SPRINGER TEXTS IN STATISTICS

All of Statistics

A Concise Course in Statistical Inference

Larry Wasserman

Springer

Larry Wasserman

All of Statistics

A Concise Course in Statistical Inference

With 95 Figures

Springer

Larry Wasserman
Department of Statistics
Carnegie Mellon University
Baker Hall 228A
Pittsburgh, PA 15213-3890
USA
larry@stat.cmu.edu

Library of Congress Cataloging-in-Publication Data
Wasserman, Larry A. (Larry Alan), 1959–
All of statistics: a concise course in statistical inference / Larry a. Wasserman.
p. cm. — (Springer texts in statistics)
Includes bibliographical references and index.
ISBN 0-387-40272-1 (alk. paper)
1. Mathematical statistics. I. Title. II. Series.
QA276.12.W37 2003
519.5—dc21 2003062209

ISBN 0-387-40272-1 Printed on acid-free paper.

Printed in the United States of America. (MVY)

9 8 7 6 5 4 3 (Corrected second printing, 2005)

springeronline.com

Contents

I Probability

II Statistical Inference

III Statistical Models and Methods

Part I

Probability

1 Probability

1.1 Introduction

Probability is a mathematical language for quantifying uncertainty. In this chapter we introduce the basic concepts underlying probability theory. We begin with the sample space, which is the set of possible outcomes.

1.2 Sample Spaces and Events

The **sample space** Ω is the set of possible outcomes of an experiment. Points ω in Ω are called **sample outcomes**, **realizations**, or **elements**. Subsets of Ω are called **Events**.

1.1 Example. If we toss a coin twice then $\Omega = \{HH, HT, TH, TT\}$. The event that the first toss is heads is $A = \{HH, HT\}$. ■

1.2 Example. Let ω be the outcome of a measurement of some physical quantity, for example, temperature. Then $\Omega = \mathbb{R} = (-\infty, \infty)$. One could argue that taking $\Omega = \mathbb{R}$ is not accurate since temperature has a lower bound. But there is usually no harm in taking the sample space to be larger than needed. The event that the measurement is larger than 10 but less than or equal to 23 is $A = (10, 23]$. ■

1.3 Example. If we toss a coin forever, then the sample space is the infinite set

$$\Omega = \Big\{\omega = (\omega_1, \omega_2, \omega_3, \ldots,) : \ \omega_i \in \{H, T\}\Big\}.$$

Let E be the event that the first head appears on the third toss. Then

$$E = \Big\{(\omega_1, \omega_2, \omega_3, \ldots,) : \ \omega_1 = T, \omega_2 = T, \omega_3 = H, \ \omega_i \in \{H, T\} \text{ for } i > 3\Big\}. \ \blacksquare$$

Given an event A, let $A^c = \{\omega \in \Omega : \ \omega \notin A\}$ denote the complement of A. Informally, A^c can be read as "not A." The complement of Ω is the empty set $\emptyset$. The union of events A and B is defined

$$A \bigcup B = \{\omega \in \Omega : \ \omega \in A \text{ or } \omega \in B \text{ or } \omega \in \text{both}\}$$

which can be thought of as "A or B." If $A_1, A_2, \ldots$ is a sequence of sets then

$$\bigcup_{i=1}^{\infty} A_i = \Big\{\omega \in \Omega : \ \omega \in A_i \text{ for at least one i}\Big\}.$$

The intersection of A and B is

$$A \bigcap B = \{\omega \in \Omega : \ \omega \in A \text{ and } \omega \in B\}$$

read "A and B." Sometimes we write $A \bigcap B$ as AB or (A, B). If $A_1, A_2, \ldots$ is a sequence of sets then

$$\bigcap_{i=1}^{\infty} A_i = \Big\{\omega \in \Omega : \ \omega \in A_i \text{ for all i}\Big\}.$$

The set difference is defined by $A - B = \{\omega : \ \omega \in A, \omega \notin B\}$. If every element of A is also contained in B we write $A \subset B$ or, equivalently, $B \supset A$. If A is a finite set, let $|A|$ denote the number of elements in A. See the following table for a summary.

Summary of Terminology

Ω	sample space
ω	outcome (point or element)
A	event (subset of Ω)
A^c	complement of A (not A)
$A \bigcup B$	union (A or B)
$A \bigcap B$ or AB	intersection (A and B)
$A - B$	set difference (ω in A but not in B)
$A \subset B$	set inclusion
$\emptyset$	null event (always false)
Ω	true event (always true)

We say that $A_1, A_2, \ldots$ are **disjoint** or are **mutually exclusive** if $A_i \bigcap A_j = \emptyset$ whenever $i \neq j$. For example, $A_1 = [0,1), A_2 = [1,2), A_3 = [2,3), \ldots$ are disjoint. A **partition** of Ω is a sequence of disjoint sets $A_1, A_2, \ldots$ such that $\bigcup_{i=1}^{\infty} A_i = \Omega$. Given an event A, define the **indicator function of** A by

$$I_A(\omega) = I(\omega \in A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

A sequence of sets $A_1, A_2, \ldots$ is **monotone increasing** if $A_1 \subset A_2 \subset \cdots$ and we define $\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$. A sequence of sets $A_1, A_2, \ldots$ is **monotone decreasing** if $A_1 \supset A_2 \supset \cdots$ and then we define $\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i$. In either case, we will write $A_n \to A$.

1.4 Example. Let $\Omega = \mathbb{R}$ and let $A_i = [0, 1/i)$ for $i = 1, 2, \ldots$. Then $\bigcup_{i=1}^{\infty} A_i = [0,1)$ and $\bigcap_{i=1}^{\infty} A_i = \{0\}$. If instead we define $A_i = (0, 1/i)$ then $\bigcup_{i=1}^{\infty} A_i = (0,1)$ and $\bigcap_{i=1}^{\infty} A_i = \emptyset$. ■

1.3 Probability

We will assign a real number $\mathbb{P}(A)$ to every event A, called the **probability** of A. [1] We also call $\mathbb{P}$ a **probability distribution** or a **probability measure.** To qualify as a probability, $\mathbb{P}$ must satisfy three axioms:

1.5 Definition. *A function* $\mathbb{P}$ *that assigns a real number* $\mathbb{P}(A)$ *to each event* A *is a* **probability distribution** *or a* **probability measure** *if it satisfies the following three axioms:*

Axiom 1: $\mathbb{P}(A) \geq 0$ *for every* A

Axiom 2: $\mathbb{P}(\Omega) = 1$

Axiom 3: *If* $A_1, A_2, \ldots$ *are disjoint then*

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

[1] It is not always possible to assign a probability to every event A if the sample space is large, such as the whole real line. Instead, we assign probabilities to a limited class of set called a σ-field. See the appendix for details.

There are many interpretations of $\mathbb{P}(A)$. The two common interpretations are frequencies and degrees of beliefs. In the frequency interpretation, $\mathbb{P}(A)$ is the long run proportion of times that A is true in repetitions. For example, if we say that the probability of heads is $1/2$, we mean that if we flip the coin many times then the proportion of times we get heads tends to $1/2$ as the number of tosses increases. An infinitely long, unpredictable sequence of tosses whose limiting proportion tends to a constant is an idealization, much like the idea of a straight line in geometry. The degree-of-belief interpretation is that $\mathbb{P}(A)$ measures an observer's strength of belief that A is true. In either interpretation, we require that Axioms 1 to 3 hold. The difference in interpretation will not matter much until we deal with statistical inference. There, the differing interpretations lead to two schools of inference: the frequentist and the Bayesian schools. We defer discussion until Chapter 11.

One can derive many properties of $\mathbb{P}$ from the axioms, such as:

$$\begin{array}{rcl}
\mathbb{P}(\emptyset) & = & 0 \\
A \subset B & \Longrightarrow & \mathbb{P}(A) \leq \mathbb{P}(B) \\
0 \leq & \mathbb{P}(A) & \leq 1 \\
\mathbb{P}(A^c) & = & 1 - \mathbb{P}(A) \\
A \bigcap B = \emptyset & \Longrightarrow & \mathbb{P}\left(A \bigcup B\right) = \mathbb{P}(A) + \mathbb{P}(B).
\end{array} \tag{1.1}$$

A less obvious property is given in the following Lemma.

1.6 Lemma. *For any events A and B,*

$$\mathbb{P}\left(A \bigcup B\right) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB).$$

PROOF. Write $A \bigcup B = (AB^c) \bigcup (AB) \bigcup (A^cB)$ and note that these events are disjoint. Hence, making repeated use of the fact that $\mathbb{P}$ is additive for disjoint events, we see that

$$\begin{array}{rcl}
\mathbb{P}\left(A \bigcup B\right) & = & \mathbb{P}\left((AB^c) \bigcup (AB) \bigcup (A^cB)\right) \\
& = & \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^cB) \\
& = & \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^cB) + \mathbb{P}(AB) - \mathbb{P}(AB) \\
& = & P\left((AB^c) \bigcup (AB)\right) + \mathbb{P}\left((A^cB) \bigcup (AB)\right) - \mathbb{P}(AB) \\
& = & \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB). \ \blacksquare
\end{array}$$

1.7 Example. Two coin tosses. Let H_1 be the event that heads occurs on toss 1 and let H_2 be the event that heads occurs on toss 2. If all outcomes are these objects is $n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$. For convenience, we define $0! = 1$. We also define

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \tag{1.2}$$

read "n choose k", which is the number of distinct ways of choosing k objects from n. For example, if we have a class of 20 people and we want to select a committee of 3 students, then there are

$$\binom{20}{3} = \frac{20!}{3!17!} = \frac{20\times 19\times 18}{3\times 2\times 1} = 1140$$

possible committees. We note the following properties:

$$\binom{n}{0} = \binom{n}{n} = 1 \quad \text{and} \quad \binom{n}{k} = \binom{n}{n-k}.$$

1.5 Independent Events

If we flip a fair coin twice, then the probability of two heads is $\frac{1}{2}\times\frac{1}{2}$. We multiply the probabilities because we regard the two tosses as independent. The formal definition of independence is as follows:

1.9 Definition. *Two events A and B are* **independent** *if*

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B) \tag{1.3}$$

and we write $A \amalg B$. A set of events $\{A_i :\ i \in I\}$ is independent if

$$\mathbb{P}\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J} \mathbb{P}(A_i)$$

for every finite subset J of I. If A and B are not independent, we write

$$A \frown\!\!\!\!\circ\!\circ\!\circ\!\circ\!\!\!\!\frown B$$

Independence can arise in two distinct ways. Sometimes, we explicitly **assume** that two events are independent. For example, in tossing a coin twice, we usually assume the tosses are independent which reflects the fact that the coin has no memory of the first toss. In other instances, we **derive** independence by verifying that $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$ holds. For example, in tossing a fair die, let $A = \{2,4,6\}$ and let $B = \{1,2,3,4\}$. Then, $A \bigcap B = \{2,4\}$,

Summary of Independence

1. A and B are independent if and only if $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$.
2. Independence is sometimes assumed and sometimes derived.
3. Disjoint events with positive probability are not independent.

