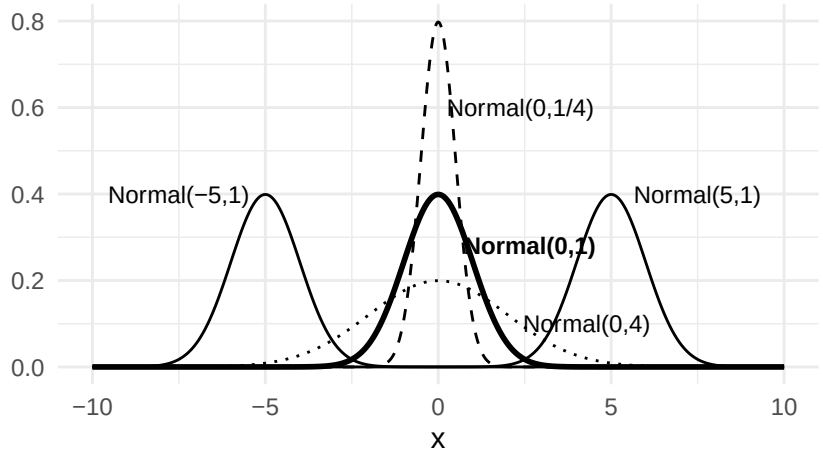

Fig. 7.6. Normal pdf, various means and variances.

The Normal distribution has the property that if $X \sim \text{Normal}(\mu, \sigma^2)$, then

$$aX + b \;\sim\; \text{Normal}(a\mu + b, a^2\sigma^2)\,. \tag{7.15}$$

The formulas for the mean and variance of $aX + b$ in (7.15) hold for all random variables with mean μ and variance σ^2; the important part in (7.15) is that the distribution itself doesn't change under the variable transformation if the distribution is Normal. An important application of this result is the fact that

$$\frac{X - \mu}{\sigma} \;\sim\; \text{Normal}(0, 1)\,.$$

Since the Normal distribution is symmetric, it should be no surprise that its skewness is zero. We will show in a later chapter that the kurtosis of the Normal distribution is equal to 3.

Example 7.14 The data set `earnings2019.csv` contains data on on almost 5000 U.S. individuals surveyed in 2019 (these individuals are part of the 2019 wave of the University of Michigan Panel Survey of Income Dynamics). The survey collected a wide variable of information from the surveyed individuals, including average hourly earnings (*earn*) in the previous year. Fig. 7.7 shows a *histogram estimate* of the distribution of $\ln earn$. The horizontal axis is divided into bins, and the frequency with which observations of $\ln earn$ falls into each bin is noted. The rectangles are then scaled so that their areas sum to one. The pdf of a Normal distribution, with mean and variance estimated from the data (see next section on how to do this) is drawn over the histogram estimate.

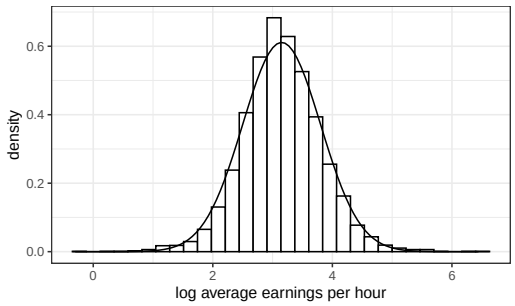

Fig. 7.7. Histogram of $\ln earn$ with Normal pdf superimposed.

It seems reasonable to model the observations of $\ln earn$ as realizations of a Normal random variable, since the distribution of the data matches the Normal pdf quite closely. We will take a closer look at this assertion in Ex. 7.15 and Ex. 7.32.

The cdf of the Normal distribution is

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(s-\mu)^2}{2\sigma^2}\right\} ds\,,\ x \in \mathbb{R}\,. \quad (7.16)$$

The Normal cdf does not have a "closed form", meaning that it cannot be expressed using a finite set of basic functions connected by arithmetic operators or powers. Nonetheless, it can be computed with high precision using numerical methods. The cdf of a Standard Normal random variable is denoted $\Phi(x)$ and is shown in Fig. 7.8.

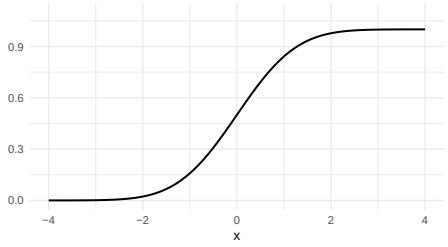

Fig. 7.8. Standard Normal cdf.

It is sometimes useful to express the pdf and cdf of a (non-Standard) Normal distribution in terms of the pdf and cdf of a Standard Normal distribution. This is easily done given that linear transformations do not change the distributional form of a Normal random variable. If $X \sim N(\mu, \sigma^2)$, then

its cdf can be written as

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

since

$$F_X(x) = \Pr(X \leq x) = \Pr\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Since the pdf is the derivative of the cdf, we have

$$f_X(x) = \frac{d}{dx}\Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).$$

The `R` function for the value of the cdf of a Normal variate with mean μ and standard deviation σ is `pnorm()`.

```
cat("pnorm(1,1,2) =", pnorm(1, mean=1, sd=2))
cat("\npnorm(2,1,2) =", pnorm(2, 1, 2))
cat("\npnorm(0,0,1) =", pnorm(0, 0, 1))
cat("\npnorm(-1.96,0,1) =", pnorm(-1.96, 0, 1))
cat("\npnorm(1.96,0,1) =", pnorm(1.96, 0, 1))
```

```
pnorm(1,1,2) = 0.5
pnorm(2,1,2) = 0.6914625
pnorm(0,0,1) = 0.5
pnorm(-1.96,0,1) = 0.0249979
pnorm(1.96,0,1) = 0.9750021
```

To get the value of q such that $\Pr(X \leq q) = 0.025$, use `qnorm()`:

```
cat("qnorm(0.025,0,1) =", qnorm(0.025, mean=0, sd=1))
```

```
qnorm(0.025,0,1) = -1.959964
```

A random variable X has the **Log-Normal distribution** with parameters μ and σ^2 if $\ln X \sim \text{Normal}(\mu, \sigma^2)$. We show in Ex. 7.8 that the Log-normal pdf with parameters μ and σ^2 is

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}, \; x \in (0, \infty). \tag{7.17}$$

It can be shown that if $X \sim \text{Log-Normal}(\mu, \sigma^2)$, then

$$E(X) = \exp\left\{\mu + \frac{\sigma^2}{2}\right\} \quad \text{and} \quad Var(X) = (\exp\{\sigma^2\} - 1)\exp\{2\mu + \sigma^2\}.$$

The Log-normal cdf does not have a closed form expression, but as with the Normal distribution, it can be computed with high precision using numerical methods. Fig. 7.9 shows the pdf and cdf of the Log-Normal distribution with $\mu = 1$ and $\sigma^2 = 1/4$.

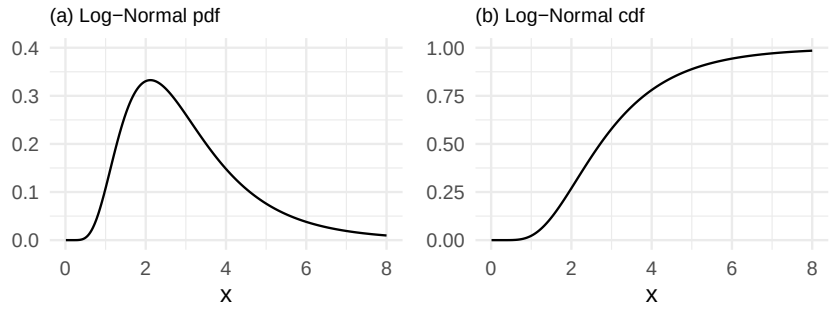

Fig. 7.9. The Log-Normal pdf and cdf with $\mu = 1$, $\sigma^2 = 1/4$.

7.1.7 *Exercises*

Ex. 7.1 Prove Theorem 7.1.

Ex. 7.2 Suppose you are told only that there are 100 balls in a container, and that 95 of the balls are either green or blue, and 90 of the balls are either red or green. *There may be balls of other colors as well.* Give a lower bound on the number of green balls in the container. Find the exact number of green balls if you are also told that there are *only* red, green and blue balls in the container? *Hint: Use part (d) of Theorem 7.1 and let* $A = \{green, blue\}$ *and* $B = \{green, red\}$.

Ex. 7.3 We proved in the text that if A and B are independent events, then A and B^c are also independent events. Show that if A and B are independent events, then A^c and B are independent, and A^c and B^c are independent.

Ex. 7.4 We can model subjective beliefs using probability distributions, and use Bayes' Theorem to update beliefs on arrival of new information. Suppose an instructor gives a student a multiple choice question with four answer options. If the student knows the answer, he chooses the correct option, otherwise he guesses from the answer options randomly.

We model the instructor's *prior belief* as a probability distribution over the space $\{knows, guesses\}$. In particular, suppose that the instructor's prior belief that the student knows the answer is $\Pr(knows) = p$, and her prior belief that the student doesn't know the answer and guesses is $\Pr(guesses) = 1 - p$. If the student answers the question correctly, how should the instructor update her beliefs regarding whether or not the student knows the answer? That is, find the probability $\Pr(knows \mid correct)$ using Bayes' Theorem in terms of p. Calculate this probability when (a) $p = 0$, (b) $p = 0.5$, and (c) $p = 1$.

Ex. 7.5 Prove that if $X \sim \text{Geometric}(p)$, then $Var(X) = \dfrac{1-p}{p^2}$.

Ex. 7.6 Prove, without invoking Jensen's Inequality, that $E(\sqrt{X}) \leq \sqrt{E(X)}$ for any random variable with range $X \geq 0$.

Ex. 7.7 Use the cdf technique to find the pdf of $Y = X^2$ if $X \sim \text{Uniform}(0, 1)$.

Ex. 7.8 If $Y = g(X)$ is increasing, then

$$F_Y(y) = \Pr(Y \leq y) = \Pr(g(X) \leq y) = \Pr(X \leq g^{-1}(y)) = F(g^{-1}(y)) .$$

If $Y = g(X)$ is decreasing, then

$$F_Y(y) = \Pr(Y \leq y) = \Pr(g(X) \leq y) = \Pr(X \geq g^{-1}(y)) = 1 - F(g^{-1}(y)) .$$

Use these results to show that for monotonic g, we have

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| .$$

Use this result to show that if $Y = \ln X \sim \text{Normal}(\mu, \sigma^2)$, then X has the log-normal pdf (7.17).

Ex. 7.9 Show that if $X \sim \text{Uniform}(0, 1)$ and F is some cdf, then $Y = F^{-1}(X)$ has pdf $f(y)$, where f is the derivative of F. Use this result to simulate 400 random numbers with distribution Normal$(1, 2)$ using the Uniform random number generator `runif()` and Normal quantile function `qnorm()`. Plot a histogram estimate of the distribution of your random numbers to verify that you have successfully simulated Normal$(1, 2)$ random numbers.

Ex. 7.10 Follow the steps below to prove **Jensen's Inequality**, which says that

$$E(g(X)) \geq g(E(X) \quad \text{for any} \quad g(\cdot) \quad \text{convex.}$$

Step 1: Let $l(x) = ax + b$ be the tangent line of the convex function $g(x)$ at the point $(E(X), g(E(X)))$, i.e., $l(x) = ax + b$ satisfies

$$l(x) = ax + b \leq g(x) \quad \text{and} \quad l(E(X)) = aE(X) + b = g(E(X)) .$$

Step 2: Prove Jensen's Inequality by using the fact that if $f_1(x) \geq f_2(x)$, then $E(f_1(X)) \geq E(f_2(X))$.

7.2 Statistics

7.2.1 *Sampling*

Statistics is about learning about a population using information contained in a sample. First we elaborate on the concepts of population and sample, and on sampling, i.e., the process of obtaining a sample from a population.

A **population** is a set of entities about which we wish to make an inference. The population might be the residents of a country, and perhaps you want to estimate the proportion of those residents who are smokers. The population could be the set of all households in a country, and perhaps you wish to estimate their total food expenditure in the previous year. The population might be the set of all "non-institutional civilians aged 16 or over" in a country, meaning all individuals in the country aged 16 and over who are not in the military, and who are not in "institutions" (jail, nursing homes and long-term care facilities). The inference of interest might be

regarding how hourly earnings are related with characteristics such as years of schooling, work experience and age.

A population might be *tangible* or *conceptual*. A tangible population refers to one that actually exists at some point in time. The examples described in the previous paragraph are tangible populations. Conceptual populations refer to the hypothetical outcomes of some activity or process. It might be the set of outcomes of an infinite number of potential tosses of a coin, the inference of interest being to test whether or not the coin is fair. The population of interest might be the set of quality measurements of all of the goods a machine can (potentially) produce, the inference of interest being whether or not the machine can meet certain quality standards. Another example of a conceptual population is all of the possible outcomes of a certain process or system that develops over time (like the aggregate levels of output, prices, interest rates, employment and other aspects of an economy). The inference might be regarding the inter-temporal relationships between these measurements.

Broadly speaking, the population of interest, tangible or conceptual, is described using a probability model where some aspect of the model is unknown, reflecting the inference to be made regarding the population. The probability model might be as simple as a pdf, or something more elaborate. The objective is to learn about these unknowns in the model/population, given a sample from the population.

The best way to sample from the population depends on the population. Even if the sampling has already been done for us, as in Example 7.14, it is still useful to understand and appreciate the principles. The aim ultimately is to get a sample that is representative of the population, in a form that allows you to make the inferences you want to make, as accurately and as precisely as possible, in the most cost-effective manner.

Consider first the tangible population case. The basic sampling design is the **simple random sample**, where each member of the population has an equal probability of being selected. How to do this depends on the particular population. For balls in a container, this can be done by mixing the balls well, and not looking when drawing out a ball. For a population of households or individuals, this is considerably harder to do. The rough idea is to assign a number to each of the N members making up the population, then using a discrete uniform random number generator to give you a set of n random integers between 1 and N, and finally surveying the households or individuals corresponding to the n random integers. From this process you obtain a random sample of the population

$$\{Y_1, Y_2, \dots, Y_n\}$$

where each Y_i is a vector of measurements (income, expenditures, …) associated with entity i. We call this a **cross-sectional sample**.

Randomness in the sampling process is key because it ensures that your sample is representative of your population. Perhaps your objective is to estimate the ratio of smokers in the population, and suppose that young males from families with lower socio-economic status are more likely to smoke. If your sampling process over- or under-samples from this category, your sample may over- or understate the percentage of smokers. Random sampling ensures that all factors that affect an individual's decision to take up smoking are balanced in the right proportions in your sample, including factors that you may not be aware of!

Getting a random sample by surveying a population is often much easier said than done, and often costly. There are issues such as how to phrase the questions and other matters of questionnaire design, how to deal with non-respondents, how to do the sampling with minimum cost, training interviewers, and so on.

There are also other sampling designs that may be useful in certain situations, including *stratified sampling*, *clustered sampling*, and *systematic sampling*. For an introduction to survey sampling methods, see Scheaffer, Mendenhall, and Ott (2006).

Populations can change over time, and it might be of interest to track these changes. If the population is revisited in a subsequent period and re-sampled, then we have a **pooled cross sectional sample**. If the population is revisited but not re-sampled, with the same individuals observed in time $t+1$ as at time t, then we get a **panel data set**.

A random sample is often, additionally, also taken to mean one that is **independently and identically distributed**. Consider drawing a sample of two balls from a box with 30 balls labelled "0" and 70 balls labelled "1". Suppose the two balls are drawn *with replacement*, meaning that you draw one ball, note its value, and return it to the box before randomly drawing a second ball. Let X_1 be the value of the first ball drawn, and X_2 be the value of the second ball drawn. Both are, of course, random variables. In this case, X_1 and X_2 will have the same distribution, i.e., they are **identically distributed**. Furthermore, the probability of drawing any particular value of X_2 is the same regardless of what value of X_1 was drawn. We say that X_1 and X_2 are **independent** random variables. We use the abbreviation *iid* for *independently and identically distributed*. We will explore the concept of independent random variables more closely later in the chapter.

Now consider drawing two balls from the box *without replacement*. In this case, X_1 and X_2 will no longer be identically distributed nor independent. The first draw X_1 will still be Bernoulli with parameter p:

$$f_{X_1}(x) = \Pr(X_1 = x) = \begin{cases} 0.7 & \text{for } x = 0 \\ 0.3 & \text{for } x = 1\,. \end{cases}$$

However, the distribution of X_2 will be different from that of X_1, and will depend on which ball was drawn in the first draw. For example, if you drew a "1-ball" in the first draw, then there are only 99 balls left in the box, with 29 1-balls, and the **conditional distribution** of X_2 will be

$$f_{X_2|X_1=1}(x) = \Pr(X_2 = x \mid X_1 = 1) = \begin{cases} 70/99 & \text{for } x = 0 \\ 29/99 & \text{for } x = 1\,. \end{cases}$$

If a 0-ball was drawn in the first draw, then the conditional distribution of X_2 will be

$$f_{X_2|X_1=0}(x) = \Pr(X_2 = x \mid X_1 = 0) = \begin{cases} 69/99 & \text{for } x = 0 \\ 30/99 & \text{for } x = 1\,. \end{cases}$$

Of course, if the population is very large, then X_1 and X_2 will be approximately iid. We can reasonably talk about an iid sample of size n from a population of size N if N is very large relative to n, even if sampling is done without replacement.

