To Isa

Preface

Taken literally, the title "All of Statistics" is an exaggeration. But in spirit, the title is apt, as the book does cover a much broader range of topics than a typical introductory book on mathematical statistics.

This book is for people who want to learn probability and statistics quickly. It is suitable for graduate or advanced undergraduate students in computer science, mathematics, statistics, and related disciplines. The book includes modern topics like nonparametric curve estimation, bootstrapping, and classification, topics that are usually relegated to follow-up courses. The reader is presumed to know calculus and a little linear algebra. No previous knowledge of probability and statistics is required.

Statistics, **data mining**, and **machine learning** are all concerned with collecting and analyzing data. For some time, statistics research was conducted in statistics departments while data mining and machine learning research was conducted in computer science departments. Statisticians thought that computer scientists were reinventing the wheel. Computer scientists thought that statistical theory didn't apply to their problems.

Things are changing. Statisticians now recognize that computer scientists are making novel contributions while computer scientists now recognize the generality of statistical theory and methodology. Clever data mining algorithms are more scalable than statisticians ever thought possible. Formal statistical theory is more pervasive than computer scientists had realized.

Students who analyze data, or who aspire to develop new methods for analyzing data, should be well grounded in basic probability and mathematical statistics. Using fancy tools like neural nets, boosting, and support vector

machines without understanding basic statistics is like doing brain surgery before knowing how to use a band-aid.

But where can students learn basic probability and statistics quickly? Nowhere. At least, that was my conclusion when my computer science colleagues kept asking me: "Where can I send my students to get a good understanding of modern statistics quickly?" The typical mathematical statistics course spends too much time on tedious and uninspiring topics (counting methods, two dimensional integrals, etc.) at the expense of covering modern concepts (bootstrapping, curve estimation, graphical models, etc.). So I set out to redesign our undergraduate honors course on probability and mathematical statistics. This book arose from that course. Here is a summary of the main features of this book.

1. The book is suitable for graduate students in computer science and honors undergraduates in math, statistics, and computer science. It is also useful for students beginning graduate work in statistics who need to fill in their background on mathematical statistics.

2. I cover advanced topics that are traditionally not taught in a first course. For example, nonparametric regression, bootstrapping, density estimation, and graphical models.

3. I have omitted topics in probability that do not play a central role in statistical inference. For example, counting methods are virtually absent.

4. Whenever possible, I avoid tedious calculations in favor of emphasizing concepts.

5. I cover nonparametric inference before parametric inference.

6. I abandon the usual "First Term = Probability" and "Second Term = Statistics" approach. Some students only take the first half and it would be a crime if they did not see any statistical theory. Furthermore, probability is more engaging when students can see it put to work in the context of statistics. An exception is the topic of stochastic processes which is included in the later material.

7. The course moves very quickly and covers much material. My colleagues joke that I cover all of statistics in this course and hence the title. The course is demanding but I have worked hard to make the material as intuitive as possible so that the material is very understandable despite the fast pace.

8. Rigor and clarity are not synonymous. I have tried to strike a good balance. To avoid getting bogged down in uninteresting technical details, many results are stated without proof. The bibliographic references at the end of each chapter point the student to appropriate sources.

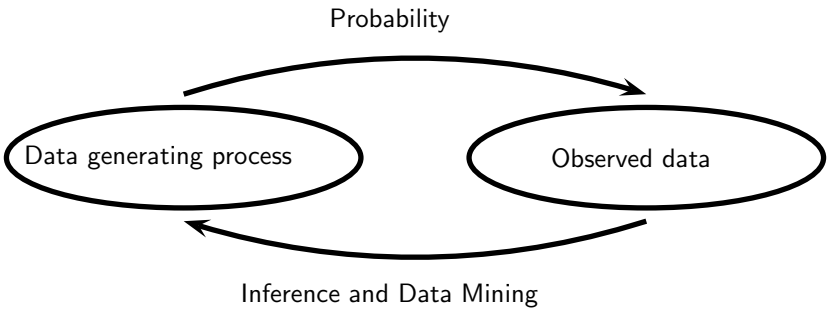


FIGURE 1. Probability and inference.

9. On my website are files with R code which students can use for doing all the computing. The website is:

   http://www.stat.cmu.edu/∼larry/all-of-statistics

   However, the book is not tied to R and any computing language can be used.

Part I of the text is concerned with probability theory, the formal language of uncertainty which is the basis of statistical inference. The basic problem that we study in probability is:

> Given a data generating process, what are the properties of the outcomes?

Part II is about statistical inference and its close cousins, data mining and machine learning. The basic problem of statistical inference is the inverse of probability:

> Given the outcomes, what can we say about the process that generated the data?

These ideas are illustrated in Figure 1. Prediction, classification, clustering, and estimation are all special cases of statistical inference. Data analysis, machine learning and data mining are various names given to the practice of statistical inference, depending on the context.

Part III applies the ideas from Part II to specific problems such as regression, graphical models, causation, density estimation, smoothing, classification, and simulation. Part III contains one more chapter on probability that covers stochastic processes including Markov chains.

I have drawn on other books in many places. Most chapters contain a section called Bibliographic Remarks which serves both to acknowledge my debt to other authors and to point readers to other useful references. I would especially like to mention the books by DeGroot and Schervish (2002) and Grimmett and Stirzaker (1982) from which I adapted many examples and exercises.

As one develops a book over several years it is easy to lose track of where presentation ideas and, especially, homework problems originated. Some I made up. Some I remembered from my education. Some I borrowed from other books. I hope I do not offend anyone if I have used a problem from their book and failed to give proper credit. As my colleague Mark Schervish wrote in his book (Schervish (1995)),

> "... the problems at the ends of each chapter have come from many sources. ... These problems, in turn, came from various sources unknown to me ... If I have used a problem without giving proper credit, please take it as a compliment."

I am indebted to many people without whose help I could not have written this book. First and foremost, the many students who used earlier versions of this text and provided much feedback. In particular, Liz Prather and Jennifer Bakal read the book carefully. Rob Reeder valiantly read through the entire book in excruciating detail and gave me countless suggestions for improvements. Chris Genovese deserves special mention. He not only provided helpful ideas about intellectual content, but also spent many, many hours writing LaTeXcode for the book. The best aspects of the book's layout are due to his hard work; any stylistic deficiencies are due to my lack of expertise. David Hand, Sam Roweis, and David Scott read the book very carefully and made numerous suggestions that greatly improved the book. John Lafferty and Peter Spirtes also provided helpful feedback. John Kimmel has been supportive and helpful throughout the writing process. Finally, my wife Isabella Verdinelli has been an invaluable source of love, support, and inspiration.

Larry Wasserman
Pittsburgh, Pennsylvania
July 2003

Statistics/Data Mining Dictionary

Statisticians and computer scientists often use different language for the same thing. Here is a dictionary that the reader may want to return to throughout the course.

Statistics	Computer Science	Meaning
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete Y from X
clustering	unsupervised learning	putting data into groups
data	training sample	$(X_1, Y_1), \ldots, (X_n, Y_n)$
covariates	features	the X_i's
classifier	hypothesis	a map from covariates to outcomes
hypothesis	—	subset of a parameter space Θ
confidence interval	—	interval that contains an unknown quantity with given frequency
directed acyclic graph	Bayes net	multivariate distribution with given conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update beliefs
frequentist inference	—	statistical methods with guaranteed frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

Contents

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

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

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

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

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 $A_1, A_2, \dots$ 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 $A_1, A_2, \dots$ 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:

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

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

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

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

equally likely, then $\mathbb{P}(H_1 \bigcup H_2) = \mathbb{P}(H_1)+\mathbb{P}(H_2)-\mathbb{P}(H_1H_2) = \frac{1}{2}+\frac{1}{2}-\frac{1}{4} = 3/4$. ∎

1.8 Theorem (Continuity of Probabilities). *If $A_n \to A$ then*

$$\mathbb{P}(A_n) \to \mathbb{P}(A)$$

as $n \to \infty$.

PROOF. Suppose that A_n is monotone increasing so that $A_1 \subset A_2 \subset \cdots$. Let $A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$. Define $B_1 = A_1$, $B_2 = \{\omega \in \Omega : \ \omega \in A_2, \omega \notin A_1\}$, $B_3 = \{\omega \in \Omega : \ \omega \in A_3, \omega \notin A_2, \omega \notin A_1\}, \ldots$ It can be shown that $B_1, B_2, \ldots$ are disjoint, $A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$ for each n and $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$. (See exercise 1.) From Axiom 3,

$$\mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} \mathbb{P}(B_i)$$

and hence, using Axiom 3 again,

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \mathbb{P}(A). \ \blacksquare$$

1.4 Probability on Finite Sample Spaces

Suppose that the sample space $\Omega = \{\omega_1, \ldots, \omega_n\}$ is finite. For example, if we toss a die twice, then Ω has 36 elements: $\Omega = \{(i, j);\ i, j \in \{1, \ldots 6\}\}$. If each outcome is equally likely, then $\mathbb{P}(A) = |A|/36$ where $|A|$ denotes the number of elements in A. The probability that the sum of the dice is 11 is 2/36 since there are two outcomes that correspond to this event.

If Ω is finite and if each outcome is equally likely, then

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|},$$

which is called the **uniform probability distribution.** To compute probabilities, we need to count the number of points in an event A. Methods for counting points are called combinatorial methods. We needn't delve into these in any great detail. We will, however, need a few facts from counting theory that will be useful later. Given n objects, the number of ways of ordering

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

$$A \not\amalg B$$

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, $A\bigcap B = \{2,4\}$,

$\mathbb{P}(AB) = 2/6 = \mathbb{P}(A)\mathbb{P}(B) = (1/2) \times (2/3)$ and so A and B are independent. In this case, we didn't assume that A and B are independent — it just turned out that they were.

Suppose that A and B are disjoint events, each with positive probability. Can they be independent? No. This follows since $\mathbb{P}(A)\mathbb{P}(B) > 0$ yet $\mathbb{P}(AB) = \mathbb{P}(\emptyset) = 0$. Except in this special case, there is no way to judge independence by looking at the sets in a Venn diagram.

1.10 Example. Toss a fair coin 10 times. Let $A =$"at least one head." Let T_j be the event that tails occurs on the j^{th} toss. Then

$$\begin{aligned}
\mathbb{P}(A) &= 1 - \mathbb{P}(A^c) \\
&= 1 - \mathbb{P}(\text{all tails}) \\
&= 1 - \mathbb{P}(T_1 T_2 \cdots T_{10}) \\
&= 1 - \mathbb{P}(T_1)\mathbb{P}(T_2)\cdots\mathbb{P}(T_{10}) \quad \text{using independence} \\
&= 1 - \left(\frac{1}{2}\right)^{10} \approx .999. \quad \blacksquare
\end{aligned}$$

1.11 Example. Two people take turns trying to sink a basketball into a net. Person 1 succeeds with probability 1/3 while person 2 succeeds with probability 1/4. What is the probability that person 1 succeeds before person 2? Let E denote the event of interest. Let A_j be the event that the first success is by person 1 and that it occurs on trial number j. Note that $A_1, A_2, \ldots$ are disjoint and that $E = \bigcup_{j=1}^{\infty} A_j$. Hence,

$$\mathbb{P}(E) = \sum_{j=1}^{\infty} \mathbb{P}(A_j).$$

Now, $\mathbb{P}(A_1) = 1/3$. A_2 occurs if we have the sequence person 1 misses, person 2 misses, person 1 succeeds. This has probability $\mathbb{P}(A_2) = (2/3)(3/4)(1/3) = (1/2)(1/3)$. Following this logic we see that $\mathbb{P}(A_j) = (1/2)^{j-1}(1/3)$. Hence,

$$\mathbb{P}(E) = \sum_{j=1}^{\infty} \frac{1}{3}\left(\frac{1}{2}\right)^{j-1} = \frac{1}{3}\sum_{j=1}^{\infty}\left(\frac{1}{2}\right)^{j-1} = \frac{2}{3}.$$

Here we used that fact that, if $0 < r < 1$ then $\sum_{j=k}^{\infty} r^j = r^k/(1-r)$. $\blacksquare$

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

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

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 $i = 1, \ldots, k$,*

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

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, $\mathbb{P}(B|A_1) = .9$, $\mathbb{P}(B|A_2) = .01$, $\mathbb{P}(B|A_1) = .01$. (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).

3. Let Ω be a sample space and let $A_1, A_2, \ldots,$ be events. Define $B_n = \bigcup_{i=n}^{\infty} A_i$ and $C_n = \bigcap_{i=n}^{\infty} A_i$.

   (a) Show that $B_1 \supset B_2 \supset \cdots$ and that $C_1 \subset C_2 \subset \cdots$.

   (b) Show that $\omega \in \bigcap_{n=1}^{\infty} B_n$ if and only if ω belongs to an infinite number of the events $A_1, A_2, \ldots$.

   (c) Show that $\omega \in \bigcup_{n=1}^{\infty} C_n$ if and only if ω belongs to all the events $A_1, A_2, \ldots$ except possibly a finite number of those events.

4. Let $\{A_i : i \in I\}$ be a collection of events where I is an arbitrary index set. Show that

$$\left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c \quad \text{and} \quad \left(\bigcap_{i\in I} A_i\right)^c = \bigcup_{i\in I} A_i^c$$

   Hint: First prove this for $I = \{1, \ldots, n\}$.

5. Suppose we toss a fair coin until we get exactly two heads. Describe the sample space S. What is the probability that exactly k tosses are required?

6. Let $\Omega = \{0, 1, \ldots, \}$. Prove that there does not exist a uniform distribution on Ω (i.e., if $\mathbb{P}(A) = \mathbb{P}(B)$ whenever $|A| = |B|$, then $\mathbb{P}$ cannot satisfy the axioms of probability).

7. Let $A_1, A_2, \ldots$ be events. Show that

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

   Hint: Define $B_n = A_n - \bigcup_{i=1}^{n-1} A_i$. Then show that the B_n are disjoint and that $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$.

8. Suppose that $\mathbb{P}(A_i) = 1$ for each i. Prove that

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

9. For fixed B such that $\mathbb{P}(B) > 0$, show that $\mathbb{P}(\cdot|B)$ satisfies the axioms of probability.

10. You have probably heard it before. Now you can solve it rigorously. It is called the "Monty Hall Problem." A prize is placed at random

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

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

18. Suppose k events form a partition of the sample space Ω, i.e., they are disjoint and $\bigcup_{i=1}^{k} A_i = \Omega$. Assume that $\mathbb{P}(B) > 0$. Prove that if $\mathbb{P}(A_1|B) < \mathbb{P}(A_1)$ then $\mathbb{P}(A_i|B) > \mathbb{P}(A_i)$ for some $i = 2, \ldots, k$.

19. Suppose that 30 percent of computer owners use a Macintosh, 50 percent use Windows, and 20 percent use Linux. Suppose that 65 percent of the Mac users have succumbed to a computer virus, 82 percent of the Windows users get the virus, and 50 percent of the Linux users get the virus. We select a person at random and learn that her system was infected with the virus. What is the probability that she is a Windows user?

20. A box contains 5 coins and each has a different probability of showing heads. Let $p_1, \ldots, p_5$ denote the probability of heads on each coin. Suppose that

$$p_1 = 0,\ p_2 = 1/4,\ p_3 = 1/2,\ p_4 = 3/4 \ \text{ and } \ p_5 = 1.$$

Let H denote "heads is obtained" and let C_i denote the event that coin i is selected.

(a) Select a coin at random and toss it. Suppose a head is obtained. What is the posterior probability that coin i was selected ($i = 1, \ldots, 5$)? In other words, find $\mathbb{P}(C_i|H)$ for $i = 1, \ldots, 5$.

(b) Toss the coin again. What is the probability of another head? In other words find $\mathbb{P}(H_2|H_1)$ where $H_j =$ "heads on toss j."

Now suppose that the experiment was carried out as follows: We select a coin at random and toss it until a head is obtained.

(c) Find $\mathbb{P}(C_i|B_4)$ where $B_4 =$ "first head is obtained on toss 4."

21. (Computer Experiment.) Suppose a coin has probability p of falling heads up. If we flip the coin many times, we would expect the proportion of heads to be near p. We will make this formal later. Take $p = .3$ and $n = 1,000$ and simulate n coin flips. Plot the proportion of heads as a function of n. Repeat for $p = .03$.

22. (Computer Experiment.) Suppose we flip a coin n times and let p denote the probability of heads. Let X be the number of heads. We call X a binomial random variable, which is discussed in the next chapter. Intuition suggests that X will be close to $n\,p$. To see if this is true, we can repeat this experiment many times and average the X values. Carry

out a simulation and compare the average of the X's to $n\,p$. 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*

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

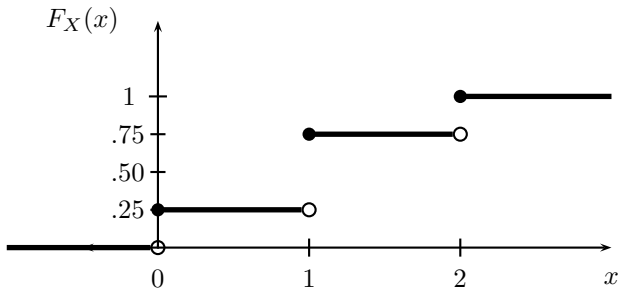


FIGURE 2.1. CDF for flipping a coin twice (Example 2.6.)

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

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

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 $F(x_1) \leq F(x_2)$.

(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.

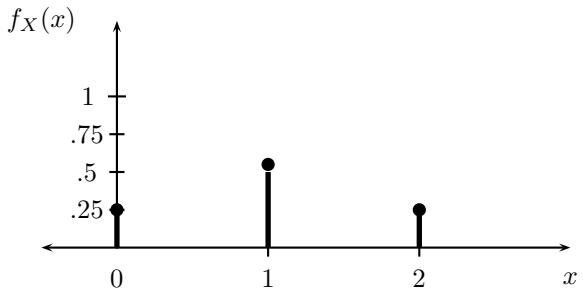


FIGURE 2.2. Probability function for flipping a coin twice (Example 2.6).

2.11 Definition. *A random variable X is* **continuous** *if there exists a function f_X such that $f_X(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and for every $a \leq b$,*

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx. \tag{2.2}$$

The function f_X is called the **probability density function (**PDF**).** *We have that*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

and $f_X(x) = F'_X(x)$ at all points x at which F_X is differentiable.

Sometimes we write $\int f(x)dx$ or $\int f$ to mean $\int_{-\infty}^{\infty} f(x)dx$.

2.12 Example. Suppose that X has PDF

$$f_X(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, $f_X(x) \geq 0$ and $\int f_X(x)dx = 1$. A random variable with this density is said to have a Uniform (0,1) distribution. This is meant to capture the idea of choosing a point at random between 0 and 1. The CDF is given by

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1. \end{cases}$$

See Figure 2.3. ■

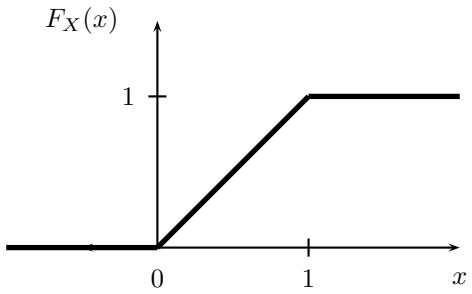


FIGURE 2.3. CDF for Uniform (0,1).

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* $F(x^-) = \lim_{y \uparrow x} F(y)$;

2. $\mathbb{P}(x < X \leq y) = F(y) - F(x)$*;*

3. $\mathbb{P}(X > x) = 1 - F(x)$*;*

4. If X *is continuous then*

$$\begin{aligned} F(b) - F(a) &= \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) \\ &= \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b). \end{aligned}$$

It is also useful to define the inverse CDF (or quantile function).

2.16 Definition. *Let* X *be a random variable with* CDF F*. The* **inverse CDF** *or* **quantile function** *is defined by*[4]

$$F^{-1}(q) = \inf\Big\{x : \ F(x) > q\Big\}$$

for $q \in [0, 1]$*. If* F *is strictly increasing and continuous then* $F^{-1}(q)$ *is the unique real number* x *such that* $F(x) = q$*.*

We call $F^{-1}(1/4)$ the **first quartile**, $F^{-1}(1/2)$ the **median** (or second quartile), and $F^{-1}(3/4)$ the **third quartile**.

Two random variables X and Y are **equal in distribution** — written $X \stackrel{d}{=} Y$ — if $F_X(x) = F_Y(x)$ for all x. This does not mean that X and Y are equal. Rather, it means that all probability statements about X and Y will be the same. For example, suppose that $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. Let $Y = -X$. Then $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = 1/2$ and so $X \stackrel{d}{=} Y$. But X and Y are not equal. In fact, $\mathbb{P}(X = Y) = 0$.

2.3 Some Important Discrete Random Variables

Warning About Notation! It is traditional to write $X \sim F$ to indicate that X has distribution F. This is unfortunate notation since the symbol $\sim$ is also used to denote an approximation. The notation $X \sim F$ is so pervasive that we are stuck with it. Read $X \sim F$ as "X has distribution F" **not** as "X is approximately F".

[4] If you are unfamiliar with "inf", just think of it as the minimum.

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 $X_2 \sim \text{Binomial}(n_2, p)$ 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

$$\mathbb{P}(X = k) = p(1-p)^{k-1}, \quad k \geq 1.$$

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 $X_2 \sim \text{Poisson}(\lambda_2)$ 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

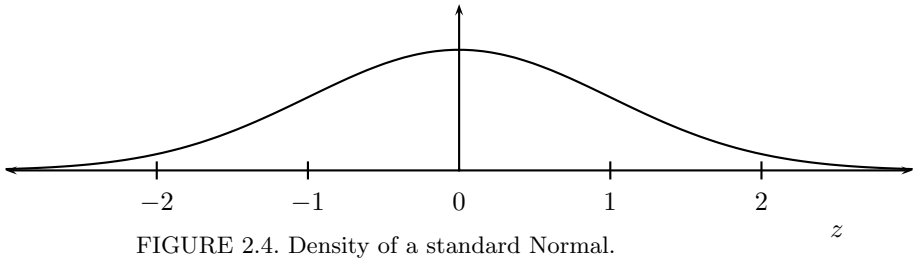


FIGURE 2.4. Density of a standard Normal.

compute $\Phi(z)$ and $\Phi^{-1}(q)$. Most statistics texts, including 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

β, denoted by $X \sim \text{Gamma}(\alpha, \beta)$, if

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

where $\alpha, \beta > 0$. The exponential distribution is just a Gamma$(1, \beta)$ distribution. If $X_i \sim \text{Gamma}(\alpha_i, \beta)$ are independent, then $\sum_{i=1}^n X_i \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$.

THE BETA DISTRIBUTION. X has a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$, denoted by $X \sim \text{Beta}(\alpha, \beta)$, if

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1.$$

t AND CAUCHY DISTRIBUTION. X has a t distribution with ν degrees of freedom — written $X \sim t_\nu$ — if

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}.$$

The t distribution is similar to a Normal but it has thicker tails. In fact, the Normal corresponds to a t with $\nu = \infty$. The Cauchy distribution is a special case of the t distribution corresponding to $\nu = 1$. The density is

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

To see that this is indeed a density:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{d\tan^{-1}(x)}{dx} \\ &= \frac{1}{\pi}\left[\tan^{-1}(\infty) - \tan^{-1}(-\infty)\right] = \frac{1}{\pi}\left[\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right] = 1. \end{aligned}$$

THE χ^2 DISTRIBUTION. X has a χ^2 distribution with p degrees of freedom — written $X \sim \chi^2_p$ — if

$$f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, \quad x > 0.$$

If $Z_1, \ldots, Z_p$ are independent standard Normal random variables then $\sum_{i=1}^p Z_i^2 \sim \chi^2_p$.

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

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

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, $\mathbb{P}(X < 1/2, Y < 1/2) = 1/4$. ■

2.21 Example. Let (X, Y) have density

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

Then

$$\begin{aligned} \int_0^1 \int_0^1 (x+y)dxdy &= \int_0^1 \left[\int_0^1 x\,dx\right] dy + \int_0^1 \left[\int_0^1 y\,dx\right] dy \\ &= \int_0^1 \frac{1}{2} dy + \int_0^1 y\,dy = \frac{1}{2} + \frac{1}{2} = 1 \end{aligned}$$

which verifies that this is a PDF ■

2.22 Example. If the distribution is defined over a non-rectangular region, then the calculations are a bit more complicated. Here is an example which I borrowed from DeGroot and Schervish (2002). Let (X, Y) have density

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

Note first that $-1 \le x \le 1$. Now let us find the value of c. The trick here is to be careful about the range of integration. We pick one variable, x say, and let it range over its values. Then, for each fixed value of x, we let y vary over its range, which is $x^2 \le y \le 1$. It may help if you look at Figure 2.5. Thus,

$$\begin{aligned} 1 &= \int\int f(x,y)dydx = c\int_{-1}^1 \int_{x^2}^1 x^2 y\,dy\,dx \\ &= c\int_{-1}^1 x^2 \left[\int_{x^2}^1 y\,dy\right] dx = c\int_{-1}^1 x^2 \frac{1-x^4}{2} dx = \frac{4c}{21}. \end{aligned}$$

Hence, $c = 21/4$. Now let us compute $\mathbb{P}(X \ge Y)$. This corresponds to the set $A = \{(x,y); 0 \le x \le 1, x^2 \le y \le x\}$. (You can see this by drawing a diagram.) So,

$$\begin{aligned} \mathbb{P}(X \ge Y) &= \frac{21}{4}\int_0^1 \int_{x^2}^x x^2\,y\,dy\,dx = \frac{21}{4}\int_0^1 x^2 \left[\int_{x^2}^x y\,dy\right] dx \\ &= \frac{21}{4}\int_0^1 x^2 \left(\frac{x^2 - x^4}{2}\right) dx = \frac{3}{20}. \quad ■ \end{aligned}$$

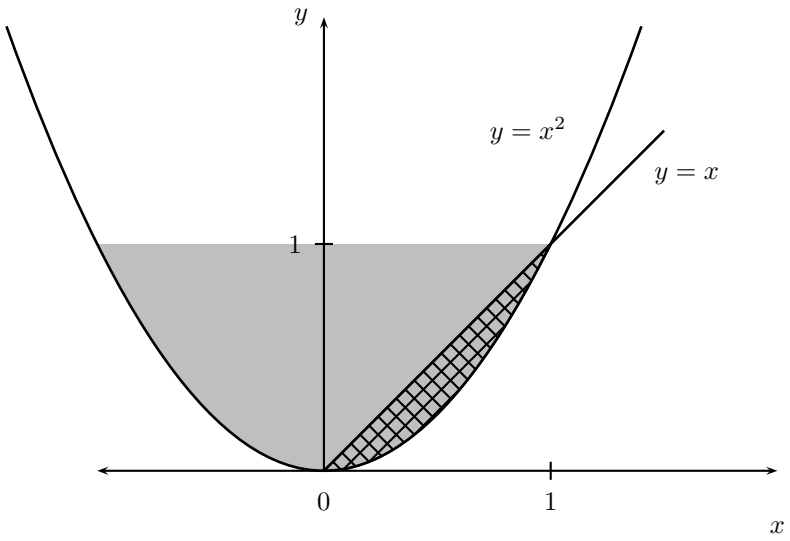


FIGURE 2.5. The light shaded region is $x^2 \leq y \leq 1$. The density is positive over this region. The hatched region is the event $X \geq Y$ intersected with $x^2 \leq y \leq 1$.