1.6 Conditional Probability

Assuming that $\mathbb{P}(B) > 0$, we define the conditional probability of A given that B has occurred as follows:

1.12 Definition. *If* $\mathbb{P}(B) > 0$ *then the* **conditional probability** *of* A *given* B *is*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}. \tag{1.4}$$

Think of $\mathbb{P}(A|B)$ as the fraction of times A occurs among those in which B occurs. For any fixed B such that $\mathbb{P}(B) > 0$, $\mathbb{P}(\cdot|B)$ is a probability (i.e., it satisfies the three axioms of probability). In particular, $\mathbb{P}(A|B) \geq 0$, $\mathbb{P}(\Omega|B) = 1$ and if $A_1, A_2, \ldots$ are disjoint then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$. But it is in general **not** true that $\mathbb{P}(A|B \bigcup C) = \mathbb{P}(A|B) + \mathbb{P}(A|C)$. The rules of probability apply to events on the left of the bar. In general it is **not** the case that $\mathbb{P}(A|B) = \mathbb{P}(B|A)$. People get this confused all the time. For example, the probability of spots given you have measles is 1 but the probability that you have measles given that you have spots is not 1. In this case, the difference between $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ is obvious but there are cases where it is less obvious. This mistake is made often enough in legal cases that it is sometimes called the prosecutor's fallacy.

1.13 Example. A medical test for a disease D has outcomes $+$ and $-$. The probabilities are:

	D	D^c
$+$	.009	.099
$-$	.001	.891

From the definition of conditional probability,

$$\mathbb{P}(+|D) = \frac{\mathbb{P}(+\bigcap D)}{\mathbb{P}(D)} = \frac{.009}{.009 + .001} = .9$$

and

$$\mathbb{P}(-|D^c) = \frac{\mathbb{P}(-\bigcap D^c)}{\mathbb{P}(D^c)} = \frac{.891}{.891 + .099} = .9.$$

Apparently, the test is fairly accurate. Sick people yield a positive 90 percent of the time and healthy people yield a negative about 90 percent of the time. Suppose you go for a test and get a positive. What is the probability you have the disease? Most people answer .90. The correct answer is

$$\mathbb{P}(D|+) = \frac{\mathbb{P}(+\bigcap D)}{\mathbb{P}(+)} = \frac{.009}{.009 + .099} \approx .08.$$

The lesson here is that you need to compute the answer numerically. Don't trust your intuition. ■

The results in the next lemma follow directly from the definition of conditional probability.

1.14 Lemma. *If A and B are independent events then $\mathbb{P}(A|B) = \mathbb{P}(A)$. Also, for any pair of events A and B,*

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

From the last lemma, we see that another interpretation of independence is that knowing B doesn't change the probability of A. The formula $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B|A)$ is sometimes helpful for calculating probabilities.

1.15 Example. Draw two cards from a deck, without replacement. Let A be the event that the first draw is the Ace of Clubs and let B be the event that the second draw is the Queen of Diamonds. Then $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B|A) = (1/52) \times (1/51)$. ■

Summary of Conditional Probability

1. If $\mathbb{P}(B) > 0$, then
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}.$$

2. $\mathbb{P}(\cdot|B)$ satisfies the axioms of probability, for fixed B. In general, $\mathbb{P}(A|\cdot)$ does not satisfy the axioms of probability, for fixed A.

3. In general, $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.

4. A and B are independent if and only if $\mathbb{P}(A|B) = \mathbb{P}(A)$.

1.7 Bayes' Theorem

Bayes' theorem is the basis of "expert systems" and "Bayes' nets," which are discussed in Chapter 17. First, we need a preliminary result.

1.16 Theorem (The Law of Total Probability). *Let $A_1, \ldots, A_k$ be a partition of Ω. Then, for any event B,*

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

PROOF. Define $C_j = BA_j$ and note that $C_1, \ldots, C_k$ are disjoint and that $B = \bigcup_{j=1}^{k} C_j$. Hence,

$$\mathbb{P}(B) = \sum_j \mathbb{P}(C_j) = \sum_j \mathbb{P}(BA_j) = \sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)$$

since $\mathbb{P}(BA_j) = \mathbb{P}(B|A_j)\mathbb{P}(A_j)$ from the definition of conditional probability. ■

1.17 Theorem (Bayes' Theorem). *Let $A_1, \ldots, A_k$ be a partition of Ω such that $\mathbb{P}(A_i) > 0$ for each i. If $\mathbb{P}(B) > 0$ then, for each $i = 1, \ldots, k$,*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}. \qquad (1.5)$$

1.18 Remark. We call $\mathbb{P}(A_i)$ the **prior probability of** A and $\mathbb{P}(A_i|B)$ the **posterior probability of** A.

PROOF. We apply the definition of conditional probability twice, followed by the law of total probability:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_iB)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}. \quad ■$$

1.19 Example. I divide my email into three categories: $A_1 =$ "spam," $A_2 =$ "low priority" and $A_3 =$ "high priority." From previous experience I find that

$\mathbb{P}(A_1) = .7$, $\mathbb{P}(A_2) = .2$ and $\mathbb{P}(A_3) = .1$. Of course, $.7 + .2 + .1 = 1$. Let B be the event that the email contains the word "free." From previous experience, $\mathbb{P}(B|A_1) = .9$, $\mathbb{P}(B|A_2) = .01$, $\mathbb{P}(B|A_3) = .01$. (Note: $.9 + .01 + .01 \neq 1$.) I receive an email with the word "free." What is the probability that it is spam? Bayes' theorem yields,

$$\mathbb{P}(A_1|B) = \frac{.9 \times .7}{(.9 \times .7) + (.01 \times .2) + (.01 \times .1)} = .995. \quad \blacksquare$$

1.8 Bibliographic Remarks

The material in this chapter is standard. Details can be found in any number of books. At the introductory level, there is DeGroot and Schervish (2002); at the intermediate level, Grimmett and Stirzaker (1982) and Karr (1993); at the advanced level there are Billingsley (1979) and Breiman (1992). I adapted many examples and exercises from DeGroot and Schervish (2002) and Grimmett and Stirzaker (1982).

1.9 Appendix

Generally, it is not feasible to assign probabilities to all subsets of a sample space Ω. Instead, one restricts attention to a set of events called a **σ-algebra** or a **σ-field** which is a class $\mathcal{A}$ that satisfies:

(i) $\emptyset \in \mathcal{A}$,

(ii) if $A_1, A_2, \ldots, \in \mathcal{A}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ and

(iii) $A \in \mathcal{A}$ implies that $A^c \in \mathcal{A}$.

The sets in $\mathcal{A}$ are said to be **measurable**. We call $(\Omega, \mathcal{A})$ a **measurable space.** If $\mathbb{P}$ is a probability measure defined on $\mathcal{A}$, then $(\Omega, \mathcal{A}, \mathbb{P})$ is called a **probability space.** When Ω is the real line, we take $\mathcal{A}$ to be the smallest σ-field that contains all the open subsets, which is called the **Borel σ-field.**

1.10 Exercises

1. Fill in the details of the proof of Theorem 1.8. Also, prove the monotone decreasing case.

2. Prove the statements in equation (1.1).

behind one of three doors. You pick a door. To be concrete, let's suppose you always pick door 1. Now Monty Hall chooses one of the other two doors, opens it and shows you that it is empty. He then gives you the opportunity to keep your door or switch to the other unopened door. Should you stay or switch? Intuition suggests it doesn't matter. The correct answer is that you should switch. Prove it. It will help to specify the sample space and the relevant events carefully. Thus write $\Omega = \{(\omega_1, \omega_2) : \omega_i \in \{1, 2, 3\}\}$ where ω_1 is where the prize is and ω_2 is the door Monty opens.

11. Suppose that A and B are independent events. Show that A^c and B^c are independent events.

12. There are three cards. The first is green on both sides, the second is red on both sides and the third is green on one side and red on the other. We choose a card at random and we see one side (also chosen at random). If the side we see is green, what is the probability that the other side is also green? Many people intuitively answer 1/2. Show that the correct answer is 2/3.

13. Suppose that a fair coin is tossed repeatedly until both a head and tail have appeared at least once.

    (a) Describe the sample space Ω.

    (b) What is the probability that three tosses will be required?

14. Show that if $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$ then A is independent of every other event. Show that if A is independent of itself then $\mathbb{P}(A)$ is either 0 or 1.

15. The probability that a child has blue eyes is 1/4. Assume independence between children. Consider a family with 3 children.

    (a) If it is known that at least one child has blue eyes, what is the probability that at least two children have blue eyes?

    (b) If it is known that the youngest child has blue eyes, what is the probability that at least two children have blue eyes?

16. Prove Lemma 1.14.

17. Show that

$$\mathbb{P}(ABC) = \mathbb{P}(A|BC)\mathbb{P}(B|C)\mathbb{P}(C).$$

out a simulation and compare the average of the X's to np. Try this for $p = .3$ and $n = 10$, $n = 100$, and $n = 1,000$.

23. (Computer Experiment.) Here we will get some experience simulating conditional probabilities. Consider tossing a fair die. Let $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$. Then, $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 2/3$ and $\mathbb{P}(AB) = 1/3$. Since $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$, the events A and B are independent. Simulate draws from the sample space and verify that $\widehat{\mathbb{P}}(AB) = \widehat{\mathbb{P}}(A)\widehat{\mathbb{P}}(B)$ where $\widehat{\mathbb{P}}(A)$ is the proportion of times A occurred in the simulation and similarly for $\widehat{\mathbb{P}}(AB)$ and $\widehat{\mathbb{P}}(B)$. Now find two events A and B that are not independent. Compute $\widehat{\mathbb{P}}(A), \widehat{\mathbb{P}}(B)$ and $\widehat{\mathbb{P}}(AB)$. Compare the calculated values to their theoretical values. Report your results and interpret.