The notation $f_{X_2|X_1=0}$ indicates that the pdf is a conditional pdf. It is the pdf of X_2 under the condition that $X_1 = 0$. We will look at conditional pdfs more closely in the next section.

Obtaining an iid sample from a conceptual population depends on the population. For tosses of a coin, a random sample can be obtained by simply carrying out the coin tosses, but doing so properly. Lazily flipping the coin back and forth results in a dependent series of observations, since there is a strong tendency to alternate between heads and tails. Damaging the coin while tossing (don't ask us how) might result in the probability of heads to change across the sampling process. Willfully controlling the toss (some people can do this) will result in a non-iid non-representative sample of the population. Basically the idea is to ensure that each instance of the "experiment" is carried out under identical conditions, with previous experiments not affecting subsequent ones, and in a way that doesn't bias toward one result or another.

What about **time series** data such as the price of a stock, or the inflation, output and interest rate series for an economy? This is data of the form

$$Y_1, Y_2, \dots, Y_t, \dots$$

where again each Y_t might be a vector of different measurements, such as

$$Y_t = (cpi_t\,,\, gdp_t\,,\, unemp_t\,, \dots)\,.$$

For such data, there isn't really any sampling to be done on the part of the researcher, apart from deciding on the frequency of the data (5-minute, daily, monthly, quarterly, ...), or the sample period. For time series like inflation and interest rates, there may be the question of which price series

or interest rate series to use, since there are typically so many in any given economy. Furthermore, each Y_t is a result of decisions made by many economic entities and their interactions. For such data sets, each Y_t should be expected to depend in some way on previous outcomes, i.e., we should expect dependencies across time. In this case, the issue is not how to get iid observations, but how to deal with the dependencies in the data. Often the dependencies are themselves the object of inference.

7.2.2 *Estimation*

We present an example to illustrate statistical estimation, and demonstrate how a small sample can give you results that are accurate and precise. We make use of the following two important properties of expectations and variances. Their proofs and generalizations will be given later. If X_1 and X_2 are random variables, then

$$E(a_1X_1 + a_2X_2) = a_1E(X_1) + a_2E(X_2)\,. \tag{7.18}$$

In addition, if X_1 and X_2 are independent, then

$$Var(a_1X_1 + a_2X_2) = a_1^2\,Var(X_1) + a_2^2\,Var(X_2)\,. \tag{7.19}$$

Property (7.18) holds for all pairs of random variables, regardless of whether they are independent or identically distributed. Property (7.19) holds if X_1 and X_2 are independent, and may or may not hold if they are dependent. We will generalize Property (7.19) later. Both properties extend in the obvious way to sums of n random variables.

Example 7.15 Suppose a box contains $N = 1000$ balls each labeled with a number from 1 to 8, with 50 *1*- and *8*-balls, 100 *2*- and *7*-balls, 150 *3*- and *6*-balls, and 200 *4*- and *5*-balls. Suppose you know that there are $N = 1000$ balls in the box, but you do not know the proportions of each numbered ball. In fact, we go further and assume that *you don't even know what numbers are on the balls*. Your job is to estimate the sum total of the numbers on the $N = 1000$ balls, which is $S = 50 \cdot 1 + 100 \cdot 2 + \cdots + 100 \cdot 7 + 50 \cdot 8 = 4500$. You are allowed to sample just $n = 10$ balls, with replacement, from the container. Can you estimate the total?

Let $\{b_1, b_2, b_3, \ldots, b_N\}$ be the set of values of each of the N balls in the container and let your sample be $\{X_i\}_{i=1}^n$. Since the total value S is the average value of the N balls times N, let's try estimating S by multiplying the *sample* average by N, i.e., using the **estimator**

$$\hat{S} = \frac{N}{n}\sum_{i=1}^{n} X_i = N\overline{X}\,, \quad \text{where } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,. \tag{7.20}$$

Of course, if you wanted to estimate the population average instead of the population total, you could have used the sample mean $\overline{X}$.

We use the following code to simulate the act of drawing from the container. It uses exactly the method you would use to get a random sample from a population, with replacement. Your sample is

```
set.seed(207)
n = 10
box <- c(rep(1,50), rep(2,100), rep(3,150), rep(4, 200),
         rep(5,200), rep(6,150), rep(7,100), rep(8,50))
sample1 <- sample(box, size=n, replace=TRUE)
sample1
```

```
[1] 2 4 5 4 6 7 8 4 4 6
```

Using the estimator (7.20), your estimate of the total sum would be

```
(1000/n)*sum(sample1)
```

```
[1] 5000
```

which isn't too bad, given that your sample size was so small, and you did not even know what values were on the balls! Why does the estimator $\hat{S}$ work?

Let X be the value of a randomly selected ball from the container. We note first that the population mean of X is

$$E(X) = \sum_{i=1}^{8} i \Pr(X = i) = \frac{1}{N} \sum_{j=1}^{N} b_i = \frac{S}{N}$$

Since your sample $\{X_i\}_{i=1}^{n}$ is a random draw from the population, we know that $E(X_i) = S/N$ for each X_i, $i = 1, 2, \ldots, n$. Using the fact that the expectation of a sum is the sum of the expectations, we have

$$E(\hat{S}) = E\left(\frac{N}{n} \sum_{i=1}^{n} X_i\right) = \frac{N}{n} \sum_{i=1}^{n} E(X_i) = \frac{N}{n} \sum_{i=1}^{n} \left(\frac{S}{N}\right) = \frac{1}{n} \sum_{i=1}^{n} S = S\,.$$

We say that $\hat{S}$ is an **unbiased** estimator for S. This means that if you use this estimator, you will not systematically over- or under-estimate the total — "on average" you will get it right. To continue with this example, suppose 11 other people carried out the same exercise as you, each sampling 10 balls without replacement from the container and estimating the total in the same way as you. Suppose your collective samples are

```
set.seed(207)
R = 12
samples <- data.frame(matrix(ncol=R, nrow = n))
colnames(samples) <- paste0("S",1:R)
for (i in 1:R){samples[,i] <- sample(box, size=n, replace=TRUE)}
```

```
samples
```

```
   S1 S2 S3 S4 S5 S6 S7 S8 S9 S10 S11 S12
1   2  4  2  5  5  7  5  4  6   1   6   1
2   4  7  3  3  5  5  2  7  3   5   8   3
3   5  3  5  1  5  4  1  1  8   3   3   1
4   4  4  6  6  4  3  5  2  3   6   5   5
5   6  4  5  6  4  5  3  4  3   2   5   4
6   7  5  6  5  6  1  7  3  5   3   1   7
7   8  4  4  1  4  6  3  4  3   2   3   2
8   4  4  2  5  4  5  8  8  2   7   7   4
9   4  5  3  5  6  7  4  5  5   7   1   7
10  6  4  7  7  2  7  5  6  4   6   4   4
```

Your respective estimates of the total are then

```
hat_S <- (1000/n)*colSums(samples,dim=1)
hat_S
```

```
  S1   S2   S3   S4   S5   S6   S7   S8   S9  S10  S11  S12
5000 4400 4300 4400 4500 5000 4300 4400 4200 4200 4300 3800
```

Some underestimate and some overestimate, but the distribution of the estimates is centered around the true total.

The following is an application of the ideas in Example 7.15 to a "real-world" situation.

Example 7.16 How much did Singaporean households spend on food on average each month over 2017/18? How much did US households spend in total on their pets in 2021? How would you find out? It would be quite costly to ask every household, even in a small country like Singapore. Imagine trying to do the same for a larger country like the US. Surveying a small sample of households, if done properly, can nonetheless allow you to estimate total expenditure on pets quite accurately. The problem is entirely analogous to estimating the population total or population average of the value on the balls in Example 7.15. By randomly sampling the population of households in the US, the US Bureau of Labor Statistics estimated via its Consumer Expenditure Survey that in 2021 pet expenditure was US$102.8 billion in the US.[6] The Singapore Department of Statistics used a similar method in its Household Expenditure Survey to estimate that Singapore households spent an average of approximately S$1200 per month on food in 2017/18.[7] These surveys also estimate expenditures on various other categories of goods and services. Among other things, this information is used in the construction of the Consumer Price Index, and in policy deliberations.

[6] See Bureau of Labor Statistics (2023).
[7] See Department of Statistics (2019).

We focus now on the problem of estimating a population mean using the sample mean. The theory will apply to a wide range of problems, including the population mean of both tangible and conceptual populations. If the population is finite and tangible, and interest is in the population total, then the sample can be multiplied by the population size N to get an estimate of the population total.

We begin by modelling the population as represented by the random variable X with probability distribution function $f_X(x)$, and with "population mean" $E(X) = \mu$ and "population variance" $Var(X) = \sigma^2$. For example,

- if $X \sim \text{Bernoulli}(p)$, then $\mu = p$ and $\sigma^2 = p(1-p)$.
- if $X \sim \text{Geometric}(p)$, then $\mu = \dfrac{1-p}{p}$ and $\sigma^2 = \dfrac{1-p}{p^2}$.
- if $X \sim \text{Uniform}(a, b)$, then $\mu = \dfrac{a+b}{2}$ and $\sigma^2 = \dfrac{(b-a)^2}{12}$.

The theory we present applies to all of the above, the Normal and Log-Normal distributions, as well as the many distributions that we will discuss later in the chapter. In fact, most of the theory does not even require us to specify what $f_X(x)$ is, and only requires that X has a mean and variance.

We assume that you have a representative random sample $\{X_i\}_{i=1}^n$. This means that the sample is iid with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for all i. Then the sample mean

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} \tag{7.21}$$

is an unbiased estimator of the population mean:

$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu\,. \tag{7.22}$$

That is, the distribution of your estimator is centered about the population mean. You can interpret this to mean that by following the estimation rule $\hat{\mu} = \overline{X}$ you will not be systematically over- or under-estimating your target parameter.

How big of an error can you expect to make? We can answer this question by looking at the variance of the estimator. Remember that the estimator is a random variable, with a mean and variance. If its mean is centered at the true value of the parameter you are estimating, it is said to be **unbiased**. The variance measures the spread of the estimator distribution around the mean, so it is a measure of how "precise" the estimator is.

Assuming we have an iid sample, we have

$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}\,. \quad (7.23)$$

The square root of the estimator variance (7.23) is called the **standard error**. It can be viewed as a measure of the potential size of the estimation error.

The variance of the sample mean is in most applications unknown, because σ^2 is unknown. Nonetheless, (7.23) tells us that we can estimate the population mean more precisely if we have larger sample sizes. If we want a numerical estimate of the standard error, we need to estimate $Var(X) = \sigma^2$. How do we do that? Since

$$Var(X) = E\left((X - E(X))^2\right)\,,$$

one obvious suggestion is to use the estimator

$$\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\,. \quad (7.24)$$

Then the variance of the estimator can be estimated as

$$\widetilde{Var(\hat{\mu})} = \frac{\widetilde{\sigma^2}}{n}\,. \quad (7.25)$$

However, (7.24) turns out to be a biased estimator for σ^2. We can show this using the fact that

$$E(X_i^2) = Var(X_i) + E(X_i)^2 = \sigma^2 + \mu^2$$

$$\text{and} \quad E(\overline{X}^2) = Var(\overline{X}) + E(\overline{X})^2 = \sigma^2/n + \mu^2\,.$$

We have

$$E(\widetilde{\sigma^2}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2) = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2.$$

The estimator (7.24) therefore systematically under-estimates σ^2. It follows that (7.25) systematically underestimates the variance of the sample mean. If your sample size n is large, the bias may be negligible for all intents and purposes in which case there shouldn't be any problem using (7.24). Nonetheless, it is easy to derive an unbiased estimator for σ^2, namely

$$\widehat{\sigma^2} = \frac{n}{n-1}\widetilde{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \quad (7.26)$$

The expressions in (7.24) and (7.26) are both called the **sample variance** of the observations. To distinguish between the two, the former can be referred to as the "uncorrected" sample variance, whereas the latter can be called the "corrected" or "unbiased" sample variance. The intuition for why the divisor in (7.26) has to be $n-1$ instead of n is that the deviations from the sample mean always sum to zero. This means that there are only $n-1$ "free" deviations from the sample mean. For example, given $\sum_{i=1}^{n}(X_i - \overline{X})$ and the first $n-1$ deviations $(X_i - \overline{X})$, $i = 1, 2, ..., n-1$, you can determine the nth deviation as $(X_i - \overline{X}) = -\sum_{i=1}^{n-1}(X_i - \overline{X})$. One "degree of freedom" was lost because we had to use the observations to compute the sample mean in order to compute the deviations from the sample mean.

In summary, if X has mean $E(X) = \mu$ and variance $Var(X) = \sigma^2$ and you have a representative iid sample $\{X_i\}_{i=1}^{n}$ from the population represented by X, then the sample mean $\hat{\mu} = \overline{X}$ is an unbiased estimator for μ. The variance of this estimator is $Var(\hat{\mu}) = \sigma^2/n$, where σ^2 can be estimated using (7.26). If the sample size is large, often (7.24) is also used.

Example 7.17 Coins can be weighted so that one side shows more frequently than the other in tosses of the coin. Let $X = 1$ if heads shows, and $X = 0$ if tails shows and let p be the probability of obtaining heads. Then $X \sim$ Bernoulli(p). Suppose you toss the coin n times and record the outcomes, giving you an iid sample $\{X_i\}_{i=1}^{n}$ such that $E(X_i) = p$ and $Var(X_i) = p(1-p)$ for all i. Then p can be estimated using the sample mean $\hat{p} = \overline{X}$ which is equal to the proportion of 1s in the sample. The variance of the estimator is

$$Var(\hat{p}) = \frac{p(1-p)}{n}$$

which can be estimated by

$$\widehat{Var(\hat{p})} = \frac{\widehat{\sigma^2}}{n} \quad \text{where } \widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 .$$

Take the square root of $\widehat{Var(\hat{p})}$ to get the sample standard error of the estimator.

Example 7.17 is an example of estimating the population mean of an infinite conceptual population. It is nonetheless mathematically identical to the next example, where we are estimating the population mean of a finite tangible population.

Example 7.18 Suppose you are interested in estimating the proportion p of smokers in a large population of size N. If X is a random draw from this population (with "smoker" = 1, "non-smoker"=0), then X is

Bernoulli(p). Suppose you randomly sample n people and ask if they are smokers. Your sample is small relative to the population, and you did the appropriate randomization in selecting your sample, so that you can consider your sample $\{X_i\}_{i=1}^n$ to be a representative iid draw from the population.

As in the coin toss example, $E(X) = p$ so an unbiased estimator for p is the sample mean, which is the proportion of smokers in your sample. For example, if you sample $n = 20$ people, and 14 are smokers, then your estimate of p is $\widehat{p} = 0.7$, i.e., you estimate that 70% of the population are smokers. Of course, this is just an estimate. If you did the same exercise again, you might get a sample in which there are 12 smokers out of 20, in which case your estimate of p would be $\widehat{p} = 0.6$, i.e., you would say that an estimated 60% of the population are smokers. If you drew a third sample, you might get 15 smokers out of 20 to get $\widehat{p} = 0.75$. You do not know the true proportion of smokers (otherwise you would not need to be estimating it) but the unbiasedness of your estimator tells you that on average you will not over- or under-estimate the proportion p. You can also estimate the standard error of the estimator as in the previous example.

Observe that the maximum value of the variance $Var(\widehat{p}) = p(1-p)/n$ occurs at $p = 0.5$. That is, for fixed n, the largest value of the standard error is

$$\sqrt{\frac{0.5(1-0.5)}{n}} = \sqrt{\frac{0.25}{n}}\,.$$

How many samples observations would you need to ensure that the standard error is less than 0.01? We have

$$\sqrt{\frac{0.25}{n}} < 0.01 \Rightarrow \frac{0.25}{n} < 0.0001 \Rightarrow n > 2500\,.$$

7.2.3 *Hypothesis Testing*

To test if the population mean is equal to some specific value μ_0, we check if the sample mean is "improbably far" from μ_0 when $\mu = \mu_0$ is assumed to be true. If it is, we construe this as evidence that the "null hypothesis" $H_0 : E(X) = \mu_0$ is false, and reject it in favor of the alternative $H_1 : E(X) \neq \mu_0$. But how far is "improbably far"? To provide an answer to this question, we need to derive the distribution of the sample mean when $\mu = \mu_0$, and to do so we need to know the distribution of X. If all you know is that $E(X) = \mu$ and $Var(X) = \sigma^2$, then you do not have enough information to derive the distribution of the sample mean. We will explain how to find an approximation to the finite sample distribution of $\overline{X}$ in this situation later in the chapter.