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*

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

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

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

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,*

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

and we write $X \amalg Y$. Otherwise we say that X and Y are **dependent** *and we write $X \wavy Y$.*

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

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

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

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

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

Now,

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

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

2.36 Definition. *For continuous random variables, the* **conditional probability density function** *is*

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

assuming that $f_Y(y) > 0$. *Then,*

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx.$$

2.37 Example. Let X and Y have a joint uniform distribution on the unit square. Thus, $f_{X|Y}(x|y) = 1$ for $0 \leq x \leq 1$ and 0 otherwise. Given $Y = y$, X is Uniform$(0, 1)$. We can write this as $X|Y = y \sim$ Uniform$(0, 1)$. ■

From the definition of the conditional density, we see that $f_{X,Y}(x, y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x)$. This can sometimes be useful as in example 2.39.

2.38 Example. Let

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

Let us find $\mathbb{P}(X < 1/4|Y = 1/3)$. In example 2.27 we saw that $f_Y(y) = y + (1/2)$. Hence,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{x+y}{y+\frac{1}{2}}.$$

So,

$$\begin{aligned} P\left(X < \frac{1}{4} \,\middle|\, Y = \frac{1}{3}\right) &= \int_0^{1/4} f_{X|Y}\left(x \,\middle|\, \frac{1}{3}\right) dx \\ &= \int_0^{1/4} \frac{x + \frac{1}{3}}{\frac{1}{3} + \frac{1}{2}} dx = \frac{\frac{1}{32} + \frac{1}{12}}{\frac{1}{3} + \frac{1}{2}} = \frac{11}{80}. \quad \blacksquare \end{aligned}$$

2.39 Example. Suppose that $X \sim$ Uniform$(0, 1)$. After obtaining a value of X we generate $Y|X = x \sim$ Uniform$(x, 1)$. What is the marginal distribution

---

avoid this problem by defining things in terms of the PDF. The fact that this leads to a well-defined theory is proved in more advanced courses. Here, we simply take it as a definition.

of Y? First note that,

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

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 $x^2 \leq y \leq 1$. Earlier, we saw that $f_X(x) = (21/8)x^2(1-x^4)$. Hence, for $x^2 \leq y \leq 1$,

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

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

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

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

2.41 Definition. *If $X_1, \ldots, X_n$ are independent and each has the same marginal distribution with* CDF *F, we say that $X_1, \ldots, X_n$ are* IID *(independent and identically distributed) and we write*

$$X_1, \ldots X_n \sim F.$$

If F has density f we also write $X_1, \ldots X_n \sim f$. We also call $X_1, \ldots, X_n$ a **random sample of size n from F.**

Much of statistical theory and practice begins with IID observations and we shall study this case in detail when we discuss statistics.

2.10 Two Important Multivariate Distributions

MULTINOMIAL. The multivariate version of a Binomial is called a Multinomial. Consider drawing a ball from an urn which has balls with k different colors labeled "color 1, color 2, ..., color k." Let $p = (p_1, \ldots, p_k)$ where $p_j \geq 0$ and $\sum_{j=1}^k p_j = 1$ and suppose that p_j is the probability of drawing a ball of color j. Draw n times (independent draws with replacement) and let $X = (X_1, \ldots, X_k)$ where X_j is the number of times that color j appears. Hence, $n = \sum_{j=1}^k X_j$. We say that X has a Multinomial (n,p) distribution written $X \sim \text{Multinomial}(n, p)$. The probability function is

$$f(x) = \binom{n}{x_1 \ldots x_k} p_1^{x_1} \cdots p_k^{x_k} \tag{2.9}$$

where

$$\binom{n}{x_1 \ldots x_k} = \frac{n!}{x_1! \cdots x_k!}.$$

2.42 Lemma. *Suppose that $X \sim \text{Multinomial}(n, p)$ where $X = (X_1, \ldots, X_k)$ and $p = (p_1, \ldots, p_k)$. The marginal distribution of X_j is Binomial (n,p_j).*

MULTIVARIATE NORMAL. The univariate Normal has two parameters, μ and σ. In the multivariate version, μ is a vector and σ is replaced by a matrix Σ. To begin, let

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_k \end{pmatrix}$$

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]

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

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

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

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,

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

Now comes the hard part: finding A_y. First suppose that $0 < y \leq 1$. 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,

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

By differentiation, the PDF is

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

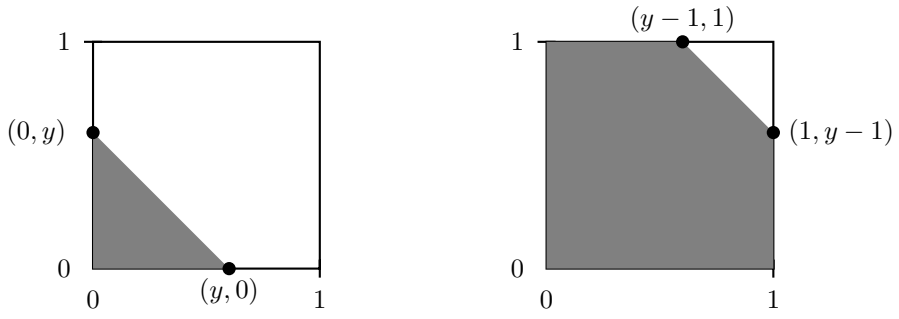
2.13 Appendix

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, $\{\omega : \ X(\omega) \leq x\} \in \mathcal{A}$.

2.14 Exercises

1. Show that

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



This is the case $0 \leq y < 1$. This is the case $1 \leq y \leq 2$.

FIGURE 2.6. The set A_y for example 2.48. A_y consists of all points (x_1, x_2) in the square below the line $x_2 = y - x_1$.

2. Let X be such that $\mathbb{P}(X = 2) = \mathbb{P}(X = 3) = 1/10$ and $\mathbb{P}(X = 5) = 8/10$. Plot the CDF F. Use F to find $\mathbb{P}(2 < X \leq 4.8)$ and $\mathbb{P}(2 \leq X \leq 4.8)$.

3. Prove Lemma 2.15.

4. Let X have probability density function

$$f_X(x) = \begin{cases} 1/4 & 0 < x < 1 \\ 3/8 & 3 < x < 5 \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Find the cumulative distribution function of X.

   (b) Let $Y = 1/X$. Find the probability density function $f_Y(y)$ for Y. Hint: Consider three cases: $\frac{1}{5} \leq y \leq \frac{1}{3}$, $\frac{1}{3} \leq y \leq 1$, and $y \geq 1$.

5. Let X and Y be discrete random variables. Show that X and Y are independent if and only if $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ for all x and y.

6. Let X have distribution F and density function f and let A be a subset of the real line. Let $I_A(x)$ be the indicator function for A:

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases}$$

   Let $Y = I_A(X)$. Find an expression for the cumulative distribution of Y. (Hint: first find the probability mass function for Y.)

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 $\pi = \lambda/(\lambda + \mu)$.

Hint 1: You may use the following fact: If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$, and X and Y are independent, then $X+Y \sim \text{Poisson}(\mu+\lambda)$.

Hint 2: Note that $\{X = x,\ X + Y = n\} = \{X = x,\ Y = n - x\}$.

17. Let
$$f_{X,Y}(x,y) = \begin{cases} c(x+y^2) & 0 \le x \le 1 \text{ and } 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$
Find $P\left(X < \frac{1}{2} \mid Y = \frac{1}{2}\right)$.

18. Let $X \sim N(3, 16)$. Solve the following using the Normal table and using a computer package.
    (a) Find $\mathbb{P}(X < 7)$.
    (b) Find $\mathbb{P}(X > -2)$.
    (c) Find x such that $\mathbb{P}(X > x) = .05$.
    (d) Find $\mathbb{P}(0 \le X < 4)$.
    (e) Find x such that $\mathbb{P}(|X| > |x|) = .05$.

19. Prove formula (2.12).

20. Let $X, Y \sim \text{Uniform}(0,1)$ be independent. Find the PDF for $X - Y$ and X/Y.

21. Let $X_1, \ldots, X_n \sim \text{Exp}(\beta)$ be IID. Let $Y = \max\{X_1, \ldots, X_n\}$. Find the PDF of Y. Hint: $Y \le y$ if and only if $X_i \le y$ for $i = 1, \ldots, n$.

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, $\mathbb{E}(X) = \int x d\,F_X(x) = \int xf_X(x)dx = \frac{1}{4}\int_{-1}^{3} x\,dx = 1$. ■

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

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

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

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 $\mathbb{E}(X) = \mathbb{E}(\sum_i X_i) = \sum_i \mathbb{E}(X_i) = np$. $\blacksquare$

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.

3.14 Definition. *Let X be a random variable with mean μ. The* **variance** *of X — denoted by σ^2 or σ_X^2 or $\mathbb{V}(X)$ or $\mathbb{V}X$ — is defined by*

$$\sigma^2 = \mathbb{E}(X-\mu)^2 = \int (x-\mu)^2 dF(x) \tag{3.7}$$

assuming this expectation exists. The **standard deviation** *is* $\text{sd}(X) = \sqrt{\mathbb{V}(X)}$ *and is also denoted by σ and σ_X.*

3.15 Theorem. *Assuming the variance is well defined, it has the following properties:*

1. $\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2$.
2. *If a and b are constants then* $\mathbb{V}(aX+b) = a^2\mathbb{V}(X)$.
3. *If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants, then*

$$\mathbb{V}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i). \tag{3.8}$$

3.16 Example. Let $X \sim \text{Binomial}(n,p)$. We write $X = \sum_i X_i$ where $X_i = 1$ if toss i is heads and $X_i = 0$ otherwise. Then $X = \sum_i X_i$ and the random variables are independent. Also, $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1-p$. Recall that

$$\mathbb{E}(X_i) = \Big(p \times 1\Big) + \Big((1-p) \times 0\Big) = p.$$

Now,

$$\mathbb{E}(X_i^2) = \Big(p \times 1^2\Big) + \Big((1-p) \times 0^2\Big) = p.$$

Therefore, $\mathbb{V}(X_i) = \mathbb{E}(X_i^2) - p^2 = p - p^2 = p(1-p)$. Finally, $\mathbb{V}(X) = \mathbb{V}(\sum_i X_i) = \sum_i \mathbb{V}(X_i) = \sum_i p(1-p) = np(1-p)$. Notice that $\mathbb{V}(X) = 0$ if $p = 1$ or $p = 0$. Make sure you see why this makes intuitive sense. ∎

If $X_1, \ldots, X_n$ are random variables then we define the **sample mean** to be

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \tag{3.9}$$

and the **sample variance** to be

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2. \tag{3.10}$$

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:

Distribution	Mean	Variance
Point mass at a	a	0
Bernoulli(p)	p	$p(1-p)$
Binomial(n,p)	np	$np(1-p)$
Geometric(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform(a,b)	$(a+b)/2$	$(b-a)^2/12$
Normal(μ,σ^2)	μ	σ^2
Exponential(β)	β	β^2
Gamma(α,β)	$\alpha\beta$	$\alpha\beta^2$
Beta(α,β)	$\alpha/(\alpha+\beta)$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$
t_ν	0 (if $\nu > 1$)	$\nu/(\nu-2)$ (if $\nu > 2$)
χ^2_p	p	$2p$
Multinomial(n,p)	np	see below
Multivariate Normal(μ,Σ)	μ	Σ

We derived $\mathbb{E}(X)$ and $\mathbb{V}(X)$ for the Binomial in the previous section. The calculations for some of the others are in the exercises.

The last two entries in the table are multivariate models which involve a random vector X of the form

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}.$$

The mean of a random vector X is defined by

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_k) \end{pmatrix}.$$

The **variance-covariance matrix** Σ is defined to be

$$\mathbb{V}(X) = \begin{bmatrix} \mathbb{V}(X_1) & \mathsf{Cov}(X_1, X_2) & \cdots & \mathsf{Cov}(X_1, X_k) \\ \mathsf{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathsf{Cov}(X_2, X_k) \\ \vdots & \vdots & \vdots & \vdots \\ \mathsf{Cov}(X_k, X_1) & \mathsf{Cov}(X_k, X_2) & \cdots & \mathbb{V}(X_k) \end{bmatrix}.$$

If $X \sim \text{Multinomial}(n,p)$ then $\mathbb{E}(X) = np = n(p_1, \ldots, p_k)$ and

$$\mathbb{V}(X) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_k \\ -np_2p_1 & np_2(1-p_2) & \cdots & -np_2p_k \\ \vdots & \vdots & \vdots & \vdots \\ -np_kp_1 & -np_kp_2 & \cdots & np_k(1-p_k) \end{pmatrix}.$$

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*

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

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

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

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 $f(x,y) = f(x)f(y|x)$,

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

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,

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

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, $\mathbb{V}\mathbb{E}(X|Q) = \mathbb{V}(nQ) = n^2\mathbb{V}(Q) = n^2\int(q - (1/2))^2dq = n^2/12$. 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.

By taking k derivatives we conclude that $\psi^{(k)}(0) = \mathbb{E}(X^k)$. This gives us a method for computing the moments of a distribution.

3.30 Example. Let $X \sim \text{Exp}(1)$. For any $t < 1$,

$$\psi_X(t) = \mathbb{E}e^{tX} = \int_0^\infty e^{tx}e^{-x}dx = \int_0^\infty e^{(t-1)x}dx = \frac{1}{1-t}.$$

The integral is divergent if $t \geq 1$. So, $\psi_X(t) = 1/(1-t)$ for all $t < 1$. Now, $\psi'(0) = 1$ and $\psi''(0) = 2$. Hence, $\mathbb{E}(X) = 1$ and $\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2 = 2 - 1 = 1$. ■

3.31 Lemma. *Properties of the* MGF.

(1) If $Y = aX + b$, then $\psi_Y(t) = e^{bt}\psi_X(at)$.

(2) If $X_1, \ldots, X_n$ are independent and $Y = \sum_i X_i$, then $\psi_Y(t) = \prod_i \psi_i(t)$ where ψ_i is the MGF *of X_i.*

3.32 Example. Let $X \sim \text{Binomial}(n, p)$. We know that $X = \sum_{i=1}^n X_i$ where $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1 - p$. Now $\psi_i(t) = \mathbb{E}e^{X_i t} = (p \times e^t) + ((1 - p)) = pe^t + q$ where $q = 1 - p$. Thus, $\psi_X(t) = \prod_i \psi_i(t) = (pe^t + q)^n$. ■

Recall that X and Y are equal in distribution if they have the same distribution function and we write $X \stackrel{d}{=} Y$.

3.33 Theorem. *Let X and Y be random variables. If $\psi_X(t) = \psi_Y(t)$ for all t in an open interval around 0, then $X \stackrel{d}{=} Y$.*

3.34 Example. Let $X_1 \sim \text{Binomial}(n_1, p)$ and $X_2 \sim \text{Binomial}(n_2, p)$ be independent. Let $Y = X_1 + X_2$. Then,

$$\psi_Y(t) = \psi_1(t)\psi_2(t) = (pe^t + q)^{n_1}(pe^t + q)^{n_2} = (pe^t + q)^{n_1+n_2}$$

and we recognize the latter as the MGF of a Binomial$(n_1 + n_2, p)$ distribution. Since the MGF characterizes the distribution (i.e., there can't be another random variable which has the same MGF) we conclude that $Y \sim \text{Binomial}(n_1 + n_2, p)$. ■

Moment Generating Functions for Some Common Distributions

Distribution	MGF $\psi(t)$
Bernoulli(p)	$pe^t + (1-p)$
Binomial(n,p)	$(pe^t + (1-p))^n$
Poisson(λ)	$e^{\lambda(e^t-1)}$
Normal(μ,σ)	$\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$
Gamma(α,β)	$\left(\frac{1}{1-\beta t}\right)^\alpha$ for $t < 1/\beta$

3.35 Example. Let $Y_1 \sim \text{Poisson}(\lambda_1)$ and $Y_2 \sim \text{Poisson}(\lambda_2)$ be independent. The moment generating function of $Y = Y_1 + Y + 2$ is $\psi_Y(t) = \psi_{Y_1}(t)\psi_{Y_2}(t) = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$ which is the moment generating function of a Poisson($\lambda_1 + \lambda_2$). We have thus proved that the sum of two independent Poisson random variables has a Poisson distribution. ∎

3.7 Appendix

Expectation as an Integral. The integral of a measurable function $r(x)$ is defined as follows. First suppose that r is simple, meaning that it takes finitely many values $a_1, \ldots, a_k$ over a partition $A_1, \ldots, A_k$. Then define

$$\int r(x)dF(x) = \sum_{i=1}^{k} a_i \, \mathbb{P}(r(X) \in A_i).$$

The integral of a positive measurable function r is defined by $\int r(x)dF(x) = \lim_i \int r_i(x)dF(x)$ where r_i is a sequence of simple functions such that $r_i(x) \leq r(x)$ and $r_i(x) \to r(x)$ as $i \to \infty$. This does not depend on the particular sequence. The integral of a measurable function r is defined to be $\int r(x)dF(x) = \int r^+(x)dF(x) - \int r^-(x)dF(x)$ assuming both integrals are finite, where $r^+(x) = \max\{r(x), 0\}$ and $r^-(x) = -\min\{r(x), 0\}$.

3.8 Exercises

1. Suppose we play a game where we start with c dollars. On each play of the game you either double or halve your money, with equal probability. What is your expected fortune after n trials?

2. Show that $\mathbb{V}(X) = 0$ if and only if there is a constant c such that $P(X = c) = 1$.

3. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$ and let $Y_n = \max\{X_1, \ldots, X_n\}$. Find $\mathbb{E}(Y_n)$.

4. A particle starts at the origin of the real line and moves along the line in jumps of one unit. For each jump the probability is p that the particle will jump one unit to the left and the probability is $1-p$ that the particle will jump one unit to the right. Let X_n be the position of the particle after n units. Find $\mathbb{E}(X_n)$ and $\mathbb{V}(X_n)$. (This is known as a **random walk**.)

5. A fair coin is tossed until a head is obtained. What is the expected number of tosses that will be required?

6. Prove Theorem 3.6 for discrete random variables.

7. Let X be a continuous random variable with CDF F. Suppose that $P(X > 0) = 1$ and that $\mathbb{E}(X)$ exists. Show that $\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > x)dx$.

   Hint: Consider integrating by parts. The following fact is helpful: if $\mathbb{E}(X)$ exists then $\lim_{x\to\infty} x[1 - F(x)] = 0$.

8. Prove Theorem 3.17.

9. (Computer Experiment.) Let $X_1, X_2, \ldots, X_n$ be $N(0, 1)$ random variables and let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Plot $\overline{X}_n$ versus n for $n = 1, \ldots, 10,000$. Repeat for $X_1, X_2, \ldots, X_n \sim$ Cauchy. Explain why there is such a difference.

10. Let $X \sim N(0, 1)$ and let $Y = e^X$. Find $\mathbb{E}(Y)$ and $\mathbb{V}(Y)$.

11. (Computer Experiment: Simulating the Stock Market.) Let $Y_1, Y_2, \ldots$ be independent random variables such that $P(Y_i = 1) = P(Y_i = -1) = 1/2$. Let $X_n = \sum_{i=1}^n Y_i$. Think of $Y_i = 1$ as "the stock price increased by one dollar", $Y_i = -1$ as "the stock price decreased by one dollar", and X_n as the value of the stock on day n.

    (a) Find $\mathbb{E}(X_n)$ and $\mathbb{V}(X_n)$.

    (b) Simulate X_n and plot X_n versus n for $n = 1, 2, \ldots, 10,000$. Repeat the whole simulation several times. Notice two things. First, it's easy to "see" patterns in the sequence even though it is random. Second,

you will find that the four runs look very different even though they were generated the same way. How do the calculations in (a) explain the second observation?

12. Prove the formulas given in the table at the beginning of Section 3.4 for the Bernoulli, Poisson, Uniform, Exponential, Gamma, and Beta. Here are some hints. For the mean of the Poisson, use the fact that $e^a = \sum_{x=0}^{\infty} a^x/x!$. To compute the variance, first compute $\mathbb{E}(X(X-1))$. For the mean of the Gamma, it will help to multiply and divide by $\Gamma(\alpha+1)/\beta^{\alpha+1}$ and use the fact that a Gamma density integrates to 1. For the Beta, multiply and divide by $\Gamma(\alpha+1)\Gamma(\beta)/\Gamma(\alpha+\beta+1)$.

13. Suppose we generate a random variable X in the following way. First we flip a fair coin. If the coin is heads, take X to have a Unif(0,1) distribution. If the coin is tails, take X to have a Unif(3,4) distribution.

    (a) Find the mean of X.

    (b) Find the standard deviation of X.

14. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be random variables and let $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ be constants. Show that

$$\mathsf{Cov}\left(\sum_{i=1}^{m} a_i X_i,\ \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \mathsf{Cov}(X_i, Y_j).$$

15. Let

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{3}(x+y) & 0 \le x \le 1,\ 0 \le y \le 2 \\ 0 & \text{otherwise.} \end{cases}$$

    Find $\mathbb{V}(2X - 3Y + 8)$.

16. Let r(x) be a function of x and let s(y) be a function of y. Show that

$$\mathbb{E}(r(X)s(Y)|X) = r(X)\mathbb{E}(s(Y)|X).$$

    Also, show that $\mathbb{E}(r(X)|X) = r(X)$.

17. Prove that

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

    Hint: Let $m = \mathbb{E}(Y)$ and let $b(x) = \mathbb{E}(Y|X = x)$. Note that $\mathbb{E}(b(X)) = \mathbb{E}\mathbb{E}(Y|X) = \mathbb{E}(Y) = m$. Bear in mind that b is a function of x. Now write $\mathbb{V}(Y) = \mathbb{E}(Y-m)^2 = \mathbb{E}((Y - b(X)) + (b(X) - m))^2$. Expand the

square and take the expectation. You then have to take the expectation of three terms. In each case, use the rule of the iterated expectation: $\mathbb{E}(\text{stuff}) = \mathbb{E}(\mathbb{E}(\text{stuff}|X))$.

18. Show that if $\mathbb{E}(X|Y = y) = c$ for some constant c, then X and Y are uncorrelated.

19. This question is to help you understand the idea of a **sampling distribution.** Let $X_1, \ldots, X_n$ be IID with mean μ and variance σ^2. Let $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then $\overline{X}_n$ is a **statistic**, that is, a function of the data. Since $\overline{X}_n$ is a random variable, it has a distribution. This distribution is called the *sampling distribution of the statistic.* Recall from Theorem 3.17 that $\mathbb{E}(\overline{X}_n) = \mu$ and $\mathbb{V}(\overline{X}_n) = \sigma^2/n$. Don't confuse the distribution of the data f_X and the distribution of the statistic $f_{\overline{X}_n}$. To make this clear, let $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$. Let f_X be the density of the Uniform$(0, 1)$. Plot f_X. Now let $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Find $\mathbb{E}(\overline{X}_n)$ and $\mathbb{V}(\overline{X}_n)$. Plot them as a function of n. Interpret. Now simulate the distribution of $\overline{X}_n$ for $n = 1, 5, 25, 100$. Check that the simulated values of $\mathbb{E}(\overline{X}_n)$ and $\mathbb{V}(\overline{X}_n)$ agree with your theoretical calculations. What do you notice about the sampling distribution of $\overline{X}_n$ as n increases?

20. Prove Lemma 3.21.

21. Let X and Y be random variables. Suppose that $\mathbb{E}(Y|X) = X$. Show that $\mathsf{Cov}(X, Y) = \mathbb{V}(X)$.

22. Let $X \sim \text{Uniform}(0, 1)$. Let $0 < a < b < 1$. Let

$$Y = \begin{cases} 1 & 0 < x < b \\ 0 & \text{otherwise} \end{cases}$$

and let

$$Z = \begin{cases} 1 & a < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Are Y and Z independent? Why/Why not?

(b) Find $\mathbb{E}(Y|Z)$. Hint: What values z can Z take? Now find $\mathbb{E}(Y|Z = z)$.

23. Find the moment generating function for the Poisson, Normal, and Gamma distributions.

24. Let $X_1, \ldots, X_n \sim \text{Exp}(\beta)$. Find the moment generating function of X_i. Prove that $\sum_{i=1}^n X_i \sim \text{Gamma}(n, \beta)$.

4
Inequalities

4.1 Probability Inequalities

Inequalities are useful for bounding quantities that might otherwise be hard to compute. They will also be used in the theory of convergence which is discussed in the next chapter. Our first inequality is Markov's inequality.