2
Random Variables

2.1 Introduction

Statistics and data mining are concerned with data. How do we link sample spaces and events to data? The link is provided by the concept of a random variable.

2.1 Definition. *A* **random variable** *is a mapping*[1]

$$X : \Omega \to \mathbb{R}$$

that assigns a real number $X(\omega)$ to each outcome ω.

At a certain point in most probability courses, the sample space is rarely mentioned anymore and we work directly with random variables. But you should keep in mind that the sample space is really there, lurking in the background.

2.2 Example. Flip a coin ten times. Let $X(\omega)$ be the number of heads in the sequence ω. For example, if $\omega = HHTHHTHHTT$, then $X(\omega) = 6$. ■

[1] Technically, a random variable must be measurable. See the appendix for details.

2.3 Example. Let $\Omega = \left\{(x,y);\ x^2+y^2 \leq 1\right\}$ be the unit disk. Consider drawing a point at random from Ω. (We will make this idea more precise later.) A typical outcome is of the form $\omega = (x,y)$. Some examples of random variables are $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x+y$, and $W(\omega) = \sqrt{x^2+y^2}$. ■

Given a random variable X and a subset A of the real line, define $X^{-1}(A) = \{\omega \in \Omega :\ X(\omega) \in A\}$ and let

$$\begin{aligned} \mathbb{P}(X \in A) &= \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega;\ X(\omega) \in A\}) \\ \mathbb{P}(X = x) &= \mathbb{P}(X^{-1}(x)) = \mathbb{P}(\{\omega \in \Omega;\ X(\omega) = x\}). \end{aligned}$$

Notice that X denotes the random variable and x denotes a particular value of X.

2.4 Example. Flip a coin twice and let X be the number of heads. Then, $\mathbb{P}(X = 0) = \mathbb{P}(\{TT\}) = 1/4$, $\mathbb{P}(X = 1) = \mathbb{P}(\{HT,TH\}) = 1/2$ and $\mathbb{P}(X = 2) = \mathbb{P}(\{HH\}) = 1/4$. The random variable and its distribution can be summarized as follows:

ω	$\mathbb{P}(\{\omega\})$	$X(\omega)$
TT	1/4	0
TH	1/4	1
HT	1/4	1
HH	1/4	2

x	$\mathbb{P}(X = x)$
0	1/4
1	1/2
2	1/4

Try generalizing this to n flips. ■

2.2 Distribution Functions and Probability Functions

Given a random variable X, we define the cumulative distribution function (or distribution function) as follows.

2.5 Definition. *The* **cumulative distribution function**, *or* CDF, *is the function* $F_X : \mathbb{R} \to [0,1]$ *defined by*

$$F_X(x) = \mathbb{P}(X \leq x). \tag{2.1}$$

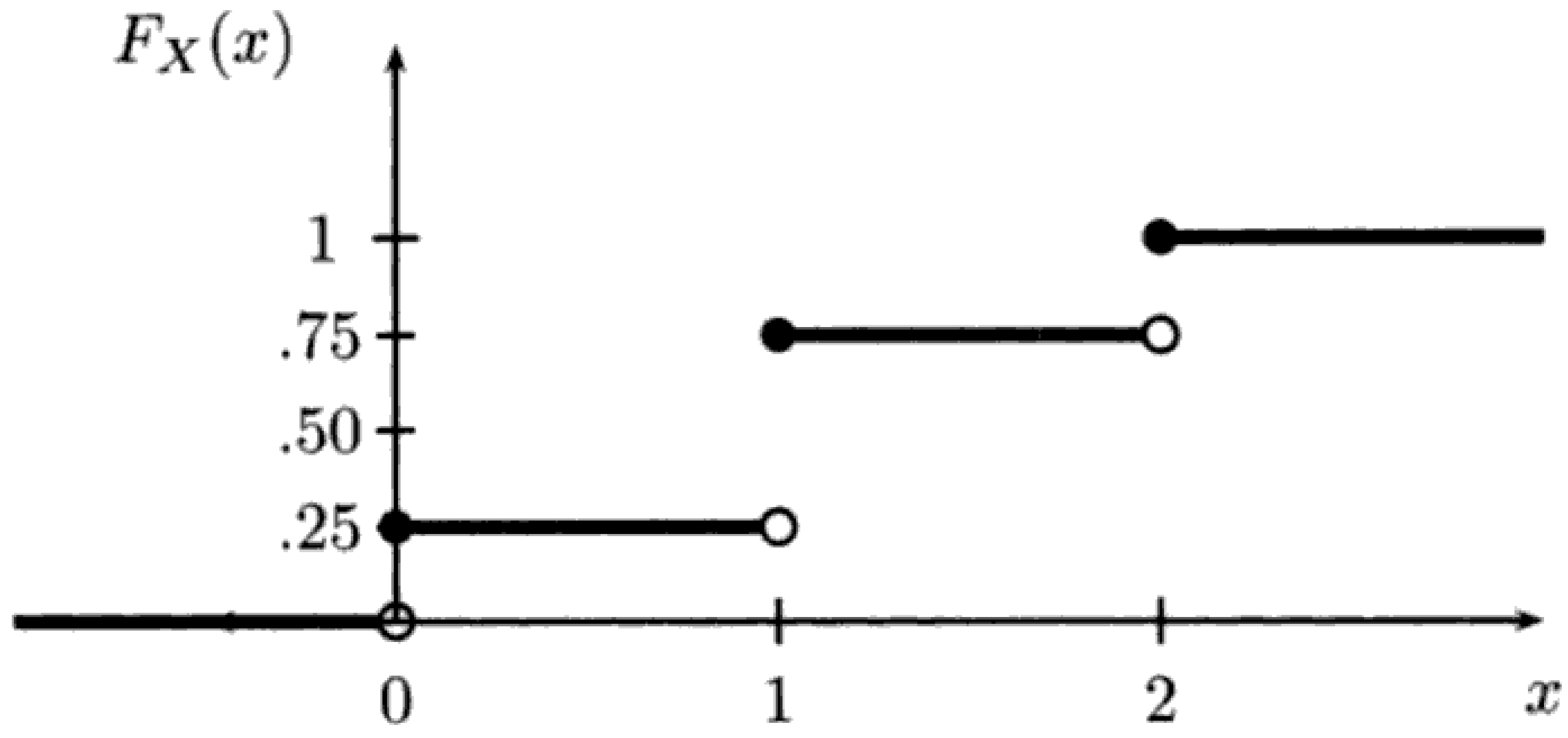

FIGURE 2.1. CDF for flipping a coin twice (Example 2.6.)

We will see later that the CDF effectively contains all the information about the random variable. Sometimes we write the CDF as F instead of F_X.

2.6 Example. Flip a fair coin twice and let X be the number of heads. Then $\mathbb{P}(X = 0) = \mathbb{P}(X = 2) = 1/4$ and $\mathbb{P}(X = 1) = 1/2$. The distribution function is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \le x < 1 \\ 3/4 & 1 \le x < 2 \\ 1 & x \ge 2. \end{cases}$$

The CDF is shown in Figure 2.1. Although this example is simple, study it carefully. CDF's can be very confusing. Notice that the function is right continuous, non-decreasing, and that it is defined for all x, even though the random variable only takes values $0, 1$, and 2. Do you see why $F_X(1.4) = .75$? ■

The following result shows that the CDF completely determines the distribution of a random variable.

2.7 Theorem. *Let X have* CDF *F and let Y have* CDF *G. If $F(x) = G(x)$ for all x, then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all A.* [2]

2.8 Theorem. *A function F mapping the real line to $[0, 1]$ is a* CDF *for some probability $\mathbb{P}$ if and only if F satisfies the following three conditions:*

(i) F is non-decreasing: $x_1 < x_2$ implies that $F(x_1) \le F(x_2)$.

(ii) F is normalized:

$$\lim_{x \to -\infty} F(x) = 0$$

[2] Technically, we only have that $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for every measurable event A.

and

$$\lim_{x\to\infty} F(x) = 1.$$

(iii) F is right-continuous: $F(x) = F(x^+)$ for all x, where

$$F(x^+) = \lim_{\substack{y\to x\\ y>x}} F(y).$$

PROOF. Suppose that F is a CDF. Let us show that (iii) holds. Let x be a real number and let $y_1, y_2, \ldots$ be a sequence of real numbers such that $y_1 > y_2 > \cdots$ and $\lim_i y_i = x$. Let $A_i = (-\infty, y_i]$ and let $A = (-\infty, x]$. Note that $A = \bigcap_{i=1}^{\infty} A_i$ and also note that $A_1 \supset A_2 \supset \cdots$. Because the events are monotone, $\lim_i \mathbb{P}(A_i) = \mathbb{P}(\bigcap_i A_i)$. Thus,

$$F(x) = \mathbb{P}(A) = \mathbb{P}\left(\bigcap_i A_i\right) = \lim_i \mathbb{P}(A_i) = \lim_i F(y_i) = F(x^+).$$

Showing (i) and (ii) is similar. Proving the other direction — namely, that if F satisfies (i), (ii), and (iii) then it is a CDF for some random variable — uses some deep tools in analysis. ■

2.9 Definition. *X is* **discrete** *if it takes countably*[3] *many values $\{x_1, x_2, \ldots\}$. We define the* **probability function** *or* **probability mass function** *for X by $f_X(x) = \mathbb{P}(X = x)$.*

Thus, $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ and $\sum_i f_X(x_i) = 1$. Sometimes we write f instead of f_X. The CDF of X is related to f_X by

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i).$$

2.10 Example. The probability function for Example 2.6 is

$$f_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise.} \end{cases}$$

See Figure 2.2. ■

[3] A set is countable if it is finite or it can be put in a one-to-one correspondence with the integers. The even numbers, the odd numbers, and the rationals are countable; the set of real numbers between 0 and 1 is not countable.

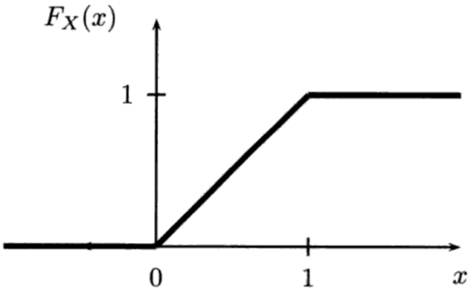

FIGURE 2.3. CDF for Uniform (0,1).