In the case of the coin toss example, the structure of the problem does provide us with enough information to derive the finite sample distribution of the sample mean. Let Y be the number of heads out of n independent

tosses $\{X_i\}_{i=1}^n$ of a coin with probability of heads p, i.e., $Y = \sum_{i=1}^n X_i$, where $X_i \sim$ Bernoulli(p). Then the probability of obtaining k heads is

$$\Pr(Y = k) = \binom{n}{k} p^i (1-p)^{n-k}, \ k = 0, 1, 2, ..., n \tag{7.27}$$

since there are $\binom{n}{k}$ ways that k heads can appear in a sequence of n coin tosses. We say that Y has a **Binomial distribution** with parameters n and p and call Y a **Binomial random variable**. We write $Y \sim$ Binomial(n, p). Then the sample mean has possible values k/n, $k = 0, 1, ..., n$, with corresponding probability

$$\Pr\left(\overline{X} = \frac{k}{n}\right) = \binom{n}{k} p^i (1-p)^{n-k}, \ k = 0, 1, 2, ..., n. \tag{7.28}$$

We can use (7.28) to help us decide whether or not to reject the hypothesis that the coin is fair.[8]

Example 7.19 Suppose we have a sample of 20 coin tosses, and suppose that the coin is in fact fair, i.e., that p is indeed equal to 0.5. The following is the pdf of the sample mean $f(i/n) = \Pr(\overline{X} = i/n), i = 0, 1, ..., n$ calculated using (7.28) with $p = 0.5$ and displayed in Fig. 7.10.

```
p <- 0.5
n <- 20
i <- 0:n       # i integers from 0 to 20
phat <- 0:n/n  # possible values of sample means
Pr_phat <- choose(n,i)*p^i*(1-p)^(n-i)
dim(Pr_phat) <- c(1,n+1) # make into row vector for presentation
colnames(Pr_phat) = paste0("p_hat=",i/n)
rownames(Pr_phat) = "Prob"
noquote(format(Pr_phat, scientific=T,digits=6)) # another way to print
```

```
     p_hat=0     p_hat=0.05  p_hat=0.1   p_hat=0.15  p_hat=0.2   p_hat=0.25
Prob 9.53674e-07 1.90735e-05 1.81198e-04 1.08719e-03 4.62055e-03 1.47858e-02
     p_hat=0.3   p_hat=0.35  p_hat=0.4   p_hat=0.45  p_hat=0.5   p_hat=0.55
Prob 3.69644e-02 7.39288e-02 1.20134e-01 1.60179e-01 1.76197e-01 1.60179e-01
     p_hat=0.6   p_hat=0.65  p_hat=0.7   p_hat=0.75  p_hat=0.8   p_hat=0.85
Prob 1.20134e-01 7.39288e-02 3.69644e-02 1.47858e-02 4.62055e-03 1.08719e-03
     p_hat=0.9   p_hat=0.95  p_hat=1
Prob 1.81198e-04 1.90735e-05 9.53674e-07
```

Notice that there are non-zero probabilities on every possible outcome of the sample mean. This means that any reasonable decision rule that we use to reject, or not reject, the null hypothesis will have a non-zero probability of rejecting H_0 even when H_0 is true (we call this a "Type I

[8]Since $E(\overline{X}_n) = p$ and $Var(\overline{X}_n) = p(1-p)/n$ and $Y = n\overline{X}$, it follows that if $Y \sim$ Binomial(n, p), then $E(Y) = np$ and $Var(Y) = n^2 p(1-p)/n = np(1-p)$. Incidentally, the Bernoulli(p) distribution is a special case of the Binomial(n,p) distribution with $n = 1$.

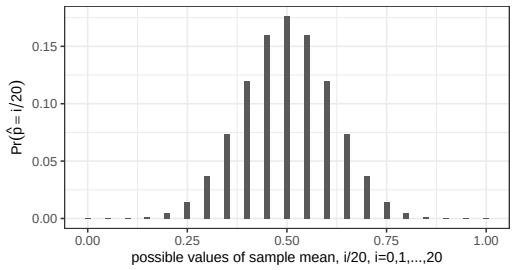

Fig. 7.10. Distribution of sample mean, n=20, p=0.5.

error"). For example, suppose we use the rule "Reject $H_0 : p = 0.5$ in favor of the alternative $H_1 : p \neq 0.5$ if the frequency of heads $\hat{p}$ is less than 0.3 or greater than 0.7", which seems not unreasonable. We can calculate from the table above that by using this rule, there is a probability of approximately

```
round(sum(Pr_phat[i/n<0.3])+sum(Pr_phat[i/n>0.7]),4)
```

```
[1] 0.0414
```

that we reject the null even though p is in fact equal to 0.5. We can reduce the probability of Type I error by allowing for a larger range for $\hat{p}$ (perhaps reject if $\hat{p} < 0.05$ or $\hat{p} > 0.95$), but then the test loses power to reject a false hypothesis (i.e., the probability of failing to reject a wrong hypothesis — a "Type II error" — increases). In practice, researchers usually opt for decision rules such that the probability of an incorrect rejection of a true null hypothesis (a value also known as the **level of significance** of the test, is around 0.01, or 0.05, or 0.10.

7.2.4 *Exercises*

Ex. 7.11 Suppose you wish to estimate the population mean, and you use the following "silly" estimator $\tilde{\mu} = X_1$ regardless of sample size. That is, you pick the first sampled observation as your estimator of the population mean, and discard the rest. Show that this is an unbiased estimator.

Ex. 7.12 We showed in (7.23) that the variance of the sample mean is $Var(\overline{X}) = \sigma^2/n$. What happens to $Var(\overline{X})$ when $n \to \infty$, i.e., as you use larger and larger sample sizes? How would you interpret this result?

Ex. 7.13 The variance of $\hat{p}$ in Example 7.17 is

$$Var(\hat{p}) = \frac{p(1-p)}{n}$$

which can be estimated by

$$\widehat{Var(\hat{p})} = \frac{\widehat{\sigma^2}}{n} \quad \text{where } \widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \,.$$

Another possibility is to use

$$\widetilde{Var(\overline{X})} = \frac{\hat{p}(1-\hat{p})}{n} \,.$$

Show that this is equivalent to using

$$\widetilde{Var(\hat{p})} = \frac{\widetilde{\sigma^2}}{n} \quad \text{where } \widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \,.$$

Show that $\widehat{Var(\overline{X})} = \dfrac{\hat{p}(1-\hat{p})}{n-1}$.

Ex. 7.14 The sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$$

is an unweighted average of the sample observations. Consider now a weighted average

$$\widetilde{X} = \sum_{i=1}^{n} w_i X_i = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n \,.$$

Show that $\widetilde{X}$ is an unbiased estimator for the population mean if $\sum_{i=1}^{n} w_i = 1$. Continuing with the assumption that $\sum_{i=1}^{n} w_i = 1$, show that

$$Var(\widetilde{X}) \geq Var(\overline{X}) \,.$$

Ex. 7.15 Use the data in `earnings2019.csv` to answer the following questions.

(a) Estimate $E(\ln earn) = \mu$ and $E(earn)$ by taking the sample means of $\ln earn_i$ and $earn_i$ respectively. Estimate also the variances of the two sample means.

(b) Find the natural exponent of the sample mean of $\ln earn_i$, and compare this value with the sample mean of $earn_i$. What might explain the discrepancy?

(c) If we assume that $\ln earn \sim \text{Normal}(\mu, \sigma^2)$, then $earn \sim \text{Log-Normal}(\mu, \sigma^2)$, with expected value $E(earn) = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}$. Use this to suggest an alternative way to estimate $E(earn)$ from the sample mean and sample variance of $\ln earn_i$.

(d) Given an iid sample $\{X_i\}_{i=1}^n$ of a random variable X, we can use the **sample skewness** and **sample kurtosis**, defined as

$$\widehat{S} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{3/2}} \quad \text{and} \quad \widehat{K} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{2}}$$

to estimate the population skewness and kurtosis of X. Compute the sample skewness and kurtosis of $\ln(earn_i)$. Is it appropriate to assume that $\ln(earn) \sim \text{Normal}(\mu, \sigma^2)$?

Ex. 7.16 A measure of the quality of an estimator $\hat{\theta}$ for a parameter θ is the **mean squared estimation error**

$$MSE(\hat{\theta}) = E((\theta - \hat{\theta})^2)\,.$$

Show that $MSE(\hat{\theta}) = Bias(\hat{\theta})^2 + Var(\hat{\theta})$ where $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Ex. 7.17 It can be shown that if Y_i, $i = 1, 2, ..., n$ are iid draws from a Normal distribution with mean μ and variance σ^2, then the variance of the unbiased variance estimator $\widehat{\sigma^2}$ defined in (7.26) is

$$Var(\widehat{\sigma^2}) = \frac{2\sigma^4}{n-1}\,.$$

Because $\widehat{\sigma^2}$ is an unbiased estimator, its MSE is also $\dfrac{2\sigma^4}{n-1}$.

(a) Show that the biased estimator $\widetilde{\sigma^2}$ defined in (7.24) has a smaller variance than $\widehat{\sigma^2}$.

(b) Show that

$$MSE(\widetilde{\sigma^2}) = \frac{2n-1}{n^2}\sigma^4\,.$$

(c) Show that $MSE(\widetilde{\sigma^2}) < MSE(\widehat{\sigma^2})$.

This is an example where the MSE of an estimator can be improved by trading off some bias for a reduced variance. Note that the arguments here have assumed that Y is normally distributed.

7.3 Joint and Conditional Probabilities

We model the joint behavior of two random variable using a **joint probability distribution function**. We will use a simple example with two discrete random variables to illustrate the main ideas.

7.3.1 *Joint and Marginal Distributions*

Suppose X and Y are discrete random variables with ranges $x = 1, 2, 3, 4, 5$ and $y = 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0$. Their joint pdf $f_{X,Y}(x, y)$ gives you the probability of events of the form $X = x$ *and* $Y = y$, i.e.,

$$f_{X,Y}(x, y) = \Pr(X = x, Y = y)\,.$$

Suppose the joint pdf of X and Y is as given below.

	6	0	0	0	0	$\frac{1}{20}$
	5.5	0	0	0	$\frac{1}{20}$	$\frac{2}{20}$
	5	0	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$
y	4.5	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0
	4	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0	0
	3.5	$\frac{2}{20}$	$\frac{1}{20}$	0	0	0
	3	$\frac{1}{20}$	0	0	0	0
		1	2	3	4	5
				x		

(7.29)

So we have

$$\Pr(X = 1, Y = 3) = f_{X,Y}(1, 3) = \frac{1}{20},$$
$$\Pr(X = 3, Y = 3) = 0,$$
$$\Pr(X \geq 4, Y \geq 5) = \frac{7}{20}$$

and so on.

What is the probability of observing $X = 1$ (regardless of the value of the accompanying Y value)? To find $\Pr(X = 1)$, add up all the probabilities of events where $X = 1$, i.e.,

$$\begin{aligned}\Pr(X = 1) &= \Pr(Y = 3, X = 1) + \Pr(Y = 3.5, X = 1) + \cdots + \Pr(Y = 6, X = 1) \\ &= \frac{1}{20} + \frac{2}{20} + \frac{1}{20} \\ &= 0.2\end{aligned}$$

This is just an application of the law of total probabilities. You can repeat this calculation for $\Pr(X = 2)$, $\Pr(X = 3)$, $\Pr(X = 4)$, $\Pr(X = 5)$. You should find that:

x	1	2	3	4	5
$\Pr(X = x)$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$

This is the "**marginal**" (or "**unconditional**") pdf of X. It turns out, in this example, that X is uniformly distributed over the values $X = 1, 2, ..., 5$. Similar calculations will give you the marginal distribution of Y.

	y	x: 1	2	3	4	5
	6	0	0	0	0	$\frac{1}{20}$
	5.5	0	0	0	$\frac{1}{20}$	$\frac{2}{20}$
	5	0	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$
	4.5	0	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0
	4	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	0	0
	3.5	$\frac{2}{20}$	$\frac{1}{20}$	0	0	0
	3	$\frac{1}{20}$	0	0	0	0

$\rightarrow$

y	$\Pr(Y=y)$
6	$\frac{1}{20}$
5.5	$\frac{3}{20}$
5	$\frac{4}{20}$
4.5	$\frac{4}{20}$
4	$\frac{4}{20}$
3.5	$\frac{3}{20}$
3	$\frac{1}{20}$

$E(Y) = 4.5$
$Var(Y) = 0.625$

x	1	2	3	4	5
$\Pr(X=x)$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$

$\rightarrow$ $E(X) = 3$, $Var(X) = 2$

The marginal distribution of Y in our example is somewhat "bell-shaped". You can calculate the (unconditional) means and variances of X and Y from their marginal pdfs using the usual formulas.

For discrete random variables, the marginal pdf of X is computed as $f_X(x) = \sum_y f_{X,Y}(x,y)$ where $\sum_y$ indicates summation over the possible values of Y. Likewise, the marginal pdf of Y is computed as $f_Y(y) = \sum_x f_{X,Y}(x,y)$. For continuous random variables, the marginals are computed as integrals: $f_X(x) = \int_y f_{X,Y}(x,y)\,dy$ and $f_Y(y) = \int_x f_{X,Y}(x,y)\,dx$. We can extend the joint pdf concept to more than two variables, e.g., $f_{X,Y,Z}(x,y,z)$, and so on.

7.3.2 *Covariance and Correlation*

It seems clear that in (7.29) there is a positive relationship between X and Y. One way to describe the relationship between the random variables is to calculate the **covariance** between X and Y, defined as

$$\sigma_{XY} = Cov(X,Y) = E\left((X - E(X))(Y - E(Y))\right).$$

In our example, we have

$$\begin{aligned}
Cov(X,Y) = {} & (5-3)(6.0-4.5)\tfrac{1}{20} + \\
& (4-3)(5.5-4.5)\tfrac{1}{20} + (5-3)(5.5-4.5)\tfrac{2}{20} + \\
& (3-3)(5.0-4.5)\tfrac{1}{20} + (4-3)(5.0-4.5)\tfrac{2}{20} + (5-3)(5.0-4.5)\tfrac{1}{20} + \\
& (2-3)(4.5-4.5)\tfrac{1}{20} + (3-3)(4.5-4.5)\tfrac{2}{20} + (4-3)(4.5-4.5)\tfrac{1}{20} + \\
& (1-3)(4.0-4.5)\tfrac{1}{20} + (2-3)(4.0-4.5)\tfrac{2}{20} + (3-3)(4.0-4.5)\tfrac{1}{20} + \\
& (1-3)(3.5-4.5)\tfrac{2}{20} + (2-3)(3.5-4.5)\tfrac{1}{20} + \\
& (1-3)(3.0-4.5)\tfrac{1}{20} \\
= {} & 1
\end{aligned}$$

One problem with the covariance measure is that it is not invariant to scale. For instance, suppose X is currently measured in thousands of dollars.

If we re-scale to dollars by multiplying X by 1000, then the covariance becomes

$$\begin{aligned} Cov(1000X, Y) &= E((1000X - E(1000X))(Y - E(Y))) \\ &= 1000E((X - E(X))(Y - E(Y))) \\ &= 1000\,Cov(X, Y). \end{aligned}$$

For this reason, the **correlation**

$$\rho_{XY} = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}},$$

which is invariant to scale and always lies between -1 and 1, is more informative. If $\rho_{XY} = 0$, then X and Y are said to be **uncorrelated**.

Given a sample $\{X_i, Y_i\}_{i=1}^n$ from a joint pdf $f_{X,Y}(x, y)$, we can estimate the covariance using the **sample covariance**

$$\hat{\sigma}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

where $\overline{X}$ and $\overline{Y}$ are the sample means of X_i and Y_i respectively. To estimate the correlation, we can divide the sample covariance by the sample standard deviations to get the **sample correlation**

$$\hat{\rho}_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}\ .$$

The following properties of means, variances and covariances are easy to show: if a and b are constants, we have

(a) $E(aX + bY) = aE(X) + bE(Y)$,

(b) $Var(aX + bY) = a^2\,Var(X) + b^2\,Var(Y) + 2ab\,Cov(X, Y)$,

(c) $Cov(X, Y) = E(XY) - E(X)E(Y)$,

(d) $Cov(X, X) = Var(X)$.

From (b), we see that the variance of a sum is the sum of the variances only if the variables are uncorrelated. From (c) we see that $Cov(X, Y) = E(XY)$ if either X or Y has mean zero.

7.3.3 *Conditional Distributions*

Another way of describing the relationship between two random variables is via conditional distributions. These describe the behavior of one variable when the other takes various values. For instance, if we observe $X = 1$ but do not observe the Y realization, what can we predict about the behavior of Y? For the joint pdf in (7.29), we know that only three values

of Y are possible when $X = 1$, with $Y = 3$ and $Y = 4$ equally likely, and $Y = 3.5$ twice as likely as either of these. Other values of Y have probability zero. Total probabilities must sum to one, so we divide each of these probabilities by their total sum, i.e., by $\Pr(X = 1)$, to obtain the conditional probabilities:

$$\Pr(Y = 3 \mid X = 1) = \frac{1/20}{4/20} = \frac{1}{4}\,,$$
$$\Pr(Y = 3.5 \mid X = 1) = \frac{2/20}{4/20} = \frac{1}{2}\,,\ \Pr(Y = 4 \mid X = 1) = \frac{1/20}{4/20} = \frac{1}{4}\,,$$
$$\Pr(Y = y \mid X = 1) = 0 \ \text{ for } \ y = 4.5, 5, 5.5 \text{ and } 6.$$

This collection of probabilities make up the **conditional pdf** of Y given $X = 1$. Making these calculations for each value of X gives

		$\Pr(Y \mid X)$				
	6	0	0	0	0	1/4
	5.5	0	0	0	1/4	1/2
	5	0	0	1/4	1/2	1/4
Y	4.5	0	1/4	1/2	1/4	0
	4	1/4	1/2	1/4	0	0
	3.5	1/2	1/4	0	0	0
	3	1/4	0	0	0	0
		1	2	3	4	5
				X		

(7.30)

Each column of (7.30) represents a complete pdf, so we have a collection of five pdfs, one for each possible value of X.