4.1 Theorem (Markov's inequality)**.** *Let X be a non-negative random variable and suppose that $\mathbb{E}(X)$ exists. For any $t > 0$,*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}. \tag{4.1}$$

PROOF. Since $X > 0$,

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty x f(x)dx = \int_0^t x f(x)dx + \int_t^\infty x f(x)dx \\ &\geq \int_t^\infty x f(x)dx \geq t \int_t^\infty f(x)dx = t\,\mathbb{P}(X > t) \quad \blacksquare \end{aligned}$$

4.2 Theorem (Chebyshev's inequality). *Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathbb{V}(X)$. Then,*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \text{and} \quad \mathbb{P}(|Z| \geq k) \leq \frac{1}{k^2} \tag{4.2}$$

where $Z = (X - \mu)/\sigma$. In particular, $\mathbb{P}(|Z| > 2) \leq 1/4$ and $\mathbb{P}(|Z| > 3) \leq 1/9$.

PROOF. We use Markov's inequality to conclude that

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(|X - \mu|^2 \geq t^2) \leq \frac{\mathbb{E}(X - \mu)^2}{t^2} = \frac{\sigma^2}{t^2}.$$

The second part follows by setting $t = k\sigma$. ∎

4.3 Example. Suppose we test a prediction method, a neural net for example, on a set of n new test cases. Let $X_i = 1$ if the predictor is wrong and $X_i = 0$ if the predictor is right. Then $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ is the observed error rate. Each X_i may be regarded as a Bernoulli with unknown mean p. We would like to know the true — but unknown — error rate p. Intuitively, we expect that $\overline{X}_n$ should be close to p. How likely is $\overline{X}_n$ to not be within ϵ of p? We have that $\mathbb{V}(\overline{X}_n) = \mathbb{V}(X_1)/n = p(1-p)/n$ and

$$\mathbb{P}(|\overline{X}_n - p| > \epsilon) \leq \frac{\mathbb{V}(\overline{X}_n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$

since $p(1-p) \leq \frac{1}{4}$ for all p. For $\epsilon = .2$ and $n = 100$ the bound is .0625. ∎

Hoeffding's inequality is similar in spirit to Markov's inequality but it is a sharper inequality. We present the result here in two parts.

4.4 Theorem (Hoeffding's Inequality). *Let $Y_1, \ldots, Y_n$ be independent observations such that $\mathbb{E}(Y_i) = 0$ and $a_i \leq Y_i \leq b_i$. Let $\epsilon > 0$. Then, for any $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n Y_i \geq \epsilon\right) \leq e^{-t\epsilon} \prod_{i=1}^n e^{t^2(b_i - a_i)^2/8}. \tag{4.3}$$

4.5 Theorem. *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(|\overline{X}_n - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2} \tag{4.4}$$

where $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$.

4.6 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Let $n = 100$ and $\epsilon = .2$. We saw that Chebyshev's inequality yielded

$$\mathbb{P}(|\overline{X}_n - p| > \epsilon) \leq .0625.$$

According to Hoeffding's inequality,

$$\mathbb{P}(|\overline{X}_n - p| > .2) \leq 2e^{-2(100)(.2)^2} = .00067$$

which is much smaller than .0625. ■

Hoeffding's inequality gives us a simple way to create a **confidence interval** for a binomial parameter p. We will discuss confidence intervals in detail later (see Chapter 6) but here is the basic idea. Fix $\alpha > 0$ and let

$$\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}.$$

By Hoeffding's inequality,

$$\mathbb{P}\left(|\overline{X}_n - p| > \epsilon_n\right) \leq 2e^{-2n\epsilon_n^2} = \alpha.$$

Let $C = (\overline{X}_n - \epsilon_n, \overline{X}_n + \epsilon_n)$. Then, $\mathbb{P}(p \notin C) = \mathbb{P}(|\overline{X}_n - p| > \epsilon_n) \leq \alpha$. Hence, $\mathbb{P}(p \in C) \geq 1 - \alpha$, that is, the random interval C traps the true parameter value p with probability $1 - \alpha$; we call C a $1 - \alpha$ confidence interval. More on this later.

The following inequality is useful for bounding probability statements about Normal random variables.

4.7 Theorem (Mill's Inequality). *Let $Z \sim N(0, 1)$. Then,*

$$\mathbb{P}(|Z| > t) \leq \sqrt{\frac{2}{\pi}}\frac{e^{-t^2/2}}{t}.$$

4.2 Inequalities For Expectations

This section contains two inequalities on expected values.

4.8 Theorem (Cauchy-Schwartz inequality). *If X and Y have finite variances then*

$$\mathbb{E}\,|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}. \tag{4.5}$$

Recall that a function g is **convex** if for each x, y and each $\alpha \in [0,1]$,

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

If g is twice differentiable and $g''(x) \geq 0$ for all x, then g is convex. It can be shown that if g is convex, then g lies above any line that touches g at some point, called a tangent line. A function g is **concave** if $-g$ is convex. Examples of convex functions are $g(x) = x^2$ and $g(x) = e^x$. Examples of concave functions are $g(x) = -x^2$ and $g(x) = \log x$.

4.9 Theorem (Jensen's inequality). *If g is convex, then*

$$\mathbb{E}g(X) \geq g(\mathbb{E}X). \tag{4.6}$$

If g is concave, then

$$\mathbb{E}g(X) \leq g(\mathbb{E}X). \tag{4.7}$$

PROOF. Let $L(x) = a + bx$ be a line, tangent to $g(x)$ at the point $\mathbb{E}(X)$. Since g is convex, it lies above the line $L(x)$. So,

$$\mathbb{E}g(X) \geq \mathbb{E}L(X) = \mathbb{E}(a+bX) = a + b\mathbb{E}(X) = L(\mathbb{E}(X)) = g(\mathbb{E}X). \quad \blacksquare$$

From Jensen's inequality we see that $\mathbb{E}(X^2) \geq (\mathbb{E}X)^2$ and if X is positive, then $\mathbb{E}(1/X) \geq 1/\mathbb{E}(X)$. Since log is concave, $\mathbb{E}(\log X) \leq \log \mathbb{E}(X)$.

4.3 Bibliographic Remarks

Devroye et al. (1996) is a good reference on probability inequalities and their use in statistics and pattern recognition. The following proof of Hoeffding's inequality is from that text.

4.4 Appendix

PROOF OF HOEFFDING'S INEQUALITY. We will make use of the exact form of Taylor's theorem: if g is a smooth function, then there is a number $\xi \in (0,u)$ such that $g(u) = g(0) + ug'(0) + \frac{u^2}{2}g''(\xi)$.

PROOF of Theorem 4.4. For any $t > 0$, we have, from Markov's inequality, that

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n Y_i \geq \epsilon\right) &= \mathbb{P}\left(t\sum_{i=1}^n Y_i \geq t\epsilon\right) = \mathbb{P}\left(e^{t\sum_{i=1}^n Y_i} \geq e^{t\epsilon}\right) \\ &\leq e^{-t\epsilon}\mathbb{E}\left(e^{t\sum_{i=1}^n Y_i}\right) = e^{-t\epsilon}\prod_i \mathbb{E}(e^{tY_i}). \qquad (4.8)\end{aligned}$$

Since $a_i \leq Y_i \leq b_i$, we can write Y_i as a convex combination of a_i and b_i, namely, $Y_i = \alpha b_i + (1-\alpha)a_i$ where $\alpha = (Y_i - a_i)/(b_i - a_i)$. So, by the convexity of e^{ty} we have

$$e^{tY_i} \leq \frac{Y_i - a_i}{b_i - a_i}e^{tb_i} + \frac{b_i - Y_i}{b_i - a_i}e^{ta_i}.$$

Take expectations of both sides and use the fact that $\mathbb{E}(Y_i) = 0$ to get

$$\mathbb{E}e^{tY_i} \leq -\frac{a_i}{b_i - a_i}e^{tb_i} + \frac{b_i}{b_i - a_i}e^{ta_i} = e^{g(u)} \qquad (4.9)$$

where $u = t(b_i - a_i)$, $g(u) = -\gamma u + \log(1 - \gamma + \gamma e^u)$ and $\gamma = -a_i/(b_i - a_i)$.

Note that $g(0) = g'(0) = 0$. Also, $g''(u) \leq 1/4$ for all $u > 0$. By Taylor's theorem, there is a $\xi \in (0,u)$ such that

$$\begin{aligned}g(u) &= g(0) + ug'(0) + \frac{u^2}{2}g''(\xi) \\ &= \frac{u^2}{2}g''(\xi) \leq \frac{u^2}{8} = \frac{t^2(b_i - a_i)^2}{8}.\end{aligned}$$

Hence,

$$\mathbb{E}e^{tY_i} \leq e^{g(u)} \leq e^{t^2(b_i - a_i)^2/8}.$$

The result follows from (4.8). ∎

PROOF of Theorem 4.5. Let $Y_i = (1/n)(X_i - p)$. Then $\mathbb{E}(Y_i) = 0$ and $a \leq Y_i \leq b$ where $a = -p/n$ and $b = (1-p)/n$. Also, $(b-a)^2 = 1/n^2$. Applying Theorem 4.4 we get

$$\mathbb{P}(\overline{X}_n - p > \epsilon) = \mathbb{P}\left(\sum_i Y_i > \epsilon\right) \leq e^{-t\epsilon}e^{t^2/(8n)}.$$

The above holds for any $t > 0$. In particular, take $t = 4n\epsilon$ and we get $\mathbb{P}(\overline{X}_n - p > \epsilon) \leq e^{-2n\epsilon^2}$. By a similar argument we can show that $\mathbb{P}(\overline{X}_n - p < -\epsilon) \leq e^{-2n\epsilon^2}$. Putting these together we get $\mathbb{P}\left(|\overline{X}_n - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$. ∎

4.5 Exercises

1. Let $X \sim \text{Exponential}(\beta)$. Find $\mathbb{P}(|X - \mu_X| \geq k\sigma_X)$ for $k > 1$. Compare this to the bound you get from Chebyshev's inequality.

2. Let $X \sim \text{Poisson}(\lambda)$. Use Chebyshev's inequality to show that $\mathbb{P}(X \geq 2\lambda) \leq 1/\lambda$.

3. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Bound $\mathbb{P}(|\overline{X}_n - p| > \epsilon)$ using Chebyshev's inequality and using Hoeffding's inequality. Show that, when n is large, the bound from Hoeffding's inequality is smaller than the bound from Chebyshev's inequality.

4. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$.

   (a) Let $\alpha > 0$ be fixed and define

   $$\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}.$$

   Let $\widehat{p}_n = n^{-1}\sum_{i=1}^n X_i$. Define $C_n = (\widehat{p}_n - \epsilon_n,\ \widehat{p}_n + \epsilon_n)$. Use Hoeffding's inequality to show that

   $$\mathbb{P}(C_n \text{ contains } p) \geq 1 - \alpha.$$

   In practice, we truncate the interval so it does not go below 0 or above 1.

   (b) (Computer Experiment.) Let's examine the properties of this confidence interval. Let $\alpha = 0.05$ and $p = 0.4$. Conduct a simulation study to see how often the interval contains p (called the coverage). Do this for various values of n between 1 and 10000. Plot the coverage versus n.

   (c) Plot the length of the interval versus n. Suppose we want the length of the interval to be no more than .05. How large should n be?

5. Prove Mill's inequality, Theorem 4.7. Hint. Note that $\mathbb{P}(|Z| > t) = 2\mathbb{P}(Z > t)$. Now write out what $\mathbb{P}(Z > t)$ means and note that $x/t > 1$ whenever $x > t$.

6. Let $Z \sim N(0,1)$. Find $\mathbb{P}(|Z| > t)$ and plot this as a function of t. From Markov's inequality, we have the bound $\mathbb{P}(|Z| > t) \leq \frac{\mathbb{E}|Z|^k}{t^k}$ for any $k > 0$. Plot these bounds for $k = 1, 2, 3, 4, 5$ and compare them to the true value of $\mathbb{P}(|Z| > t)$. Also, plot the bound from Mill's inequality.

7. Let $X_1, \ldots, X_n \sim N(0,1)$. Bound $\mathbb{P}(|\overline{X}_n| > t)$ using Mill's inequality, where $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Compare to the Chebyshev bound.

5
Convergence of Random Variables

5.1 Introduction

The most important aspect of probability theory concerns the behavior of sequences of random variables. This part of probability is called **large sample theory**, or **limit theory**, or **asymptotic theory**. The basic question is this: what can we say about the limiting behavior of a sequence of random variables $X_1, X_2, X_3, \ldots$? Since statistics and data mining are all about gathering data, we will naturally be interested in what happens as we gather more and more data.

In calculus we say that a sequence of real numbers x_n converges to a limit x if, for every $\epsilon > 0$, $|x_n - x| < \epsilon$ for all large n. In probability, convergence is more subtle. Going back to calculus for a moment, suppose that $x_n = x$ for all n. Then, trivially, $\lim_{n\to\infty} x_n = x$. Consider a probabilistic version of this example. Suppose that $X_1, X_2, \ldots$ is a sequence of random variables which are independent and suppose each has a $N(0,1)$ distribution. Since these all have the same distribution, we are tempted to say that X_n "converges" to $X \sim N(0,1)$. But this can't quite be right since $\mathbb{P}(X_n = X) = 0$ for all n. (Two continuous random variables are equal with probability zero.)

Here is another example. Consider $X_1, X_2, \ldots$ where $X_i \sim N(0, 1/n)$. Intuitively, X_n is very concentrated around 0 for large n so we would like to say that X_n converges to 0. But $\mathbb{P}(X_n = 0) = 0$ for all n. Clearly, we need to

develop some tools for discussing convergence in a rigorous way. This chapter develops the appropriate methods.

There are two main ideas in this chapter which we state informally here:

1. The **law of large numbers** says that the sample average $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ **converges in probability** to the expectation $\mu = \mathbb{E}(X_i)$. This means that $\overline{X}_n$ is close to μ with high probability.

2. The **central limit theorem** says that $\sqrt{n}(\overline{X}_n - \mu)$ **converges in distribution** to a Normal distribution. This means that the sample average has approximately a Normal distribution for large n.

5.2 Types of Convergence

The two main types of convergence are defined as follows.

5.1 Definition. *Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable. Let F_n denote the* CDF *of X_n and let F denote the* CDF *of X.*

1. *X_n* **converges to X in probability**, *written $X_n \xrightarrow{\text{P}} X$, if, for every $\epsilon > 0$,*
$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \tag{5.1}$$
as $n \to \infty$.

2. *X_n* **converges to X in distribution**, *written $X_n \rightsquigarrow X$, if*
$$\lim_{n\to\infty} F_n(t) = F(t) \tag{5.2}$$
at all t for which F is continuous.

When the limiting random variable is a point mass, we change the notation slightly. If $\mathbb{P}(X = c) = 1$ and $X_n \xrightarrow{\text{P}} X$ then we write $X_n \xrightarrow{\text{P}} c$. Similarly, if $X_n \rightsquigarrow X$ we write $X_n \rightsquigarrow c$.

There is another type of convergence which we introduce mainly because it is useful for proving convergence in probability.

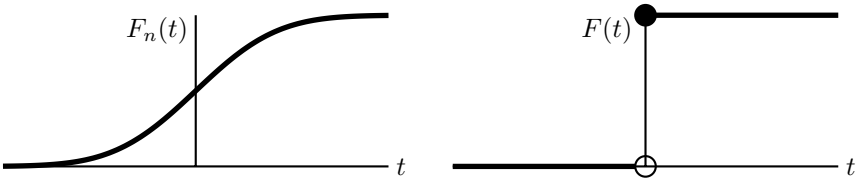


FIGURE 5.1. Example 5.3. X_n converges in distribution to X because $F_n(t)$ converges to $F(t)$ at all points except $t = 0$. Convergence is not required at $t = 0$ because $t = 0$ is not a point of continuity for F.

5.2 Definition. X_n **converges to** X **in quadratic mean** *(also called convergence in* L_2*), written* $X_n \xrightarrow{\text{qm}} X$*, if*

$$\mathbb{E}(X_n - X)^2 \to 0 \tag{5.3}$$

as $n \to \infty$.

Again, if X is a point mass at c we write $X_n \xrightarrow{\text{qm}} c$ instead of $X_n \xrightarrow{\text{qm}} X$.

5.3 Example. Let $X_n \sim N(0, 1/n)$. Intuitively, X_n is concentrating at 0 so we would like to say that X_n converges to 0. Let's see if this is true. Let F be the distribution function for a point mass at 0. Note that $\sqrt{n}X_n \sim N(0, 1)$. Let Z denote a standard normal random variable. For $t < 0$, $F_n(t) = \mathbb{P}(X_n < t) = \mathbb{P}(\sqrt{n}X_n < \sqrt{n}t) = \mathbb{P}(Z < \sqrt{n}t) \to 0$ since $\sqrt{n}t \to -\infty$. For $t > 0$, $F_n(t) = \mathbb{P}(X_n < t) = \mathbb{P}(\sqrt{n}X_n < \sqrt{n}t) = \mathbb{P}(Z < \sqrt{n}t) \to 1$ since $\sqrt{n}t \to \infty$. Hence, $F_n(t) \to F(t)$ for all $t \neq 0$ and so $X_n \rightsquigarrow 0$. Notice that $F_n(0) = 1/2 \neq F(1/2) = 1$ so convergence fails at $t = 0$. That doesn't matter because $t = 0$ is not a continuity point of F and the definition of convergence in distribution only requires convergence at continuity points. See Figure 5.1. Now consider convergence in probability. For any $\epsilon > 0$, using Markov's inequality,

$$\begin{aligned} \mathbb{P}(|X_n| > \epsilon) &= \mathbb{P}(|X_n|^2 > \epsilon^2) \\ &\leq \frac{\mathbb{E}(X_n^2)}{\epsilon^2} = \frac{\frac{1}{n}}{\epsilon^2} \to 0 \end{aligned}$$

as $n \to \infty$. Hence, $X_n \xrightarrow{\text{P}} 0$. ■

The next theorem gives the relationship between the types of convergence. The results are summarized in Figure 5.2.

5.4 Theorem. *The following relationships hold:*

(a) $X_n \xrightarrow{\text{qm}} X$ *implies that* $X_n \xrightarrow{\text{P}} X$.

(b) $X_n \xrightarrow{\text{P}} X$ *implies that* $X_n \rightsquigarrow X$.

(c) If $X_n \rightsquigarrow X$ *and if* $\mathbb{P}(X = c) = 1$ *for some real number* c*, then* $X_n \xrightarrow{\text{P}} X$.

In general, none of the reverse implications hold except the special case in (c).

PROOF. We start by proving (a). Suppose that $X_n \xrightarrow{\text{qm}} X$. Fix $\epsilon > 0$. Then, using Markov's inequality,

$$\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|X_n - X|^2 > \epsilon^2) \leq \frac{\mathbb{E}|X_n - X|^2}{\epsilon^2} \to 0.$$

Proof of (b). This proof is a little more complicated. You may skip it if you wish. Fix $\epsilon > 0$ and let x be a continuity point of F. Then

$$\begin{aligned} F_n(x) &= \mathbb{P}(X_n \leq x) = \mathbb{P}(X_n \leq x, X \leq x + \epsilon) + \mathbb{P}(X_n \leq x, X > x + \epsilon) \\ &\leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \\ &= F(x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon). \end{aligned}$$

Also,

$$\begin{aligned} F(x - \epsilon) &= \mathbb{P}(X \leq x - \epsilon) = \mathbb{P}(X \leq x - \epsilon, X_n \leq x) + \mathbb{P}(X \leq x - \epsilon, X_n > x) \\ &\leq F_n(x) + \mathbb{P}(|X_n - X| > \epsilon). \end{aligned}$$

Hence,

$$F(x - \epsilon) - \mathbb{P}(|X_n - X| > \epsilon) \leq F_n(x) \leq F(x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon).$$

Take the limit as $n \to \infty$ to conclude that

$$F(x - \epsilon) \leq \liminf_{n \to \infty} F_n(x) \leq \limsup_{n \to \infty} F_n(x) \leq F(x + \epsilon).$$

This holds for all $\epsilon > 0$. Take the limit as $\epsilon \to 0$ and use the fact that F is continuous at x and conclude that $\lim_n F_n(x) = F(x)$.

Proof of (c). Fix $\epsilon > 0$. Then,

$$\begin{aligned} \mathbb{P}(|X_n - c| > \epsilon) &= \mathbb{P}(X_n < c - \epsilon) + \mathbb{P}(X_n > c + \epsilon) \\ &\leq \mathbb{P}(X_n \leq c - \epsilon) + \mathbb{P}(X_n > c + \epsilon) \\ &= F_n(c - \epsilon) + 1 - F_n(c + \epsilon) \\ &\to F(c - \epsilon) + 1 - F(c + \epsilon) \\ &= 0 + 1 - 1 = 0. \end{aligned}$$

Let us now show that the reverse implications do not hold.

CONVERGENCE IN PROBABILITY DOES NOT IMPLY CONVERGENCE IN QUADRATIC MEAN. Let $U \sim \text{Unif}(0, 1)$ and let $X_n = \sqrt{n} I_{(0,1/n)}(U)$. Then $\mathbb{P}(|X_n| > \epsilon) =$

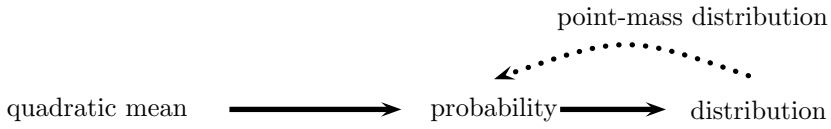


FIGURE 5.2. Relationship between types of convergence.

$\mathbb{P}(\sqrt{n}I_{(0,1/n)}(U) > \epsilon) = \mathbb{P}(0 \leq U < 1/n) = 1/n \to 0$. Hence, $X_n \xrightarrow{\text{P}} 0$. But $\mathbb{E}(X_n^2) = n\int_0^{1/n} du = 1$ for all n so X_n does not converge in quadratic mean.

CONVERGENCE IN DISTRIBUTION DOES NOT IMPLY CONVERGENCE IN PROBABILITY. Let $X \sim N(0,1)$. Let $X_n = -X$ for $n = 1, 2, 3, \ldots$; hence $X_n \sim N(0,1)$. X_n has the same distribution function as X for all n so, trivially, $\lim_n F_n(x) = F(x)$ for all x. Therefore, $X_n \rightsquigarrow X$. But $\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|2X| > \epsilon) = \mathbb{P}(|X| > \epsilon/2) \neq 0$. So X_n does not converge to X in probability. ■

Warning! One might conjecture that if $X_n \xrightarrow{\text{P}} b$, then $\mathbb{E}(X_n) \to b$. This is not[1] true. Let X_n be a random variable defined by $\mathbb{P}(X_n = n^2) = 1/n$ and $\mathbb{P}(X_n = 0) = 1 - (1/n)$. Now, $\mathbb{P}(|X_n| < \epsilon) = \mathbb{P}(X_n = 0) = 1 - (1/n) \to 1$. Hence, $X_n \xrightarrow{\text{P}} 0$. However, $\mathbb{E}(X_n) = [n^2 \times (1/n)] + [0 \times (1 - (1/n))] = n$. Thus, $\mathbb{E}(X_n) \to \infty$.

Summary. Stare at Figure 5.2.

Some convergence properties are preserved under transformations.

5.5 Theorem. *Let X_n, X, Y_n, Y be random variables. Let g be a continuous function.*

(a) If $X_n \xrightarrow{\text{P}} X$ and $Y_n \xrightarrow{\text{P}} Y$, then $X_n + Y_n \xrightarrow{\text{P}} X + Y$.
(b) If $X_n \xrightarrow{\text{qm}} X$ and $Y_n \xrightarrow{\text{qm}} Y$, then $X_n + Y_n \xrightarrow{\text{qm}} X + Y$.
(c) If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $X_n + Y_n \rightsquigarrow X + c$.
(d) If $X_n \xrightarrow{\text{P}} X$ and $Y_n \xrightarrow{\text{P}} Y$, then $X_n Y_n \xrightarrow{\text{P}} XY$.
(e) If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $X_n Y_n \rightsquigarrow cX$.
(f) If $X_n \xrightarrow{\text{P}} X$, then $g(X_n) \xrightarrow{\text{P}} g(X)$.
(g) If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$.

Parts (c) and (e) are know as **Slutzky's theorem**. It is worth noting that $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ does not in general imply that $X_n + Y_n \rightsquigarrow X + Y$.

[1]We can conclude that $\mathbb{E}(X_n) \to b$ if X_n is uniformly integrable. See the appendix.

5.3 The Law of Large Numbers

Now we come to a crowning achievement in probability, the law of large numbers. This theorem says that the mean of a large sample is close to the mean of the distribution. For example, the proportion of heads of a large number of tosses is expected to be close to 1/2. We now make this more precise.

Let $X_1, X_2, \ldots$ be an IID sample, let $\mu = \mathbb{E}(X_1)$ and [2] $\sigma^2 = \mathbb{V}(X_1)$. Recall that the sample mean is defined as $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ and that $\mathbb{E}(\overline{X}_n) = \mu$ and $\mathbb{V}(\overline{X}_n) = \sigma^2/n$.

5.6 Theorem (The Weak Law of Large Numbers (WLLN)). [3]
If $X_1, \ldots, X_n$ are IID*, then $\overline{X}_n \xrightarrow{\text{P}} \mu$.*

Interpretation of the WLLN: The distribution of $\overline{X}_n$ becomes more concentrated around μ as n gets large.

PROOF. Assume that $\sigma < \infty$. This is not necessary but it simplifies the proof. Using Chebyshev's inequality,

$$\mathbb{P}\left(|\overline{X}_n - \mu| > \epsilon\right) \leq \frac{\mathbb{V}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

which tends to 0 as $n \to \infty$. ■

5.7 Example. Consider flipping a coin for which the probability of heads is p. Let X_i denote the outcome of a single toss (0 or 1). Hence, $p = P(X_i = 1) = E(X_i)$. The fraction of heads after n tosses is $\overline{X}_n$. According to the law of large numbers, $\overline{X}_n$ converges to p in probability. This does not mean that $\overline{X}_n$ will numerically equal p. It means that, when n is large, the distribution of $\overline{X}_n$ is tightly concentrated around p. Suppose that $p = 1/2$. How large should n be so that $P(.4 \leq \overline{X}_n \leq .6) \geq .7$? First, $\mathbb{E}(\overline{X}_n) = p = 1/2$ and $\mathbb{V}(\overline{X}_n) = \sigma^2/n = p(1-p)/n = 1/(4n)$. From Chebyshev's inequality,

$$\begin{aligned}
\mathbb{P}(.4 \leq \overline{X}_n \leq .6) &= \mathbb{P}(|\overline{X}_n - \mu| \leq .1) \\
&= 1 - \mathbb{P}(|\overline{X}_n - \mu| > .1) \\
&\geq 1 - \frac{1}{4n(.1)^2} = 1 - \frac{25}{n}.
\end{aligned}$$

The last expression will be larger than .7 if $n = 84$. ■

---

[2] Note that $\mu = \mathbb{E}(X_i)$ is the same for all i so we can define $\mu = \mathbb{E}(X_i)$ for any i. By convention, we often write $\mu = \mathbb{E}(X_1)$.