2.13 Example. Suppose that X has PDF

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{(1+x)^2} & \text{otherwise.} \end{cases}$$

Since $\int f(x)dx = 1$, this is a well-defined PDF. ■

Warning! Continuous random variables can lead to confusion. First, note that if X is continuous then $\mathbb{P}(X = x) = 0$ for every x. Don't try to think of $f(x)$ as $\mathbb{P}(X = x)$. This only holds for discrete random variables. We get probabilities from a PDF by integrating. A PDF can be bigger than 1 (unlike a mass function). For example, if $f(x) = 5$ for $x \in [0, 1/5]$ and 0 otherwise, then $f(x) \geq 0$ and $\int f(x)dx = 1$ so this is a well-defined PDF even though $f(x) = 5$ in some places. In fact, a PDF can be unbounded. For example, if $f(x) = (2/3)x^{-1/3}$ for $0 < x < 1$ and $f(x) = 0$ otherwise, then $\int f(x)dx = 1$ even though f is not bounded.

2.14 Example. Let

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{(1+x)} & \text{otherwise.} \end{cases}$$

This is not a PDF since $\int f(x)dx = \int_0^\infty dx/(1+x) = \int_1^\infty du/u = \log(\infty) = \infty$. ■

2.15 Lemma. *Let F be the* CDF *for a random variable X. Then:*

1. $\mathbb{P}(X = x) = F(x) - F(x^-)$ *where* $F(x^-) = \lim_{y \uparrow x} F(y)$*;*

The Point Mass Distribution. X has a point mass distribution at a, written $X \sim \delta_a$, if $\mathbb{P}(X = a) = 1$ in which case

$$F(x) = \begin{cases} 0 & x < a \\ 1 & x \geq a. \end{cases}$$

The probability mass function is $f(x) = 1$ for $x = a$ and 0 otherwise.

The Discrete Uniform Distribution. Let $k > 1$ be a given integer. Suppose that X has probability mass function given by

$$f(x) = \begin{cases} 1/k & \text{for } x = 1, \ldots, k \\ 0 & \text{otherwise.} \end{cases}$$

We say that X has a uniform distribution on $\{1, \ldots, k\}$.

The Bernoulli Distribution. Let X represent a binary coin flip. Then $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ for some $p \in [0, 1]$. We say that X has a Bernoulli distribution written $X \sim \text{Bernoulli}(p)$. The probability function is $f(x) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$.

The Binomial Distribution. Suppose we have a coin which falls heads up with probability p for some $0 \leq p \leq 1$. Flip the coin n times and let X be the number of heads. Assume that the tosses are independent. Let $f(x) = \mathbb{P}(X = x)$ be the mass function. It can be shown that

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

A random variable with this mass function is called a Binomial random variable and we write $X \sim \text{Binomial}(n, p)$. If $X_1 \sim \text{Binomial}(n_1, p)$ and $X_2 \sim \text{Binomial}(n_2, p)$ are independent then $X_1 + X_2 \sim \text{Binomial}(n_1 + n_2, p)$.

Warning! Let us take this opportunity to prevent some confusion. X is a random variable; x denotes a particular value of the random variable; n and p are **parameters**, that is, fixed real numbers. The parameter p is usually unknown and must be estimated from data; that's what statistical inference is all about. In most statistical models, there are random variables and parameters: don't confuse them.

The Geometric Distribution. X has a geometric distribution with parameter $p \in (0, 1)$, written $X \sim \text{Geom}(p)$, if

$$\mathbb{P}(X = k) = p(1-p)^{k-1}, \quad k = 1, 2, 3, \ldots$$

We have that

$$\sum_{k=1}^{\infty} \mathbb{P}(X=k) = p \sum_{k=1}^{\infty} (1-p)^k = \frac{p}{1-(1-p)} = 1.$$

Think of X as the number of flips needed until the first head when flipping a coin.

THE POISSON DISTRIBUTION. X has a Poisson distribution with parameter λ, written $X \sim \text{Poisson}(\lambda)$ if

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad x \geq 0.$$

Note that

$$\sum_{x=0}^{\infty} f(x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1.$$

The Poisson is often used as a model for counts of rare events like radioactive decay and traffic accidents. If $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ are independent then $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Warning! We defined random variables to be mappings from a sample space Ω to $\mathbb{R}$ but we did not mention the sample space in any of the distributions above. As I mentioned earlier, the sample space often "disappears" but it is really there in the background. Let's construct a sample space explicitly for a Bernoulli random variable. Let $\Omega = [0,1]$ and define $\mathbb{P}$ to satisfy $\mathbb{P}([a,b]) = b - a$ for $0 \leq a \leq b \leq 1$. Fix $p \in [0,1]$ and define

$$X(\omega) = \begin{cases} 1 & \omega \leq p \\ 0 & \omega > p. \end{cases}$$

Then $\mathbb{P}(X=1) = \mathbb{P}(\omega \leq p) = \mathbb{P}([0,p]) = p$ and $\mathbb{P}(X=0) = 1-p$. Thus, $X \sim \text{Bernoulli}(p)$. We could do this for all the distributions defined above. In practice, we think of a random variable like a random number but formally it is a mapping defined on some sample space.

2.4 Some Important Continuous Random Variables

THE UNIFORM DISTRIBUTION. X has a Uniform(a,b) distribution, written $X \sim \text{Uniform}(a,b)$, if

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

where $a < b$. The distribution function is

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x > b. \end{cases}$$

Normal (Gaussian). X has a Normal (or Gaussian) distribution with parameters μ and σ, denoted by $X \sim N(\mu, \sigma^2)$, if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad x \in \mathbb{R} \tag{2.3}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. The parameter μ is the "center" (or mean) of the distribution and σ is the "spread" (or standard deviation) of the distribution. (The mean and standard deviation will be formally defined in the next chapter.) The Normal plays an important role in probability and statistics. Many phenomena in nature have approximately Normal distributions. Later, we shall study the Central Limit Theorem which says that the distribution of a sum of random variables can be approximated by a Normal distribution.

We say that X has a **standard Normal distribution** if $\mu = 0$ and $\sigma = 1$. Tradition dictates that a standard Normal random variable is denoted by Z. The PDF and CDF of a standard Normal are denoted by $\phi(z)$ and $\Phi(z)$. The PDF is plotted in Figure 2.4. There is no closed-form expression for Φ. Here are some useful facts:

(i) If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$.

(ii) If $Z \sim N(0, 1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.

(iii) If $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$ are independent, then

$$\sum_{i=1}^n X_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

It follows from (i) that if $X \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} \mathbb{P}(a < X < b) &= \mathbb{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \end{aligned}$$

Thus we can compute any probabilities we want as long as we can compute the CDF $\Phi(z)$ of a standard Normal. All statistical computing packages will

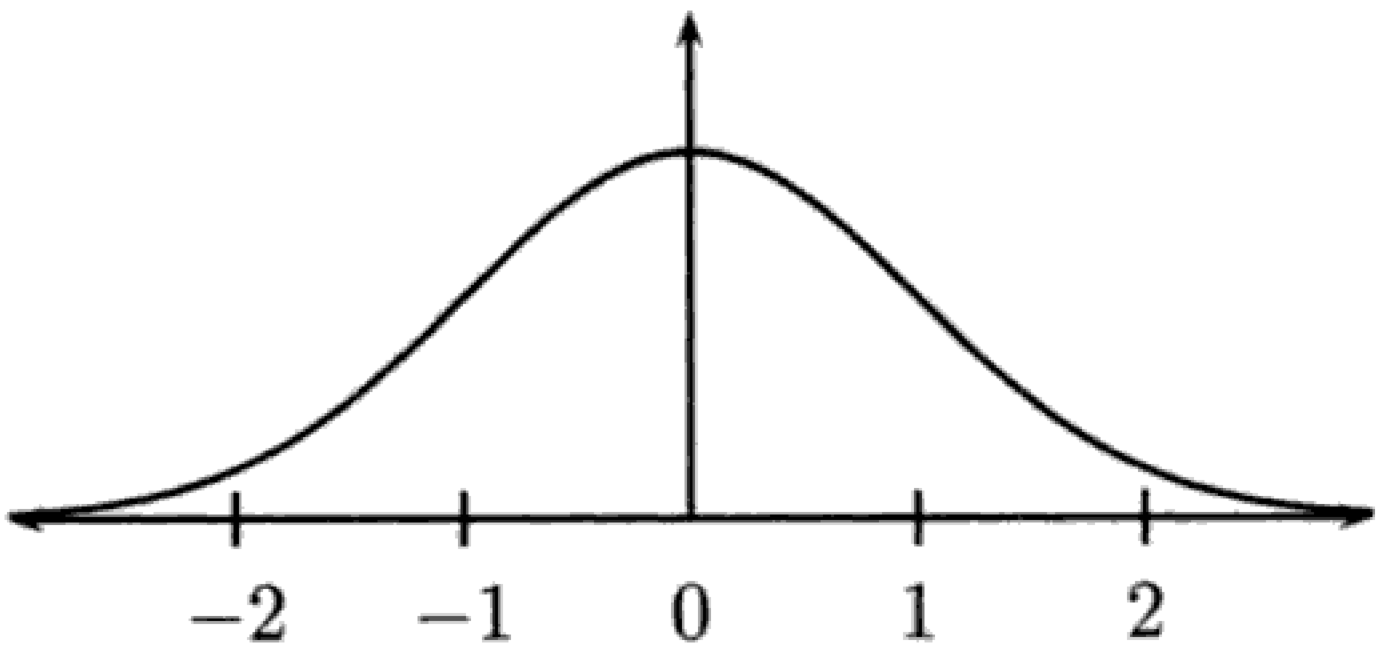

FIGURE 2.4. Density of a standard Normal.

compute $\Phi(z)$ and $\Phi^{-1}(q)$. Older statistics texts (not this one) have a table of values of $\Phi(z)$.