For any given value of $X = x$, we can use the corresponding conditional pdf to compute the conditional mean of Y given $X = x$, and the conditional variance of Y given $X = x$. For $X = 1$, we have:

$$\begin{aligned} E(Y \mid X = 1) &= 3(\tfrac{1}{4}) + 3.5(\tfrac{1}{2}) + 4(\tfrac{1}{4}) + 4.5(0) + 5(0) + 5.5(0) + 6(0) \\ &= 3.5 \end{aligned}$$
$$\begin{aligned} Var(Y \mid X = 1) &= (3 - 3.5)^2(\tfrac{1}{4}) + (3.5 - 3.5)^2(\tfrac{1}{2}) + (4 - 3.5)^2(\tfrac{1}{4}) + 0 + 0 + 0 + 0 \\ &= 0.125 \end{aligned}$$

Repeating these calculations for each value of X we get:

X	1	2	3	4	5
$E(Y \mid X)$	3.5	4	4.5	5	5.5
$Var(Y \mid X)$	0.125	0.125	0.125	0.125	0.125

(7.31)

Notice that $E(Y \mid X)$ is a function of X. In our example, the conditional mean of Y given X increases with X. In fact we have

$$E(Y \mid X) = 3 + 0.5X\,,\ X = 1, 2, 3, 4, 5\,.$$

The conditional variance in this example turns out to be constant: $Var(Y \mid X) = 0.125$ for all X. In general it will also be a function of X.

In this example, knowledge of the value of X gives us information that we can use to refine our view of the behavior of Y or to predict Y using information in X. For instance, if we know that X is small (relative to its mean), then we know that the mean of Y will also tend to be small (relative to its unconditional mean). If we know that the X is large, then we also know that the Y outcome will be large. If we do not observe X, then our view regarding the mean value of Y will have to cover all possible values of X, which is what the unconditional mean of Y does. The fact that X gives us information about Y is also reflected in the reduction in variance from $Var(Y) = 0.625$ to $Var(Y \mid X) = 0.125$. This last result does not hold in general, but it does hold when $Var(Y \mid X)$ is constant.

For two continuous random variables X and Y, the conditional distributions are defined as

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \text{ and } f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

when $f_X(x) \neq 0$ and $f_Y(y) \neq 0$. Another way of writing this is

$$f_{X,Y}(x,y) = f_{Y|X}(y \mid x) f_X(x) = f_{X|Y}(x \mid y) f_Y(y).$$

This decomposition of joint pdfs can be extended to more than two variables, e.g., we have

$$f_{X,Y,Z}(x,y,z) = f_{Z|X,Y}(z \mid x,y) f_{Y|X}(y \mid x) f_X(x)\,.$$

7.3.4 *Manipulating Conditional Moments*

Manipulation of conditional expectations and variances follows one simple principle: whatever is being conditioned on can be treated as "fixed" (i.e., like a constant) as far as that expectation or variance is concerned.

Example 7.20

(a) $E(aXY \mid X) = aXE(Y \mid X)\,,\ Var(aXY \mid X) = a^2X^2\,Var(Y \mid X)$,

(b) $E(aX \mid X) = aX$ (contrast with $E(aX) = aE(X)$, a constant),

(c) $Var(aX \mid X) = 0$ (contrast with $Var(aX) = a^2\,Var(X)$),

(d) If $Y = \beta_0 + \beta_1 X + \epsilon$ with $E(\epsilon \mid X) = 0$ and $Var(\epsilon \mid X) = \sigma^2$, then

$$E(Y \mid X) = \beta_0 + \beta_1 X \text{ and } Var(Y \mid X) = \sigma^2. \tag{7.32}$$

In linear regression analysis, we often begin with an assumption that the conditional expectation takes some form, such as (7.32), the objective being to estimate the parameters β_0 and β_1.

7.3.5 *The Law of Iterated Expectations*

Recall that for the joint pdf in (7.29), we have $E(Y) = 4.5$, and

X	1	2	3	4	5
$E(Y \mid X)$	3.5	4	4.5	5	5.5
$\Pr(X)$	0.2	0.2	0.2	0.2	0.2

While $E(Y)$ is a single number, $E(Y \mid X)$ is a random variable when considered over all possible values of X. In our example, $E(Y \mid X)$ is a (uniformly distributed) random variable with possible values 3.5, 4.0, 4.5, 5.0, and 5.5. If we calculate the mean of this random variable, we get

$$E(E(Y \mid X)) = 3.5(0.2) + 4(0.2) + 4.5(0.2) + 5(0.2) + 5.5(0.2) = 4.5\,.$$

This value turns out to be exactly the same as the value of $E(Y)$, which we calculated earlier in Section 7.3.1. The equality of $E(E(Y \mid X))$ and $E(Y)$ is not a coincidence, but an example of the **Law of Iterated Expectations**:

$$E_X(E_{Y|X}(Y \mid X)) = E_Y(Y). \tag{7.33}$$

We add the subscript to the expectation notation in (7.33) to be clear as to the probabilities over which the expectations are taken, e.g., $E_{Y|X}$ indicates that the expectation is taken over the conditional probabilities of Y given X, whereas E_Y and E_X indicate that the expectations are taken under the marginal distributions of Y and X respectively. We often drop the subscripts for cleaner exposition. The Law of Iterated Expectations is also known as the **Law of Total Expectations**.

The Law of Iterated Expectations says (roughly speaking) that we can get the 'overall' average of Y by taking the Y average for each possible value of X, and then taking the average of those averages.

More generally, we have

$$E_{X,Y}(g(X,Y)) = E_X(E_{Y|X}(g(X,Y)))$$

Proof:

$$\begin{aligned}
E_{X,Y}(g(X,Y)) &= \int_X \int_Y g(x,y) f_{X,Y}(x,y)\, dy\, dx \\
&= \int_X \int_Y g(x,y) f_{Y|X}(y \mid x) f_X(x)\, dy\, dx \\
&= \int_X \left(\int_Y g(x,y) f_{Y|X}(y \mid x)\, dy \right) f_X(x)\, dx \\
&= E_X\left(E_{Y|X}(g(X,Y) \mid X) \right)
\end{aligned}$$

If $g(X,Y) = Y$, we get the law of iterated expectations as stated in (7.33).

The Law of Iterated Expectations implies the following:

(a) If $E(Y \mid X) = c$, then $E(Y) = c$,

(b) If $E(Y \mid X) = c$, then $Cov(X, Y) = 0$.

Result (a) says that if the expected value of Y is c for every possible value of X, then the "overall" mean must be that same constant, and (b) says that $E(Y \mid X) = c$ is a sufficient condition for $Cov(X, Y) = 0$. The derivation of these results is straightforward: if $E(Y \mid X) = c$, then

$$E(Y) = E(E(Y \mid X)) = E(c) = c$$

which proves (a). For (b), we note that

$$E(YX) = E(E(YX \mid X)) = E(XE(Y \mid X)) = E(cX) = cE(X)\,,$$

therefore

$$Cov(X, Y) = E(XY) - E(X)E(Y) = cE(X) - cE(X) = 0\,.$$

Although constant conditional mean implies zero covariance, the converse does not necessarily hold. For instance, suppose X is zero mean and has a symmetric distribution (which together implies that $E(X^3) = 0$). Suppose $Y = X^2$. Then $E(Y \mid X) = X^2$ but

$$\begin{aligned} Cov(X, Y) &= E(XY) - E(X)E(Y) \\ &= E(XE(Y \mid X)) - 0E(Y) = E(X^3) = 0. \end{aligned}$$

The Law of Iterated Expectations can be extended to more than two variables. For example, for random variables W, X and Y, we have

$$E(X \mid Y) = E(E(X \mid Y, W) \mid Y).$$

You are asked in Ex. 7.26 to show the **Law of Iterated Variance** or **Law of Total Variance**:

$$Var(Y) = E(\,Var(Y \mid X)) + Var(E(Y \mid X)). \tag{7.34}$$

7.3.6 *Independent Random Variables*

Two random variables are said to be **independent** if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y). \tag{7.35}$$

For discrete random variables, this means that

$$\Pr(X = i, Y = j) = \Pr(X = i)\Pr(Y = j)$$

for all possible values of X and Y. Independence of X and Y implies

$$f_{Y|X}(y \mid x) = f_Y(y) \text{ and } f_{X|Y}(x \mid y) = f_X(x).$$

Knowledge of the realized value of one variable does not add any information regarding the probabilistic behavior of the other.

Independence implies $E(Y \mid X) = E(Y)$, $Var(Y \mid X) = Var(Y)$, and so on. Independence of X and Y implies zero covariance between the two random variables: if X and Y are independent, then

$$E(XY) = E(XE(Y \mid X)) = E(X)E(Y)\,,$$

therefore

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0\,.$$

However, zero covariance does not imply independence. Ex. 7.24 at the end of this section presents an example where X and Y have zero covariance, but where the variance of Y increases in X.

Bivariate *Normal* random variables (see next section) are an exception: if two random variables have a Bivariate Normal distribution and are uncorrelated, then they are also independent.

7.3.7 *Exercises*

Ex. 7.18 Starting from the definition $Cov(X,Y) = E((X - E(X))(Y - E(Y)))$ and using the properties of expectations, show that

$$Cov(X,Y) = E(XY) - E(X)E(Y)\,.$$

Ex. 7.19 For random variables X_1, X_2 and X_3 and constants a_1, a_2 and a_3, show that

$$Var\left(\sum_{i=1}^{3} a_i X_i\right) = \sum_{i=1}^{3}\sum_{j=1}^{3} a_i a_j Cov(X_i, X_j).$$

Remark: Since $Cov(X_i, X_i) = Var(X_i)$, we can also write the equality as

$$\begin{aligned} Var\left(\sum_{i=1}^{3} a_i X_i\right) &= a_1^2\, Var(X_1) + a_2^2\, Var(X_2) + a_3^2\, Var(X_3) \\ &\quad + 2a_1 a_2\, Cov(X_1, X_2) + 2a_1 a_3\, Cov(X_1, X_3) + 2a_2 a_3\, Cov(X_2, X_3)\,. \end{aligned}$$

Ex. 7.20 Show that

$$Cov(a_1X_1 + a_2X_2, b_1Y_1 + b_2Y_2 + b_3Y_3) = \sum_{i=1}^{2}\sum_{j=1}^{3} a_i b_j Cov(X_i, Y_j).$$

Ex. 7.21 Show for the joint pdf (7.29) that the correlation of X and Y is 0.8944.

Ex. 7.22 Explain why the correlation always lies between -1 and 1, inclusive. *Hint: For arbitrary α, we have $Var(X - \alpha Y) \geq 0$. Expand $Var(X - \alpha Y)$ and let $\alpha = Cov(X,Y)/Var(Y)$*

Ex. 7.23 For the joint pdf (7.29), find the conditional distribution of Y given $X \geq 3$, and the corresponding conditional mean and variance.

Ex. 7.24 Suppose Y and X have the following joint distribution function:

	10	0	0	0	0	0.1
	9	0	0	0	0.1	0
	8	0	0	0.1	0	0
	7	0	0.1	0	0	0
	6	0.1	0	0	0	0
Y	5	0.1	0	0	0	0
	4	0	0.1	0	0	0
	3	0	0	0.1	0	0
	2	0	0	0	0.1	0
	1	0	0	0	0	0.1
		1	2	3	4	5
				X		

(a) Find the marginal distributions of X and Y.

(b) Find the conditional distribution, conditional mean, and conditional variance of Y given X, and of X given Y.

(c) Find $Cov(X, Y)$.

Ex. 7.25 Show that if $E(Y \mid X) = a + bX$, then

$$b = \frac{Cov(X, Y)}{Var(X)} \quad \text{and} \quad a = E(Y) - bE(X) \ .$$

If you know that $E(Y \mid X) = 3 + 0.5X$ and $Var(X) = 2$, what is $Cov(X, Y)$?

Ex. 7.26 Prove (7.34). Use this relationship to show that

(a) $Var(Y) = E(\, Var(Y \mid X))$ if $E(Y \mid X)$ is constant.

(b) $Var(Y \mid X) \leq Var(Y)$ if $Var(Y \mid X)$ is constant.

Ex. 7.27 Suppose Y and X have the following joint pdf:

	5	0.01	0.04	0.03	0.01	0.01
	4	0.02	0.08	0.06	0.02	0.02
Y	3	0.04	0.16	0.12	0.04	0.04
	2	0.02	0.08	0.06	0.02	0.02
	1	0.01	0.04	0.03	0.01	0.01
		1	2	3	4	5
				X		

Are the variables independent? Are they identically distributed (i.e., do they have the same marginal distributions?) Change the probabilities in the joint pdf of X and Y so that the two variables are independently and identically distributed (but not uniformly distributed).

7.4 Distributions Related to the Normal Distribution

We briefly discuss three univariate distributions, all are related to the Normal distribution. We will not require the pdf or cdf of these distributions in this book, but the properties of these distributions should be noted. We also discuss the Bivariate Normal distribution.

7.4.1 *Chi-Square distribution*

If $X \sim \text{Normal}(0,1)$, then X^2 has the **Chi-Square distribution** with one degree of freedom. If $X_1, X_2, \ldots, X_k$ are independent Standard Normal variates, then $\sum_{i=1}^{k} X_i^2$ has a Chi-Square distribution with k degrees of freedom, denoted $\chi^2(k)$. If $X \sim \chi^2(k)$, then $E(X) = k$ and $var(X) = 2k$. The pdfs of several Chi-square distributions are shown in Fig. 7.11.

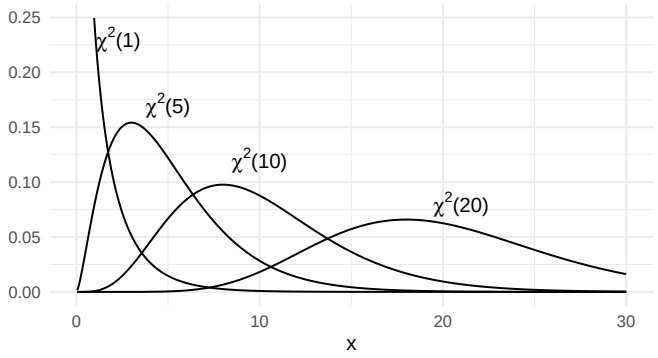

Fig. 7.11. The χ^2 pdf, various degrees of freedom.

7.4.2 *Student's t distribution*

If X and W are independent variables with $X \sim \text{Normal}(0,1)$ and $W \sim \chi^2(v)$, then

$$\frac{X}{\sqrt{W/v}} \sim t(v)$$

where $t(v)$ denotes the **Student's t-distribution**[9] (or simply **t-distribution**) with v degrees of freedom. A t-distributed random variable has zero mean and variance $\frac{v}{v-2}$ (the mean does not exist unless $v > 1$ and variance does not exist unless $v > 2$).

[9]Due to William Sealy Gosset (1876-1937), an English chemist working for Guinness Brewery. He published the t-statistic under the pen name "Student" as Guinness did not allow their scientists to publish under their own name (or to mention "beer" or "Guinness" in their papers). He later became head brewer of Guinness, but died a month after his promotion.

The t-pdf is similar to that of the Standard Normal pdf in that it is symmetrically bell-shaped and centered about zero. However, it has fatter tails than a Normal distribution (its kurtosis, when it exists, is always greater than 3). This means that a t-distributed random variable has greater probability of extreme realizations than a comparable normal variate. The t-pdf has the property that it converges to the standard normal pdf as $v \to \infty$. Fig. 7.12 shows the t-pdf with degrees-of-freedom parameter $v = 1$, 5, 10, and 20, and also the Standard Normal pdf. The $t(1)$ and $t(5)$ distributions are indicated, with the $t(10)$ and $t(20)$ distributions "between" the $t(5)$ and the Normal(0,1) pdf.

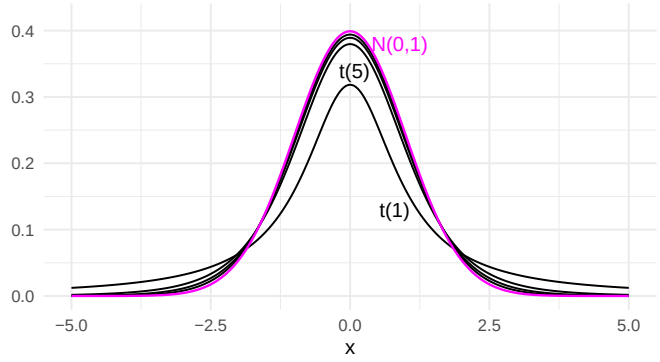

Fig. 7.12. The t distribution.