[3] There is a stronger theorem in the appendix called the strong law of large numbers.

5.4 The Central Limit Theorem

The law of large numbers says that the distribution of $\overline{X}_n$ piles up near μ. This isn't enough to help us approximate probability statements about $\overline{X}_n$. For this we need the central limit theorem.

Suppose that $X_1, \ldots, X_n$ are IID with mean μ and variance σ^2. The central limit theorem (CLT) says that $\overline{X}_n = n^{-1} \sum_i X_i$ has a distribution which is approximately Normal with mean μ and variance σ^2/n. This is remarkable since nothing is assumed about the distribution of X_i, except the existence of the mean and variance.

5.8 Theorem (The Central Limit Theorem (CLT)). *Let $X_1, \ldots, X_n$ be* IID *with mean μ and variance σ^2. Let $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then*

$$Z_n \equiv \frac{\overline{X}_n - \mu}{\sqrt{\mathbb{V}(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightsquigarrow Z$$

where $Z \sim N(0,1)$. In other words,

$$\lim_{n \to \infty} \mathbb{P}(Z_n \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Interpretation: Probability statements about $\overline{X}_n$ can be approximated using a Normal distribution. It's the probability statements that we are approximating, not the random variable itself.

In addition to $Z_n \rightsquigarrow N(0,1)$, there are several forms of notation to denote the fact that the distribution of Z_n is converging to a Normal. They all mean the same thing. Here they are:

$$\begin{aligned}
Z_n &\approx N(0,1) \\
\overline{X}_n &\approx N\left(\mu, \frac{\sigma^2}{n}\right) \\
\overline{X}_n - \mu &\approx N\left(0, \frac{\sigma^2}{n}\right) \\
\sqrt{n}(\overline{X}_n - \mu) &\approx N\left(0, \sigma^2\right) \\
\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} &\approx N(0,1).
\end{aligned}$$

5.9 Example. Suppose that the number of errors per computer program has a Poisson distribution with mean 5. We get 125 programs. Let $X_1, \ldots, X_{125}$ be

the number of errors in the programs. We want to approximate $\mathbb{P}(\overline{X}_n < 5.5)$. Let $\mu = \mathbb{E}(X_1) = \lambda = 5$ and $\sigma^2 = \mathbb{V}(X_1) = \lambda = 5$. Then,

$$\begin{aligned}\mathbb{P}(\overline{X}_n < 5.5) &= \mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} < \frac{\sqrt{n}(5.5-\mu)}{\sigma}\right) \\ &\approx \mathbb{P}(Z < 2.5) = .9938. \quad \blacksquare\end{aligned}$$

The central limit theorem tells us that $Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ is approximately N(0,1). However, we rarely know σ. Later, we will see that we can estimate σ^2 from $X_1, \ldots, X_n$ by

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

This raises the following question: if we replace σ with S_n, is the central limit theorem still true? The answer is yes.

5.10 Theorem. *Assume the same conditions as the CLT. Then,*

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \rightsquigarrow N(0,1).$$

You might wonder, how accurate the normal approximation is. The answer is given in the Berry-Essèen theorem.

5.11 Theorem (The Berry-Essèen Inequality)**.** *Suppose that* $\mathbb{E}|X_1|^3 < \infty$. *Then*

$$\sup_z |\mathbb{P}(Z_n \leq z) - \Phi(z)| \leq \frac{33}{4}\frac{\mathbb{E}|X_1 - \mu|^3}{\sqrt{n}\sigma^3}. \tag{5.4}$$

There is also a multivariate version of the central limit theorem.

5.12 Theorem (Multivariate central limit theorem)**.** *Let* $X_1, \ldots, X_n$ *be* IID *random vectors where*

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ki} \end{pmatrix}$$

with mean

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_{1i}) \\ \mathbb{E}(X_{2i}) \\ \vdots \\ \mathbb{E}(X_{ki}) \end{pmatrix}$$

and variance matrix Σ. *Let*

$$\overline{X} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_k \end{pmatrix}.$$

where $\overline{X}_j = n^{-1}\sum_{i=1}^n X_{ji}$. *Then,*

$$\sqrt{n}(\overline{X} - \mu) \rightsquigarrow N(0, \Sigma).$$

5.5 The Delta Method

If Y_n has a limiting Normal distribution then the delta method allows us to find the limiting distribution of $g(Y_n)$ where g is any smooth function.

5.13 Theorem (The Delta Method). *Suppose that*

$$\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \rightsquigarrow N(0,1)$$

and that g *is a differentiable function such that* $g'(\mu) \neq 0$. *Then*

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \rightsquigarrow N(0,1).$$

In other words,

$$Y_n \approx N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{implies that} \quad g(Y_n) \approx N\left(g(\mu), (g'(\mu))^2 \frac{\sigma^2}{n}\right).$$

5.14 Example. Let $X_1, \ldots, X_n$ be IID with finite mean μ and finite variance σ^2. By the central limit theorem, $\sqrt{n}(\overline{X}_n - \mu)/\sigma \rightsquigarrow N(0,1)$. Let $W_n = e^{\overline{X}_n}$. Thus, $W_n = g(\overline{X}_n)$ where $g(s) = e^s$. Since $g'(s) = e^s$, the delta method implies that $W_n \approx N(e^\mu, e^{2\mu}\sigma^2/n)$. ∎

There is also a multivariate version of the delta method.

5.15 Theorem (The Multivariate Delta Method). *Suppose that* $Y_n = (Y_{n1}, \ldots, Y_{nk})$ *is a sequence of random vectors such that*

$$\sqrt{n}(Y_n - \mu) \rightsquigarrow N(0, \Sigma).$$

Let $g : \mathbb{R}^k \to \mathbb{R}$ and let

$$\nabla g(y) = \begin{pmatrix} \frac{\partial g}{\partial y_1} \\ \vdots \\ \frac{\partial g}{\partial y_k} \end{pmatrix}.$$

Let ∇_μ denote $\nabla g(y)$ evaluated at $y = \mu$ and assume that the elements of ∇_μ are nonzero. Then

$$\sqrt{n}(g(Y_n) - g(\mu)) \rightsquigarrow N\left(0, \nabla_\mu^T \Sigma \nabla_\mu\right).$$

5.16 Example. Let

$$\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix}, \ \cdots, \ \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}$$

be IID random vectors with mean $\mu = (\mu_1, \mu_2)^T$ and variance Σ. Let

$$\overline{X}_1 = \frac{1}{n}\sum_{i=1}^n X_{1i}, \quad \overline{X}_2 = \frac{1}{n}\sum_{i=1}^n X_{2i}$$

and define $Y_n = \overline{X}_1 \overline{X}_2$. Thus, $Y_n = g(\overline{X}_1, \overline{X}_2)$ where $g(s_1, s_2) = s_1 s_2$. By the central limit theorem,

$$\sqrt{n}\begin{pmatrix} \overline{X}_1 - \mu_1 \\ \overline{X}_2 - \mu_2 \end{pmatrix} \rightsquigarrow N(0, \Sigma).$$

Now

$$\nabla g(s) = \begin{pmatrix} \frac{\partial g}{\partial s_1} \\ \frac{\partial g}{\partial s_2} \end{pmatrix} = \begin{pmatrix} s_2 \\ s_1 \end{pmatrix}$$

and so

$$\nabla_\mu^T \Sigma \nabla_\mu = (\mu_2 \ \ \mu_1) \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \begin{pmatrix} \mu_2 \\ \mu_1 \end{pmatrix} = \mu_2^2\sigma_{11} + 2\mu_1\mu_2\sigma_{12} + \mu_1^2\sigma_{22}.$$

Therefore,

$$\sqrt{n}(\overline{X}_1\overline{X}_2 - \mu_1\mu_2) \rightsquigarrow N\Big(0, \mu_2^2\sigma_{11} + 2\mu_1\mu_2\sigma_{12} + \mu_1^2\sigma_{22}\Big). \quad \blacksquare$$

5.6 Bibliographic Remarks

Convergence plays a central role in modern probability theory. For more details, see Grimmett and Stirzaker (1982), Karr (1993), and Billingsley (1979). Advanced convergence theory is explained in great detail in van der Vaart and Wellner (1996) and and van der Vaart (1998).

5.7 Appendix

5.7.1 Almost Sure and L_1 Convergence

We say that X_n **converges almost surely to** X, written $X_n \xrightarrow{\text{as}} X$, if

$$\mathbb{P}(\{s:\ X_n(s) \to X(s)\}) = 1.$$

We say that X_n **converges in** L_1 **to** X, written $X_n \xrightarrow{L_1} X$, if

$$\mathbb{E}|X_n - X| \to 0$$

as $n \to \infty$.

5.17 Theorem. *Let X_n and X be random variables. Then:*
(a) $X_n \xrightarrow{\text{as}} X$ implies that $X_n \xrightarrow{\text{P}} X$.
(b) $X_n \xrightarrow{\text{qm}} X$ implies that $X_n \xrightarrow{L_1} X$.
(c) $X_n \xrightarrow{L_1} X$ implies that $X_n \xrightarrow{\text{P}} X$.

The weak law of large numbers says that $\overline{X}_n$ converges to $\mathbb{E}(X_1)$ in probability. The strong law asserts that this is also true almost surely.

5.18 Theorem (The Strong Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be* IID*. If $\mu = \mathbb{E}|X_1| < \infty$ then $\overline{X}_n \xrightarrow{\text{as}} \mu$.*

A sequence X_n is **asymptotically uniformly integrable** if

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}\left(|X_n| I(|X_n| > M)\right) = 0.$$

5.19 Theorem. *If $X_n \xrightarrow{\text{P}} b$ and X_n is asymptotically uniformly integrable, then $\mathbb{E}(X_n) \to b$.*

5.7.2 Proof of the Central Limit Theorem

Recall that if X is a random variable, its moment generating function (MGF) is $\psi_X(t) = \mathbb{E}e^{tX}$. Assume in what follows that the MGF is finite in a neighborhood around $t = 0$.

5.20 Lemma. *Let $Z_1, Z_2, \ldots$ be a sequence of random variables. Let ψ_n be the* MGF *of Z_n. Let Z be another random variable and denote its* MGF *by ψ. If $\psi_n(t) \to \psi(t)$ for all t in some open interval around 0, then $Z_n \rightsquigarrow Z$.*

PROOF OF THE CENTRAL LIMIT THEOREM. Let $Y_i = (X_i - \mu)/\sigma$. Then, $Z_n = n^{-1/2}\sum_i Y_i$. Let $\psi(t)$ be the MGF of Y_i. The MGF of $\sum_i Y_i$ is $(\psi(t))^n$ and MGF of Z_n is $[\psi(t/\sqrt{n})]^n \equiv \xi_n(t)$. Now $\psi'(0) = \mathbb{E}(Y_1) = 0$, $\psi''(0) = \mathbb{E}(Y_1^2) = \mathbb{V}(Y_1) = 1$. So,

$$\begin{aligned}\psi(t) &= \psi(0) + t\psi'(0) + \frac{t^2}{2!}\psi''(0) + \frac{t^3}{3!}\psi'''(0) + \cdots \\ &= 1 + 0 + \frac{t^2}{2} + \frac{t^3}{3!}\psi'''(0) + \cdots \\ &= 1 + \frac{t^2}{2} + \frac{t^3}{3!}\psi'''(0) + \cdots\end{aligned}$$

Now,

$$\begin{aligned}\xi_n(t) &= \left[\psi\left(\frac{t}{\sqrt{n}}\right)\right]^n \\ &= \left[1 + \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}}\psi'''(0) + \cdots\right]^n \\ &= \left[1 + \frac{\frac{t^2}{2} + \frac{t^3}{3!n^{1/2}}\psi'''(0) + \cdots}{n}\right]^n \\ &\to e^{t^2/2}\end{aligned}$$

which is the MGF of a N(0,1). The result follows from the previous Theorem. In the last step we used the fact that if $a_n \to a$ then

$$\left(1 + \frac{a_n}{n}\right)^n \to e^a. \quad \blacksquare$$

5.8 Exercises

1. Let $X_1, \ldots, X_n$ be IID with finite mean $\mu = \mathbb{E}(X_1)$ and finite variance $\sigma^2 = \mathbb{V}(X_1)$. Let $\overline{X}_n$ be the sample mean and let S_n^2 be the sample variance.

   (a) Show that $\mathbb{E}(S_n^2) = \sigma^2$.

   (b) Show that $S_n^2 \xrightarrow{\text{P}} \sigma^2$. Hint: Show that $S_n^2 = c_n n^{-1}\sum_{i=1}^n X_i^2 - d_n\overline{X}_n^2$ where $c_n \to 1$ and $d_n \to 1$. Apply the law of large numbers to $n^{-1}\sum_{i=1}^n X_i^2$ and to $\overline{X}_n$. Then use part (e) of Theorem 5.5.

2. Let $X_1, X_2, \ldots$ be a sequence of random variables. Show that $X_n \xrightarrow{\text{qm}} b$ if and only if

$$\lim_{n\to\infty}\mathbb{E}(X_n) = b \quad \text{and} \quad \lim_{n\to\infty}\mathbb{V}(X_n) = 0.$$

3. Let $X_1, \ldots, X_n$ be IID and let $\mu = \mathbb{E}(X_1)$. Suppose that the variance is finite. Show that $\overline{X}_n \xrightarrow{\text{qm}} \mu$.

4. Let $X_1, X_2, \ldots$ be a sequence of random variables such that

$$\mathbb{P}\left(X_n = \frac{1}{n}\right) = 1 - \frac{1}{n^2} \quad \text{and} \quad \mathbb{P}(X_n = n) = \frac{1}{n^2}.$$

   Does X_n converge in probability? Does X_n converge in quadratic mean?

5. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Prove that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{\text{P}} p \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{\text{qm}} p.$$

6. Suppose that the height of men has mean 68 inches and standard deviation 2.6 inches. We draw 100 men at random. Find (approximately) the probability that the average height of men in our sample will be at least 68 inches.

7. Let $\lambda_n = 1/n$ for $n = 1, 2, \ldots$. Let $X_n \sim \text{Poisson}(\lambda_n)$.

   (a) Show that $X_n \xrightarrow{\text{P}} 0$.

   (b) Let $Y_n = nX_n$. Show that $Y_n \xrightarrow{\text{P}} 0$.

8. Suppose we have a computer program consisting of $n = 100$ pages of code. Let X_i be the number of errors on the i^{th} page of code. Suppose that the $X_i's$ are Poisson with mean 1 and that they are independent. Let $Y = \sum_{i=1}^{n} X_i$ be the total number of errors. Use the central limit theorem to approximate $\mathbb{P}(Y < 90)$.

9. Suppose that $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. Define

$$X_n = \begin{cases} X & \text{with probability } 1 - \frac{1}{n} \\ e^n & \text{with probability } \frac{1}{n}. \end{cases}$$

   Does X_n converge to X in probability? Does X_n converge to X in distribution? Does $\mathbb{E}(X - X_n)^2$ converge to 0?

10. Let $Z \sim N(0, 1)$. Let $t > 0$. Show that, for any $k > 0$,

$$\mathbb{P}(|Z| > t) \leq \frac{\mathbb{E}|Z|^k}{t^k}.$$

    Compare this to Mill's inequality in Chapter 4.

11. Suppose that $X_n \sim N(0, 1/n)$ and let X be a random variable with distribution $F(x) = 0$ if $x < 0$ and $F(x) = 1$ if $x \geq 0$. Does X_n converge to X in probability? (Prove or disprove). Does X_n converge to X in distribution? (Prove or disprove).

12. Let $X, X_1, X_2, X_3, \ldots$ be random variables that are positive and integer valued. Show that $X_n \rightsquigarrow X$ if and only if

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$$

for every integer k.

13. Let $Z_1, Z_2, \ldots$ be IID random variables with density f. Suppose that $\mathbb{P}(Z_i > 0) = 1$ and that $\lambda = \lim_{x\downarrow 0} f(x) > 0$. Let

$$X_n = n \min\{Z_1, \ldots, Z_n\}.$$

Show that $X_n \rightsquigarrow Z$ where Z has an exponential distribution with mean $1/\lambda$.

14. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$. Let $Y_n = \overline{X}_n^2$. Find the limiting distribution of Y_n.

15. Let

$$\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix}, \ \ldots, \ \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}$$

be IID random vectors with mean $\mu = (\mu_1, \mu_2)$ and variance Σ. Let

$$\overline{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i}, \quad \overline{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i}$$

and define $Y_n = \overline{X}_1/\overline{X}_2$. Find the limiting distribution of Y_n.

16. Construct an example where $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ but $X_n + Y_n$ does not converge in distribution to $X + Y$.

Part II

Statistical Inference

6

Models, Statistical Inference and Learning

6.1 Introduction

Statistical inference, or "learning" as it is called in computer science, is the process of using data to infer the distribution that generated the data. A typical statistical inference question is:

Given a sample $X_1, \ldots, X_n \sim F$, how do we infer F?

In some cases, we may want to infer only some feature of F such as its mean.

6.2 Parametric and Nonparametric Models

A **statistical model** $\mathfrak{F}$ is a set of distributions (or densities or regression functions). A **parametric model** is a set $\mathfrak{F}$ that can be parameterized by a finite number of parameters. For example, if we assume that the data come from a Normal distribution, then the model is

$$\mathfrak{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \quad \mu \in \mathbb{R},\ \sigma > 0 \right\}. \tag{6.1}$$

This is a two-parameter model. We have written the density as $f(x; \mu, \sigma)$ to show that x is a value of the random variable whereas μ and σ are parameters.

In general, a parametric model takes the form

$$\mathfrak{F} = \left\{ f(x; \theta) : \ \theta \in \Theta \right\} \tag{6.2}$$

where θ is an unknown parameter (or vector of parameters) that can take values in the **parameter space** Θ. If θ is a vector but we are only interested in one component of θ, we call the remaining parameters **nuisance parameters**. A **nonparametric model** is a set $\mathfrak{F}$ that cannot be parameterized by a finite number of parameters. For example, $\mathfrak{F}_{\mathsf{ALL}} = \{\text{all CDF}'s\}$ is nonparametric. [1]

6.1 Example (One-dimensional Parametric Estimation)**.** Let $X_1, \ldots, X_n$ be independent Bernoulli(p) observations. The problem is to estimate the parameter p. ■

6.2 Example (Two-dimensional Parametric Estimation)**.** Suppose that $X_1, \ldots, X_n \sim F$ and we assume that the PDF $f \in \mathfrak{F}$ where $\mathfrak{F}$ is given in (6.1). In this case there are two parameters, μ and σ. The goal is to estimate the parameters from the data. If we are only interested in estimating μ, then μ is the parameter of interest and σ is a nuisance parameter. ■

6.3 Example (Nonparametric estimation of the CDF)**.** Let $X_1, \ldots, X_n$ be independent observations from a CDF F. The problem is to estimate F assuming only that $F \in \mathfrak{F}_{\mathsf{ALL}} = \{\text{all CDF}'s\}$. ■

6.4 Example (Nonparametric density estimation)**.** Let $X_1, \ldots, X_n$ be independent observations from a CDF F and let $f = F'$ be the PDF. Suppose we want to estimate the PDF f. It is not possible to estimate f assuming only that $F \in \mathfrak{F}_{\mathsf{ALL}}$. We need to assume some smoothness on f. For example, we might assume that $f \in \mathfrak{F} = \mathfrak{F}_{\mathsf{DENS}} \bigcap \mathfrak{F}_{\mathsf{SOB}}$ where $\mathfrak{F}_{\mathsf{DENS}}$ is the set of all probability density functions and

$$\mathfrak{F}_{\mathsf{SOB}} = \left\{ f : \ \int (f''(x))^2 dx < \infty \right\}.$$

The class $\mathfrak{F}_{\mathsf{SOB}}$ is called a **Sobolev space**; it is the set of functions that are not "too wiggly." ■

6.5 Example (Nonparametric estimation of functionals)**.** Let $X_1, \ldots, X_n \sim F$. Suppose we want to estimate $\mu = \mathbb{E}(X_1) = \int x \, dF(x)$ assuming only that

[1] The distinction between parametric and nonparametric is more subtle than this but we don't need a rigorous definition for our purposes.

μ exists. The mean μ may be thought of as a function of F: we can write $\mu = T(F) = \int x\, dF(x)$. In general, any function of F is called a **statistical functional**. Other examples of functionals are the variance $T(F) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$ and the median $T(F) = F^{-1}(1/2)$. ■

6.6 Example (Regression, prediction, and classification)**.** Suppose we observe pairs of data $(X_1, Y_1), \ldots (X_n, Y_n)$. Perhaps X_i is the blood pressure of subject i and Y_i is how long they live. X is called a **predictor** or **regressor** or **feature** or **independent variable**. Y is called the **outcome** or the **response variable** or the **dependent variable**. We call $r(x) = \mathbb{E}(Y|X = x)$ the **regression function**. If we assume that $r \in \mathfrak{F}$ where $\mathfrak{F}$ is finite dimensional — the set of straight lines for example — then we have a **parametric regression model**. If we assume that $r \in \mathfrak{F}$ where $\mathfrak{F}$ is not finite dimensional then we have a **nonparametric regression model**. The goal of predicting Y for a new patient based on their X value is called **prediction**. If Y is discrete (for example, live or die) then prediction is instead called **classification.** If our goal is to estimate the function r, then we call this **regression** or **curve estimation**. Regression models are sometimes written as

$$Y = r(X) + \epsilon \tag{6.3}$$

where $\mathbb{E}(\epsilon) = 0$. We can always rewrite a regression model this way. To see this, define $\epsilon = Y - r(X)$ and hence $Y = Y + r(X) - r(X) = r(X) + \epsilon$. Moreover, $\mathbb{E}(\epsilon) = \mathbb{E}\mathbb{E}(\epsilon|X) = \mathbb{E}(\mathbb{E}(Y - r(X))|X) = \mathbb{E}(\mathbb{E}(Y|X) - r(X)) = \mathbb{E}(r(X) - r(X)) = 0$. ■

What's Next? It is traditional in most introductory courses to start with parametric inference. Instead, we will start with nonparametric inference and then we will cover parametric inference. In some respects, nonparametric inference is easier to understand and is more useful than parametric inference.

Frequentists and Bayesians. There are many approaches to statistical inference. The two dominant approaches are called **frequentist inference** and **Bayesian inference**. We'll cover both but we will start with frequentist inference. We'll postpone a discussion of the pros and cons of these two until later.

Some Notation. If $\mathfrak{F} = \{f(x;\, \theta) :\ \theta \in \Theta\}$ is a parametric model, we write $\mathbb{P}_\theta(X \in A) = \int_A f(x;\, \theta)dx$ and $\mathbb{E}_\theta(r(X)) = \int r(x) f(x;\, \theta)dx$. The subscript θ indicates that the probability or expectation is with respect to $f(x;\, \theta)$; it does not mean we are averaging over θ. Similarly, we write $\mathbb{V}_\theta$ for the variance.

6.3 Fundamental Concepts in Inference

Many inferential problems can be identified as being one of three types: estimation, confidence sets, or hypothesis testing. We will treat all of these problems in detail in the rest of the book. Here, we give a brief introduction to the ideas.

6.3.1 Point Estimation

Point estimation refers to providing a single "best guess" of some quantity of interest. The quantity of interest could be a parameter in a parametric model, a CDF F, a probability density function f, a regression function r, or a prediction for a future value Y of some random variable.

> By convention, we denote a point estimate of θ by $\widehat{\theta}$ or $\widehat{\theta}_n$. Remember that θ is a fixed, unknown quantity. The estimate $\widehat{\theta}$ depends on the data so $\widehat{\theta}$ is a random variable.

More formally, let $X_1, \ldots, X_n$ be n IID data points from some distribution F. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\widehat{\theta}_n = g(X_1, \ldots, X_n).$$

The bias of an estimator is defined by

$$\mathsf{bias}(\widehat{\theta}_n) = \mathbb{E}_\theta(\widehat{\theta}_n) - \theta. \tag{6.4}$$

We say that $\widehat{\theta}_n$ is **unbiased** if $\mathbb{E}(\widehat{\theta}_n) = \theta$. Unbiasedness used to receive much attention but these days is considered less important; many of the estimators we will use are biased. A reasonable requirement for an estimator is that it should converge to the true parameter value as we collect more and more data. This requirement is quantified by the following definition:

6.7 Definition. *A point estimator $\widehat{\theta}_n$ of a parameter θ is* **consistent** *if $\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta$.*

The distribution of $\widehat{\theta}_n$ is called the **sampling distribution**. The standard deviation of $\widehat{\theta}_n$ is called the **standard error**, denoted by se:

$$\mathsf{se} = \mathsf{se}(\widehat{\theta}_n) = \sqrt{\mathbb{V}(\widehat{\theta}_n)}. \tag{6.5}$$

Often, the standard error depends on the unknown F. In those cases, se is an unknown quantity but we usually can estimate it. The estimated standard error is denoted by $\widehat{\mathsf{se}}$.