2.17 Example. Suppose that $X \sim N(3,5)$. Find $\mathbb{P}(X > 1)$. The solution is

$$\mathbb{P}(X > 1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}\left(Z < \frac{1-3}{\sqrt{5}}\right) = 1 - \Phi(-0.8944) = 0.81.$$

Now find $q = \Phi^{-1}(0.2)$. This means we have to find q such that $\mathbb{P}(X < q) = 0.2$. We solve this by writing

$$0.2 = \mathbb{P}(X < q) = \mathbb{P}\left(Z < \frac{q-\mu}{\sigma}\right) = \Phi\left(\frac{q-\mu}{\sigma}\right).$$

From the Normal table, $\Phi(-0.8416) = 0.2$. Therefore,

$$-0.8416 = \frac{q-\mu}{\sigma} = \frac{q-3}{\sqrt{5}}$$

and hence $q = 3 - 0.8416\sqrt{5} = 1.1181$. ■

Exponential Distribution. X has an Exponential distribution with parameter β, denoted by $X \sim \text{Exp}(\beta)$, if

$$f(x) = \frac{1}{\beta}e^{-x/\beta}, \quad x > 0$$

where $\beta > 0$. The exponential distribution is used to model the lifetimes of electronic components and the waiting times between rare events.

Gamma Distribution. For $\alpha > 0$, the **Gamma function** is defined by $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$. X has a Gamma distribution with parameters α and

2.5 Bivariate Distributions

Given a pair of discrete random variables X and Y, define the **joint mass function** by $f(x, y) = \mathbb{P}(X = x \text{ and } Y = y)$. From now on, we write $\mathbb{P}(X = x \text{ and } Y = y)$ as $\mathbb{P}(X = x, Y = y)$. We write f as $f_{X,Y}$ when we want to be more explicit.

2.18 Example. Here is a bivariate distribution for two random variables X and Y each taking values 0 or 1:

	$Y = 0$	$Y = 1$	
X=0	1/9	2/9	1/3
X=1	2/9	4/9	2/3
	1/3	2/3	1

Thus, $f(1, 1) = \mathbb{P}(X = 1, Y = 1) = 4/9$. ■

2.19 Definition. *In the continuous case, we call a function $f(x, y)$ a* PDF *for the random variables (X, Y) if*

(i) $f(x, y) \geq 0$ for all (x, y),

(ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$ and,

(iii) for any set $A \subset \mathbb{R} \times \mathbb{R}$, $\mathbb{P}((X, Y) \in A) = \int \int_A f(x, y) dx dy$.

In the discrete or continuous case we define the joint CDF as $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$.

2.20 Example. Let (X, Y) be uniform on the unit square. Then,

$$f(x, y) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1,\ 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{P}(X < 1/2, Y < 1/2)$. The event $A = \{X < 1/2, Y < 1/2\}$ corresponds to a subset of the unit square. Integrating f over this subset corresponds, in this case, to computing the area of the set A which is $1/4$. So, $\mathbb{P}(X < 1/2, Y < 1/2) = 1/4$. ■

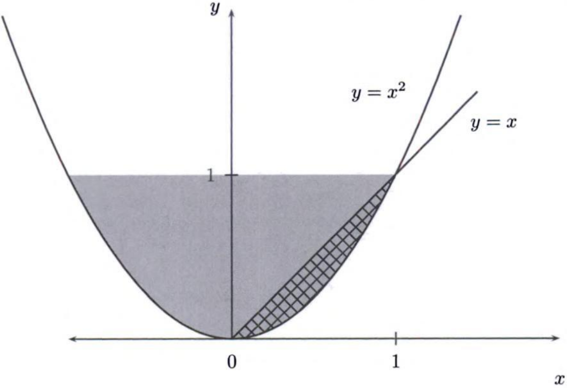

FIGURE 2.5. The light shaded region is $x^2 \le y \le 1$. The density is positive over this region. The hatched region is the event $X \ge Y$ intersected with $x^2 \le y \le 1$.

2.6 Marginal Distributions

2.23 Definition. *If (X, Y) have joint distribution with mass function $f_{X,Y}$, then the* **marginal mass function for** *X is defined by*

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f(x, y) \tag{2.4}$$

and the **marginal mass function for** *Y is defined by*

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y) = \sum_x f(x, y). \tag{2.5}$$

2.24 Example. Suppose that $f_{X,Y}$ is given in the table that follows. The marginal distribution for X corresponds to the row totals and the marginal distribution for Y corresponds to the columns totals.

	$Y = 0$	$Y = 1$	
X=0	1/10	2/10	3/10
X=1	3/10	4/10	7/10
	4/10	6/10	1

For example, $f_X(0) = 3/10$ and $f_X(1) = 7/10$. ■

2.25 Definition. *For continuous random variables, the marginal densities are*

$$f_X(x) = \int f(x,y)dy, \quad \text{and} \quad f_Y(y) = \int f(x,y)dx. \qquad (2.6)$$

The corresponding marginal distribution functions are denoted by F_X and F_Y.

2.26 Example. Suppose that

$$f_{X,Y}(x,y) = e^{-(x+y)}$$

for $x, y \geq 0$. Then $f_X(x) = e^{-x}\int_0^\infty e^{-y}dy = e^{-x}$. ■

2.27 Example. Suppose that

$$f(x,y) = \begin{cases} x+y & \text{if } 0 \leq x \leq 1,\ 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$f_Y(y) = \int_0^1 (x+y)\,dx = \int_0^1 x\,dx + \int_0^1 y\,dx = \frac{1}{2} + y. \quad ■$$

2.28 Example. Let (X,Y) have density

$$f(x,y) = \begin{cases} \frac{21}{4}x^2y & \text{if } x^2 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus,

$$f_X(x) = \int f(x,y)dy = \frac{21}{4}x^2\int_{x^2}^1 y\,dy = \frac{21}{8}x^2(1-x^4)$$

for $-1 \leq x \leq 1$ and $f_X(x) = 0$ otherwise. ■

2.7 Independent Random Variables

2.29 Definition. *Two random variables X and Y are* **independent** *if, for every A and B,*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \qquad (2.7)$$

and we write $X \amalg Y$. Otherwise we say that X and Y are **dependent** *and we write $X \not\amalg Y$.*

In principle, to check whether X and Y are independent we need to check equation (2.7) for all subsets A and B. Fortunately, we have the following result which we state for continuous random variables though it is true for discrete random variables too.

2.30 Theorem. *Let X and Y have joint* PDF $f_{X,Y}$. *Then $X \amalg Y$ if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all values x and y.* [5]

2.31 Example. Let X and Y have the following distribution:

	$Y=0$	$Y=1$	
X=0	1/4	1/4	1/2
X=1	1/4	1/4	1/2
	1/2	1/2	1

Then, $f_X(0) = f_X(1) = 1/2$ and $f_Y(0) = f_Y(1) = 1/2$. X and Y are independent because $f_X(0)f_Y(0) = f(0,0)$, $f_X(0)f_Y(1) = f(0,1)$, $f_X(1)f_Y(0) = f(1,0)$, $f_X(1)f_Y(1) = f(1,1)$. Suppose instead that X and Y have the following distribution:

	$Y=0$	$Y=1$	
X=0	1/2	0	1/2
X=1	0	1/2	1/2
	1/2	1/2	1

These are not independent because $f_X(0)f_Y(1) = (1/2)(1/2) = 1/4$ yet $f(0,1) = 0$. ■

2.32 Example. Suppose that X and Y are independent and both have the same density

$$f(x) = \begin{cases} 2x & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let us find $\mathbb{P}(X + Y \le 1)$. Using independence, the joint density is

$$f(x,y) = f_X(x)f_Y(y) = \begin{cases} 4xy & \text{if } 0 \le x \le 1, \quad 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

[5] The statement is not rigorous because the density is defined only up to sets of measure 0.

Now,

$$\begin{aligned}\mathbb{P}(X+Y\leq 1) &= \int\int_{x+y\leq 1} f(x,y)dydx \\ &= 4\int_0^1 x\left[\int_0^{1-x} ydy\right]dx \\ &= 4\int_0^1 x\frac{(1-x)^2}{2}dx = \frac{1}{6}. \quad \blacksquare\end{aligned}$$

The following result is helpful for verifying independence.

2.33 Theorem. *Suppose that the range of X and Y is a (possibly infinite) rectangle. If $f(x,y) = g(x)h(y)$ for some functions g and h (not necessarily probability density functions) then X and Y are independent.*

2.34 Example. Let X and Y have density

$$f(x,y) = \begin{cases} 2e^{-(x+2y)} & \text{if } x > 0 \text{ and } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The range of X and Y is the rectangle $(0,\infty)\times(0,\infty)$. We can write $f(x,y) = g(x)h(y)$ where $g(x) = 2e^{-x}$ and $h(y) = e^{-2y}$. Thus, $X \amalg Y$. ■

2.8 Conditional Distributions

If X and Y are discrete, then we can compute the conditional distribution of X given that we have observed $Y = y$. Specifically, $\mathbb{P}(X = x|Y = y) = \mathbb{P}(X = x, Y = y)/\mathbb{P}(Y = y)$. This leads us to define the conditional probability mass function as follows.

2.35 Definition. *The* **conditional probability mass function** *is*

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

if $f_Y(y) > 0$.

For continuous distributions we use the same definitions. [6] The interpretation differs: in the discrete case, $f_{X|Y}(x|y)$ is $\mathbb{P}(X = x|Y = y)$, but in the continuous case, we must integrate to get a probability.

[6] We are treading in deep water here. When we compute $\mathbb{P}(X \in A|Y = y)$ in the continuous case we are conditioning on the event $\{Y = y\}$ which has probability 0. We

of Y? First note that,

$$f_X(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{1-x} & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

So,

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x) = \begin{cases} \frac{1}{1-x} & \text{if } 0 < x < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The marginal for Y is

$$f_Y(y) = \int_0^y f_{X,Y}(x,y)dx = \int_0^y \frac{dx}{1-x} = -\int_1^{1-y} \frac{du}{u} = -\log(1-y)$$

for $0 < y < 1$. ■

2.40 Example. Consider the density in Example 2.28. Let's find $f_{Y|X}(y|x)$. When $X = x$, y must satisfy $x^2 \le y \le 1$. Earlier, we saw that $f_X(x) = (21/8)x^2(1-x^4)$. Hence, for $x^2 \le y \le 1$,

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{21}{4}x^2 y}{\frac{21}{8}x^2(1-x^4)} = \frac{2y}{1-x^4}.$$

Now let us compute $\mathbb{P}(Y \ge 3/4 | X = 1/2)$. This can be done by first noting that $f_{Y|X}(y|1/2) = 32y/15$. Thus,

$$\mathbb{P}(Y \ge 3/4 | X = 1/2) = \int_{3/4}^1 f(y|1/2)dy = \int_{3/4}^1 \frac{32y}{15} dy = \frac{7}{15}. \quad ■$$

2.9 Multivariate Distributions and IID Samples

Let $X = (X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are random variables. We call X a **random vector**. Let $f(x_1, \ldots, x_n)$ denote the PDF. It is possible to define their marginals, conditionals etc. much the same way as in the bivariate case. We say that $X_1, \ldots, X_n$ are independent if, for every $A_1, \ldots, A_n$,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i). \tag{2.8}$$

It suffices to check that $f(x_1, \ldots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$.

where $Z_1, \ldots, Z_k \sim N(0,1)$ are independent. The density of Z is [7]

$$\begin{aligned} f(z) &= \prod_{i=1}^{k} f(z_i) = \frac{1}{(2\pi)^{k/2}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{k} z_j^2 \right\} \\ &= \frac{1}{(2\pi)^{k/2}} \exp\left\{ -\frac{1}{2} z^T z \right\}. \end{aligned}$$

We say that Z has a standard multivariate Normal distribution written $Z \sim N(0, I)$ where it is understood that 0 represents a vector of k zeroes and I is the $k \times k$ identity matrix.