The following is a comparison of the tail probabilities of the Standard Normal and the t-distribution.

```
           N(0,1)   t(1)   t(5)  t(10)  t(20)  t(30)
P[X<-2.57] 0.0051 0.1181 0.0250 0.0139 0.0091 0.0077
P[X<-1.96] 0.0250 0.1502 0.0536 0.0392 0.0320 0.0297
P[X<-1.64] 0.0505 0.1743 0.0810 0.0660 0.0583 0.0557
```

7.4.3 *F-distribution*

If W_1 and W_2 are independent Chi-Square random variables with degrees of freedom v_1 and v_2 respectively, then

$$\frac{W_1/v_1}{W_2/v_2} \sim F(v_1, v_2)$$

where $F(v_1, v_2)$ denotes the **F-distribution** with v_1 and v_2 degrees of freedom. If $X \sim F(v_1, v_2)$, then

$$E(X) = \frac{v_2}{v_2 - 2} \quad \text{for} \quad v > 2\,,$$
$$Var(X) = 2\left(\frac{v_2}{v_2 - 2}\right)^2 \frac{v_1 + v_2 - 2}{v_1(v_2 - 4)} \quad \text{for} \quad v_2 > 4\,.$$

The F-distribution is also related to the Student-t and Chi-squared distributions in that

i. If $X \sim t(v)$, then $X^2 \sim F(1, v)$,

ii. If $X \sim F(v_1, v_2)$, then the pdf of $v_1 X$ tends to the $\chi^2(v_1)$ pdf as $v_2 \to \infty$.

Fig. 7.13 shows the $F(3, 20)$ pdf.

Fig. 7.13. The F distribution.

7.4.4 *The Bivariate Normal Distribution*

Two random variables X and Y follow the Bivariate Normal distribution if their joint pdf has the form

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1 - \rho_{XY}^2}} \exp\left\{-\frac{1}{2}\frac{\tilde{x}^2 - 2\rho_{XY}\tilde{x}\tilde{y} + \tilde{y}^2}{1 - \rho_{XY}^2}\right\} \tag{7.36}$$

where $\tilde{x} = \dfrac{x - \mu_X}{\sigma_X}$ and $\tilde{y} = \dfrac{y - \mu_Y}{\sigma_Y}$.

The Bivariate Normal distribution has five parameters μ_X, μ_Y, σ_X^2, σ_Y^2 and ρ_{XY} being the unconditional means of X and Y, the unconditional variances of X and Y, and the correlation between X and Y, respectively. We write $(X, Y) \sim \text{Normal}_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY})$

Contour plots are helpful for visualizing Bivariate Normal distributions. We show the contour plots of a bivariate Normal distribution with

$$(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY}) = (1, 0, 1, 2, 0.9).$$

in Fig. 7.14(a). The 3D plot of the Bivariate Normal joint pdf is shown in Fig. 7.14(b).

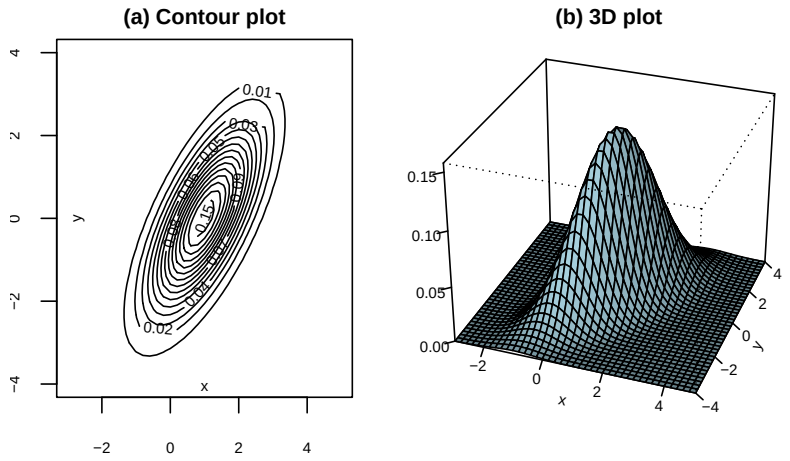

Fig. 7.14. A Bivariate Normal distribution (parameter values given in text).

The marginal and conditional distributions of Bivariate Normal random variables are also Normal. To see this, we "complete the square" on $\tilde{x}^2 - 2\rho_{XY}\tilde{x}\tilde{y} + \tilde{y}^2$ to get

$$\begin{aligned}
\tilde{x}^2 - 2\rho_{XY}\tilde{x}\tilde{y} + \tilde{y}^2 &= (\tilde{x} - \rho_{XY}\tilde{y})^2 + (1 - \rho_{XY}^2)\tilde{y}^2 \\
&= \left[\frac{x - \mu_X}{\sigma_X} - \frac{\sigma_{XY}}{\sigma_X \sigma_Y}\frac{y - \mu_Y}{\sigma_Y}\right]^2 + (1 - \rho_{XY}^2)\left(\frac{y - \mu_Y}{\sigma_Y}\right)^2 \\
&= \frac{1}{\sigma_X^2}\left[x - \mu_X - \frac{\sigma_{XY}}{\sigma_Y^2}(y - \mu_Y)\right]^2 + (1 - \rho_{XY}^2)\left(\frac{y - \mu_Y}{\sigma_Y}\right)^2 \\
&= \frac{1}{\sigma_X^2}\left[x - (\alpha_X + \beta_X y)\right]^2 + (1 - \rho_{XY}^2)\left(\frac{y - \mu_Y}{\sigma_Y}\right)^2
\end{aligned}$$

where $\alpha_X = \mu_X - \beta_X \mu_Y$ and $\beta_X = \dfrac{\sigma_{XY}}{\sigma_Y^2}$. Then the joint pdf can be written as

$$
\begin{aligned}
& f_{X,Y}(x,y) \\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \exp\left\{-\frac{1}{2}\frac{1}{1-\rho_{XY}^2}(\tilde{x}^2 - 2\rho_{XY}\tilde{x}\tilde{y} + \tilde{y}^2)\right\} \\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \\
&\quad \times \exp\left\{-\frac{1}{2}\frac{1}{1-\rho_{XY}^2}\left[\frac{1}{\sigma_X^2}\left[x-(\alpha_X+\beta_X y)\right]^2 + (1-\rho_{XY}^2)\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\} \\
&= AB
\end{aligned}
$$

$$
\text{where} \qquad A = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_X^2(1-\rho_{XY}^2)}} \exp\left\{-\frac{1}{2}\frac{\left[x-(\alpha_X+\beta_X y)\right]^2}{\sigma_X^2(1-\rho_{XY}^2)}\right\}
$$

$$
\text{and} \qquad B = \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}.
$$

If we compare expressions A and B with the expression for the Normal pdf, we see that B is a Normal pdf $f_Y(y)$ with mean μ_Y and variance σ_Y^2, and if we take y as fixed, then A is a conditional normal pdf $f_{X|Y}(x \mid y)$ with mean $\alpha + \beta y$ and variance $\sigma_X^2 - \sigma_{XY}^2/\sigma_Y^2$. That is, if X and Y have the Bivariate Normal distribution (7.36), then

- the marginal distribution of Y is Normal(μ_Y, σ_Y^2),
- the conditional distribution of X given Y is Normal$(\mu_{X|Y}, \sigma_{X|Y}^2)$ where

$$
\mu_{X|Y} = \mu_X + \frac{\sigma_{XY}}{\sigma_Y^2}(y-\mu_Y) \quad \text{and} \quad \sigma_{X|Y}^2 = \sigma_X^2 - \frac{\sigma_{XY}^2}{\sigma_Y^2}.
$$

  The conditional mean can be written as $\mu_{X|Y} = \alpha_X + \beta_X y$ where $\alpha_X = \mu_X - \beta_X\mu_X$ and $\beta_X = \dfrac{\sigma_{XY}}{\sigma_Y^2}$.

Similarly,

- the marginal distribution of X is Normal$(\mu_X\, \sigma_X^2)$,
- the conditional distribution of Y given X is Normal$(\mu_{Y|X}, \sigma_{Y|X}^2)$ where

$$
\mu_{Y|X} = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(x-\mu_X) \quad \text{and} \quad \sigma_{Y|X}^2 = \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}.
$$

The conditional mean can be written as $\mu_{Y|X} = \alpha_Y + \beta_Y x$ where $\alpha_Y = \mu_Y - \beta_Y\mu_X$ and $\beta_Y = \dfrac{\sigma_{XY}}{\sigma_X^2}$.

It follows immediately from the decomposition of the Bivariate Normal joint pdf that if X and Y are Bivariate Normal and uncorrelated, then they

are independent random variables (see Ex. 7.28). It can also be shown that if X and Y are Bivariate Normal, then

$$aX + bY \sim \text{Normal}(a\mu_X + b\mu_Y\,,\; a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY})\,.$$

We omit the proof of this last result.

7.4.5 *Exercises*

Ex. 7.28 Show that X and Y are independent if they are Bivariate Normal and uncorrelated. *Hint: show that* $f_{X,Y}(x,y) = f_Y(y)f_X(x)$ *when* $\rho_{xy} = 0$.

Ex. 7.29 The function `pnorm(x, mean, sd)` returns the Normal(mean, sd^2) cdf evaluated at `x`, i.e., it returns

$$\Pr(X \le x) = \Phi\left(\frac{x - \text{mean}}{\text{sd}}\right)$$

for a Normal(mean, sd^2) random variable X. The quantile function `qnorm(p, mean, sd)`, when evaluated at probability `p`, returns the value of x for which $\Pr(X \le x) = p$. For example:

```
pnorm(0, mean=0, sd=1)    # Pr[X <= 0] for X~N(0,1)
```

```
[1] 0.5
```

```
qnorm(0.5, mean=0, sd=1)  # c such that Pr[X<=c]=0.5 when X~N(0,1)
```

```
[1] 0
```

The corresponding functions for the t, Chi-Square, and F-distributions are

- `pt(x,df)` and `qt(p, df)`,
- `pchisq(x,df)` and `qchisq(p,df)`, and
- `pf(x,df1,df2)` and `qf(p,df1, df2)` respectively. Find:

i. $\Pr(X \le -2.5)$ when $X \sim N(0,1)$

ii. $\Pr(X \le -2.5)$ when $X \sim t(5)$

iii. c such that $\Pr(X > c) = 0.05$ when $X \sim \chi^2(5)$.

iv. $\Pr(-1.96 \le X \le 1.96)$ when $X \sim N(0,1)$

v. c such that $\Pr(-c \le X \le c) = 0.95$ when $X \sim N(0,1)$

vi. c such that $\Pr(-c \le X \le c) = 0.95$ when $X \sim t(12)$

vii. c such that $\Pr(-c \le X \le c) = 0.95$ when $X \sim t(100)$

viii. c such that $\Pr(X > c) = 0.05$ when $X \sim F(5,8)$.

ix. c such that $\Pr(X > c) = 0.05$ when $X \sim F(5,80)$.

x. c such that $\Pr(X > c) = 0.05$ when $X \sim F(5,8000)$.

7.5 Asymptotic Theory

Our discussion of estimation and hypothesis testing in Section 7.2 was for finite samples. Asymptotic analysis refers to results that apply "in the limit", i.e., as the sample size goes to infinity. It serves to approximate the finite sample properties of estimators when the sample size is reasonably large, and is especially helpful when the finite sample properties are unknown. We continue to focus on the sample mean, which we now denote as $\overline{X}_n$ to indicate the sample size used to calculate it.

7.5.1 *Consistency and the Law of Large Numbers*

We have mentioned that larger sample sizes are desirable as they lead to smaller estimator variances. For the general problem of estimating the population mean μ of a random variable X using the sample mean, we have $E(\overline{X}_n) = \mu$ for all n and $Var(\overline{X}_n) = \sigma^2/n \to 0$ as $n \to \infty$, where σ^2 is the variance of X. Since $\overline{X}_n$ is unbiased and its variance collapses to zero as the sample size tends to infinity, the estimator "converges" to the population mean as the sample size grows larger and larger.

The convergence of $\overline{X}_n$ to μ is, however, not quite the same as the convergence of, say, the deterministic sequence $1/n$ to zero. In the latter case, you know that if n is large enough, then $1/n$ will *definitely* be within a certain distance of 0. For instance, if $n > 1000$, then $1/n < 0.001$ *for sure*. In the case of $\overline{X}_n$, which is a sequence of *random variables*, we cannot make such a definite claim.

The convergence concept we use for random variables is "convergence in probability". A sequence of random variables Y_n is said to **converge in probability** to some value c as $n \to \infty$ if for any $\epsilon > 0$ (no matter how small), the *probability* that $|Y_n - c| \geq \epsilon$ tends to zero as $n \to \infty$. This allows for some probability that the distance between Y_n and c exceeds ϵ at any sample size n, but as n increases towards infinity, this probability becomes vanishingly small.[10] We write

$$\text{plim } Y_n = c \quad \text{or} \quad Y_n \xrightarrow{p} c\,.$$

We can extend this definition to "convergence in probability to a random variable": we say that

$$Y_n \xrightarrow{p} Z \quad \text{if} \quad Y_n - Z \xrightarrow{p} 0\,.$$

In the context of parameter estimation, we say that an estimator is **consistent** if it converges in probability to the true value of the parameter it is estimating as the sample size goes to infinity. The sample mean $\overline{X}_n$

[10] That $E(Y_n)$ converges to c and $Var(Y_n)$ converges to zero is sufficient to guarantee that $Y_n \xrightarrow{p} 0$.

is a consistent estimator for μ under quite general conditions. (We view $\overline{X}_n$ as a sequence of random variables in n. If certain conditions hold, this sequence converges in probability to $E(X)$ as $n \to \infty$.) This result is known as the **Law of Large Numbers**. There are several laws of large numbers, each describing a set of conditions which, if met, guarantee the consistency of the sample mean. We state one such law below:

Theorem 7.4 (Khinchine's Weak Law of Large Numbers[11], WLLN) *If* $\{X_i\}_{i=1}^n$ *are iid with* $E(X_i) = \mu < \infty$ *for all* i*, then* $\overline{X}_n \xrightarrow{p} \mu$.

There are other kinds of convergence concepts for sequences of random variables, but for the moment we consider only convergence in probability. The theorem above is referred to as a *weak* law of large numbers because the convergence concept used is convergence in probability, and there are "stronger" forms of probabilistic convergence.

The following is a result concerning convergence in probability that is used frequently:

Proposition 7.1 *If* $g(\cdot)$ *is a continuous function, then*

$$Y_n \xrightarrow{p} c \;\Rightarrow\; g(Y_n) \xrightarrow{p} g(c). \tag{7.37}$$

That is, if plim Y_n *exists, and* $g(\cdot)$ *is continuous, then* plim $g(Y_n) = g(\text{plim } Y_n)$.

For example, if Y_n converges in probability to c, then $Y_n^2 \xrightarrow{p} c^2$. Result (7.37) extends to continuous functions of multiple variables. This implies, for instance, that if $Y_n \xrightarrow{p} c_y$ and $Z_n \xrightarrow{p} c_z$, then

- $Y_n + Z_n \xrightarrow{p} c_y + c_z$,
- $Y_n Z_n \xrightarrow{p} c_y c_z$,
- $Y_n / Z_n \xrightarrow{p} c_y / c_z$, as long as c_z is non-zero.

Example 7.21 Suppose $\{X_i\}_{i=1}^n$ is an iid sample, with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$ for all i. We show that the biased estimator

$$\widetilde{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2$$

is consistent for the population variance σ^2. Since $\{X_i\}_{i=1}^n$ are iid, so are $\{X_i^2\}_{i=1}^n$. Furthermore, $E(X_i^2) = \sigma^2 + \mu^2 < \infty$, so (by the WLLN)

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{p} \sigma^2 + \mu^2 \,.$$

[11] Alexander Iacovlevich Khinchin (1894-1959)

Since $\overline{X}_n \xrightarrow{p} \mu$ and since quadratic functions are continuous functions, we have $\overline{X}_n^2 \xrightarrow{p} \mu^2$. Therefore

$$\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 \xrightarrow{p} \sigma^2 + \mu^2 - \mu^2 = \sigma^2 \,.$$

This example shows that estimators that are biased in finite samples can nonetheless be consistent.

Example 7.22 Since $\widehat{\sigma_n^2} = \frac{n}{n-1}\widetilde{\sigma_n^2}$, and because $\frac{n}{n-1} \to 1$ and $\widetilde{\sigma^2} \xrightarrow{p} \sigma^2$, we have $\widehat{\sigma_n^2} \xrightarrow{p} \sigma^2$.

Example 7.23 Since both $\widehat{\sigma_n^2}$ and $\widetilde{\sigma_n^2}$ are consistent estimators for σ^2, both $(\widehat{\sigma_n^2})^{1/2}$ and $(\widetilde{\sigma_n^2})^{1/2}$ are consistent estimators for σ.

We have seen earlier that unbiasedness, unlike consistency, generally does not carry over to non-linear functions of estimators. For instance, we saw earlier that $E(\hat{p}^2) \geq p^2$ despite $E(\hat{p}) = p$.

It may seem that unbiasedness is a more relevant way to judge an estimator than consistency since we never have infinite sample sizes, but consistency is still useful as it captures the idea of convergence to the population parameter as sample size increases. Furthermore, in more complex applications it can be difficult or impossible to find unbiased estimators, but straightforward to find consistent ones. We have also seen that it is easy to find consistent estimators of continuous functions of parameters once we have consistent estimators for the parameters.

In Example 7.19 we derived the distribution of the sample mean in the coin toss example, and derived its distribution for a fair coin and a sample size of 20. We repeat this exercise, this time for a coin with $p = 0.25$, for sample sizes 5, 10, 20, 100, 200 and 400. We present the probability distribution functions graphically in Fig. 7.15. The convergence in probability of the sample mean to the true value of p can be seen in these graphs.