6.8 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\widehat{p}_n = n^{-1}\sum_i X_i$. Then $\mathbb{E}(\widehat{p}_n) = n^{-1}\sum_i \mathbb{E}(X_i) = p$ so $\widehat{p}_n$ is unbiased. The standard error is $\mathsf{se} = \sqrt{\mathbb{V}(\widehat{p}_n)} = \sqrt{p(1-p)/n}$. The estimated standard error is $\widehat{\mathsf{se}} = \sqrt{\widehat{p}(1-\widehat{p})/n}$. ■

The quality of a point estimate is sometimes assessed by the **mean squared error**, or MSE defined by

$$\text{MSE} = \mathbb{E}_\theta(\widehat{\theta}_n - \theta)^2. \tag{6.6}$$

Keep in mind that $\mathbb{E}_\theta(\cdot)$ refers to expectation with respect to the distribution

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

that generated the data. It does not mean we are averaging over a distribution for θ.

6.9 Theorem. *The* MSE *can be written as*

$$\text{MSE} = \mathsf{bias}^2(\widehat{\theta}_n) + \mathbb{V}_\theta(\widehat{\theta}_n). \tag{6.7}$$

PROOF. Let $\overline{\theta}_n = E_\theta(\widehat{\theta}_n)$. Then

$$\begin{aligned}
\mathbb{E}_\theta(\widehat{\theta}_n - \theta)^2 &= \mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n + \overline{\theta}_n - \theta)^2 \\
&= \mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n)^2 + 2(\overline{\theta}_n - \theta)\mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n) + \mathbb{E}_\theta(\overline{\theta}_n - \theta)^2 \\
&= (\overline{\theta}_n - \theta)^2 + \mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n)^2 \\
&= \mathsf{bias}^2(\widehat{\theta}_n) + \mathbb{V}(\widehat{\theta}_n)
\end{aligned}$$

where we have used the fact that $\mathbb{E}_\theta(\widehat{\theta}_n - \overline{\theta}_n) = \overline{\theta}_n - \overline{\theta}_n = 0$. ■

6.10 Theorem. *If* $\mathsf{bias} \to 0$ *and* $\mathsf{se} \to 0$ *as* $n \to \infty$ *then* $\widehat{\theta}_n$ *is consistent, that is,* $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$.

PROOF. If $\mathsf{bias} \to 0$ and $\mathsf{se} \to 0$ then, by Theorem 6.9, $\mathsf{MSE} \to 0$. It follows that $\widehat{\theta}_n \xrightarrow{\text{qm}} \theta$. (Recall Definition 5.2.) The result follows from part (b) of Theorem 5.4. ■

6.11 Example. Returning to the coin flipping example, we have that $\mathbb{E}_p(\widehat{p}_n) = p$ so the bias $= p - p = 0$ and $\mathsf{se} = \sqrt{p(1-p)/n} \to 0$. Hence, $\widehat{p}_n \xrightarrow{\text{P}} p$, that is, $\widehat{p}_n$ is a consistent estimator. ■

Many of the estimators we will encounter will turn out to have, approximately, a Normal distribution.

6.12 Definition. *An estimator is* **asymptotically Normal** *if*

$$\frac{\widehat{\theta}_n - \theta}{\mathsf{se}} \rightsquigarrow N(0,1). \tag{6.8}$$

6.3.2 Confidence Sets

A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C_n = (a,b)$ where $a = a(X_1,\ldots,X_n)$ and $b = b(X_1,\ldots,X_n)$ are functions of the data such that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha, \quad \text{for all } \theta \in \Theta. \tag{6.9}$$

In words, (a,b) traps θ with probability $1-\alpha$. We call $1-\alpha$ the **coverage** of the confidence interval.

Warning! C_n is random and θ is fixed.

Commonly, people use 95 percent confidence intervals, which corresponds to choosing $\alpha = 0.05$. If θ is a vector then we use a **confidence set** (such as a sphere or an ellipse) instead of an interval.

Warning! There is much confusion about how to interpret a confidence interval. A confidence interval is not a probability statement about θ since θ is a fixed quantity, not a random variable. Some texts interpret confidence intervals as follows: if I repeat the experiment over and over, the interval will contain the parameter 95 percent of the time. This is correct but useless since we rarely repeat the same experiment over and over. A better interpretation is this:

> On day 1, you collect data and construct a 95 percent confidence interval for a parameter θ_1. On day 2, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_2. On day 3, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_3. You continue this way constructing confidence intervals for a sequence of unrelated parameters $\theta_1, \theta_2, \ldots$ Then 95 percent of your intervals will trap the true parameter value. There is no need to introduce the idea of repeating the same experiment over and over.

6.13 Example. Every day, newspapers report opinion polls. For example, they might say that "83 percent of the population favor arming pilots with guns." Usually, you will see a statement like "this poll is accurate to within 4 points

95 percent of the time." They are saying that 83 ± 4 is a 95 percent confidence interval for the true but unknown proportion p of people who favor arming pilots with guns. If you form a confidence interval this way every day for the rest of your life, 95 percent of your intervals will contain the true parameter. This is true even though you are estimating a different quantity (a different poll question) every day. ■

6.14 Example. The fact that a confidence interval is not a probability statement about θ is confusing. Consider this example from Berger and Wolpert (1984). Let θ be a fixed, known real number and let X_1, X_2 be independent random variables such that $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. Now define $Y_i = \theta + X_i$ and suppose that you only observe Y_1 and Y_2. Define the following "confidence interval" which actually only contains one point:

$$C = \begin{cases} \{Y_1 - 1\} & \text{if } Y_1 = Y_2 \\ \{(Y_1 + Y_2)/2\} & \text{if } Y_1 \neq Y_2. \end{cases}$$

You can check that, no matter what θ is, we have $\mathbb{P}_\theta(\theta \in C) = 3/4$ so this is a 75 percent confidence interval. Suppose we now do the experiment and we get $Y_1 = 15$ and $Y_2 = 17$. Then our 75 percent confidence interval is $\{16\}$. However, we are certain that $\theta = 16$. If you wanted to make a probability statement about θ you would probably say that $\mathbb{P}(\theta \in C | Y_1, Y_2) = 1$. There is nothing wrong with saying that $\{16\}$ is a 75 percent confidence interval. But is it not a probability statement about θ. ■

In Chapter 11 we will discuss Bayesian methods in which we treat θ as if it were a random variable and we do make probability statements about θ. In particular, we will make statements like "the probability that θ is in C_n, given the data, is 95 percent." However, these Bayesian intervals refer to degree-of-belief probabilities. These Bayesian intervals will not, in general, trap the parameter 95 percent of the time.

6.15 Example. In the coin flipping setting, let $C_n = (\widehat{p}_n - \epsilon_n,\ \widehat{p}_n + \epsilon_n)$ where $\epsilon_n^2 = \log(2/\alpha)/(2n)$. From Hoeffding's inequality (4.4) it follows that

$$\mathbb{P}(p \in C_n) \geq 1 - \alpha$$

for every p. Hence, C_n is a $1 - \alpha$ confidence interval. ■

As mentioned earlier, point estimators often have a limiting Normal distribution, meaning that equation (6.8) holds, that is, $\widehat{\theta}_n \approx N(\theta, \widehat{\mathsf{se}}^2)$. In this case we can construct (approximate) confidence intervals as follows.

6.16 Theorem (Normal-based Confidence Interval). *Suppose that* $\widehat{\theta}_n \approx N(\theta, \widehat{\mathsf{se}}^2)$. *Let* Φ *be the* CDF *of a standard Normal and let* $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$, *that is,* $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ *and* $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$ *where* $Z \sim N(0,1)$. *Let*

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}},\ \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}). \tag{6.10}$$

Then

$$\mathbb{P}_\theta(\theta \in C_n) \to 1-\alpha. \tag{6.11}$$

PROOF. Let $Z_n = (\widehat{\theta}_n - \theta)/\widehat{\mathsf{se}}$. By assumption $Z_n \rightsquigarrow Z$ where $Z \sim N(0,1)$. Hence,

$$\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}_\theta\left(\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}} < \theta < \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}\right) \\
&= \mathbb{P}_\theta\left(-z_{\alpha/2} < \frac{\widehat{\theta}_n - \theta}{\widehat{\mathsf{se}}} < z_{\alpha/2}\right) \\
&\to \mathbb{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \\
&= 1-\alpha. \quad \blacksquare
\end{aligned}$$

For 95 percent confidence intervals, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96 \approx 2$ leading to the approximate 95 percent confidence interval $\widehat{\theta}_n \pm 2\,\widehat{\mathsf{se}}$.

6.17 Example. Let $X_1, \ldots, X_n \sim$ Bernoulli(p) and let $\widehat{p}_n = n^{-1}\sum_{i=1}^n X_i$. Then $\mathbb{V}(\widehat{p}_n) = n^{-2}\sum_{i=1}^n \mathbb{V}(X_i) = n^{-2}\sum_{i=1}^n p(1-p) = n^{-2}np(1-p) = p(1-p)/n$. Hence, $\mathsf{se} = \sqrt{p(1-p)/n}$ and $\widehat{\mathsf{se}} = \sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}$. By the Central Limit Theorem, $\widehat{p}_n \approx N(p, \widehat{\mathsf{se}}^2)$. Therefore, an approximate $1-\alpha$ confidence interval is

$$\widehat{p}_n \pm z_{\alpha/2}\widehat{\mathsf{se}} = \widehat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

Compare this with the confidence interval in example 6.15. The Normal-based interval is shorter but it only has approximately (large sample) correct coverage. $\blacksquare$

6.3.3 Hypothesis Testing

In **hypothesis testing**, we start with some default theory — called a **null hypothesis** — and we ask if the data provide sufficient evidence to reject the theory. If not we retain the null hypothesis. [2]

[2] The term "retaining the null hypothesis" is due to Chris Genovese. Other terminology is "accepting the null" or "failing to reject the null."

6.18 Example (Testing if a Coin is Fair)**.** Let

$$X_1, \ldots, X_n \sim \text{Bernoulli}(p)$$

be n independent coin flips. Suppose we want to test if the coin is fair. Let H_0 denote the hypothesis that the coin is fair and let H_1 denote the hypothesis that the coin is not fair. H_0 is called the **null hypothesis** and H_1 is called the **alternative hypothesis**. We can write the hypotheses as

$$H_0 : p = 1/2 \quad \text{versus} \quad H_1 : p \neq 1/2.$$

It seems reasonable to reject H_0 if $T = |\widehat{p}_n - (1/2)|$ is large. When we discuss hypothesis testing in detail, we will be more precise about how large T should be to reject H_0. ■

6.4 Bibliographic Remarks

Statistical inference is covered in many texts. Elementary texts include DeGroot and Schervish (2002) and Larsen and Marx (1986). At the intermediate level I recommend Casella and Berger (2002), Bickel and Doksum (2000), and Rice (1995). At the advanced level, Cox and Hinkley (2000), Lehmann and Casella (1998), Lehmann (1986), and van der Vaart (1998).

6.5 Appendix

Our definition of confidence interval requires that $\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$ for all $\theta \in \Theta$. A **pointwise asymptotic** confidence interval requires that $\liminf_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$ for all $\theta \in \Theta$. A **uniform asymptotic** confidence interval requires that $\liminf_{n\to\infty} \inf_{\theta\in\Theta} \mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha$. The approximate Normal-based interval is a pointwise asymptotic confidence interval.

6.6 Exercises

1. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ and let $\widehat{\lambda} = n^{-1} \sum_{i=1}^n X_i$. Find the bias, se, and MSE of this estimator.
2. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$ and let $\widehat{\theta} = \max\{X_1, \ldots, X_n\}$. Find the bias, se, and MSE of this estimator.

3. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$ and let $\widehat{\theta} = 2\overline{X}_n$. Find the bias, se, and MSE of this estimator.

7

Estimating the CDF and Statistical Functionals

The first inference problem we will consider is nonparametric estimation of the CDF F. Then we will estimate statistical functionals, which are functions of CDF, such as the mean, the variance, and the correlation. The nonparametric method for estimating functionals is called the plug-in method.

7.1 The Empirical Distribution Function

Let $X_1, \ldots, X_n \sim F$ be an IID sample where F is a distribution function on the real line. We will estimate F with the empirical distribution function, which is defined as follows.

7.1 Definition. *The* **empirical distribution function** $\widehat{F}_n$ *is the* CDF *that puts mass* $1/n$ *at each data point* X_i*. Formally,*

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n} \tag{7.1}$$

where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x. \end{cases}$$

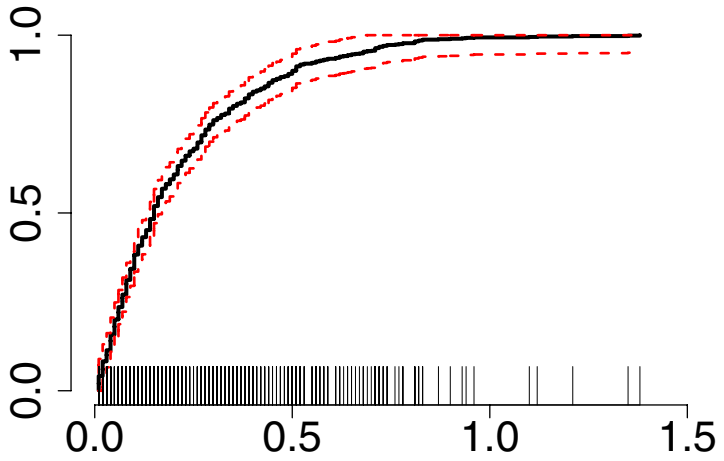


FIGURE 7.1. Nerve data. Each vertical line represents one data point. The solid line is the empirical distribution function. The lines above and below the middle line are a 95 percent confidence band.

7.2 Example (Nerve Data). Cox and Lewis (1966) reported 799 waiting times between successive pulses along a nerve fiber. Figure 7.1 shows the empirical CDF $\widehat{F}_n$. The data points are shown as small vertical lines at the bottom of the plot. Suppose we want to estimate the fraction of waiting times between .4 and .6 seconds. The estimate is $\widehat{F}_n(.6) - \widehat{F}_n(.4) = .93 - .84 = .09$. ■

7.3 Theorem. *At any fixed value of x,*

$$\begin{aligned} \mathbb{E}\left(\widehat{F}_n(x)\right) &= F(x), \\ \mathbb{V}\left(\widehat{F}_n(x)\right) &= \frac{F(x)(1-F(x))}{n}, \\ \text{MSE} &= \frac{F(x)(1-F(x))}{n} \to 0, \\ \widehat{F}_n(x) &\xrightarrow{\text{P}} F(x). \end{aligned}$$

7.4 Theorem (The Glivenko-Cantelli Theorem). *Let $X_1, \ldots, X_n \sim F$. Then* [1]

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{\text{P}} 0.$$

Now we give an inequality that will be used to construct a confidence band.

[1] More precisely, $\sup_x |\widehat{F}_n(x) - F(x)|$ converges to 0 almost surely.

7.5 Theorem (The Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality). *Let* $X_1, \ldots, X_n \sim F$. *Then, for any* $\epsilon > 0$,

$$\mathbb{P}\left(\sup_x |F(x) - \widehat{F}_n(x)| > \epsilon \right) \leq 2e^{-2n\epsilon^2}. \tag{7.2}$$

From the DKW inequality, we can construct a confidence set as follows:

A Nonparametric $1-\alpha$ Confidence Band for F

Define,

$$\begin{aligned} L(x) &= \max\{\widehat{F}_n(x) - \epsilon_n, 0\} \\ U(x) &= \min\{\widehat{F}_n(x) + \epsilon_n, 1\} \\ \text{where} \quad \epsilon_n &= \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}. \end{aligned}$$

It follows from (7.2) that for any F,

$$\mathbb{P}\Big(L(x) \leq F(x) \leq U(x) \text{ for all } x \Big) \geq 1 - \alpha. \tag{7.3}$$

7.6 Example. The dashed lines in Figure 7.1 give a 95 percent confidence band using $\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{.05}\right)} = .048$. ■

7.2 Statistical Functionals

A **statistical functional** $T(F)$ is any function of F. Examples are the mean $\mu = \int x\, dF(x)$, the variance $\sigma^2 = \int (x-\mu)^2\, dF(x)$ and the median $m = F^{-1}(1/2)$.

7.7 Definition. *The* **plug-in estimator** *of* $\theta = T(F)$ *is defined by*

$$\widehat{\theta}_n = T(\widehat{F}_n).$$

In other words, just plug in $\widehat{F}_n$ *for the unknown* F.

7.8 Definition. *If* $T(F) = \int r(x) dF(x)$ *for some function* $r(x)$ *then* T *is called a* **linear functional.**

The reason $T(F) = \int r(x)dF(x)$ is called a linear functional is because T satisfies

$$T(aF + bG) = aT(F) + bT(G),$$

hence T is linear in its arguments. Recall that $\int r(x)dF(x)$ is defined to be $\int r(x)f(x)dx$ in the continuous case and $\sum_j r(x_j)f(x_j)$ in the discrete. The empirical cdf $\widehat{F}_n(x)$ is discrete, putting mass $1/n$ at each X_i. Hence, if $T(F) = \int r(x)dF(x)$ is a linear functional then we have:

7.9 Theorem. *The plug-in estimator for linear functional* $T(F) = \int r(x)dF(x)$ *is:*

$$T(\widehat{F}_n) = \int r(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i). \tag{7.4}$$

Sometimes we can find the estimated standard error se of $T(\widehat{F}_n)$ by doing some calculations. However, in other cases it is not obvious how to estimate the standard error. In the next chapter, we will discuss a general method for finding $\widehat{\mathsf{se}}$. For now, let us just assume that somehow we can find $\widehat{\mathsf{se}}$.

In many cases, it turns out that

$$T(\widehat{F}_n) \approx N(T(F), \widehat{\mathsf{se}}^2). \tag{7.5}$$

By equation (6.11), an approximate $1-\alpha$ confidence interval for $T(F)$ is then

$$T(\widehat{F}_n) \pm z_{\alpha/2}\,\widehat{\mathsf{se}}. \tag{7.6}$$

We will call this the **Normal-based interval.** For a 95 percent confidence interval, $z_{\alpha/2} = z_{.05/2} = 1.96 \approx 2$ so the interval is

$$T(\widehat{F}_n) \pm 2\,\widehat{\mathsf{se}}.$$

7.10 Example (The mean)**.** Let $\mu = T(F) = \int x\,dF(x)$. The plug-in estimator is $\widehat{\mu} = \int x\,d\widehat{F}_n(x) = \overline{X}_n$. The standard error is $\mathsf{se} = \sqrt{\mathbb{V}(\overline{X}_n)} = \sigma/\sqrt{n}$. If $\widehat{\sigma}$ denotes an estimate of σ, then the estimated standard error is $\widehat{\sigma}/\sqrt{n}$. (In the next example, we shall see how to estimate σ.) A Normal-based confidence interval for μ is $\overline{X}_n \pm z_{\alpha/2}\,\widehat{\mathsf{se}}$. ■

7.11 Example (The Variance)**.** Let $\sigma^2 = T(F) = \mathbb{V}(X) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$. The plug-in estimator is

$$\widehat{\sigma}^2 \quad = \quad \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Another reasonable estimator of σ^2 is the sample variance

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

In practice, there is little difference between $\widehat{\sigma}^2$ and S_n^2 and you can use either one. Returning to the last example, we now see that the estimated standard error of the estimate of the mean is $\widehat{\mathsf{se}} = \widehat{\sigma}/\sqrt{n}$. ∎

7.12 Example (The Skewness)**.** Let μ and σ^2 denote the mean and variance of a random variable X. The skewness is defined to be

$$\kappa = \frac{\mathbb{E}(X-\mu)^3}{\sigma^3} = \frac{\int (x-\mu)^3 dF(x)}{\left\{\int (x-\mu)^2 dF(x)\right\}^{3/2}}.$$

The skewness measures the lack of symmetry of a distribution. To find the plug-in estimate, first recall that $\widehat{\mu} = n^{-1}\sum_i X_i$ and $\widehat{\sigma}^2 = n^{-1}\sum_i (X_i - \widehat{\mu})^2$. The plug-in estimate of κ is

$$\widehat{\kappa} = \frac{\int (x-\mu)^3 d\widehat{F}_n(x)}{\left\{\int (x-\mu)^2 d\widehat{F}_n(x)\right\}^{3/2}} = \frac{\frac{1}{n}\sum_i (X_i - \widehat{\mu})^3}{\widehat{\sigma}^3}. \quad ∎$$

7.13 Example (Correlation)**.** Let $Z = (X, Y)$ and let $\rho = T(F) = \mathbb{E}(X - \mu_X)(Y - \mu_Y)/(\sigma_x \sigma_y)$ denote the correlation between X and Y, where $F(x, y)$ is bivariate. We can write

$$T(F) = a(T_1(F), T_2(F), T_3(F), T_4(F), T_5(F))$$

where

$$T_1(F) = \int x\, dF(z), \quad T_2(F) = \int y\, dF(z), \quad T_3(F) = \int xy\, dF(z),$$
$$T_4(F) = \int x^2\, dF(z), \quad T_5(F) = \int y^2\, dF(z),$$

and

$$a(t_1, \ldots, t_5) = \frac{t_3 - t_1 t_2}{\sqrt{(t_4 - t_1^2)(t_5 - t_2^2)}}.$$

Replace F with $\widehat{F}_n$ in $T_1(F), \ldots, T_5(F)$, and take

$$\widehat{\rho} = a(T_1(\widehat{F}_n), T_2(\widehat{F}_n), T_3(\widehat{F}_n), T_4(\widehat{F}_n), T_5(\widehat{F}_n)).$$

We get

$$\widehat{\rho} = \frac{\sum_i (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_i (X_i - \overline{X}_n)^2}\sqrt{\sum_i (Y_i - \overline{Y}_n)^2}}$$

which is called the **sample correlation**. ■

7.14 Example (Quantiles)**.** Let F be strictly increasing with density f. For $0 < p < 1$, the p^{th} quantile is defined by $T(F) = F^{-1}(p)$. The estimate if $T(F)$ is $\widehat{F}_n^{-1}(p)$. We have to be a bit careful since $\widehat{F}_n$ is not invertible. To avoid ambiguity we define

$$\widehat{F}_n^{-1}(p) = \inf\{x : \ \widehat{F}_n(x) \geq p\}.$$

We call $T(\widehat{F}_n) = \widehat{F}_n^{-1}(p)$ the p^{th} **sample quantile**. ■

Only in the first example did we compute a standard error or a confidence interval. How shall we handle the other examples? When we discuss parametric methods, we will develop formulas for standard errors and confidence intervals. But in our nonparametric setting we need something else. In the next chapter, we will introduce the bootstrap for getting standard errors and confidence intervals.

7.15 Example (Plasma Cholesterol)**.** Figure 7.2 shows histograms for plasma cholesterol (in mg/dl) for 371 patients with chest pain (Scott et al. (1978)). The histograms show the percentage of patients in 10 bins. The first histogram is for 51 patients who had no evidence of heart disease while the second histogram is for 320 patients who had narrowing of the arteries. Is the mean cholesterol different in the two groups? Let us regard these data as samples from two distributions F_1 and F_2. Let $\mu_1 = \int x dF_1(x)$ and $\mu_2 = \int x dF_2(x)$ denote the means of the two populations. The plug-in estimates are $\widehat{\mu}_1 = \int x d\widehat{F}_{n,1}(x) = \overline{X}_{n,1} = 195.27$ and $\widehat{\mu}_2 = \int x d\widehat{F}_{n,2}(x) = \overline{X}_{n,2} = 216.19$. Recall that the standard error of the sample mean $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ is

$$\mathsf{se}(\widehat{\mu}) = \sqrt{\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n X_i\right)} = \sqrt{\frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(X_i)} = \sqrt{\frac{n\sigma^2}{n^2}} = \frac{\sigma}{\sqrt{n}}$$

which we estimate by

$$\widehat{\mathsf{se}}(\widehat{\mu}) = \frac{\widehat{\sigma}}{\sqrt{n}}$$

where

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

For the two groups this yields $\widehat{\mathsf{se}}(\widehat{\mu}_1) = 5.0$ and $\widehat{\mathsf{se}}(\widehat{\mu}_2) = 2.4$. Approximate 95 percent confidence intervals for μ_1 and μ_2 are $\widehat{\mu}_1 \pm 2\widehat{\mathsf{se}}(\widehat{\mu}_1) = (185, 205)$ and $\widehat{\mu}_2 \pm 2\widehat{\mathsf{se}}(\widehat{\mu}_2) = (211, 221)$.

Now, consider the functional $\theta = T(F_2) - T(F_1)$ whose plug-in estimate is $\widehat{\theta} = \widehat{\mu}_2 - \widehat{\mu}_1 = 216.19 - 195.27 = 20.92$. The standard error of $\widehat{\theta}$ is

$$\mathsf{se} = \sqrt{\mathbb{V}(\widehat{\mu}_2 - \widehat{\mu}_1)} = \sqrt{\mathbb{V}(\widehat{\mu}_2) + \mathbb{V}(\widehat{\mu}_1)} = \sqrt{(\mathsf{se}(\widehat{\mu}_1))^2 + (\mathsf{se}(\widehat{\mu}_2))^2}$$

and we estimate this by

$$\widehat{\mathsf{se}} = \sqrt{(\widehat{\mathsf{se}}(\widehat{\mu}_1))^2 + (\widehat{\mathsf{se}}(\widehat{\mu}_2))^2} = 5.55.$$

An approximate 95 percent confidence interval for θ is $\widehat{\theta} \pm 2\,\widehat{\mathsf{se}}(\widehat{\theta}_n) = (9.8, 32.0)$. This suggests that cholesterol is higher among those with narrowed arteries. We should not jump to the conclusion (from these data) that cholesterol causes heart disease. The leap from statistical evidence to causation is very subtle and is discussed in Chapter 16. ■

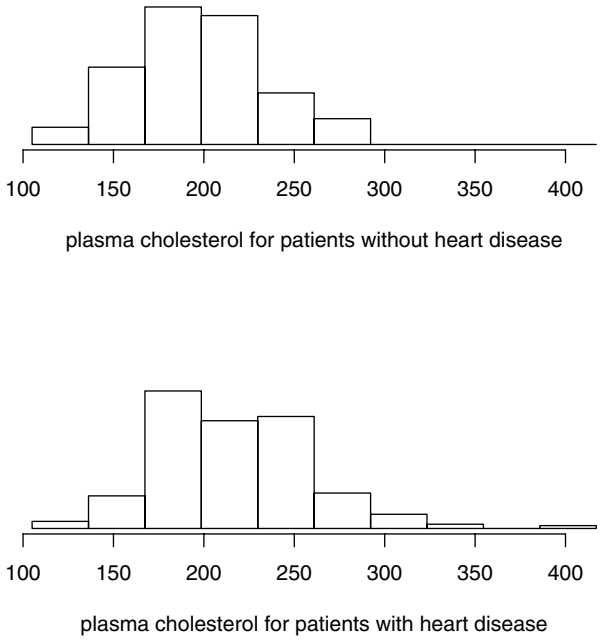


FIGURE 7.2. Plasma cholesterol for 51 patients with no heart disease and 320 patients with narrowing of the arteries.