More generally, a vector X has a multivariate Normal distribution, denoted by $X \sim N(\mu, \Sigma)$, if it has density [8]

$$f(x;\, \mu, \Sigma) = \frac{1}{(2\pi)^{k/2} |(\Sigma)|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\} \qquad (2.10)$$

where $|\Sigma|$ denotes the determinant of Σ, μ is a vector of length k and Σ is a $k \times k$ symmetric, positive definite matrix. [9] Setting $\mu = 0$ and $\Sigma = I$ gives back the standard Normal.

Since Σ is symmetric and positive definite, it can be shown that there exists a matrix $\Sigma^{1/2}$ — called the square root of Σ — with the following properties: (i) $\Sigma^{1/2}$ is symmetric, (ii) $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$ and (iii) $\Sigma^{1/2}\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma^{1/2} = I$ where $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$.

2.43 Theorem. *If $Z \sim N(0, I)$ and $X = \mu + \Sigma^{1/2} Z$ then $X \sim N(\mu, \Sigma)$. Conversely, if $X \sim N(\mu, \Sigma)$, then $\Sigma^{-1/2}(X - \mu) \sim N(0, I)$.*

Suppose we partition a random Normal vector X as $X = (X_a, X_b)$. We can similarly partition $\mu = (\mu_a, \mu_b)$ and

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}.$$

2.44 Theorem. *Let $X \sim N(\mu, \Sigma)$. Then:*

(1) The marginal distribution of X_a is $X_a \sim N(\mu_a, \Sigma_{aa})$.

(2) The conditional distribution of X_b given $X_a = x_a$ is

$$X_b | X_a = x_a \sim N\left(\mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a),\ \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \right).$$

(3) If a is a vector then $a^T X \sim N(a^T \mu, a^T \Sigma a)$.

(4) $V = (X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi^2_k$.

[7] If a and b are vectors then $a^T b = \sum_{i=1}^{k} a_i b_i$.

[8] Σ^{-1} is the inverse of the matrix Σ.

[9] A matrix Σ is positive definite if, for all nonzero vectors x, $x^T \Sigma x > 0$.

2.11 Transformations of Random Variables

Suppose that X is a random variable with PDF f_X and CDF F_X. Let $Y = r(X)$ be a function of X, for example, $Y = X^2$ or $Y = e^X$. We call $Y = r(X)$ a transformation of X. How do we compute the PDF and CDF of Y? In the discrete case, the answer is easy. The mass function of Y is given by

$$\begin{aligned} f_Y(y) &= \mathbb{P}(Y = y) = \mathbb{P}(r(X) = y) \\ &= \mathbb{P}(\{x;\ r(x) = y\}) = \mathbb{P}(X \in r^{-1}(y)). \end{aligned}$$

2.45 Example. Suppose that $\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = 1/4$ and $\mathbb{P}(X = 0) = 1/2$. Let $Y = X^2$. Then, $\mathbb{P}(Y = 0) = \mathbb{P}(X = 0) = 1/2$ and $\mathbb{P}(Y = 1) = \mathbb{P}(X = 1) + \mathbb{P}(X = -1) = 1/2$. Summarizing:

x	$f_X(x)$
-1	1/4
0	1/2
1	1/4

y	$f_Y(y)$
0	1/2
1	1/2

Y takes fewer values than X because the transformation is not one-to-one. ■

The continuous case is harder. There are three steps for finding f_Y:

Three Steps for Transformations

1. For each y, find the set $A_y = \{x :\ r(x) \leq y\}$.
2. Find the CDF

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(r(X) \leq y) \\ &= \mathbb{P}(\{x;\ r(x) \leq y\}) \\ &= \int_{A_y} f_X(x) dx. \end{aligned} \tag{2.11}$$

3. The PDF is $f_Y(y) = F'_Y(y)$.

2.46 Example. Let $f_X(x) = e^{-x}$ for $x > 0$. Hence, $F_X(x) = \int_0^x f_X(s)ds = 1 - e^{-x}$. Let $Y = r(X) = \log X$. Then, $A_y = \{x :\ x \leq e^y\}$ and

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(\log X \leq y) \\ &= \mathbb{P}(X \leq e^y) = F_X(e^y) = 1 - e^{-e^y}. \end{aligned}$$

Therefore, $f_Y(y) = e^y e^{-e^y}$ for $y \in \mathbb{R}$. ■

2.47 Example. Let $X \sim \text{Uniform}(-1,3)$. Find the PDF of $Y = X^2$. The density of X is

$$f_X(x) = \begin{cases} 1/4 & \text{if } -1 < x < 3 \\ 0 & \text{otherwise.} \end{cases}$$

Y can only take values in $(0,9)$. Consider two cases: (i) $0 < y < 1$ and (ii) $1 \leq y < 9$. For case (i), $A_y = [-\sqrt{y}, \sqrt{y}]$ and $F_Y(y) = \int_{A_y} f_X(x)dx = (1/2)\sqrt{y}$. For case (ii), $A_y = [-1, \sqrt{y}]$ and $F_Y(y) = \int_{A_y} f_X(x)dx = (1/4)(\sqrt{y}+1)$. Differentiating F we get

$$f_Y(y) = \begin{cases} \frac{1}{4\sqrt{y}} & \text{if } 0 < y < 1 \\ \frac{1}{8\sqrt{y}} & \text{if } 1 < y < 9 \\ 0 & \text{otherwise.} \end{cases} \blacksquare$$

When r is strictly monotone increasing or strictly monotone decreasing then r has an inverse $s = r^{-1}$ and in this case one can show that

$$f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right|. \qquad (2.12)$$

2.12 Transformations of Several Random Variables

In some cases we are interested in transformations of several random variables. For example, if X and Y are given random variables, we might want to know the distribution of X/Y, $X+Y$, $\max\{X,Y\}$ or $\min\{X,Y\}$. Let $Z = r(X,Y)$ be the function of interest. The steps for finding f_Z are the same as before:

Three Steps for Transformations

1. For each z, find the set $A_z = \{(x,y) : \ r(x,y) \leq z\}$.

2. Find the CDF

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \leq z) = \mathbb{P}(r(X,Y) \leq z) \\ &= \mathbb{P}(\{(x,y);\ r(x,y) \leq z\}) = \int\int_{A_z} f_{X,Y}(x,y)\,dx\,dy. \end{aligned}$$

3. Then $f_Z(z) = F'_Z(z)$.

2.48 Example. Let $X_1, X_2 \sim \text{Uniform}(0,1)$ be independent. Find the density of $Y = X_1 + X_2$. The joint density of (X_1, X_2) is

$$f(x_1, x_2) = \begin{cases} 1 & 0 < x_1 < 1,\ 0 < x_2 < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let $r(x_1, x_2) = x_1 + x_2$. Now,

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \le y) = \mathbb{P}(r(X_1, X_2) \le y) \\ &= \mathbb{P}(\{(x_1, x_2) :\ r(x_1, x_2) \le y\}) = \int\int_{A_y} f(x_1, x_2) dx_1 dx_2. \end{aligned}$$

Now comes the hard part: finding A_y. First suppose that $0 < y \le 1$. Then A_y is the triangle with vertices $(0,0)$, $(y,0)$ and $(0,y)$. See Figure 2.6. In this case, $\int\int_{A_y} f(x_1, x_2) dx_1 dx_2$ is the area of this triangle which is $y^2/2$. If $1 < y < 2$, then A_y is everything in the unit square except the triangle with vertices $(1, y-1), (1,1), (y-1, 1)$. This set has area $1 - (2-y)^2/2$. Therefore,

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ \frac{y^2}{2} & 0 \le y < 1 \\ 1 - \frac{(2-y)^2}{2} & 1 \le y < 2 \\ 1 & y \ge 2. \end{cases}$$

By differentiation, the PDF is

$$f_Y(y) = \begin{cases} y & 0 \le y \le 1 \\ 2 - y & 1 \le y \le 2 \\ 0 & \text{otherwise.} \ \blacksquare \end{cases}$$

2.13 Appendix

Recall that a probability measure $\mathbb{P}$ is defined on a σ-field $\mathcal{A}$ of a sample space Ω. A random variable X is a **measurable** map $X : \Omega \to \mathbb{R}$. Measurable means that, for every x, $\{\omega :\ X(\omega) \le x\} \in \mathcal{A}$.

2.14 Exercises

1. Show that

$$\mathbb{P}(X = x) = F(x^+) - F(x^-).$$

7. Let X and Y be independent and suppose that each has a Uniform$(0, 1)$ distribution. Let $Z = \min\{X, Y\}$. Find the density $f_Z(z)$ for Z. Hint: It might be easier to first find $\mathbb{P}(Z > z)$.

8. Let X have CDF F. Find the CDF of $X^+ = \max\{0, X\}$.

9. Let $X \sim \text{Exp}(\beta)$. Find $F(x)$ and $F^{-1}(q)$.

10. Let X and Y be independent. Show that $g(X)$ is independent of $h(Y)$ where g and h are functions.

11. Suppose we toss a coin once and let p be the probability of heads. Let X denote the number of heads and let Y denote the number of tails.

    (a) Prove that X and Y are dependent.

    (b) Let $N \sim \text{Poisson}(\lambda)$ and suppose we toss a coin N times. Let X and Y be the number of heads and tails. Show that X and Y are independent.