7.5.2 *Asymptotic Normality*

The distribution of the sample mean in Fig. 7.15, with $p = 0.25$, is unsurprisingly skewed in small samples because of the low probability of heads relative to tails. However, the shape of the distribution appears to quickly become quite symmetric as the sample size grows, and appears to converge to a familiar bell-shaped distribution. Of course, in the limit the distribution collapses to a degenerate one with all of the probability at $p = 0.25$, since the variance of the sample mean, $Var(\hat{p}) = p(1-p)/n$, goes to zero as $n \to \infty$. In what sense, then, can we say that the distribution of $\hat{p}$ converges to a normal distribution? Suppose we scale the sample mean (after

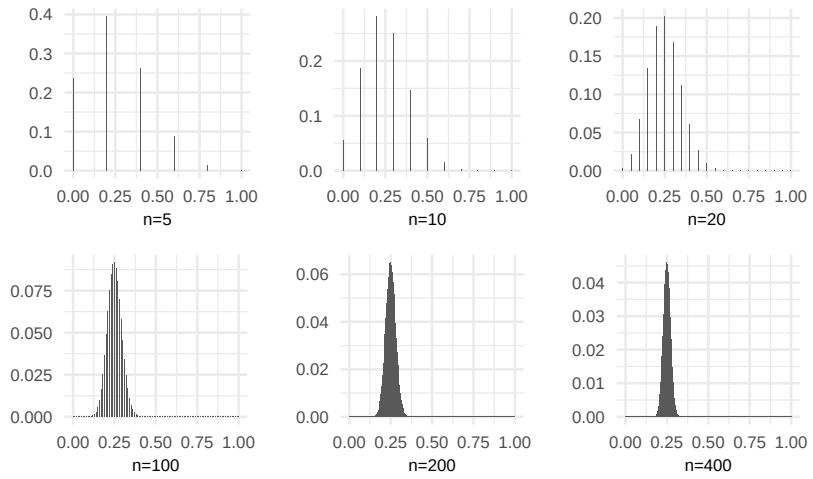

Fig. 7.15. Convergence in probability of $\hat{p}$ to $p = 0.25$.

subtracting p) by $\sqrt{n}$, and consider the distribution of

$$\sqrt{n}(\hat{p} - p). \tag{7.38}$$

This random variable has mean 0 and a non-collapsing variance $p(1-p)$, and we can speak about how the shape of the limiting distribution of (7.38) as $n \to \infty$ without the problem of its distribution collapsing to a single point.

The plots in Fig. 7.16 show the same distributions as in Fig. 7.15, but after centering and scaling as in (7.38). The distributions appear to take the shape of a Normal distribution as the sample size increases.

How can the pdf of the sample mean, which in the coin toss example is a discrete random variable, converge to the pdf of a Normal random variable, which is a continuous random variable? The notion of a discrete pdf converging to a continuous one is best thought of in terms of their cdfs. In Fig. 7.17, we juxtapose the cdfs of the distributions of $\sqrt{n}(\hat{p} - p)$ at $n = 20, 100, 400$ (these are all step functions) with the cdf of the Normal distribution with mean 0 and variance $p(1-p)$ where $p = 0.25$. At each value of $\sqrt{n}(\hat{p} - p)$, i.e., "pointwise", the step function gets closer to the Normal cdf as $n \to \infty$. We say that a sequence of random variables Y_n with corresponding cdfs $F_n(y_n)$ **converges in distribution** to a random variable Y with cdf $F(y)$ if, as $n \to \infty$, $F_n(x_n) \to F(x)$ at all points where $F(\cdot)$ is continuous. We write $Y_n \xrightarrow{d} Y$. We can also use the name of the limiting distribution in this notation, e.g., if $Y \sim \text{Normal}(0, 1)$, we can write $Y_n \xrightarrow{d} \text{Normal}(0, 1)$.

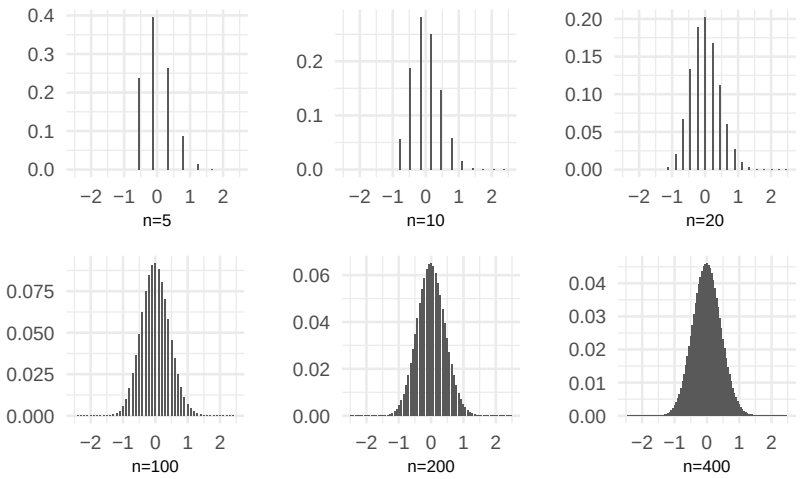

Fig. 7.16. Convergence in distribution of $\sqrt{n}(\hat{p} - 0.25)$ to Normal.

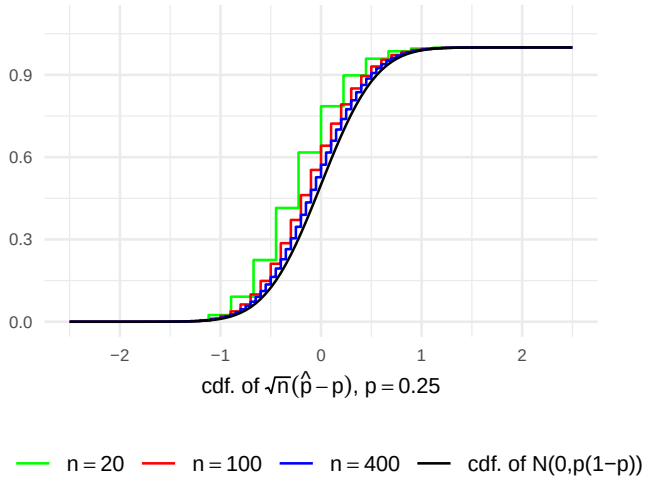

Fig. 7.17. Convergence in distribution of $\sqrt{n}(\hat{p} - 0.25)$ to Normal (cdf view).

7.5.3 *The Central Limit Theorem*

The convergence of the cdf of the (centered and scaled) sample mean in the coin toss example to a Normal cdf is an instance of the **Central Limit Theorem** (CLT), a key result in probability theory. As with the LLN, there are many CLTs, each listing a set of conditions under which convergence to normality is guaranteed. We state one such CLT:

Theorem 7.5 (Lindeberg-Levy CLT[12]) *If* $\{X_i\}_{i=1}^n$ *are i.i.d. with* $E(X_i) = \mu$ *and* $Var(X_i) = \sigma^2 < \infty$ *for all* i*, then*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \text{Normal}(0, \sigma^2)\,.$$

If the conditions for the CLT hold, we would be justified, in large samples, to say that the distribution of $\sqrt{n}(\overline{X}_n - \mu)$ is approximately Normal$(0, \sigma^2)$, or that $\overline{X}$ is approximately Normal$(\mu, \sigma^2/n)$. This last statement is sometimes written

$$\overline{X}_n \overset{a}{\sim} \text{Normal}(\mu, \sigma^2/n)\,,$$

where the "a" stands for "approximately" (some take "a" to stand for "asymptotically").

Our plots of the distribution of $\sqrt{n}(\hat{p}_n - p)$ in the coin toss example suggests convergence in distribution to Normal$(0, p(1-p))$. The sample in the coin toss example meets the requirements of the Lindeberg-Levy CLT, so in fact $\sqrt{n}(\hat{p}_n - p)$ converges in distribution to Normal$(0, p(1-p))$

Proposition 7.2 (Properties of convergence in distribution)

(a) If $g(\cdot)$ *is a continuous function and* $Y_n \xrightarrow{d} Y$*, then* $g(Y_n) \xrightarrow{d} g(Y)$*.*

(b) If $Y_n \xrightarrow{p} Y$*, then* $Y_n \xrightarrow{d} Y$*.*

(c) If $a_n \xrightarrow{p} a$ *and* $Y_n \xrightarrow{d} Y$*, then* $a_n Y_n \xrightarrow{d} aY$ *and* $a_n + Y_n \xrightarrow{d} a + Y$*.*

Example 7.24 If $Y_n \xrightarrow{d} Y \sim \text{Normal}(0, 1)$, then $Y_n^2 \xrightarrow{d} Y^2 \sim \chi(1)^2$, since the square of a Standard Normal is $\chi^2(1)$.

Example 7.25 If $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \text{Normal}(0, \sigma^2)$ and s_n^2 is any consistent estimator of σ^2, then $1/s_n = (1/s_n^2)^{1/2}$ converges in probability to $1/\sigma$, and therefore

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} = \frac{\overline{X}_n - \mu}{\sqrt{s_n^2/n}} \xrightarrow{d} \text{Normal}(0, 1). \tag{7.39}$$

Result (7.39) is useful for hypotheses testing when one is unable or unwilling to make an assumption regarding the distribution of the sample.

[12] Jarl Waldemer Lindeberg (1876-1932) and Paul Pierre Lévy (1886-1971).

Suppose $\{X_i\}_{i=1}^n$ is an iid sample with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Then the sample mean $\overline{X}_n$ is a consistent estimator for μ and

$$\widehat{\sigma_n^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is a consistent estimator for σ^2. To test the null hypothesis $H_0 : \mu = \mu_0$, we need to find the distribution of the sample mean, but you cannot do this unless you know the distribution of each X_i. Result (7.39), however, tells us that if our sample size is large enough, then under the null hypothesis, we have

$$t = \frac{\overline{X}_n - \mu_0}{\sqrt{\widehat{\sigma_n^2}/n}} \overset{a}{\sim} \text{Normal}(0,1). \tag{7.40}$$

This suggests that we use the decision rule

"reject the null if $|\,t\,| > c_\alpha$, i.e., if $t < -c_\alpha$ or $t > c_\alpha$"

where c_α, the value such that $\Pr(|\,t\,| > c_\alpha) = \alpha$, the chosen level of significance, is found from the Standard Normal distribution. For $\alpha = 0.01, 0.05, 0.10$ levels of significance, the appropriate values of c_α are, respectively, approximately

```
round(qnorm(c(0.995, 0.975, 0.95)),3)
```

```
[1] 2.576 1.960 1.645
```

The 0.05 level of significance test, in particular, says to reject $H_0 : \mu = \mu_0$ if the absolute distance from the sample mean to the hypothesized mean μ_0 is more than 1.96 (or approximately 2) sample standard errors.

A test based on the t-statistic in (7.40) with critical values based on the Normal$(0,1)$ distribution, would be an approximate test in the sense that the true significance level may not be exactly α, as intended. Nonetheless, it is a way forward in a situation where an exact test is unavailable. Even where the exact distribution is available, such as in our coin toss example, the test statistics in (7.40) is a convenient approximation.

Example 7.26 Earlier we showed an example of 20 tosses of three coins, where the true value of p for coins 1, 2, and 3 are 0.5, 0.6, and 0.9 respectively. We replicate the results below, this time also computing the corresponding t-statistics for the hypothesis $H_0 : p = 0.5$, i.e., we compute

$$t = \frac{\hat{p} - 0.5}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n-1}}}$$

where $\hat{p}$ is the sample mean of the observations, and where the denominator is the sample standard deviation of the estimator (see Ex. 7.13).

```
set.seed(13)   # Initialize random number generator for replicability
N <- 20
Coin1 <- rbinom(N,1,0.5)   # 20 tosses of a fair coin
Coin2 <- rbinom(N,1,0.6)   # 20 tosses of a slightly unfair coin
Coin3 <- rbinom(N,1,0.9)   # 20 tosses of a very warped coin!
Tosses <- rbind(Coin1, Coin2, Coin3)  # place outcomes into three rows
p_hat <- apply(Tosses,1,mean) # apply mean function to each row of Tosses
var_p_hat <- p_hat*(1-p_hat)/(N-1) # as per formula derived in notes
se_p_hat <- sqrt(var_p_hat)
t_stat <- (p_hat-0.5)/se_p_hat
p_val <- 2*(1-pnorm(t_stat,0,1))
d <- cbind(p_hat,se_p_hat,t_stat,p_val)
round(d,4)
```

```
      p_hat se_p_hat t_stat  p_val
Coin1  0.65   0.1094 1.3708 0.1704
Coin2  0.70   0.1051 1.9024 0.0571
Coin3  0.85   0.0819 4.2726 0.0000
```

Using the asymptotic tests, we make the following conclusions:

- we (correctly) retain (i.e., do not reject) the hypothesis that Coin 1 is fair at any of the usual levels of significance.
- we (correctly) reject H_0 for Coin 2 at 0.1 level of significance, but (incorrectly) retain H_0 at the 0.05 level of significance.
- we resoundingly (and correctly) reject fairness of Coin 3 at all conventional levels of significance.

The result for Coin 2 illustrates the fact that it can be hard to reject a mildly incorrect hypothesis. All tests have poor power in such cases. The results for Coin 1 and Coin 3 turned out to be correct in this example, but you should remember that there were non-zero probabilities of rejecting fairness for Coin 1, and not rejecting fairness for Coin 3.

In addition to the t-statistic, we also compute the **p-value**, defined as the probability that the t-statistic, prior to realization, would exceed the realized t-statistic in absolute terms if H_0 were true. We reject a null hypothesis at α level of significance if the corresponding p-value of the test is smaller than α.

The test statistic in (7.40) is often called a t-statistic because its exact distribution turns out to be the t-distribution with degrees of freedom $n-1$ if the iid sample observations are Normally distributed, i.e., if it is the case that $X_i \sim \text{Normal}(\mu, \sigma^2)$ for all i, then

$$t = \frac{\overline{X}_n - \mu}{\sqrt{\widehat{\sigma_n^2}/n}} \quad \sim \quad t(n-1)\,.$$

Since X_i is not normal in the coin toss example, this result does not apply, and we rely on the asymptotic result. Of course, the t-distribution is itself

approximately Standard Normal when n is large, so if $X \sim \text{Normal}(\mu, \sigma^2)$, the t-statistic converges in distribution to a Standard Normal for *two* reasons: because of the CLT, and because the t-distribution anyway converges to the Standard Normal as the degrees of freedom parameter goes to infinity. The usefulness of result (7.40) is in its applicability regardless of the distribution of the sample, when sample sizes are large enough, as long as the conditions of the CLT hold.

7.5.4 *Exercises*

Ex. 7.30 Suppose you want to estimate the mean and variance of a Log-Normal random variable X. You don't have the data, but you have the sample mean and sample variance of $\ln X_i$ calculated from a random sample of size n. Suggest consistent estimators for $E(X)$ and $Var(X)$. How do you know that these estimators are consistent?

Ex. 7.31 Use the random number generator in R to produce 400 separate iid samples each of size 50 observations from a $\chi(1)^2$ distribution (you can imagine there are 400 people carrying out this task, each drawing an iid sample of size 50 from a Chi-Square distribution with 1 degree of freedom). Label the samples

$$\{X_{i,j}\}_{j=1}^{400} \quad \text{for} \quad i = 1, 2, ..., 50\,.$$

Compute

$$\left\{\sqrt{n}(\overline{X}_{i,n} - 1)\right\}_{i=1}^{400} \text{ where } \overline{X}_{i,n} = \frac{1}{n}\sum_{j=1}^{n} X_{i,j} \text{ for } n = 5, 10, 20, 50\,.$$

Calculate the sample mean, variance, skewness and kurtosis of $\{\sqrt{n}(\overline{X}_{i,n} - 1)\}_{i=1}^{400}$ for $n = 25, 50, 75, 100$. Do these estimates conform with the prediction of the Central Limit Theorem?

7.6 The Bootstrap

We mentioned in Chapter 1 that statistical analyses are relying more and more on computational algorithms. In this section we discuss one such class of algorithms, called the **bootstrap**, for computing standard errors of estimators.

We have shown that the standard error of the sample mean calculated from an iid sample $\{Y_i\}_{i=1}^n$ with population mean $E(Y_i) = \mu$ and variance $Var(Y_i) = \sigma^2$ is

$$\text{s.e.}(\overline{Y}) = \sqrt{Var(\overline{Y})} = \sqrt{\frac{\sigma^2}{n}}\,,$$

which we can estimate with

$$\widehat{\text{s.e.}(\overline{Y})} = \sqrt{\frac{\widehat{\sigma^2}}{n}}\,, \text{ where } \widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2. \tag{7.41}$$

Example 7.27 Suppose Y is Chi-Square with two degrees of freedom, so $\mu = E(Y) = 2$ and $\sigma^2 = Var(Y) = 4$. Suppose *you do not know this*, and you only have the following sample of 16 observations of Y:

```
set.seed(123)  # seed is arbitrarily chosen
ysmp <- round(rchisq(16,2),2)
cat(ysmp[1:8],"\n")
cat(ysmp[9:16])
```

```
0.36 3.39 3.24 4.9 1.76 5.33 7.77 1.93
1.37 0.07 1.83 0.5 0.58 0.23 3.09 0.8
```

Your estimate of μ and the (estimated) standard error is therefore

```
n <- length(ysmp)
muhat <- sum(ysmp)/n                # you can also use mean(ysmp)
sg2 <- sum((ysmp-muhat)^2)/(n-1)    # you can also use var(ysmp)
cat("sample mean: ", muhat, ", std. err.: ", sqrt(sg2/n))
```

```
sample mean:  2.321875 , std. err.:  0.5459474
```

Since $Var(Y) = \sigma^2 = 4$, the actual standard error of the sample mean is

$$\sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{4}{16}} = 0.5$$

but as σ^2 is unknown to us, we can only provide an estimate of it.