7.3 Bibliographic Remarks

The Glivenko-Cantelli theorem is the tip of the iceberg. The theory of distribution functions is a special case of what are called empirical processes which underlie much of modern statistical theory. Some references on empirical processes are Shorack and Wellner (1986) and van der Vaart and Wellner (1996).

7.4 Exercises

1. Prove Theorem 7.3.

2. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $Y_1, \ldots, Y_m \sim \text{Bernoulli}(q)$. Find the plug-in estimator and estimated standard error for p. Find an approximate 90 percent confidence interval for p. Find the plug-in estimator and estimated standard error for $p - q$. Find an approximate 90 percent confidence interval for $p - q$.

3. (Computer Experiment.) Generate 100 observations from a N(0,1) distribution. Compute a 95 percent confidence band for the CDF F (as described in the appendix). Repeat this 1000 times and see how often the confidence band contains the true distribution function. Repeat using data from a Cauchy distribution.

4. Let $X_1, \ldots, X_n \sim F$ and let $\widehat{F}_n(x)$ be the empirical distribution function. For a fixed x, use the central limit theorem to find the limiting distribution of $\widehat{F}_n(x)$.

5. Let x and y be two distinct points. Find $\text{Cov}(\widehat{F}_n(x), \widehat{F}_n(y))$.

6. Let $X_1, \ldots, X_n \sim F$ and let $\widehat{F}$ be the empirical distribution function. Let $a < b$ be fixed numbers and define $\theta = T(F) = F(b) - F(a)$. Let $\widehat{\theta} = T(\widehat{F}_n) = \widehat{F}_n(b) - \widehat{F}_n(a)$. Find the estimated standard error of $\widehat{\theta}$. Find an expression for an approximate $1 - \alpha$ confidence interval for θ.

7. Data on the magnitudes of earthquakes near Fiji are available on the website for this book. Estimate the CDF $F(x)$. Compute and plot a 95 percent confidence envelope for F (as described in the appendix). Find an approximate 95 percent confidence interval for $F(4.9) - F(4.3)$.

8. Get the data on eruption times and waiting times between eruptions of the Old Faithful geyser from the website. Estimate the mean waiting time and give a standard error for the estimate. Also, give a 90 percent confidence interval for the mean waiting time. Now estimate the median waiting time. In the next chapter we will see how to get the standard error for the median.

9. 100 people are given a standard antibiotic to treat an infection and another 100 are given a new antibiotic. In the first group, 90 people recover; in the second group, 85 people recover. Let p_1 be the probability of recovery under the standard treatment and let p_2 be the probability of recovery under the new treatment. We are interested in estimating $\theta = p_1 - p_2$. Provide an estimate, standard error, an 80 percent confidence interval, and a 95 percent confidence interval for θ.

10. In 1975, an experiment was conducted to see if cloud seeding produced rainfall. 26 clouds were seeded with silver nitrate and 26 were not. The decision to seed or not was made at random. Get the data from

    http://lib.stat.cmu.edu/DASL/Stories/CloudSeeding.html

    Let θ be the difference in the mean precipitation from the two groups. Estimate θ. Estimate the standard error of the estimate and produce a 95 percent confidence interval.

8
The Bootstrap

The **bootstrap** is a method for estimating standard errors and computing confidence intervals. Let $T_n = g(X_1, \ldots, X_n)$ be a **statistic**, that is, T_n is any function of the data. Suppose we want to know $\mathbb{V}_F(T_n)$, the variance of T_n. We have written $\mathbb{V}_F$ to emphasize that the variance usually depends on the unknown distribution function F. For example, if $T_n = \overline{X}_n$ then $\mathbb{V}_F(T_n) = \sigma^2/n$ where $\sigma^2 = \int (x-\mu)^2 dF(x)$ and $\mu = \int x dF(x)$. Thus the variance of T_n is a function of F. The bootstrap idea has two steps:

Step 1: Estimate $\mathbb{V}_F(T_n)$ with $\mathbb{V}_{\widehat{F}_n}(T_n)$.

Step 2: Approximate $\mathbb{V}_{\widehat{F}_n}(T_n)$ using simulation.

For $T_n = \overline{X}_n$, we have for Step 1 that $\mathbb{V}_{\widehat{F}_n}(T_n) = \widehat{\sigma}^2/n$ where $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)$. In this case, Step 1 is enough. However, in more complicated cases we cannot write down a simple formula for $\mathbb{V}_{\widehat{F}_n}(T_n)$ which is why we need Step 2. Before proceeding, let us discuss the idea of simulation.

8.1 Simulation

Suppose we draw an IID sample $Y_1, \ldots, Y_B$ from a distribution G. By the law of large numbers,

$$\overline{Y}_n = \frac{1}{B}\sum_{j=1}^{B} Y_j \xrightarrow{\text{P}} \int y\, dG(y) = \mathbb{E}(Y)$$

as $B \to \infty$. So if we draw a large sample from G, we can use the sample mean $\overline{Y}_n$ to approximate $\mathbb{E}(Y)$. In a simulation, we can make B as large as we like, in which case, the difference between $\overline{Y}_n$ and $\mathbb{E}(Y)$ is negligible. More generally, if h is any function with finite mean then

$$\frac{1}{B}\sum_{j=1}^{B} h(Y_j) \xrightarrow{\text{P}} \int h(y) dG(y) = \mathbb{E}(h(Y))$$

as $B \to \infty$. In particular,

$$\begin{aligned} \frac{1}{B}\sum_{j=1}^{B}(Y_j - \overline{Y})^2 &= \frac{1}{B}\sum_{j=1}^{B} Y_j^2 - \left(\frac{1}{B}\sum_{j=1}^{B} Y_j\right)^2 \\ &\xrightarrow{\text{P}} \int y^2 dF(y) - \left(\int y dF(y)\right)^2 = \mathbb{V}(Y). \end{aligned}$$

Hence, we can use the sample variance of the simulated values to approximate $\mathbb{V}(Y)$.

8.2 Bootstrap Variance Estimation

According to what we just learned, we can approximate $\mathbb{V}_{\widehat{F}_n}(T_n)$ by simulation. Now $\mathbb{V}_{\widehat{F}_n}(T_n)$ means "the variance of T_n if the distribution of the data is $\widehat{F}_n$." How can we simulate from the distribution of T_n when the data are assumed to have distribution $\widehat{F}_n$? The answer is to simulate $X_1^*, \ldots, X_n^*$ from $\widehat{F}_n$ and then compute $T_n^* = g(X_1^*, \ldots, X_n^*)$. This constitutes one draw from the distribution of T_n. The idea is illustrated in the following diagram:

$$\begin{array}{rccccc} \text{Real world} & F & \Longrightarrow & X_1, \ldots, X_n & \Longrightarrow & T_n = g(X_1, \ldots, X_n) \\ \text{Bootstrap world} & \widehat{F}_n & \Longrightarrow & X_1^*, \ldots, X_n^* & \Longrightarrow & T_n^* = g(X_1^*, \ldots, X_n^*) \end{array}$$

How do we simulate $X_1^*, \ldots, X_n^*$ from $\widehat{F}_n$? Notice that $\widehat{F}_n$ puts mass 1/n at each data point $X_1, \ldots, X_n$. Therefore,

drawing an observation from $\widehat{F}_n$ is equivalent to drawing one point at random from the original data set.

Thus, to simulate $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$ it suffices to draw n observations with replacement from $X_1, \ldots, X_n$. Here is a summary:

Bootstrap Variance Estimation

1. Draw $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$.
2. Compute $T_n^* = g(X_1^*, \ldots, X_n^*)$.
3. Repeat steps 1 and 2, B times, to get $T_{n,1}^*, \ldots, T_{n,B}^*$.
4. Let

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*\right)^2. \tag{8.1}$$

8.1 Example. The following pseudocode shows how to use the bootstrap to estimate the standard error of the median.

Bootstrap for The Median

```
Given data X = (X(1), ..., X(n)):

T <- median(X)
Tboot <- vector of length B
for(i in 1:B){
    Xstar <- sample of size n from X (with replacement)
    Tboot[i] <- median(Xstar)
    }
se <- sqrt(variance(Tboot))
```

The following schematic diagram will remind you that we are using two approximations:

$$\mathbb{V}_F(T_n) \overbrace{\approx}^{\text{not so small}} \mathbb{V}_{\widehat{F}_n}(T_n) \overbrace{\approx}^{\text{small}} v_{boot}.$$

8.2 Example. Consider the nerve data. Let $\theta = T(F) = \int (x-\mu)^3 dF(x)/\sigma^3$ be the skewness. The skewness is a measure of asymmetry. A Normal distribution,

for example, has skewness 0. The plug-in estimate of the skewness is

$$\widehat{\theta} = T(\widehat{F}_n) = \frac{\int (x-\mu)^3 d\widehat{F}_n(x)}{\widehat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^3}{\widehat{\sigma}^3} = 1.76.$$

To estimate the standard error with the bootstrap we follow the same steps as with the median example except we compute the skewness from each bootstrap sample. When applied to the nerve data, the bootstrap, based on $B = 1,000$ replications, yields a standard error for the estimated skewness of .16. ■

8.3 Bootstrap Confidence Intervals

There are several ways to construct bootstrap confidence intervals. Here we discuss three methods.

Method 1: The Normal Interval. The simplest method is the Normal interval

$$T_n \pm z_{\alpha/2}\, \widehat{\mathsf{se}}_{\text{boot}} \tag{8.2}$$

where $\widehat{\mathsf{se}}_{\text{boot}} = \sqrt{v_{\text{boot}}}$ is the bootstrap estimate of the standard error. This interval is not accurate unless the distribution of T_n is close to Normal.

Method 2: Pivotal Intervals. Let $\theta = T(F)$ and $\widehat{\theta}_n = T(\widehat{F}_n)$ and define the **pivot** $R_n = \widehat{\theta}_n - \theta$. Let $\widehat{\theta}^*_{n,1}, \ldots, \widehat{\theta}^*_{n,B}$ denote bootstrap replications of $\widehat{\theta}_n$. Let $H(r)$ denote the CDF of the pivot:

$$H(r) = \mathbb{P}_F(R_n \leq r). \tag{8.3}$$

Define $C^\star_n = (a, b)$ where

$$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \quad \text{and} \quad b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right). \tag{8.4}$$

It follows that

$$\begin{aligned}
\mathbb{P}(a \leq \theta \leq b) &= \mathbb{P}(a - \widehat{\theta}_n \leq \theta - \widehat{\theta}_n \leq b - \widehat{\theta}_n) \\
&= \mathbb{P}(\widehat{\theta}_n - b \leq \widehat{\theta}_n - \theta \leq \widehat{\theta}_n - a) \\
&= \mathbb{P}(\widehat{\theta}_n - b \leq R_n \leq \widehat{\theta}_n - a) \\
&= H(\widehat{\theta}_n - a) - H(\widehat{\theta}_n - b) \\
&= H\left(H^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.
\end{aligned}$$

Hence, $C_n^\star$ is an exact $1-\alpha$ confidence interval for θ. Unfortunately, a and b depend on the unknown distribution H but we can form a bootstrap estimate of H:

$$\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R_{n,b}^* \leq r) \tag{8.5}$$

where $R_{n,b}^* = \widehat{\theta}_{n,b}^* - \widehat{\theta}_n$. Let r_β^* denote the β sample quantile of $(R_{n,1}^*, \ldots, R_{n,B}^*)$ and let θ_β^* denote the β sample quantile of $(\widehat{\theta}_{n,1}^*, \ldots, \widehat{\theta}_{n,B}^*)$. Note that $r_\beta^* = \theta_\beta^* - \widehat{\theta}_n$. It follows that an approximate $1-\alpha$ confidence interval is $C_n = (\widehat{a}, \widehat{b})$ where

$$\begin{aligned} \widehat{a} &= \widehat{\theta}_n - \widehat{H}^{-1}\left(1-\frac{\alpha}{2}\right) = \widehat{\theta}_n - r_{1-\alpha/2}^* = 2\widehat{\theta}_n - \theta_{1-\alpha/2}^* \\ \widehat{b} &= \widehat{\theta}_n - \widehat{H}^{-1}\left(\frac{\alpha}{2}\right) \quad = \widehat{\theta}_n - r_{\alpha/2}^* \quad = 2\widehat{\theta}_n - \theta_{\alpha/2}^*. \end{aligned}$$

In summary, the $1-\alpha$ **bootstrap pivotal confidence** interval is

$$C_n = \left(2\widehat{\theta}_n - \widehat{\theta}_{1-\alpha/2}^*,\ 2\widehat{\theta}_n - \widehat{\theta}_{\alpha/2}^*\right). \tag{8.6}$$

8.3 Theorem. *Under weak conditions on $T(F)$,*

$$\mathbb{P}_F(T(F) \in C_n) \to 1-\alpha$$

as $n \to \infty$, where C_n is given in (8.6).

Method 3: Percentile Intervals. The **bootstrap percentile interval** is defined by

$$C_n = \left(\theta_{\alpha/2}^*,\ \theta_{1-\alpha/2}^*\right).$$

The justification for this interval is given in the appendix.

8.4 Example. For estimating the skewness of the nerve data, here are the various confidence intervals.

Method	95% Interval
Normal	(1.44, 2.09)
Pivotal	(1.48, 2.11)
Percentile	(1.42, 2.03)

All these confidence intervals are approximate. The probability that $T(F)$ is in the interval is not exactly $1-\alpha$. All three intervals have the same level of accuracy. There are more accurate bootstrap confidence intervals but they are more complicated and we will not discuss them here.

8.5 Example (The Plasma Cholesterol Data)**.** Let us return to the cholesterol data. Suppose we are interested in the difference of the medians. Pseudocode for the bootstrap analysis is as follows:

```
x1 <- first sample
x2 <- second sample
n1 <- length(x1)
n2 <- length(x2)
th.hat <- median(x2) - median(x1)
B <- 1000
Tboot <- vector of length B
for(i in 1:B){
     xx1 <- sample of size n1 with replacement from x1
     xx2 <- sample of size n2 with replacement from x2
     Tboot[i] <- median(xx2) - median(xx1)
     }
se <- sqrt(variance(Tboot))
Normal     <- (th.hat - 2*se, th.hat + 2*se)
percentile <- (quantile(Tboot,.025), quantile(Tboot,.975))
pivotal    <- ( 2*th.hat-quantile(Tboot,.975),
                  2*th.hat-quantile(Tboot,.025) )
```

The point estimate is 18.5, the bootstrap standard error is 7.42 and the resulting approximate 95 percent confidence intervals are as follows:

Method	95% Interval
Normal	(3.7, 33.3)
Pivotal	(5.0, 34.0)
Percentile	(5.0, 33.3)

Since these intervals exclude 0, it appears that the second group has higher cholesterol although there is considerable uncertainty about how much higher as reflected in the width of the intervals. ■

The next two examples are based on small sample sizes. In practice, statistical methods based on very small sample sizes might not be reliable. We include the examples for their pedagogical value but we do want to sound a note of caution about interpreting the results with some skepticism.

8.6 Example. Here is an example that was one of the first used to illustrate the bootstrap by Bradley Efron, the inventor of the bootstrap. The data are LSAT scores (for entrance to law school) and GPA.

```
LSAT   576   635   558   578   666   580   555   661
       651   605   653   575   545   572   594

GPA    3.39  3.30  2.81  3.03  3.44  3.07  3.00  3.43
       3.36  3.13  3.12  2.74  2.76  2.88  3.96
```

Each data point is of the form $X_i = (Y_i, Z_i)$ where $Y_i = \text{LSAT}_i$ and $Z_i = \text{GPA}_i$. The law school is interested in the correlation

$$\theta = \frac{\int\int (y - \mu_Y)(z - \mu_Z) dF(y, z)}{\sqrt{\int (y - \mu_Y)^2 dF(y) \int (z - \mu_Z)^2 dF(z)}}.$$

The plug-in estimate is the sample correlation

$$\widehat{\theta} = \frac{\sum_i (Y_i - \overline{Y})(Z_i - \overline{Z})}{\sqrt{\sum_i (Y_i - \overline{Y})^2 \sum_i (Z_i - \overline{Z})^2}}.$$

The estimated correlation is $\widehat{\theta} = .776$. The bootstrap based on $B = 1000$ gives $\widehat{\mathsf{se}} = .137$. Figure 8.1 shows the data and a histogram of the bootstrap replications $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_B$. This histogram is an approximation to the sampling distribution of $\widehat{\theta}$. The Normal-based 95 percent confidence interval is $.78 \pm 2\widehat{\mathsf{se}} = (.51, 1.00)$ while the percentile interval is (.46,.96). In large samples, the two methods will show closer agreement. ■

8.7 Example. This example is from Efron and Tibshirani (1993). When drug companies introduce new medications, they are sometimes required to show bioequivalence. This means that the new drug is not substantially different than the current treatment. Here are data on eight subjects who used medical patches to infuse a hormone into the blood. Each subject received three treatments: placebo, old-patch, new-patch.

subject	placebo	old	new	old − placebo	new − old
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
4	13357	21816	23798	8459	1982
5	9055	13850	12560	4795	-1290
6	6290	9806	10157	3516	351
7	12412	17208	16570	4796	-638
8	18806	29044	26325	10238	-2719

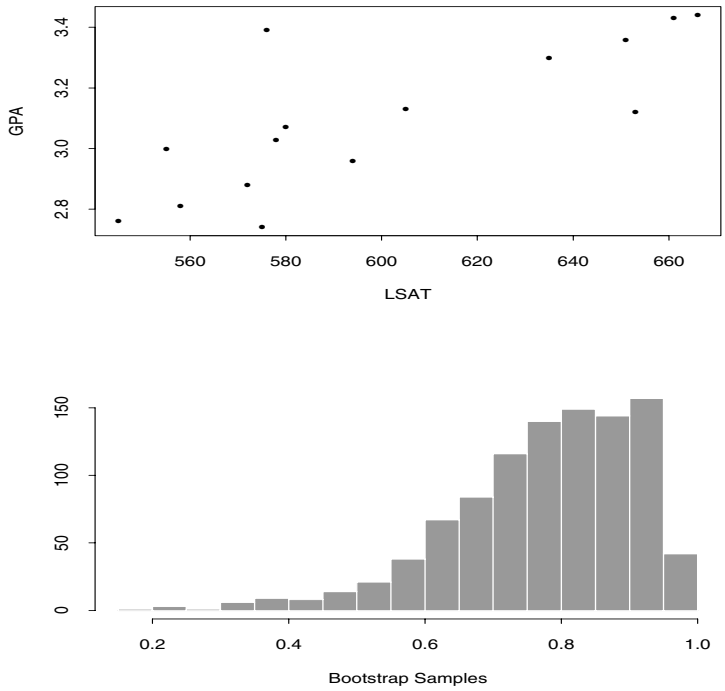


FIGURE 8.1. Law school data. The top panel shows the raw data. The bottom panel is a histogram of the correlations computed from each bootstrap sample.

Let $Z = \text{old} - \text{placebo}$ and $Y = \text{new} - \text{old}$. The Food and Drug Administration (FDA) requirement for bioequivalence is that $|\theta| \leq .20$ where

$$\theta = \frac{\mathbb{E}_F(Y)}{\mathbb{E}_F(Z)}.$$

The plug-in estimate of θ is

$$\widehat{\theta} = \frac{\overline{Y}}{\overline{Z}} = \frac{-452.3}{6342} = -0.0713.$$

The bootstrap standard error is $\widehat{\mathsf{se}} = 0.105$. To answer the bioequivalence question, we compute a confidence interval. From $B = 1000$ bootstrap replications we get the 95 percent interval (-0.24,0.15). This is not quite contained

in (-0.20,0.20) so at the 95 percent level we have not demonstrated bioequivalence. Figure 8.2 shows the histogram of the bootstrap values. ■

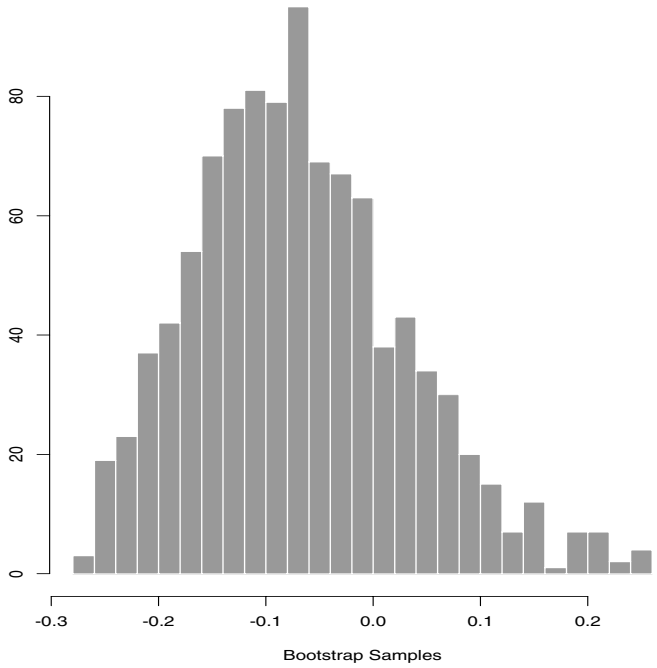


FIGURE 8.2. Patch data.

8.4 Bibliographic Remarks

The bootstrap was invented by Efron (1979). There are several books on these topics including Efron and Tibshirani (1993), Davison and Hinkley (1997), Hall (1992) and Shao and Tu (1995). Also, see section 3.6 of van der Vaart and Wellner (1996).

8.5 Appendix

8.5.1 The Jackknife

There is another method for computing standard errors called the **jackknife**, due to Quenouille (1949). It is less computationally expensive than the boot-

strap but is less general. Let $T_n = T(X_1, \ldots, X_n)$ be a statistic and $T_{(-i)}$ denote the statistic with the i^{th} observation removed. Let $\overline{T}_n = n^{-1}\sum_{i=1}^n T_{(-i)}$. The jackknife estimate of $\text{var}(T_n)$ is

$$v_{\text{jack}} = \frac{n-1}{n}\sum_{i=1}^n (T_{(-i)} - \overline{T}_n)^2$$

and the jackknife estimate of the standard error is $\widehat{\mathsf{se}}_{jack} = \sqrt{v_{jack}}$. Under suitable conditions on T, it can be shown that v_{jack} consistently estimates $\text{var}(T_n)$ in the sense that $v_{\text{jack}}/\text{var}(\text{T}_\text{n}) \xrightarrow{\text{P}} 1$. However, unlike the bootstrap, the jackknife does not produce consistent estimates of the standard error of sample quantiles.

8.5.2 Justification For The Percentile Interval

Suppose there exists a monotone transformation $U = m(T)$ such that $U \sim N(\phi, c^2)$ where $\phi = m(\theta)$. We do not suppose we know the transformation, only that one exists. Let $U_b^* = m(\theta_{n,b}^*)$. Let u_β^* be the β sample quantile of the U_b^*'s. Since a monotone transformation preserves quantiles, we have that $u_{\alpha/2}^* = m(\theta_{\alpha/2}^*)$. Also, since $U \sim N(\phi, c^2)$, the $\alpha/2$ quantile of U is $\phi - z_{\alpha/2}c$. Hence $u_{\alpha/2}^* = \phi - z_{\alpha/2}c$. Similarly, $u_{1-\alpha/2}^* = \phi + z_{\alpha/2}c$. Therefore,

$$\begin{aligned}
\mathbb{P}(\theta_{\alpha/2}^* \leq \theta \leq \theta_{1-\alpha/2}^*) &= \mathbb{P}(m(\theta_{\alpha/2}^*) \leq m(\theta) \leq m(\theta_{1-\alpha/2}^*)) \\
&= \mathbb{P}(u_{\alpha/2}^* \leq \phi \leq u_{1-\alpha/2}^*) \\
&= \mathbb{P}(U - cz_{\alpha/2} \leq \phi \leq U + cz_{\alpha/2}) \\
&= \mathbb{P}(-z_{\alpha/2} \leq \frac{U-\phi}{c} \leq z_{\alpha/2}) \\
&= 1 - \alpha.
\end{aligned}$$

An exact normalizing transformation will rarely exist but there may exist approximate normalizing transformations.

8.6 Exercises

1. Consider the data in Example 8.6. Find the plug-in estimate of the correlation coefficient. Estimate the standard error using the bootstrap. Find a 95 percent confidence interval using the Normal, pivotal, and percentile methods.

2. (Computer Experiment.) Conduct a simulation to compare the various bootstrap confidence interval methods. Let $n = 50$ and let $T(F) = \int (x-\mu)^3 dF(x)/\sigma^3$ be the skewness. Draw $Y_1, \ldots, Y_n \sim N(0,1)$ and set $X_i = e^{Y_i}$, $i = 1, \ldots, n$. Construct the three types of bootstrap 95 percent intervals for $T(F)$ from the data $X_1, \ldots, X_n$. Repeat this whole thing many times and estimate the true coverage of the three intervals.

3. Let
$$X_1, \ldots, X_n \sim t_3$$
where $n = 25$. Let $\theta = T(F) = (q_{.75} - q_{.25})/1.34$ where q_p denotes the p^{th} quantile. Do a simulation to compare the coverage and length of the following confidence intervals for θ: (i) Normal interval with standard error from the bootstrap, (ii) bootstrap percentile interval, and (iii) pivotal bootstrap interval.

4. Let $X_1, \ldots, X_n$ be distinct observations (no ties). Show that there are
$$\binom{2n-1}{n}$$
distinct bootstrap samples.

   Hint: Imagine putting n balls into n buckets.