12. Prove Theorem 2.33.

13. Let $X \sim N(0, 1)$ and let $Y = e^X$.

    (a) Find the PDF for Y. Plot it.

    (b) (**Computer Experiment.**) Generate a vector $x = (x_1, \ldots, x_{10,000})$ consisting of 10,000 random standard Normals. Let $y = (y_1, \ldots, y_{10,000})$ where $y_i = e^{x_i}$. Draw a histogram of y and compare it to the PDF you found in part (a).

14. Let (X, Y) be uniformly distributed on the unit disk $\{(x, y) : x^2 + y^2 \leq 1\}$. Let $R = \sqrt{X^2 + Y^2}$. Find the CDF and PDF of R.

15. (**A universal random number generator.**) Let X have a continuous, strictly increasing CDF F. Let $Y = F(X)$. Find the density of Y. This is called the probability integral transform. Now let $U \sim \text{Uniform}(0, 1)$ and let $X = F^{-1}(U)$. Show that $X \sim F$. Now write a program that takes Uniform (0,1) random variables and generates random variables from an Exponential (β) distribution.

16. Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ and assume that X and Y are independent. Show that the distribution of X given that $X + Y = n$ is Binomial(n, π) where $\pi = \lambda/(\lambda + \mu)$.

3

Expectation

3.1 Expectation of a Random Variable

The mean, or expectation, of a random variable X is the average value of X.

3.1 Definition. *The* **expected value,** *or* **mean,** *or* **first moment,** *of X is defined to be*

$$\mathbb{E}(X) = \int x\, dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous} \end{cases} \tag{3.1}$$

assuming that the sum (or integral) is well defined. We use the following notation to denote the expected value of X:

$$\mathbb{E}(X) = \mathbb{E}X = \int x\, dF(x) = \mu = \mu_X. \tag{3.2}$$

The expectation is a one-number summary of the distribution. Think of $\mathbb{E}(X)$ as the average $\sum_{i=1}^n X_i/n$ of a large number of IID draws $X_1, \ldots, X_n$. The fact that $\mathbb{E}(X) \approx \sum_{i=1}^n X_i/n$ is actually more than a heuristic; it is a theorem called the law of large numbers that we will discuss in Chapter 5.

The notation $\int x\, dF(x)$ deserves some comment. We use it merely as a convenient unifying notation so we don't have to write $\sum_x x f(x)$ for discrete random variables and $\int xf(x)dx$ for continuous random variables, but you should be aware that $\int x\,dF(x)$ has a precise meaning that is discussed in real analysis courses.

To ensure that $\mathbb{E}(X)$ is well defined, we say that $\mathbb{E}(X)$ exists if $\int_x |x|dF_X(x) < \infty$. Otherwise we say that the expectation does not exist.

3.2 Example. Let $X \sim \text{Bernoulli}(p)$. Then $\mathbb{E}(X) = \sum_{x=0}^{1} xf(x) = (0 \times (1-p)) + (1 \times p) = p$. ■

3.3 Example. Flip a fair coin two times. Let X be the number of heads. Then, $\mathbb{E}(X) = \int xdF_X(x) = \sum_x xf_X(x) = (0 \times f(0)) + (1 \times f(1)) + (2 \times f(2)) = (0 \times (1/4)) + (1 \times (1/2)) + (2 \times (1/4)) = 1$. ■

3.4 Example. Let $X \sim \text{Uniform}(-1, 3)$. Then, $\mathbb{E}(X) = \int xd\,F_X(x) = \int xf_X(x)dx = \frac{1}{4}\int_{-1}^{3} x\,dx = 1$. ■

3.5 Example. Recall that a random variable has a Cauchy distribution if it has density $f_X(x) = \{\pi(1+x^2)\}^{-1}$. Using integration by parts, (set $u = x$ and $v = \tan^{-1} x$),

$$\int |x|dF(x) = \frac{2}{\pi}\int_0^\infty \frac{x\,dx}{1+x^2} = \left[x\ \tan^{-1}(x)\right]_0^\infty - \int_0^\infty \tan^{-1} x\,dx = \infty$$

so the mean does not exist. If you simulate a Cauchy distribution many times and take the average, you will see that the average never settles down. This is because the Cauchy has thick tails and hence extreme observations are common. ■

From now on, whenever we discuss expectations, we implicitly assume that they exist.

Let $Y = r(X)$. How do we compute $\mathbb{E}(Y)$? One way is to find $f_Y(y)$ and then compute $\mathbb{E}(Y) = \int yf_Y(y)dy$. But there is an easier way.

3.6 Theorem (The Rule of the Lazy Statistician). *Let $Y = r(X)$. Then*

$$\mathbb{E}(Y) = \mathbb{E}(r(X)) = \int r(x)dF_X(x). \tag{3.3}$$

This result makes intuitive sense. Think of playing a game where we draw X at random and then I pay you $Y = r(X)$. Your average income is $r(x)$ times the chance that $X = x$, summed (or integrated) over all values of x. Here is a special case. Let A be an event and let $r(x) = I_A(x)$ where $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ if $x \notin A$. Then

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

In other words, probability is a special case of expectation.

3.7 Example. Let $X \sim \text{Unif}(0,1)$. Let $Y = r(X) = e^X$. Then,

$$\mathbb{E}(Y) = \int_0^1 e^x f(x) dx = \int_0^1 e^x dx = e - 1.$$

Alternatively, you could find $f_Y(y)$ which turns out to be $f_Y(y) = 1/y$ for $1 < y < e$. Then, $\mathbb{E}(Y) = \int_1^e y\, f(y) dy = e - 1$. ■

3.8 Example. Take a stick of unit length and break it at random. Let Y be the length of the longer piece. What is the mean of Y? If X is the break point then $X \sim \text{Unif}(0,1)$ and $Y = r(X) = \max\{X, 1 - X\}$. Thus, $r(x) = 1 - x$ when $0 < x < 1/2$ and $r(x) = x$ when $1/2 \leq x < 1$. Hence,

$$\mathbb{E}(Y) = \int r(x) dF(x) = \int_0^{1/2} (1 - x) dx + \int_{1/2}^1 x\, dx = \frac{3}{4}. \quad ■$$

Functions of several variables are handled in a similar way. If $Z = r(X, Y)$ then

$$\mathbb{E}(Z) = \mathbb{E}(r(X, Y)) = \int\int r(x, y) dF(x, y). \tag{3.4}$$

3.9 Example. Let (X, Y) have a jointly uniform distribution on the unit square. Let $Z = r(X, Y) = X^2 + Y^2$. Then,

$$\begin{aligned} \mathbb{E}(Z) &= \int\int r(x, y) dF(x, y) = \int_0^1 \int_0^1 (x^2 + y^2)\, dx dy \\ &= \int_0^1 x^2\, dx + \int_0^1 y^2\, dy = \frac{2}{3}. \quad ■ \end{aligned}$$

The k^{th} **moment** of X is defined to be $\mathbb{E}(X^k)$ assuming that $\mathbb{E}(|X|^k) < \infty$.

3.10 Theorem. *If the k^{th} moment exists and if $j < k$ then the j^{th} moment exists.*

PROOF. We have

$$\mathbb{E}|X|^j = \int_{-\infty}^{\infty} |x|^j f_X(x) dx$$

$$
\begin{aligned}
&= \int_{|x|\le 1} |x|^j f_X(x)dx + \int_{|x|>1} |x|^j f_X(x)dx \\
&\le \int_{|x|\le 1} f_X(x)dx + \int_{|x|>1} |x|^k f_X(x)dx \\
&\le 1 + \mathbb{E}(|X|^k) < \infty. \quad \blacksquare
\end{aligned}
$$

The k^{th} central moment is defined to be $\mathbb{E}((X-\mu)^k)$.

3.2 Properties of Expectations

3.11 Theorem. *If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then*

$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i \mathbb{E}(X_i). \tag{3.5}$$

3.12 Example. Let $X \sim \text{Binomial}(n,p)$. What is the mean of X? We could try to appeal to the definition:

$$\mathbb{E}(X) = \int x\, dF_X(x) = \sum_x x f_X(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$$

but this is not an easy sum to evaluate. Instead, note that $X = \sum_{i=1}^n X_i$ where $X_i = 1$ if the i^{th} toss is heads and $X_i = 0$ otherwise. Then $\mathbb{E}(X_i) = (p \times 1) + ((1-p)\times 0) = p$ and $\mathbb{E}(X) = \mathbb{E}(\sum_i X_i) = \sum_i \mathbb{E}(X_i) = np$. ■

3.13 Theorem. *Let $X_1, \ldots, X_n$ be independent random variables. Then,*

$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_i \mathbb{E}(X_i). \tag{3.6}$$

Notice that the summation rule does not require independence but the multiplication rule does.

3.3 Variance and Covariance

The variance measures the "spread" of a distribution. [1]

[1] We can't use $\mathbb{E}(X-\mu)$ as a measure of spread since $\mathbb{E}(X-\mu) = \mathbb{E}(X) - \mu = \mu - \mu = 0$. We can and sometimes do use $\mathbb{E}|X-\mu|$ as a measure of spread but more often we use the variance.

3.17 Theorem. *Let $X_1, \ldots, X_n$ be* IID *and let $\mu = \mathbb{E}(X_i)$, $\sigma^2 = \mathbb{V}(X_i)$. Then*

$$\mathbb{E}(\overline{X}_n) = \mu, \quad \mathbb{V}(\overline{X}_n) = \frac{\sigma^2}{n} \quad \text{and} \quad \mathbb{E}(S_n^2) = \sigma^2.$$

If X and Y are random variables, then the covariance and correlation between X and Y measure how strong the linear relationship is between X and Y.

3.18 Definition. *Let X and Y be random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y. Define the* **covariance** *between X and Y by*

$$\mathsf{Cov}(X, Y) = \mathbb{E}\Big((X - \mu_X)(Y - \mu_Y)\Big) \tag{3.11}$$

and the **correlation** *by*

$$\rho = \rho_{X,Y} = \rho(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{3.12}$$

3.19 Theorem. *The covariance satisfies:*

$$\mathsf{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

The correlation satisfies:

$$-1 \leq \rho(X, Y) \leq 1.$$

If $Y = aX + b$ for some constants a and b then $\rho(X, Y) = 1$ if $a > 0$ and $\rho(X, Y) = -1$ if $a < 0$. If X and Y are independent, then $\mathsf{Cov}(X, Y) = \rho = 0$. The converse is not true in general.