The bootstrap provides an alternative way of estimating the standard error of an estimator, and is based on the following idea. *If* we could obtain B samples, each of size n, of the variable Y:

$$\begin{gathered}\{Y_1^{(1)}, Y_2^{(1)}, \dots, Y_n^{(1)}\} \\ \{Y_1^{(2)}, Y_2^{(2)}, \dots, Y_n^{(2)}\} \\ \vdots \\ \{Y_1^{(B)}, Y_2^{(B)}, \dots, Y_n^{(B)}\}\end{gathered}$$

then we can calculate a sample mean for each these B samples, which we denote

$$\hat{\mu}^{(b)} = \overline{Y}^{(b)}\,,\ b = 1, 2, ..., B\,.$$

This is essentially a sample of size B from the distribution of the sample mean obtained from samples of Y of size n. We would then be able to estimate the variance of $\hat{\mu} = \overline{Y}$ using

$$\widetilde{Var(\hat{\mu})} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\mu}^{(b)} - \bar{\hat{\mu}})^2 \quad \text{where} \quad \bar{\hat{\mu}} = (1/B)\sum_{b=1}^{B}\hat{\mu}^{(b)}\,.$$

However, you only have the one sample $\{Y_1, Y_2, \ldots, Y_n\}$. The idea of the bootstrap is to treat this one sample as the population, and then draw samples of size n with replacement from $\{Y_1, Y_2, \ldots, Y_n\}$, putting equal probability on each of the n observations. Doing this B times gives

$$\begin{gathered}\{Y_1^{*(1)}, Y_2^{*(1)}, \ldots, Y_n^{*(1)}\} \\ \{Y_1^{*(2)}, Y_2^{*(2)}, \ldots, Y_n^{*(2)}\} \\ \vdots \\ \{Y_1^{*(B)}, Y_2^{*(B)}, \ldots, Y_n^{*(B)}\}\end{gathered}$$

where each of the $Y_i^{*(b)}$ is a value from $\{Y_1, Y_2, \ldots, Y_n\}$. (There will be repeated values and missing Y_i values in each of the B "bootstrap" samples). From these B samples we calculate sample mean $\hat{\mu}^{*(b)}$, $b = 1, 2, ..., B$. Finally, we estimate

$$\widetilde{Var(\hat{\mu})} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\mu}^{*(b)} - \overline{\hat{\mu}^*})^2 \quad \text{where} \quad \overline{\hat{\mu}^*} = (1/B)\sum_{b=1}^{B}\hat{\mu}^{*(b)}\,. \tag{7.42}$$

Example 7.28 We use the bootstrap approach to estimate the standard error of the sample mean in Example 7.27. We use the `sample()` function in `R` to draw $B = 200$ samples of size 16 from the original sample. We show a few of these samples below. We then use the $B = 200$ bootstrapped sample means to calculate the standard error of the sample mean.

```
set.seed(456)
B <- 200                ## Bootstrap replication sample
bmeans <- rep(NA, B)    ## To store the bootstrapped means
for (b in 1:B){
  ysmpb <- sample(ysmp, 16, replace=T)
  bmeans[b] <- mean(ysmpb)
  if (b %in% c(1,2, 200)){
    cat("sample", b, ":", ysmpb[1:8], "\n")
    cat("         :", ysmpb[9:16], "\n")
  }
}
cat("Std. err. for sample mean using", B,
           "bootstrap samples:",sqrt(var(bmeans)))
```

```
sample 1 : 0.58 1.76 3.24 5.33 1.76 0.5 1.83 1.37
         : 0.23 3.09 0.07 1.37 3.09 0.07 3.24 1.83
sample 2 : 7.77 1.93 0.23 0.23 0.58 1.76 4.9 0.23
         : 3.09 1.37 0.36 5.33 1.93 1.76 1.37 3.39
sample 200 : 0.58 0.58 1.76 0.8 1.93 1.93 3.09 0.23
         : 1.37 1.37 0.5 7.77 0.5 0.07 7.77 1.76
Std. err. for sample mean using 200 bootstrap samples: 0.5268645
```

Why would we ever use the bootstrap method rather than (7.41) to estimate standard errors? For the sample mean, we were able to easily derive an expression for its standard error, and a way to estimate it. However, the standard error for many other statistics are much harder to derive, in many cases unknown, or perhaps known only under assumptions which may or may not hold. In many of these cases, the bootstrap approach will works; you merely have to replace $\widehat{\mu}^{*(b)}$ in (7.42) with the statistic for which you wish to estimate its standard error.

Example 7.29 The sample median of the sample in Example 7.27 is

```
medhat <- median(ysmp)
cat("Median:", medhat)
```

```
Median: 1.795
```

What is the standard error of the median? One quick way to estimate it is to use the bootstrap approach, which we do below:

```
set.seed(789)
B <- 200                 ## Bootstrap replication sample
bmeds <- rep(NA, 200)   ## To store the bootstrapped means
for (b in 1:B){
  ysmpb <- sample(ysmp, 16, replace=T)
  bmeds[b] <- median(ysmpb)
}
cat("Std. err. for sample median using", B,
         "bootstrap samples:", sqrt(var(bmeds)))
```

```
Std. err. for sample median using 200 bootstrap samples: 0.7163798
```

Of course, there are details which we have not been able to cover in our very short introduction to the bootstrap. For a more thorough introduction to the bootstrap, see Efron and Tibshirani (1993).[13]

7.6.1 *Exercises*

Ex. 7.32 In Ex. 7.15, you calculated the sample skewness and kurtosis coefficients for the observations $\log(earn_i)$ from the data set `earnings2019.csv`. Use the bootstrap (with 200 bootstrap samples) to calculate standard errors of your estimates.

[13]This name "bootstrap" comes from the phrase "to pull yourself up by your bootstraps". In footwear, bootstraps are straps attached to the sides or back of boots that you can pull on to help you put on the boots. The phrase was intended as a snarky remark to indicate something you *cannot* do, which is to haul yourself out of a hole by pulling on your own bootstraps. In statistics, we use the term "bootstrap" to refer to resampling from a sample to obtain new samples which are then used to calculate certain properties of estimators. This might seem impossible at first, since we are "reusing a sample", but as illustrated in our example, this method actually works quite well.

7.7 Solutions to Exercises

Ex. 7.1: We prove (c) first. Since $A \cap A^C = \emptyset$ and $A \cup A^c = \Omega$, we have

$$1 = \Pr(\Omega) = \Pr(A \cup A^c) = \Pr(A) + \Pr(A^c)\,,$$

which proves that $\Pr(A^c) = 1 - \Pr(A)$. Since $\Pr(A^c) \geq 0$, (c) implies (a). Since B is the union of the disjoint sets $A \cap B$ and $B \cap A^c$, (b) holds. Part (d) follows from part (b) and the fact that $A \cup B$ is the union of the disjoint sets A and $B \cap A^c$. For part (e): If $A \subset B$, then B is the union of the disjoint sets $B - A$ and A, so that $\Pr(B) = \Pr(B - A) + \Pr(A)$. Since $\Pr(B - A) \geq 0$, part (e) follows.

Ex. 7.2: Consider drawing a ball at random from this container. Let $A = \{green, blue\}$, the event that either a green or blue ball is drawn, and $B = \{green, red\}$, the event that either a green or red ball is drawn. The probability that a green ball is drawn is then $\Pr(A \cap B)$. From

$$1 \geq \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

we have

$$\Pr(\{green\}) = \Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1 = 0.95 + 0.90 - 1 = 0.85.$$

There are at least 85 green balls. If there are only red, green and blue balls in the container, then $\Pr(A \cup B) = \Pr(\Omega) = 1$. It follows that there are exactly 85 green balls in the container.

Ex. 7.3: If $\Pr(A \cap B) = \Pr(A)\Pr(B)$, then

$$\begin{aligned}\Pr(A^c \cap B) &= \Pr(B) - \Pr(A \cap B)\\ &= \Pr(B) - \Pr(A)\Pr(B) = \Pr(B)(1 - \Pr(A)) = \Pr(B)\Pr(A^c)\end{aligned}$$

and

$$\begin{aligned}\Pr(A^c \cap B^c) &= \Pr((A \cup B)^c) = 1 - \Pr(A \cup B)\\ &= 1 - [\Pr(A) + \Pr(B) - \Pr(A)\Pr(B)]\\ &= (1 - \Pr(A))(1 - \Pr(B)) = \Pr(A^c)\Pr(B^c)\,.\end{aligned}$$

Ex. 7.4: We have

$$\begin{aligned}&\Pr(knows \mid correct)\\ &= \frac{\Pr(correct \mid knows)\Pr(knows)}{\Pr(correct \mid knows)\Pr(knows) + \Pr(correct \mid guesses)\Pr(guesses)}\\ &= \frac{1 \cdot p}{1 \cdot p + (1/4)(1 - p)} = \frac{4p}{1 + 3p}\,.\end{aligned}$$

If $p = 0$, then $\Pr(knows \mid correct) = 0$. That is, if the instructor is sure that the student does not know the answer, then she won't change her mind even if the student answers the question correctly. If $p = 0.5$, then $\Pr(knows \mid correct) = 0.8$. If $p = 1$, then $\Pr(knows \mid correct) = 1$.

Ex. 7.5: We showed in the text that if $X \sim \text{Geometric}(p)$, then $E(X) = (1-p)/p$ and $E(X^2) = (1-p)(2-p)/p^2$. Therefore

$$Var(X) = E(X^2) - E(X)^2 = \frac{(1-p)(2-p)}{p^2} - \frac{(1-p)^2}{p^2} = \frac{1-p}{p^2}\,.$$

Ex. 7.6: We can use the fact that $Var(\sqrt{X}) = E(X) - E(\sqrt{X})^2 \geq 0$.

Ex. 7.7: We have $F_Y(y) = \Pr(X^2 \leq y) = \Pr(X \leq \sqrt{y}) = \sqrt{y}$ for all $y \in (0,1)$. Since $X \sim \text{Uniform}(0,1)$. Differentiating $F_Y(y)$ gives $f_Y(y) = \dfrac{1}{2\sqrt{y}}$, $y \in (0,1)$.

Ex. 7.8: For $Y = g(X)$ increasing, differentiating $F_Y(y) = F(g^{-1}(y))$ gives $f_Y(y) = f(g^{-1}(y))\frac{dg^{-1}(y)}{dy} = f(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|$. The last equality holds since $\frac{dg^{-1}(y)}{dy}$ is positive. For $Y = g(X)$ decreasing, differentiating $F_Y(y) = 1 - F(g^{-1}(y))$ gives $f_Y(y) = -f(g^{-1}(y))\frac{dg^{-1}(y)}{dy} = f(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|$. The last equality holds since $\frac{dg^{-1}(y)}{dy}$ is negative.

Now let $Y = g^{-1}(X) = \ln X$, $X > 0$, so that $|dg^{-1}(x)/dx| = 1/x$. Since

$$p_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\},$$

the pdf of $X = g(Y)$ is

$$p_X(x) = \left|\frac{dg^{-1}(x)}{dx}\right| f_Y(g^{-1}(x)) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}.$$

Ex. 7.9: If $X \sim \text{Uniform}[0,1]$, then $f_X(x) = 1$ for all $x \in [0,1]$. Let $y = F^{-1}(x) = g(x)$. Then $x = g^{-1}(y) = F(y)$, and

$$\left|\frac{dg^{-1}(y)}{dy}\right| = f(y)\,.$$

Therefore $f_Y(y) = f(g^{-1}(y))\left|\dfrac{dg^{-1}(y)}{dy}\right| = f(y)$.

```
library(tidyverse)
library(patchwork)
library(latex2exp)
set.seed(123)
df <- data.frame(x=runif(400,0,1)) %>%    # Simulate 400 Uniform[0,1]
  mutate(y=qnorm(x, mean=1, sd=sqrt(2)))  # Convert to Normal(1,2)
p1 <- ggplot(df, aes(x = x)) +
  geom_histogram(aes(y = ..density..), fill = "white", color="black",
                 breaks=seq(0,1,0.1)) +
  scale_x_continuous("X") + theme_bw()
p2 <- ggplot(df, aes(x = y)) +
  geom_histogram(aes(y = ..density..), fill = "white", color="black") +
  stat_function(fun = dnorm, args = with(df, c(mean=1, sd = sqrt(2)))) +
  scale_x_continuous(TeX("$Y=\\Phi^{-1}(1,2)$")) + theme_bw()
p1 | p2
```

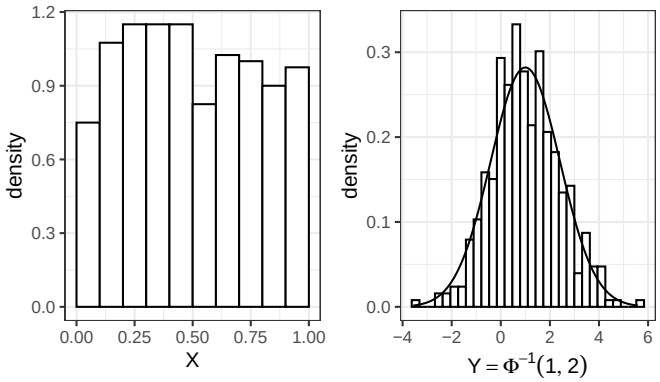

Fig. 7.18. Histogram density estimate of simulated Normal(1,2).

Ex. 7.10: Following the hints given, Jensen's Inequality comes from taking expectations on both sides of the inequality $g(X) \geq l(X) = aX + b$:

$$E(g(X)) \geq E(l(X)) = E(aX + b) = aE(X) + b = g(E(X))\,.$$

Ex. 7.11: The "silly" estimator is unbiased, since $E(\tilde{\mu}) = E(X_1) = \mu$. *Remark: Of course the reason for not throwing away observations is that averaging several observations allows positive and negative errors to cancel, resulting in more precise estimators. This is reflected in the smaller variance of the sample mean σ^2/n. The variance going from one to two observations alone halves the variance.*

Ex. 7.12: $\lim_{n\to\infty} Var(\overline{X}) = \lim_{n\to\infty} \sigma^2/n = 0$. The sample mean tends to get closer to the true mean as sample size increases.

Ex. 7.13: In this example, X_i are ones and zeroes, so $X_i = X_i^2$. Therefore

$$\begin{aligned}\widetilde{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i - \overline{X}^2 = \overline{X}(1-\overline{X}) = \hat{p}(1-\hat{p})\,.\end{aligned}$$

That is,

$$\widetilde{Var(\overline{X})} = \frac{\hat{p}(1-\hat{p})}{n} = \frac{\widetilde{\sigma^2}}{n} \quad \text{where } \widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\,.$$

Since $\widehat{\sigma^2} = \dfrac{n}{n-1}\widetilde{\sigma^2}$, we have

$$\widehat{Var(\hat{p})} = \frac{\widehat{\sigma^2}}{n} = \frac{n}{n-1}\frac{\widetilde{\sigma^2}}{n} = \frac{\hat{p}(1-\hat{p})}{n-1}.$$

Ex. 7.14: The mean of $\widetilde{X} = \sum_{i=1}^{n} w_i X_i$ is

$$E(\widetilde{X}) = \sum_{i=1}^{n} w_i E(X_i) = \mu \sum_{i=1}^{n} w_i = \mu$$

if $\sum_{i=1}^{n} w_i = 1$. Continuing with the assumption that $\sum_{i=1}^{n} w_i = 1$, let

$$w_i = (1/n) + a_i \quad \text{where} \quad \sum_{i=1}^{n} a_i = 0\,.$$

Then

$$\begin{aligned} Var(\widetilde{X}) = Var\left(\sum_{i=1}^{n} w_i X_i\right) &= \sum_{i=1}^{n} w_i^2\, Var(X_i) = \sigma^2 \sum_{i=1}^{n} w_i^2 \\ &= \sigma^2 \sum_{i=1}^{n} (a_i + 1/n)^2 = \sigma^2 \sum_{i=1}^{n} (a_i^2 - 2a_i/n + 1/n^2) \\ &= \sigma^2 \left(\sum_{i=1}^{n} a_i^2 - (2/n)\sum_{i=1}^{n} a_i + 1/n\right) = \sigma^2 \sum_{i=1}^{n} a_i^2 + \sigma^2/n \\ &= \sigma^2 \sum_{i=1}^{n} a_i^2 + Var(\overline{X}) \geq Var(\overline{X})\,. \end{aligned}$$

Ex. 7.15: (a) Calculate sample mean and variance of $earn_i$ and $\ln(earn_i)$:

```
library(tidyverse)
df2 <- read_csv("data\\earnings2019.csv", show_col_types=FALSE)
earnstats <- df2 %>%
  summarise(earn_mean = mean(earn),
            logearn_mean = mean(log(earn)),
            earn_var = var(earn), logearn_var = var(log(earn)))
earnstats
```

```
# A tibble: 1 x 4
  earn_mean logearn_mean earn_var logearn_var
      <dbl>        <dbl>    <dbl>       <dbl>
1      29.2         3.15     671.       0.426
```

(b) Compute exponential of sample mean of $\ln(earn_i)$:

```
exp(earnstats$logearn_mean)
```

```
[1] 23.36102
```

The natural exponent of the sample mean of $\ln(earn_i)$ is considerably lower than the sample mean of $earn_i$. This is a consequence of Jensen's Inequality.