5. Let $X_1, \ldots, X_n$ be distinct observations (no ties). Let $X_1^*, \ldots, X_n^*$ denote a bootstrap sample and let $\overline{X}_n^* = n^{-1} \sum_{i=1}^n X_i^*$. Find: $\mathbb{E}(\overline{X}_n^* | X_1, \ldots, X_n)$, $\mathbb{V}(\overline{X}_n^* | X_1, \ldots, X_n)$, $\mathbb{E}(\overline{X}_n^*)$ and $\mathbb{V}(\overline{X}_n^*)$.

6. (Computer Experiment.) Let $X_1, ..., X_n$ Normal$(\mu, 1)$. Let $\theta = e^\mu$ and let $\widehat{\theta} = e^{\overline{X}}$. Create a data set (using $\mu = 5$) consisting of n=100 observations.

   (a) Use the bootstrap to get the se and 95 percent confidence interval for θ.

   (b) Plot a histogram of the bootstrap replications. This is an estimate of the distribution of $\widehat{\theta}$. Compare this to the true sampling distribution of $\widehat{\theta}$.

7. Let $X_1, ..., X_n \sim$ Uniform$(0, \theta)$. Let $\widehat{\theta} = X_{max} = \max\{X_1, ..., X_n\}$. Generate a data set of size 50 with $\theta = 1$.

   (a) Find the distribution of $\widehat{\theta}$. Compare the true distribution of $\widehat{\theta}$ to the histograms from the bootstrap.

(b) This is a case where the bootstrap does very poorly. In fact, we can prove that this is the case. Show that $P(\widehat{\theta} = \widehat{\theta}) = 0$ and yet $P(\widehat{\theta}^* = \widehat{\theta}) \approx .632$. Hint: show that, $P(\widehat{\theta}^* = \widehat{\theta}) = 1 - (1 - (1/n))^n$ then take the limit as n gets large.

8. Let $T_n = \overline{X}_n^2$, $\mu = \mathbb{E}(X_1)$, $\alpha_k = \int |x-\mu|^k dF(x)$ and $\widehat{\alpha}_k = n^{-1}\sum_{i=1}^n |X_i - \overline{X}_n|^k$. Show that

$$v_{\text{boot}} = \frac{4\overline{X}_n^2\widehat{\alpha}_2}{n} + \frac{4\overline{X}_n\widehat{\alpha}_3}{n^2} + \frac{\widehat{\alpha}_4}{n^3}.$$

9
Parametric Inference

We now turn our attention to parametric models, that is, models of the form

$$\mathfrak{F} = \left\{ f(x;\theta) : \ \theta \in \Theta \right\} \tag{9.1}$$

where the $\Theta \subset \mathbb{R}^k$ is the parameter space and $\theta = (\theta_1, \ldots, \theta_k)$ is the parameter. The problem of inference then reduces to the problem of estimating the parameter θ.

Students learning statistics often ask: how would we ever know that the distribution that generated the data is in some parametric model? This is an excellent question. Indeed, we would rarely have such knowledge which is why nonparametric methods are preferable. Still, studying methods for parametric models is useful for two reasons. First, there are some cases where background knowledge suggests that a parametric model provides a reasonable approximation. For example, counts of traffic accidents are known from prior experience to follow approximately a Poisson model. Second, the inferential concepts for parametric models provide background for understanding certain nonparametric methods.

We begin with a brief discussion about parameters of interest and nuisance parameters in the next section, then we will discuss two methods for estimating θ, the method of moments and the method of maximum likelihood.

9.1 Parameter of Interest

Often, we are only interested in some function $T(\theta)$. For example, if $X \sim N(\mu, \sigma^2)$ then the parameter is $\theta = (\mu, \sigma)$. If our goal is to estimate μ then $\mu = T(\theta)$ is called the **parameter of interest** and σ is called a **nuisance parameter**. The parameter of interest might be a complicated function of θ as in the following example.

9.1 Example. Let $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$. The parameter is $\theta = (\mu, \sigma)$ and the parameter space is $\Theta = \{(\mu, \sigma) : \ \mu \in \mathbb{R}, \ \sigma > 0\}$. Suppose that X_i is the outcome of a blood test and suppose we are interested in τ, the fraction of the population whose test score is larger than 1. Let Z denote a standard Normal random variable. Then

$$\begin{aligned} \tau &= \mathbb{P}(X > 1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}\left(\frac{X-\mu}{\sigma} < \frac{1-\mu}{\sigma}\right) \\ &= 1 - \mathbb{P}\left(Z < \frac{1-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{1-\mu}{\sigma}\right). \end{aligned}$$

The parameter of interest is $\tau = T(\mu, \sigma) = 1 - \Phi((1-\mu)/\sigma)$. ■

9.2 Example. Recall that X has a Gamma(α, β) distribution if

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

where $\alpha, \beta > 0$ and

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

is the Gamma function. The parameter is $\theta = (\alpha, \beta)$. The Gamma distribution is sometimes used to model lifetimes of people, animals, and electronic equipment. Suppose we want to estimate the mean lifetime. Then $T(\alpha, \beta) = \mathbb{E}_\theta(X_1) = \alpha\beta$. ■

9.2 The Method of Moments

The first method for generating parametric estimators that we will study is called the method of moments. We will see that these estimators are not optimal but they are often easy to compute. They are are also useful as starting values for other methods that require iterative numerical routines.

Suppose that the parameter $\theta = (\theta_1, \ldots, \theta_k)$ has k components. For $1 \leq j \leq k$, define the j^{th} **moment**

$$\alpha_j \equiv \alpha_j(\theta) = \mathbb{E}_\theta(X^j) = \int x^j dF_\theta(x) \tag{9.2}$$

and the j^{th} **sample moment**

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^n X_i^j. \tag{9.3}$$

9.3 Definition. *The* **method of moments estimator** $\widehat{\theta}_n$ *is defined to be the value of* θ *such that*

$$\begin{aligned} \alpha_1(\widehat{\theta}_n) &= \widehat{\alpha}_1 \\ \alpha_2(\widehat{\theta}_n) &= \widehat{\alpha}_2 \\ &\vdots \quad \vdots \quad \vdots \\ \alpha_k(\widehat{\theta}_n) &= \widehat{\alpha}_k. \end{aligned} \tag{9.4}$$

Formula (9.4) defines a system of k equations with k unknowns.

9.4 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Then $\alpha_1 = \mathbb{E}_p(X) = p$ and $\widehat{\alpha}_1 = n^{-1}\sum_{i=1}^n X_i$. By equating these we get the estimator

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i. \quad \blacksquare$$

9.5 Example. Let $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$. Then, $\alpha_1 = \mathbb{E}_\theta(X_1) = \mu$ and $\alpha_2 = \mathbb{E}_\theta(X_1^2) = \mathbb{V}_\theta(X_1) + (\mathbb{E}_\theta(X_1))^2 = \sigma^2 + \mu^2$. We need to solve the equations[1]

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n}\sum_{i=1}^n X_i \\ \widehat{\sigma}^2 + \widehat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^n X_i^2. \end{aligned}$$

This is a system of 2 equations with 2 unknowns. The solution is

$$\widehat{\mu} = \overline{X}_n$$

[1] Recall that $\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$. Hence, $\mathbb{E}(X^2) = \mathbb{V}(X) + (\mathbb{E}(X))^2$.

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

9.6 Theorem. *Let $\widehat{\theta}_n$ denote the method of moments estimator. Under appropriate conditions on the model, the following statements hold:*

1. *The estimate $\widehat{\theta}_n$ exists with probability tending to 1.*
2. *The estimate is consistent: $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$.*
3. *The estimate is asymptotically Normal:*

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N(0, \Sigma)$$

where

$$\Sigma = g\mathbb{E}_\theta(YY^T)g^T,$$

$Y = (X, X^2, \ldots, X^k)^T$, $g = (g_1, \ldots, g_k)$ and $g_j = \partial\alpha_j^{-1}(\theta)/\partial\theta$.

The last statement in the theorem above can be used to find standard errors and confidence intervals. However, there is an easier way: the bootstrap. We defer discussion of this until the end of the chapter.

9.3 Maximum Likelihood

The most common method for estimating parameters in a parametric model is the **maximum likelihood method**. Let $X_1, \ldots, X_n$ be IID with PDF $f(x;\theta)$.

9.7 Definition. *The* **likelihood function** *is defined by*

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta). \tag{9.5}$$

The **log-likelihood function** *is defined by $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$.*

The likelihood function is just the joint density of the data, except that we **treat it is a function of the parameter θ.** Thus, $\mathcal{L}_n : \Theta \to [0, \infty)$. The likelihood function is not a density function: in general, it is **not** true that $\mathcal{L}_n(\theta)$ integrates to 1 (with respect to θ).

9.8 Definition. *The* **maximum likelihood estimator** MLE*, denoted by $\widehat{\theta}_n$, is the value of θ that maximizes $\mathcal{L}_n(\theta)$.*

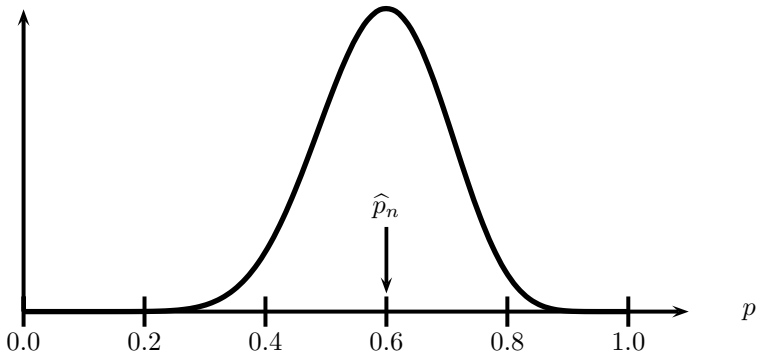


FIGURE 9.1. Likelihood function for Bernoulli with $n = 20$ and $\sum_{i=1}^n X_i = 12$. The MLE is $\widehat{p}_n = 12/20 = 0.6$.

The maximum of $\ell_n(\theta)$ occurs at the same place as the maximum of $\mathcal{L}_n(\theta)$, so maximizing the log-likelihood leads to the same answer as maximizing the likelihood. Often, it is easier to work with the log-likelihood.

9.9 Remark. If we multiply $\mathcal{L}_n(\theta)$ by any positive constant c (not depending on θ) then this will not change the MLE. Hence, we shall often drop constants in the likelihood function.

9.10 Example. Suppose that $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. The probability function is $f(x; p) = p^x(1-p)^{1-x}$ for $x = 0, 1$. The unknown parameter is p. Then,

$$\mathcal{L}_n(p) = \prod_{i=1}^n f(X_i; p) = \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

where $S = \sum_i X_i$. Hence,

$$\ell_n(p) = S \log p + (n - S) \log(1 - p).$$

Take the derivative of $\ell_n(p)$, set it equal to 0 to find that the MLE is $\widehat{p}_n = S/n$. See Figure 9.1. ■

9.11 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. The parameter is $\theta = (\mu, \sigma)$ and the likelihood function (ignoring some constants) is:

$$\begin{aligned} \mathcal{L}_n(\mu, \sigma) &= \prod_i \frac{1}{\sigma} \exp\left\{-\frac{1}{2\sigma^2}(X_i - \mu)^2\right\} \\ &= \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (X_i - \mu)^2\right\} \end{aligned}$$

$$= \quad \sigma^{-n} \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}\right\}$$

where $\overline{X} = n^{-1}\sum_i X_i$ is the sample mean and $S^2 = n^{-1}\sum_i (X_i - \overline{X})^2$. The last equality above follows from the fact that $\sum_i (X_i - \mu)^2 = nS^2 + n(\overline{X} - \mu)^2$ which can be verified by writing $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \overline{X} + \overline{X} - \mu)^2$ and then expanding the square. The log-likelihood is

$$\ell(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X}-\mu)^2}{2\sigma^2}.$$

Solving the equations

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = 0,$$

we conclude that $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma} = S$. It can be verified that these are indeed global maxima of the likelihood. ■

9.12 Example (A Hard Example)**.** Here is an example that many people find confusing. Let $X_1, \ldots, X_n \sim Unif(0, \theta)$. Recall that

$$f(x; \theta) = \begin{cases} 1/\theta & 0 \leq x \leq \theta \\ 0 & \text{otherwise.} \end{cases}$$

Consider a fixed value of θ. Suppose $\theta < X_i$ for some i. Then, $f(X_i; \theta) = 0$ and hence $\mathcal{L}_n(\theta) = \prod_i f(X_i; \theta) = 0$. It follows that $\mathcal{L}_n(\theta) = 0$ if any $X_i > \theta$. Therefore, $\mathcal{L}_n(\theta) = 0$ if $\theta < X_{(n)}$ where $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Now consider any $\theta \geq X_{(n)}$. For every X_i we then have that $f(X_i; \theta) = 1/\theta$ so that $\mathcal{L}_n(\theta) = \prod_i f(X_i; \theta) = \theta^{-n}$. In conclusion,

$$\mathcal{L}_n(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \theta \geq X_{(n)} \\ 0 & \theta < X_{(n)}. \end{cases}$$

See Figure 9.2. Now $\mathcal{L}_n(\theta)$ is strictly decreasing over the interval $[X_{(n)}, \infty)$. Hence, $\widehat{\theta}_n = X_{(n)}$. ■

The maximum likelihood estimators for the multivariate Normal and the multinomial can be found in Theorems 14.5 and 14.3.

9.4 Properties of Maximum Likelihood Estimators

Under certain conditions on the model, the maximum likelihood estimator $\widehat{\theta}_n$ possesses many properties that make it an appealing choice of estimator. The main properties of the MLE are:

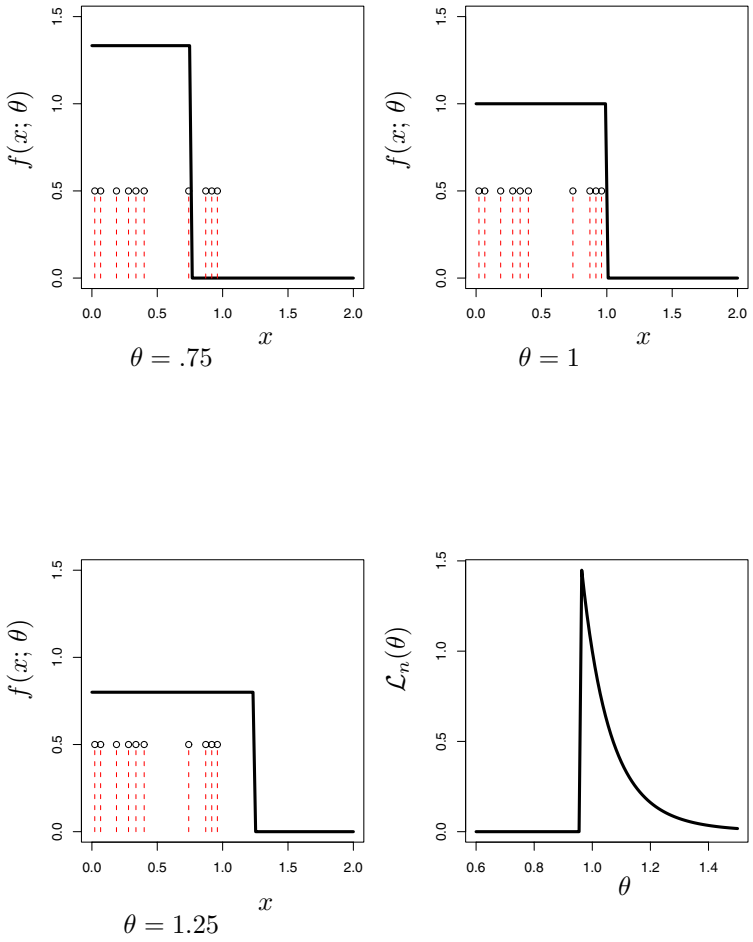


FIGURE 9.2. Likelihood function for Uniform $(0, \theta)$. The vertical lines show the observed data. The first three plots show $f(x;\theta)$ for three different values of θ. When $\theta < X_{(n)} = \max\{X_1, \ldots, X_n\}$, as in the first plot, $f(X_{(n)};\theta) = 0$ and hence $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i;\theta) = 0$. Otherwise $f(X_i;\,\theta) = 1/\theta$ for each i and hence $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i;\theta) = (1/\theta)^n$. The last plot shows the likelihood function.

1. The MLE is **consistent**: $\widehat{\theta}_n \xrightarrow{\text{P}} \theta_\star$ where $\theta_\star$ denotes the true value of the parameter θ;

2. The MLE is **equivariant**: if $\widehat{\theta}_n$ is the MLE of θ then $g(\widehat{\theta}_n)$ is the MLE of $g(\theta)$;

3. The MLE is **asymptotically Normal**: $(\widehat{\theta} - \theta_\star)/\widehat{\mathsf{se}} \rightsquigarrow N(0,1)$; also, the estimated standard error $\widehat{\mathsf{se}}$ can often be computed analytically;

4. The MLE is **asymptotically optimal** or **efficient**: roughly, this means that among all well-behaved estimators, the MLE has the smallest variance, at least for large samples;

5. The MLE is approximately the Bayes estimator. (This point will be explained later.)

We will spend some time explaining what these properties mean and why they are good things. In sufficiently complicated problems, these properties will no longer hold and the MLE will no longer be a good estimator. For now we focus on the simpler situations where the MLE works well. The properties we discuss only hold if the model satisfies certain **regularity conditions.** These are essentially smoothness conditions on $f(x;\theta)$. **Unless otherwise stated, we shall tacitly assume that these conditions hold.**

9.5 Consistency of Maximum Likelihood Estimators

Consistency means that the MLE converges in probability to the true value. To proceed, we need a definition. If f and g are PDF's, define the **Kullback-Leibler distance** [2] between f and g to be

$$D(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx. \tag{9.6}$$

It can be shown that $D(f,g) \geq 0$ and $D(f,f) = 0$. For any $\theta, \psi \in \Theta$ write $D(\theta,\psi)$ to mean $D(f(x;\theta), f(x;\psi))$.

We will say that the model $\mathfrak{F}$ is **identifiable** if $\theta \neq \psi$ implies that $D(\theta,\psi) > 0$. This means that different values of the parameter correspond to different distributions. We will assume from now on the the model is identifiable.

[2] This is not a distance in the formal sense because $D(f,g)$ is not symmetric.

Let $\theta_\star$ denote the true value of θ. Maximizing $\ell_n(\theta)$ is equivalent to maximizing

$$M_n(\theta) = \frac{1}{n}\sum_i \log \frac{f(X_i;\,\theta)}{f(X_i;\,\theta_\star)}.$$

This follows since $M_n(\theta) = n^{-1}(\ell_n(\theta) - \ell_n(\theta_\star))$ and $\ell_n(\theta_\star)$ is a constant (with respect to θ). By the law of large numbers, $M_n(\theta)$ converges to

$$\begin{aligned}
\mathbb{E}_{\theta_\star}\left(\log \frac{f(X_i;\,\theta)}{f(X_i;\,\theta_\star)}\right) &= \int \log\left(\frac{f(x;\,\theta)}{f(x;\,\theta_\star)}\right) f(x;\,\theta_\star)dx \\
&= -\int \log\left(\frac{f(x;\,\theta_\star)}{f(x;\,\theta)}\right) f(x;\,\theta_\star)dx \\
&= -D(\theta_\star,\theta).
\end{aligned}$$

Hence, $M_n(\theta) \approx -D(\theta_\star,\theta)$ which is maximized at $\theta_\star$ since $-D(\theta_\star,\theta_\star) = 0$ and $-D(\theta_\star,\theta) < 0$ for $\theta \neq \theta_\star$. Therefore, we expect that the maximizer will tend to $\theta_\star$. To prove this formally, we need more than $M_n(\theta) \xrightarrow{\text{P}} -D(\theta_\star,\theta)$. We need this convergence to be uniform over θ. We also have to make sure that the function $D(\theta_\star,\theta)$ is well behaved. Here are the formal details.

9.13 Theorem. *Let $\theta_\star$ denote the true value of θ. Define*

$$M_n(\theta) = \frac{1}{n}\sum_i \log \frac{f(X_i;\,\theta)}{f(X_i;\,\theta_\star)}$$

and $M(\theta) = -D(\theta_\star,\theta)$. Suppose that

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\text{P}} 0 \tag{9.7}$$

and that, for every $\epsilon > 0$,

$$\sup_{\theta:|\theta-\theta_\star|\geq\epsilon} M(\theta) < M(\theta_\star). \tag{9.8}$$

Let $\widehat{\theta}_n$ denote the MLE. *Then $\widehat{\theta}_n \xrightarrow{\text{P}} \theta_\star$.*

The proof is in the appendix.

9.6 Equivariance of the MLE

9.14 Theorem. *Let $\tau = g(\theta)$ be a function of θ. Let $\widehat{\theta}_n$ be the* MLE *of θ. Then $\widehat{\tau}_n = g(\widehat{\theta}_n)$ is the* MLE *of τ.*

PROOF. Let $h = g^{-1}$ denote the inverse of g. Then $\widehat{\theta}_n = h(\widehat{\tau}_n)$. For any τ, $\mathcal{L}(\tau) = \prod_i f(x_i; h(\tau)) = \prod_i f(x_i; \theta) = \mathcal{L}(\theta)$ where $\theta = h(\tau)$. Hence, for any τ, $\mathcal{L}_n(\tau) = \mathcal{L}(\theta) \leq \mathcal{L}(\widehat{\theta}) = \mathcal{L}_n(\widehat{\tau})$. ■

9.15 Example. Let $X_1, \ldots, X_n \sim N(\theta, 1)$. The MLEfor θ is $\widehat{\theta}_n = \overline{X}_n$. Let $\tau = e^\theta$. Then, the MLE for τ is $\widehat{\tau} = e^{\widehat{\theta}} = e^{\overline{X}}$. ■

9.7 Asymptotic Normality

It turns out that the distribution of $\widehat{\theta}_n$ is approximately Normal and we can compute its approximate variance analytically. To explore this, we first need a few definitions.

9.16 Definition. *The* **score function** *is defined to be*

$$s(X; \theta) = \frac{\partial \log f(X; \theta)}{\partial \theta}. \tag{9.9}$$

The **Fisher information** *is defined to be*

$$\begin{aligned} I_n(\theta) &= \mathbb{V}_\theta \left(\sum_{i=1}^n s(X_i; \theta) \right) \\ &= \sum_{i=1}^n \mathbb{V}_\theta \left(s(X_i; \theta) \right). \end{aligned} \tag{9.10}$$

For $n = 1$ we will sometimes write $I(\theta)$ instead of $I_1(\theta)$. It can be shown that $\mathbb{E}_\theta(s(X; \theta)) = 0$. It then follows that $\mathbb{V}_\theta(s(X; \theta)) = \mathbb{E}_\theta(s^2(X; \theta))$. In fact, a further simplification of $I_n(\theta)$ is given in the next result.

9.17 Theorem. *$I_n(\theta) = nI(\theta)$. Also,*

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta \left(\frac{\partial^2 \log f(X; \theta)}{\partial \theta^2} \right) \\ &= -\int \left(\frac{\partial^2 \log f(x; \theta)}{\partial \theta^2} \right) f(x;\, \theta) dx. \end{aligned} \tag{9.11}$$

9.18 Theorem (Asymptotic Normality of the MLE). *Let* $\mathsf{se} = \sqrt{\mathbb{V}(\widehat{\theta}_n)}$. *Under appropriate regularity conditions, the following hold:*

1. $\mathsf{se} \approx \sqrt{1/I_n(\theta)}$ *and*

$$\frac{(\widehat{\theta}_n - \theta)}{\mathsf{se}} \rightsquigarrow N(0,1). \tag{9.12}$$

2. *Let* $\widehat{\mathsf{se}} = \sqrt{1/I_n(\widehat{\theta}_n)}$. *Then,*

$$\frac{(\widehat{\theta}_n - \theta)}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1). \tag{9.13}$$

The proof is in the appendix. The first statement says that $\widehat{\theta}_n \approx N(\theta, \mathsf{se})$ where the approximate standard error of $\widehat{\theta}_n$ is $\mathsf{se} = \sqrt{1/I_n(\theta)}$. The second statement says that this is still true even if we replace the standard error by its estimated standard error $\widehat{\mathsf{se}} = \sqrt{1/I_n(\widehat{\theta}_n)}$.

Informally, the theorem says that the distribution of the MLE can be approximated with $N(\theta, \widehat{\mathsf{se}}^2)$. From this fact we can construct an (asymptotic) confidence interval.

9.19 Theorem. *Let*

$$C_n = \left(\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}},\ \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}\right).$$

Then, $\mathbb{P}_\theta(\theta \in C_n) \to 1-\alpha$ *as* $n \to \infty$.

PROOF. Let Z denote a standard normal random variable. Then,

$$\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}_\theta\left(\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}} \le \theta \le \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}\right) \\
&= \mathbb{P}_\theta\left(-z_{\alpha/2} \le \frac{\widehat{\theta}_n - \theta}{\widehat{\mathsf{se}}} \le z_{\alpha/2}\right) \\
&\to \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha. \quad \blacksquare
\end{aligned}$$

For $\alpha = .05$, $z_{\alpha/2} = 1.96 \approx 2$, so:

$$\widehat{\theta}_n \pm 2\ \widehat{\mathsf{se}} \tag{9.14}$$

is an approximate 95 percent confidence interval.

When you read an opinion poll in the newspaper, you often see a statement like: the poll is accurate to within one point, 95 percent of the time. They are simply giving a 95 percent confidence interval of the form $\widehat{\theta}_n \pm 2\,\widehat{\mathsf{se}}$.