3.20 Theorem. $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathsf{Cov}(X, Y)$ *and* $\mathbb{V}(X - Y) = \mathbb{V}(X) + \mathbb{V}(Y) - 2\mathsf{Cov}(X, Y)$. *More generally, for random variables $X_1, \ldots, X_n$,*

$$\mathbb{V}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \mathbb{V}(X_i) + 2 \sum\sum_{i<j} a_i a_j \mathsf{Cov}(X_i, X_j).$$

3.4 Expectation and Variance of Important Random Variables

Here we record the expectation of some important random variables:

To see this, note that the marginal distribution of any one component of the vector $X_i \sim \text{Binomial}(n, p_i)$. Thus, $\mathbb{E}(X_i) = np_i$ and $\mathbb{V}(X_i) = np_i(1-p_i)$. Note also that $X_i + X_j \sim \text{Binomial}(n, p_i + p_j)$. Thus, $\mathbb{V}(X_i + X_j) = n(p_i + p_j)(1 - [p_i + p_j])$. On the other hand, using the formula for the variance of a sum, we have that $\mathbb{V}(X_i + X_j) = \mathbb{V}(X_i) + \mathbb{V}(X_j) + 2\mathsf{Cov}(X_i, X_j) = np_i(1-p_i) + np_j(1-p_j) + 2\mathsf{Cov}(X_i, X_j)$. If we equate this formula with $n(p_i + p_j)(1 - [p_i + p_j])$ and solve, we get $\mathsf{Cov}(X_i, X_j) = -np_ip_j$.

Finally, here is a lemma that can be useful for finding means and variances of linear combinations of multivariate random vectors.

3.21 Lemma. *If a is a vector and X is a random vector with mean μ and variance Σ, then $\mathbb{E}(a^T X) = a^T \mu$ and $\mathbb{V}(a^T X) = a^T \Sigma a$. If A is a matrix then $\mathbb{E}(AX) = A\mu$ and $\mathbb{V}(AX) = A\Sigma A^T$.*

3.5 Conditional Expectation

Suppose that X and Y are random variables. What is the mean of X among those times when $Y = y$? The answer is that we compute the mean of X as before but we substitute $f_{X|Y}(x|y)$ for $f_X(x)$ in the definition of expectation.

3.22 Definition. *The conditional expectation of X given $Y = y$ is*

$$\mathbb{E}(X|Y=y) = \begin{cases} \sum x\, f_{X|Y}(x|y) & \text{discrete case} \\ \int x\, f_{X|Y}(x|y)\, dx & \text{continuous case.} \end{cases} \tag{3.13}$$

If $r(x, y)$ is a function of x and y then

$$\mathbb{E}(r(X,Y)|Y=y) = \begin{cases} \sum r(x,y)\, f_{X|Y}(x|y) & \text{discrete case} \\ \int r(x,y)\, f_{X|Y}(x|y)\, dx & \text{continuous case.} \end{cases} \tag{3.14}$$

Warning! Here is a subtle point. Whereas $\mathbb{E}(X)$ is a number, $\mathbb{E}(X|Y = y)$ is a function of y. Before we observe Y, we don't know the value of $\mathbb{E}(X|Y = y)$ so it is a random variable which we denote $\mathbb{E}(X|Y)$. In other words, $\mathbb{E}(X|Y)$ is the random variable whose value is $\mathbb{E}(X|Y = y)$ when $Y = y$. Similarly, $\mathbb{E}(r(X, Y)|Y)$ is the random variable whose value is $\mathbb{E}(r(X, Y)|Y = y)$ when $Y = y$. This is a very confusing point so let us look at an example.

3.23 Example. Suppose we draw $X \sim \text{Unif}(0, 1)$. After we observe $X = x$, we draw $Y|X = x \sim \text{Unif}(x, 1)$. Intuitively, we expect that $\mathbb{E}(Y|X = x) =$

$(1+x)/2$. In fact, $f_{Y|X}(y|x) = 1/(1-x)$ for $x < y < 1$ and

$$\mathbb{E}(Y|X=x) = \int_x^1 y\, f_{Y|X}(y|x)dy = \frac{1}{1-x}\int_x^1 y\, dy = \frac{1+x}{2}$$

as expected. Thus, $\mathbb{E}(Y|X) = (1+X)/2$. Notice that $\mathbb{E}(Y|X) = (1+X)/2$ is a random variable whose value is the number $\mathbb{E}(Y|X=x) = (1+x)/2$ once $X = x$ is observed. ■

3.24 Theorem (The Rule of Iterated Expectations). *For random variables X and Y, assuming the expectations exist, we have that*

$$\mathbb{E}\left[\mathbb{E}(Y|X)\right] = \mathbb{E}(Y) \quad \text{and} \quad \mathbb{E}\left[\mathbb{E}(X|Y)\right] = \mathbb{E}(X). \tag{3.15}$$

More generally, for any function $r(x,y)$ we have

$$\mathbb{E}\left[\mathbb{E}(r(X,Y)|X)\right] = \mathbb{E}(r(X,Y)). \tag{3.16}$$

PROOF. We'll prove the first equation. Using the definition of conditional expectation and the fact that $f(x,y) = f(x)f(y|x)$,

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}(Y|X)\right] &= \int \mathbb{E}(Y|X=x)f_X(x)dx = \int\int yf(y|x)dyf(x)dx \\ &= \int\int yf(y|x)f(x)dxdy = \int\int yf(x,y)dxdy = \mathbb{E}(Y). \quad \blacksquare\end{aligned}$$

3.25 Example. Consider example 3.23. How can we compute $\mathbb{E}(Y)$? One method is to find the joint density $f(x,y)$ and then compute $\mathbb{E}(Y) = \int\int yf(x,y)dxdy$. An easier way is to do this in two steps. First, we already know that $\mathbb{E}(Y|X) = (1+X)/2$. Thus,

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}\mathbb{E}(Y|X) = \mathbb{E}\left(\frac{(1+X)}{2}\right) \\ &= \frac{(1+\mathbb{E}(X))}{2} = \frac{(1+(1/2))}{2} = 3/4. \quad \blacksquare\end{aligned}$$

3.26 Definition. *The* **conditional variance** *is defined as*

$$\mathbb{V}(Y|X=x) = \int (y-\mu(x))^2 f(y|x)dy \tag{3.17}$$

where $\mu(x) = \mathbb{E}(Y|X=x)$.

3.27 Theorem. *For random variables X and Y,*

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

3.28 Example. Draw a county at random from the United States. Then draw n people at random from the county. Let X be the number of those people who have a certain disease. If Q denotes the proportion of people in that county with the disease, then Q is also a random variable since it varies from county to county. Given $Q = q$, we have that $X \sim \text{Binomial}(n, q)$. Thus, $\mathbb{E}(X|Q = q) = nq$ and $\mathbb{V}(X|Q = q) = nq(1 - q)$. Suppose that the random variable Q has a Uniform (0,1) distribution. A distribution that is constructed in stages like this is called a **hierarchical model** and can be written as

$$\begin{aligned} Q &\sim \text{Uniform}(0,1) \\ X|Q = q &\sim \text{Binomial}(n, q). \end{aligned}$$

Now, $\mathbb{E}(X) = \mathbb{E}\mathbb{E}(X|Q) = \mathbb{E}(nQ) = n\mathbb{E}(Q) = n/2$. Let us compute the variance of X. Now, $\mathbb{V}(X) = \mathbb{E}\mathbb{V}(X|Q) + \mathbb{V}\mathbb{E}(X|Q)$. Let's compute these two terms. First, $\mathbb{E}\mathbb{V}(X|Q) = \mathbb{E}[nQ(1 - Q)] = n\mathbb{E}(Q(1 - Q)) = n \int q(1 - q)f(q)dq = n \int_0^1 q(1 - q)dq = n/6$. Next, $\mathbb{V}\mathbb{E}(X|Q) = \mathbb{V}(nQ) = n^2\mathbb{V}(Q) = n^2 \int (q - (1/2))^2 dq = n^2/12$. Hence, $\mathbb{V}(X) = (n/6) + (n^2/12)$. ■

3.6 Moment Generating Functions

Now we will define the moment generating function which is used for finding moments, for finding the distribution of sums of random variables and which is also used in the proofs of some theorems.

3.29 Definition. *The* **moment generating function** MGF, *or* **Laplace transform**, *of X is defined by*

$$\psi_X(t) = \mathbb{E}(e^{tX}) = \int e^{tx} dF(x)$$

where t varies over the real numbers.

In what follows, we assume that the MGF is well defined for all t in some open interval around $t = 0$. [2]

When the MGF is well defined, it can be shown that we can interchange the operations of differentiation and "taking expectation." This leads to

$$\psi'(0) = \left[\frac{d}{dt}\mathbb{E}e^{tX}\right]_{t=0} = \mathbb{E}\left[\frac{d}{dt}e^{tX}\right]_{t=0} = \mathbb{E}\left[Xe^{tX}\right]_{t=0} = \mathbb{E}(X).$$

[2] A related function is the characteristic function, defined by $\mathbb{E}(e^{itX})$ where $i = \sqrt{-1}$. This function is always well defined for all t.

This book is for people who want to learn probability and statistics quickly. It brings together many of the main ideas in modern statistics in one place. The book is suitable for students and researchers in statistics, computer science, data mining, and machine learning.

This book covers a much wider range of topics than a typical introductory text on mathematical statistics. It includes modern topics like nonparametric curve estimation, bootstrapping, and classification, topics that are usually relegated to follow-up courses. The reader is assumed to know calculus and a little linear algebra. No previous knowledge of probability and statistics is required. The text can be used at the advanced undergraduate and graduate levels.

Larry Wasserman is Professor of Statistics at Carnegie Mellon University. He is also a member of the Center for Automated Learning and Discovery in the School of Computer Science. His research areas include nonparametric inference, asymptotic theory, causality, and applications to astrophysics, bioinformatics, and genetics. He is the 1999 winner of the Committee of Presidents of Statistical Societies Presidents' Award and the 2002 winner of the Centre de recherches mathématiques de Montreal–Statistical Society of Canada Prize in statistics. He is Associate Editor of *The Journal of the American Statistical Association* and *The Annals of Statistics*. He is a fellow of the American Statistical Association and of the Institute of Mathematical Statistics.

ISBN 0-387-40272-1

EAN

9 780387 402727 >

springeronline.com