(c) We can estimate the mean of $earn$ using $\exp\left\{\hat{\mu} + \frac{\widehat{\sigma^2}}{2}\right\}$ where $\hat{\mu}$ and $\widehat{\sigma^2}$ are the sample mean and sample variance of $\ln(earn)$. We get

```
exp(earnstats$logearn_mean + earnstats$logearn_var/2)
```

```
[1] 28.90756
```

The adjusted estimate is more in line with the sample mean of $earn_i$.

(d) Calculate sample skewness and kurtosis coefficients to check appropriateness of the Normal assumption for $\ln(earn_i)$:

```
skew_kurt <- function(x){
  n <- length(x)
  s <- (1/n)*sum((x-mean(x))^3)/(sd(x)^3)
  k <- (1/n)*sum((x-mean(x))^4)/(var(x)^2)
  return(c(skewness = s, kurtosis=k))
}
print(skew_kurt(log(df2$earn)))
```

```
  skewness   kurtosis
0.05108766 4.12674254
```

The skewness coefficient is close to zero. The kurtosis coefficient is larger than three, though only slightly. To form an opinion on whether or not the Normality assumption is appropriate, it would be better if we can test if the skewness and kurtosis coefficients are different from 0 and 3 respectively. Nonetheless, Normality assumption appears to be reasonable.

Ex. 7.16: (a) Since $\widetilde{\sigma^2} = \dfrac{n-1}{n}\widehat{\sigma^2}$, we have

$$Var(\widetilde{\sigma^2}) = \frac{(n-1)^2}{n^2} Var(\widehat{\sigma^2}) < Var(\widehat{\sigma^2}) .$$

(b) An expression for $Var(\widetilde{\sigma^2})$ is

$$Var(\widetilde{\sigma^2}) = \frac{(n-1)^2}{n^2} Var(\widehat{\sigma^2}) = \frac{2(n-1)}{n^2}\sigma^4.$$

Furthermore, $E(\widetilde{\sigma^2}) = \dfrac{n-1}{n}E(\widehat{\sigma^2}) = \dfrac{n-1}{n}\sigma^2$, therefore the bias of $\widetilde{\sigma^2}$ is

$$Bias(\widetilde{\sigma^2}) = \sigma^2 - \frac{n-1}{n}\sigma^2 = \frac{1}{n}\sigma^2 .$$

Therefore $MSE(\widetilde{\sigma^2}) = Var(\widetilde{\sigma^2}) + Bias(\widetilde{\sigma^2})^2 = \dfrac{2n-1}{n^2}\sigma^4$.

(c) We have $MSE(\widehat{\sigma^2}) = \dfrac{2\sigma^4}{n-1} = \dfrac{2n\sigma^4}{n^2-n} > \dfrac{(2n-1)\sigma^4}{n^2} = MSE(\widetilde{\sigma^2})$.

Ex. 7.18: We have

$$\begin{aligned} Cov(X,Y) &= E((X - E(X)(Y - E(Y)) \\ &= E((XY - E(X)Y - E(Y)X + E(X)E(Y))) \\ &= E(XY) - E(E(X)Y) - E(E(Y)X) + E(E(X)E(Y)) \\ &= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) . \end{aligned}$$

Ex. 7.19: We have

$$\begin{aligned} Var(\Sigma_{i=1}^3 a_i X_i) &= E((\Sigma_{i=1}^3 a_i X_i)^2) - (E(\Sigma_{i=1}^3 a_i X_i)^2 \\ &= E(\Sigma_{i=1}^3 \Sigma_{j=1}^3 a_i a_j X_i X_j) - (\Sigma_{i=1}^3 a_i E(X_i))^2 \\ &= \Sigma_{i=1}^3 \Sigma_{j=1}^3 a_i a_j E(X_i X_j) - \Sigma_{i=1}^3 \Sigma_{j=1}^3 a_i a_j E(X_i) E(X_j) \\ &= \Sigma_{i=1}^3 \Sigma_{j=1}^3 a_i a_j (E(X_i X_j) - E(X_i) E(X_j)) \\ &= \Sigma_{i=1}^3 \Sigma_{j=1}^3 a_i a_j Cov(X_i, X_j)\,. \end{aligned}$$

Ex. 7.20: The proof is similar to that of Ex. 7.19.

$$Cov(a_1 X_1 + a_2 X_2, b_1 Y_1 + b_2 Y_2 + b_3 Y_3) = \sum_{i=1}^{2} \sum_{j=1}^{3} a_i b_j Cov(X_i, Y_j).$$

Ex. 7.21: We showed in the text that $Cov(X,Y) = 1$, $Var(X) = 2$ and $Var(Y) = 0.625$. Therefore

$$Cov(X,Y) = \frac{Var(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{1}{\sqrt{2}\sqrt{0.625}} = 0.8944\,.$$

Ex. 7.22: We have

$$Var(X - \alpha Y) = Var(X) + \alpha^2 Var(Y) - 2\alpha\ Cov(X,Y) \geq 0 \ \text{ for all } \ \alpha\,.$$

Setting $\alpha = Cov(X,Y)/Var(Y)$ gives

$$Var(X) + \frac{Cov(X,Y)^2}{Var(Y)} - 2\frac{Cov(X,Y)^2}{Var(Y)} \geq 0.$$

Rearranging and taking square roots gives the desired result. *Remark: See Ex. 4.8 for the sample correlation coefficient version of this result, which is a consequence of the Cauchy-Schwarz Inequality for vectors with a finite number of terms. The result proved in Ex. 7.22 is a consequence of a more general version of the Cauchy-Schwarz Inequality.*

Ex. 7.23: We have

$$\begin{aligned} \Pr(Y_i \mid X \geq 3) &= \frac{\Pr(Y = i, X \geq 3)}{\Pr(X \geq 3)} \\ &= \frac{\Pr(Y = i, X_3 = 3) + \Pr(Y = i, X = 4) + \Pr(Y - i, X = 5)}{\Pr(X = 3) + \Pr(X = 4) + \Pr(X = 5)}\,. \end{aligned}$$

Since $\Pr(X \geq 3) = 0.6$, we have

y	3	3.5	4	4.5	5	5.5	6
$\Pr(Y = y \mid X \geq 3)$	0	0	$\frac{1}{12}$	$\frac{3}{12}$	$\frac{4}{20}$	$\frac{3}{20}$	$\frac{1}{20}$

with $E(Y \mid X \geq 3) = 5$ and $Var(Y \mid X \geq 3) = \frac{7}{24}$.

Ex. 7.24: (a) X and Y both have uniform marginal distributions:

y	3	3.5	4	4.5	5	5.5	6
$\Pr(Y=y)$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

and

x	1	2	3	4	5
$\Pr(X=x)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

with means and variances $E(Y) = 4.5$, $E(X) = 3$, $Var(Y) = 0.7$ and $Var(X) = 2$.

(b) The conditional distribution of Y given X is as shown below

$\Pr(Y = y \mid X = x)$

y \ x	1	2	3	4	5
10	0	0	0	0	0.5
9	0	0	0	0.5	0
8	0	0	0.5	0	0
7	0	0.5	0	0	0
6	0.5	0	0	0	0
5	0.5	0	0	0	0
4	0	0.5	0	0	0
3	0	0	0.5	0	0
2	0	0	0	0.5	0
1	0	0	0	0	0.5

with $E(Y \mid X = x) = 0.55$ for all x, and $Var(Y \mid X = x) = (x - 0.5)^2$. As for the conditional distribution of X given Y, we see that $X = |y - 5.5| + 0.5$ with probability 1 when $Y = y$ for $y = 1, 2, 3, \ldots, 10$.

(c) We know $Cov(X, Y) = 0$ since $E(Y \mid X)$ does not vary with X. Note here that X and Y are clearly "related" despite $Cov(X, Y) = 0$, e.g., the conditional variance of Y is increasing in X. Furthermore, note that here $E(X \mid Y)$ *does* vary with Y, despite $Cov(X, Y) = 0$.

Ex. 7.25: Taking expectation over X on both sides of $E(Y \mid X) = a + bX$ gives

$$E(E(Y \mid X)) = E(Y) = a + bE(X)\,.$$

We also have

$$E(YX) = E(E(YX \mid X)) = E(XE(Y \mid X)) = E(aX + bX^2) = aE(X) + bE(X^2)\,.$$

Therefore

$$Cov(X, Y) = E(YX) - E(X)E(Y) = b(E(X^2) - E(X)^2) = b\,Var(X)$$

and $b = Cov(Y, X)/\,Var(X)$ follows. If you know that $E(Y \mid X) = 3 + 0.5X$ and $Var(X) = 2$, then $Cov(X, Y) = 1$ (compare with covariance calculation made for the joint pdf (7.29)).

Ex. 7.26: The relationship follows from

$$\begin{aligned} Var(E(Y \mid X)) &= E(E(Y \mid X)^2) - E(E(Y \mid X))^2 \\ &= E(E(Y^2 \mid X) - Var(Y \mid X)) - E(E(Y \mid X))^2 \\ &= E(Y^2) - E(\,Var(Y \mid X)) - E(Y)^2 \\ &= Var(Y) - E(\,Var(Y \mid X))\,. \end{aligned}$$

(a) If $E(Y \mid X)$ is constant, then $Var(E(Y \mid X)) = 0$, so $Var(Y) = E(Var(Y \mid X))$.
(b) If $Var(Y \mid X)$ is constant, then $E(Var(Y \mid X)) = Var(Y \mid X)$, so we have

$$Var(E(Y \mid X)) = Var(Y) - Var(Y \mid X) \geq 0$$

since the LHS is cannot be negative.

Ex. 7.27: The variables X and Y in this question are independent (you can verify that the conditional distribution of Y does not change with X, nor does the conditional distribution of X given Y). But they are not identical: the marginal distribution of X and Y are

x	1	2	3	4	5
$\Pr(X = x)$	0.1	0.4	0.3	0.1	0.1

and

y	1	2	3	4	5
$\Pr(Y = y)$	0.1	0.2	0.4	0.2	0.1

They would be independent and identically distributed if, for example, their joint pdf is

	5	0.01	0.04	0.03	0.01	0.01
	4	0.01	0.04	0.03	0.01	0.01
Y	3	0.03	0.12	0.09	0.03	0.03
	2	0.04	0.16	0.12	0.04	0.04
	1	0.01	0.04	0.03	0.01	0.01
		1	2	3	4	5
				X		

Ex. 7.28: We showed that the expression for the bivariate Normal pdf is

$$f_{X,Y}(x,y) = \underbrace{\frac{1}{\sqrt{2\pi}\sqrt{\sigma_X^2(1-\rho_{XY}^2)}} \exp\left\{-\frac{1}{2}\frac{[x-(\alpha+\beta y)]^2}{\sigma_X^2(1-\rho_{XY}^2)}\right\}}_{A} \underbrace{\frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}}_{B}$$

where $\alpha = \mu_X - \beta\mu_Y$ and $\beta = \dfrac{\sigma_{XY}}{\sigma_Y^2}$. If $\rho_{XY} = 0$, expression (A) becomes

$$\frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}\right\}.$$

This is just the marginal pdf of X and (B) is the marginal pdf of Y. The joint pdf of X and Y is the product of the marginals, so X and Y are independent.

Ex. 7.29:

i. $\Pr(X \leq -2.5)$ when $X \sim N(0,1)$. Ans: 0.0062097.
ii. $\Pr(X \leq -2.5)$ when $X \sim t(5)$. Ans: 0.027245.
iii. c such that $\Pr(X > c) = 0.05$ when $X \sim \chi^2(5)$. Ans: 11.0704977.
iv. $\Pr(-1.96 \leq X \leq 1.96)$ when $X \sim N(0,1)$. Ans: 0.9500042.
v. c such that $\Pr(-c \leq X \leq c) = 0.95$ when $X \sim N(0,1)$. -1.959964.
vi. c such that $\Pr(-c \leq X \leq c) = 0.95$ when $X \sim t(12)$. Ans: -2.1788128.
vii. c such that $\Pr(-c \leq X \leq c) = 0.95$ when $X \sim t(100)$. Ans: -1.9839715.

viii. c such that $\Pr(X > c) = 0.05$ when $X \sim F(5, 8)$. Ans: 3.6874987.

ix. c such that $\Pr(X > c) = 0.05$ when $X \sim F(5, 80)$. Ans: 2.3287206.

x. c such that $\Pr(X > c) = 0.05$ when $X \sim F(5, 8000)$. Ans: 2.2152166

Compare the output for viii, ix and x with iii. The $df_1 \times F(df_1, df_2)$ distribution converges to the $\chi(df_1)^2$ distribution. This is the reason 5 times `qf(0.95, df1=5, df2)` converges to `qchisq(0.95, df1=5)` as df_2 gets larger.

Ex. 7.30: We have $X \sim$ Log-Normal(μ, σ^2). We wish to estimate $E(X) = \exp\{\mu + \sigma^2/2\}$ from sample mean and variance calculated from a random sample of $Y_i = \ln X_i$, $i = 1, 2, \ldots, n$. Since $Y_i \sim$ Normal(μ, σ^2), we know that $\overline{Y} \xrightarrow{p} \mu$ and $\widehat{\sigma^2} \xrightarrow{p} \sigma^2$, where $\widehat{\sigma^2}$ is $\frac{1}{n-1}\Sigma_{i=1}^{n}(Y_i - \overline{Y})^2$. Therefore

$$\exp\left\{\overline{Y} + \frac{\widehat{\sigma^2}}{2}\right\} \xrightarrow{p} \exp\left\{\mu + \frac{\sigma^2}{2}\right\}.$$

Ex. 7.31: We calculate the skewness and kurtosis coefficients using the `skew_kurt` function from the answer in Ex. 7.15. The following code calculates the bootstrap standard errors.

```
set.seed(8)
X <- matrix(rchisq(20000, df=1), ncol = 400)  # 400 samples, 50 obs each
mu <- cbind(sqrt(5)*(colMeans(X[1:5,]) - 1),
            sqrt(10)*(colMeans(X[1:10,]) - 1),
            sqrt(20)*(colMeans(X[1:20,]) - 1),
            sqrt(50)*(colMeans(X[1:50,]) - 1))
results <- rbind(
    c(mean=mean(mu[,1]), variance=var(mu[,1]), skew_kurt(mu[,1])),
    c(mean=mean(mu[,2]), variance=var(mu[,2]), skew_kurt(mu[,2])),
    c(mean=mean(mu[,3]), variance=var(mu[,3]), skew_kurt(mu[,3])),
    c(mean=mean(mu[,4]), variance=var(mu[,4]), skew_kurt(mu[,4])))
rownames(results) <- c("n:5", "n:10", "n:20", "n:50")
results
```

```
           mean variance  skewness kurtosis
n:5   0.08128763 1.830255 1.1923771 6.605085
n:10  0.01968846 1.774488 0.7683145 4.018212
n:20  0.01203637 1.960076 0.6708130 4.451202
n:50 -0.02505807 1.937929 0.4067155 3.135972
```

If X_i is iid with mean μ and variance σ^2, then $\sqrt{n}(\overline{X}_n - \mu)$ has mean 0 and variance σ^2. Furthermore, the Central Limit Theorem predicts that

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \text{Normal}(\mu, \sigma^2).$$

In our simulation, $X_i \sim \chi^2(1)$, so $\mu = 1$ and $\sigma^2 = 2$. Furthermore, the $\chi^2(1)$ distribution is positively skewed, so the distribution of the mean should also be positively skewed and non-Normal. However, the CLT says that the distribution of the sample mean should be approximately Normal for large sample sizes. This is reflected in our results, where the mean and variances are close to 0 and

2 respectively. Furthermore, our skewness and kurtosis are closer to 0 and 3 respectively for larger sample sizes.

Ex. 7.32: See the following code. We assume you have the data stored in `df2` and have available the function `skew_kurt()` as in the answer to Ex. 7.15. We calculate our bootstrap standard errors based on 200 bootstrap samples.

```
## earnings data in df2$earn
skreps <- data.frame(skewness_se=rep(NA, 200), kurtosis_se=rep(NA, 200))
n <- length(df2$earn)
for (i in 1:200){
  x <- sample(log(df2$earn), n, replace=TRUE)
  skreps[i,] <- skew_kurt(x)
}
sk <- rbind(skew_kurt(log(df2$earn)), sapply(skreps, sd))
rownames(sk) <- c("Est.", "Std.err.")
print(sk)
```

```
          skewness  kurtosis
Est.     0.05108766 4.1267425
Std.err. 0.06022960 0.1881308
```

We do not reject skewness $= 0$ 0.05 level of significance but we would reject kurtosis $= 3$ at 0.05 level of significance. The distribution is slightly "fat-tailed".