9.20 Example. Let $X_1, \ldots, X_n \sim$ Bernoulli(p). The MLE is $\widehat{p}_n = \sum_i X_i/n$ and $f(x;p) = p^x(1-p)^{1-x}$, $\log f(x;p) = x \log p + (1-x)\log(1-p)$,

$$s(X;p) = \frac{X}{p} - \frac{1-X}{1-p},$$

and

$$-s'(X;p) = \frac{X}{p^2} + \frac{1-X}{(1-p)^2}.$$

Thus,

$$I(p) = \mathbb{E}_p(-s'(X;p)) = \frac{p}{p^2} + \frac{(1-p)}{(1-p)^2} = \frac{1}{p(1-p)}.$$

Hence,

$$\widehat{\mathsf{se}} = \frac{1}{\sqrt{I_n(\widehat{p}_n)}} = \frac{1}{\sqrt{nI(\widehat{p}_n)}} = \left\{ \frac{\widehat{p}(1-\widehat{p})}{n} \right\}^{1/2}.$$

An approximate 95 percent confidence interval is

$$\widehat{p}_n \pm 2 \left\{ \frac{\widehat{p}_n(1-\widehat{p}_n)}{n} \right\}^{1/2}. \quad \blacksquare$$

9.21 Example. Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ where σ^2 is known. The score function is $s(X;\theta) = (X-\theta)/\sigma^2$ and $s'(X;\theta) = -1/\sigma^2$ so that $I_1(\theta) = 1/\sigma^2$. The MLE is $\widehat{\theta}_n = \overline{X}_n$. According to Theorem 9.18, $\overline{X}_n \approx N(\theta, \sigma^2/n)$. In this case, the Normal approximation is actually exact. $\blacksquare$

9.22 Example. Let $X_1, \ldots, X_n \sim$ Poisson(λ). Then $\widehat{\lambda}_n = \overline{X}_n$ and some calculations show that $I_1(\lambda) = 1/\lambda$, so

$$\widehat{\mathsf{se}} = \frac{1}{\sqrt{nI(\widehat{\lambda}_n)}} = \sqrt{\frac{\widehat{\lambda}_n}{n}}.$$

Therefore, an approximate $1-\alpha$ confidence interval for λ is $\widehat{\lambda}_n \pm z_{\alpha/2}\sqrt{\widehat{\lambda}_n/n}$. $\blacksquare$

9.8 Optimality

Suppose that $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. The MLE is $\widehat{\theta}_n = \overline{X}_n$. Another reasonable estimator of θ is the sample median $\widetilde{\theta}_n$. The MLE satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N(0, \sigma^2).$$

It can be proved that the median satisfies

$$\sqrt{n}(\widetilde{\theta}_n - \theta) \rightsquigarrow N\left(0, \sigma^2 \frac{\pi}{2}\right).$$

This means that the median converges to the right value but has a larger variance than the MLE.

More generally, consider two estimators T_n and U_n and suppose that

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, t^2),$$

and that

$$\sqrt{n}(U_n - \theta) \rightsquigarrow N(0, u^2).$$

We define the asymptotic relative efficiency of U to T by $\text{ARE}(U, T) = t^2/u^2$. In the Normal example, $\text{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) = 2/\pi = .63$. The interpretation is that if you use the median, you are effectively using only a fraction of the data.

9.23 Theorem. *If $\widehat{\theta}_n$ is the* MLE *and $\widetilde{\theta}_n$ is any other estimator then* [3]

$$\text{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) \leq 1.$$

Thus, the MLE *has the smallest (asymptotic) variance and we say that the* MLE *is* **efficient** *or* **asymptotically optimal.**

This result is predicated upon the assumed model being correct. If the model is wrong, the MLE may no longer be optimal. We will discuss optimality in more generality when we discuss decision theory in Chapter 12.

9.9 The Delta Method

Let $\tau = g(\theta)$ where g is a smooth function. The maximum likelihood estimator of τ is $\widehat{\tau} = g(\widehat{\theta})$. Now we address the following question: what is the distribution of $\widehat{\tau}$?

9.24 Theorem (The Delta Method). *If $\tau = g(\theta)$ where g is differentiable and $g'(\theta) \neq 0$ then*

$$\frac{(\widehat{\tau}_n - \tau)}{\widehat{\mathsf{se}}(\widehat{\tau})} \rightsquigarrow N(0, 1) \tag{9.15}$$

[3] The result is actually more subtle than this but the details are too complicated to consider here.

where $\widehat{\tau}_n = g(\widehat{\theta}_n)$ *and*

$$\widehat{\mathsf{se}}(\widehat{\tau}_n) = |g'(\widehat{\theta})|\, \widehat{\mathsf{se}}(\widehat{\theta}_n) \tag{9.16}$$

Hence, if

$$C_n = \left(\widehat{\tau}_n - z_{\alpha/2}\, \widehat{\mathsf{se}}(\widehat{\tau}_n),\ \widehat{\tau}_n + z_{\alpha/2}\, \widehat{\mathsf{se}}(\widehat{\tau}_n) \right) \tag{9.17}$$

then $\mathbb{P}_\theta(\tau \in C_n) \to 1 - \alpha$ *as* $n \to \infty$.

9.25 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\psi = g(p) = \log(p/(1-p))$. The Fisher information function is $I(p) = 1/(p(1-p))$ so the estimated standard error of the MLE $\widehat{p}_n$ is

$$\widehat{\mathsf{se}} = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

The MLE of ψ is $\widehat{\psi} = \log \widehat{p}/(1-\widehat{p})$. Since, $g'(p) = 1/(p(1-p))$, according to the delta method

$$\widehat{\mathsf{se}}(\widehat{\psi}_n) = |g'(\widehat{p}_n)| \widehat{\mathsf{se}}(\widehat{p}_n) = \frac{1}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}.$$

An approximate 95 percent confidence interval is

$$\widehat{\psi}_n \pm \frac{2}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}. \quad \blacksquare$$

9.26 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Suppose that μ is known, σ is unknown and that we want to estimate $\psi = \log \sigma$. The log-likelihood is $\ell(\sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2$. Differentiate and set equal to 0 and conclude that

$$\widehat{\sigma}_n = \sqrt{\frac{\sum_i (X_i - \mu)^2}{n}}.$$

To get the standard error we need the Fisher information. First,

$$\log f(X; \sigma) = -\log \sigma - \frac{(X-\mu)^2}{2\sigma^2}$$

with second derivative

$$\frac{1}{\sigma^2} - \frac{3(X-\mu)^2}{\sigma^4},$$

and hence

$$I(\sigma) = -\frac{1}{\sigma^2} + \frac{3\sigma^2}{\sigma^4} = \frac{2}{\sigma^2}.$$

Therefore, $\widehat{\mathsf{se}} = \widehat{\sigma}_n/\sqrt{2n}$. Let $\psi = g(\sigma) = \log\sigma$. Then, $\widehat{\psi}_n = \log\widehat{\sigma}_n$. Since $g' = 1/\sigma$,

$$\widehat{\mathsf{se}}(\widehat{\psi}_n) = \frac{1}{\widehat{\sigma}_n}\frac{\widehat{\sigma}_n}{\sqrt{2n}} = \frac{1}{\sqrt{2n}},$$

and an approximate 95 percent confidence interval is $\widehat{\psi}_n \pm 2/\sqrt{2n}$. ∎

9.10 Multiparameter Models

These ideas can directly be extended to models with several parameters. Let $\theta = (\theta_1, \ldots, \theta_k)$ and let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ be the MLE. Let $\ell_n = \sum_{i=1}^n \log f(X_i; \theta)$,

$$H_{jj} = \frac{\partial^2 \ell_n}{\partial \theta_j^2} \quad \text{and} \quad H_{jk} = \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta_k}.$$

Define the **Fisher Information Matrix** by

$$I_n(\theta) = -\begin{bmatrix} \mathbb{E}_\theta(H_{11}) & \mathbb{E}_\theta(H_{12}) & \cdots & \mathbb{E}_\theta(H_{1k}) \\ \mathbb{E}_\theta(H_{21}) & \mathbb{E}_\theta(H_{22}) & \cdots & \mathbb{E}_\theta(H_{2k}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{E}_\theta(H_{k1}) & \mathbb{E}_\theta(H_{k2}) & \cdots & \mathbb{E}_\theta(H_{kk}) \end{bmatrix}. \tag{9.18}$$

Let $J_n(\theta) = I_n^{-1}(\theta)$ be the inverse of I_n.

9.27 Theorem. *Under appropriate regularity conditions,*

$$(\widehat{\theta} - \theta) \approx N(0, J_n).$$

Also, if $\widehat{\theta}_j$ is the j^{th} component of $\widehat{\theta}$, then

$$\frac{(\widehat{\theta}_j - \theta_j)}{\widehat{\mathsf{se}}_j} \rightsquigarrow N(0,1) \tag{9.19}$$

where $\widehat{\mathsf{se}}_j^2 = J_n(j,j)$ is the j^{th} diagonal element of J_n. The approximate covariance of $\widehat{\theta}_j$ and $\widehat{\theta}_k$ is $\mathsf{Cov}(\widehat{\theta}_j, \widehat{\theta}_k) \approx J_n(j,k)$.

There is also a multiparameter delta method. Let $\tau = g(\theta_1, \ldots, \theta_k)$ be a function and let

$$\nabla g = \begin{pmatrix} \frac{\partial g}{\partial \theta_1} \\ \vdots \\ \frac{\partial g}{\partial \theta_k} \end{pmatrix}$$

be the gradient of g.

9.28 Theorem (Multiparameter delta method). *Suppose that ∇g evaluated at $\widehat{\theta}$ is not 0. Let $\widehat{\tau} = g(\widehat{\theta})$. Then*

$$\frac{(\widehat{\tau} - \tau)}{\widehat{\mathsf{se}}(\widehat{\tau})} \rightsquigarrow N(0,1)$$

where

$$\widehat{\mathsf{se}}(\widehat{\tau}) = \sqrt{(\widehat{\nabla} g)^T \widehat{J}_n (\widehat{\nabla} g)}, \tag{9.20}$$

$\widehat{J}_n = J_n(\widehat{\theta}_n)$ *and* $\widehat{\nabla} g$ *is* ∇g *evaluated at* $\theta = \widehat{\theta}$.

9.29 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Let $\tau = g(\mu, \sigma) = \sigma/\mu$. In Excercise 8 you will show that

$$I_n(\mu, \sigma) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix}.$$

Hence,

$$J_n = I_n^{-1}(\mu, \sigma) = \frac{1}{n} \begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{bmatrix}.$$

The gradient of g is

$$\nabla g = \begin{pmatrix} -\frac{\sigma}{\mu^2} \\ \frac{1}{\mu} \end{pmatrix}.$$

Thus,

$$\widehat{\mathsf{se}}(\widehat{\tau}) = \sqrt{(\widehat{\nabla} g)^T \widehat{J}_n (\widehat{\nabla} g)} = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{\widehat{\mu}^4} + \frac{\widehat{\sigma}^2}{2\widehat{\mu}^2}}. \quad \blacksquare$$

9.11 The Parametric Bootstrap

For parametric models, standard errors and confidence intervals may also be estimated using the bootstrap. There is only one change. In the nonparametric bootstrap, we sampled $X_1^*, \ldots, X_n^*$ from the empirical distribution $\widehat{F}_n$. In the parametric bootstrap we sample instead from $f(x;\ \widehat{\theta}_n)$. Here, $\widehat{\theta}_n$ could be the MLE or the method of moments estimator.

9.30 Example. Consider example 9.29. To get the bootstrap standard error, simulate $X_1, \ldots, X_n^* \sim N(\widehat{\mu}, \widehat{\sigma}^2)$, compute $\widehat{\mu}^* = n^{-1} \sum_i X_i^*$ and $\widehat{\sigma}^{2*} = n^{-1} \sum_i (X_i^* - \widehat{\mu}^*)^2$. Then compute $\widehat{\tau}^* = g(\widehat{\mu}^*, \widehat{\sigma}^*) = \widehat{\sigma}^*/\widehat{\mu}^*$. Repeating this B times yields bootstrap replications

$$\widehat{\tau}_1^*, \ldots, \widehat{\tau}_B^*$$

and the estimated standard error is

$$\widehat{\mathsf{se}}_{\text{boot}} = \sqrt{\frac{\sum_{b=1}^{B}(\widehat{\tau}_b^* - \widehat{\tau})^2}{B}}. \quad \blacksquare$$

The bootstrap is much easier than the delta method. On the other hand, the delta method has the advantage that it gives a closed form expression for the standard error.

9.12 Checking Assumptions

If we assume the data come from a parametric model, then it is a good idea to check that assumption. One possibility is to check the assumptions informally by inspecting plots of the data. For example, if a histogram of the data looks very bimodal, then the assumption of Normality might be questionable. A formal way to test a parametric model is to use a **goodness-of-fit test.** See Section 10.8.

9.13 Appendix

9.13.1 Proofs

PROOF OF THEOREM 9.13. Since $\widehat{\theta}_n$ maximizes $M_n(\theta)$, we have $M_n(\widehat{\theta}_n) \geq M_n(\theta_\star)$. Hence,

$$\begin{aligned} M(\theta_\star) - M(\widehat{\theta}_n) &= M_n(\theta_\star) - M(\widehat{\theta}_n) + M(\theta_\star) - M_n(\theta_\star) \\ &\leq M_n(\widehat{\theta}_n) - M(\widehat{\theta}_n) + M(\theta_\star) - M_n(\theta_\star) \\ &\leq \sup_\theta |M_n(\theta) - M(\theta)| + M(\theta_\star) - M_n(\theta_\star) \\ &\xrightarrow{\text{P}} 0 \end{aligned}$$

where the last line follows from (9.7). It follows that, for any $\delta > 0$,

$$\mathbb{P}\left(M(\widehat{\theta}_n) < M(\theta_\star) - \delta\right) \to 0.$$

Pick any $\epsilon > 0$. By (9.8), there exists $\delta > 0$ such that $|\theta - \theta_\star| \geq \epsilon$ implies that $M(\theta) < M(\theta_\star) - \delta$. Hence,

$$\mathbb{P}(|\widehat{\theta}_n - \theta_\star| > \epsilon) \leq \mathbb{P}\left(M(\widehat{\theta}_n) < M(\theta_\star) - \delta\right) \to 0. \quad \blacksquare$$

Next we want to prove Theorem 9.18. First we need a lemma.

9.31 Lemma. *The score function satisfies*

$$\mathbb{E}_\theta\left[s(X;\theta)\right] = 0.$$

PROOF. Note that $1 = \int f(x;\theta)dx$. Differentiate both sides of this equation to conclude that

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int f(x;\theta)dx = \int \frac{\partial}{\partial\theta} f(x;\theta)dx \\ &= \int \frac{\frac{\partial f(x;\theta)}{\partial\theta}}{f(x;\theta)} f(x;\theta)dx = \int \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx \\ &= \int s(x;\theta)f(x;\theta)dx = \mathbb{E}_\theta s(X;\theta). \quad \blacksquare \end{aligned}$$

PROOF OF THEOREM 9.18. Let $\ell(\theta) = \log \mathcal{L}(\theta)$. Then,

$$0 = \ell'(\widehat{\theta}) \approx \ell'(\theta) + (\widehat{\theta} - \theta)\ell''(\theta).$$

Rearrange the above equation to get $\widehat{\theta} - \theta = -\ell'(\theta)/\ell''(\theta)$ or, in other words,

$$\sqrt{n}(\widehat{\theta} - \theta) = \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)} \equiv \frac{\text{TOP}}{\text{BOTTOM}}.$$

Let $Y_i = \partial \log f(X_i;\theta)/\partial\theta$. Recall that $\mathbb{E}(Y_i) = 0$ from the previous lemma and also $\mathbb{V}(Y_i) = I(\theta)$. Hence,

$$\text{TOP} = n^{-1/2}\sum_i Y_i = \sqrt{n}\overline{Y} = \sqrt{n}(\overline{Y} - 0) \rightsquigarrow W \sim N(0, I(\theta))$$

by the central limit theorem. Let $A_i = -\partial^2 \log f(X_i;\theta)/\partial\theta^2$. Then $\mathbb{E}(A_i) = I(\theta)$ and

$$\text{BOTTOM} = \overline{A} \xrightarrow{\text{P}} I(\theta)$$

by the law of large numbers. Apply Theorem 5.5 part (e), to conclude that

$$\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow \frac{W}{I(\theta)} \stackrel{d}{=} N\left(0, \frac{1}{I(\theta)}\right).$$

Assuming that $I(\theta)$ is a continuous function of θ, it follows that $I(\widehat{\theta}_n) \xrightarrow{\text{P}} I(\theta)$. Now

$$\begin{aligned} \frac{\widehat{\theta}_n - \theta}{\widehat{\mathsf{se}}} &= \sqrt{n}I^{1/2}(\widehat{\theta}_n)(\widehat{\theta}_n - \theta) \\ &= \left\{\sqrt{n}I^{1/2}(\theta)(\widehat{\theta}_n - \theta)\right\}\sqrt{\frac{I(\widehat{\theta}_n)}{I(\theta)}}. \end{aligned}$$

The first term tends in distribution to N(0,1). The second term tends in probability to 1. The result follows from Theorem 5.5 part (e). ■

Outline of Proof of Theorem 9.24. Write

$$\widehat{\tau}_n = g(\widehat{\theta}_n) \approx g(\theta) + (\widehat{\theta}_n - \theta)g'(\theta) = \tau + (\widehat{\theta}_n - \theta)g'(\theta).$$

Thus,

$$\sqrt{n}(\widehat{\tau}_n - \tau) \approx \sqrt{n}(\widehat{\theta}_n - \theta)g'(\theta),$$

and hence

$$\frac{\sqrt{nI(\theta)}(\widehat{\tau}_n - \tau)}{g'(\theta)} \approx \sqrt{nI(\theta)}(\widehat{\theta}_n - \theta).$$

Theorem 9.18 tells us that the right-hand side tends in distribution to a N(0,1). Hence,

$$\frac{\sqrt{nI(\theta)}(\widehat{\tau}_n - \tau)}{g'(\theta)} \rightsquigarrow N(0,1)$$

or, in other words,

$$\widehat{\tau}_n \approx N\left(\tau, \mathsf{se}^2(\widehat{\tau}_n)\right),$$

where

$$\mathsf{se}^2(\widehat{\tau}_n) = \frac{(g'(\theta))^2}{nI(\theta)}.$$

The result remains true if we substitute $\widehat{\theta}_n$ for θ by Theorem 5.5 part (e). ■

9.13.2 Sufficiency

A **statistic** is a function $T(X^n)$ of the data. A sufficient statistic is a statistic that contains all the information in the data. To make this more formal, we need some definitions.

9.32 Definition. *Write $x^n \leftrightarrow y^n$ if $f(x^n;\theta) = c\, f(y^n;\theta)$ for some constant c that might depend on x^n and y^n but not θ. A statistic $T(x^n)$ is* **sufficient** *if $T(x^n) \leftrightarrow T(y^n)$ implies that $x^n \leftrightarrow y^n$.*

Notice that if $x^n \leftrightarrow y^n$, then the likelihood function based on x^n has the same shape as the likelihood function based on y^n. Roughly speaking, a statistic is sufficient if we can calculate the likelihood function knowing only $T(X^n)$.

9.33 Example. Let $X_1, \ldots, X_n \sim$ Bernoulli(p). Then $\mathcal{L}(p) = p^S(1-p)^{n-S}$ where $S = \sum_i X_i$, so S is sufficient. ■

9.34 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ and let $T = (\overline{X}, S)$. Then

$$f(X^n; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}\right\}$$

where S^2 is the sample variance. The last expression depends on the data only through T and therefore, $T = (\overline{X}, S)$ is a sufficient statistic. Note that $U = (17\,\overline{X}, S)$ is also a sufficient statistic. If I tell you the value of U then you can easily figure out T and then compute the likelihood. Sufficient statistics are far from unique. Consider the following statistics for the $N(\mu, \sigma^2)$ model:

$$\begin{aligned} T_1(X^n) &= (X_1, \ldots, X_n) \\ T_2(X^n) &= (\overline{X}, S) \\ T_3(X^n) &= \overline{X} \\ T_4(X^n) &= (\overline{X}, S, X_3). \end{aligned}$$

The first statistic is just the whole data set. This is sufficient. The second is also sufficient as we proved above. The third is not sufficient: you can't compute $\mathcal{L}(\mu, \sigma)$ if I only tell you $\overline{X}$. The fourth statistic T_4 is sufficient. The statistics T_1 and T_4 are sufficient but they contain redundant information. Intuitively, there is a sense in which T_2 is a "more concise" sufficient statistic than either T_1 or T_4. We can express this formally by noting that T_2 is a function of T_1 and similarly, T_2 is a function of T_4. For example, $T_2 = g(T_4)$ where $g(a_1, a_2, a_3) = (a_1, a_2)$. ■

9.35 Definition. *A statistic T is* **minimal sufficient** *if (i) it is sufficient; and (ii) it is a function of every other sufficient statistic.*

9.36 Theorem. *T is minimal sufficient if the following is true:*

$$T(x^n) = T(y^n) \text{ if and only if } x^n \leftrightarrow y^n.$$

A statistic induces a partition on the set of outcomes. We can think of sufficiency in terms of these partitions.

9.37 Example. Let $X_1, X_2 \sim Bernoulli(\theta)$. Let $V = X_1$, $T = \sum_i X_i$ and $U = (T, X_1)$. Here is the set of outcomes and the statistics:

X_1	X_2	V	T	U
0	0	0	0	(0,0)
0	1	0	1	(1,0)
1	0	1	1	(1,1)
1	1	1	2	(2,1)

The partitions induced by these statistics are:

$$\begin{aligned} V &\longrightarrow \{(0,0),(0,1)\},\ \{(1,0),(1,1)\} \\ T &\longrightarrow \{(0,0)\},\ \{(0,1),(1,0)\},\ \{(1,1)\} \\ U &\longrightarrow \{(0,0)\},\ \{(0,1)\},\ \{(1,0)\},\ \{(1,1)\}. \end{aligned}$$

Then V is not sufficient but T and U are sufficient. T is minimal sufficient; U is not minimal since if $x^n = (1,0)$ and $y^n = (0,1)$, then $x^n \nleftrightarrow y^n$ yet $U(x^n) \neq U(y^n)$. The statistic $W = 17\,T$ generates the same partition as T. It is also minimal sufficient. ■

9.38 Example. For a $N(\mu, \sigma^2)$ model, $T = (\overline{X}, S)$ is a minimal sufficient statistic. For the Bernoulli model, $T = \sum_i X_i$ is a minimal sufficient statistic. For the Poisson model, $T = \sum_i X_i$ is a minimal sufficient statistic. Check that $T = (\sum_i X_i, X_1)$ is sufficient but not minimal sufficient. Check that $T = X_1$ is not sufficient. ■

I did not give the usual definition of sufficiency. The usual definition is this: T is sufficient if the distribution of X^n given $T(X^n) = t$ does not depend on θ. In other words, T is sufficient if $f(x_1, \ldots, x_n | t; \theta) = h(x_1, \ldots, x_n, t)$ where h is some function that does not depend on θ.

9.39 Example. Two coin flips. Let $X = (X_1, X_2) \sim \text{Bernoulli}(p)$. Then $T = X_1 + X_2$ is sufficient. To see this, we need the distribution of (X_1, X_2) given $T = t$. Since T can take 3 possible values, there are 3 conditional distributions to check. They are: (i) the distribution of (X_1, X_2) given $T = 0$:

$$P(X_1 = 0, X_2 = 0 | t = 0) = 1, P(X_1 = 0, X_2 = 1 | t = 0) = 0,$$

$$P(X_1 = 1, X_2 = 0 | t = 0) = 0, P(X_1 = 1, X_2 = 1 | t = 0) = 0;$$

(ii) the distribution of (X_1, X_2) given $T = 1$:

$$P(X_1 = 0, X_2 = 0 | t = 1) = 0, P(X_1 = 0, X_2 = 1 | t = 1) = \frac{1}{2},$$

$$P(X_1 = 1, X_2 = 0 | t = 1) = \frac{1}{2}, P(X_1 = 1, X_2 = 1 | t = 1) = 0; \text{ and}$$

(iii) the distribution of (X_1, X_2) given $T = 2$:

$$P(X_1 = 0, X_2 = 0 | t = 2) = 0, P(X_1 = 0, X_2 = 1 | t = 2) = 0,$$

$$P(X_1 = 1, X_2 = 0 | t = 2) = 0, P(X_1 = 1, X_2 = 1 | t = 2) = 1.$$

None of these depend on the parameter p. Thus, the distribution of $X_1, X_2 | T$ does not depend on θ, so T is sufficient. ■

9.40 Theorem (Factorization Theorem). *T is sufficient if and only if there are functions $g(t,\theta)$ and $h(x)$ such that $f(x^n;\theta) = g(t(x^n),\theta)h(x^n)$.*

9.41 Example. Return to the two coin flips. Let $t = x_1 + x_2$. Then

$$\begin{aligned} f(x_1,x_2;\theta) &= f(x_1;\theta)f(x_2;\theta) \\ &= \theta^{x_1}(1-\theta)^{1-x_1}\theta^{x_2}(1-\theta)^{1-x_2} \\ &= g(t,\theta)h(x_1,x_2) \end{aligned}$$

where $g(t,\theta) = \theta^t(1-\theta)^{2-t}$ and $h(x_1,x_2) = 1$. Therefore, $T = X_1 + X_2$ is sufficient. ■

Now we discuss an implication of sufficiency in point estimation. Let $\widehat{\theta}$ be an estimator of θ. The Rao-Blackwell theorem says that an estimator should only depend on the sufficient statistic, otherwise it can be improved. Let $R(\theta,\widehat{\theta}) = \mathbb{E}_\theta(\theta - \widehat{\theta})^2$ denote the MSE of the estimator.

9.42 Theorem (Rao-Blackwell). *Let $\widehat{\theta}$ be an estimator and let T be a sufficient statistic. Define a new estimator by*

$$\widetilde{\theta} = \mathbb{E}(\widehat{\theta}|T).$$

Then, for every θ, $R(\theta,\widetilde{\theta}) \leq R(\theta,\widehat{\theta})$.

9.43 Example. Consider flipping a coin twice. Let $\widehat{\theta} = X_1$. This is a well-defined (and unbiased) estimator. But it is not a function of the sufficient statistic $T = X_1 + X_2$. However, note that $\widetilde{\theta} = \mathbb{E}(X_1|T) = (X_1 + X_2)/2$. By the Rao-Blackwell Theorem, $\widetilde{\theta}$ has MSE at least as small as $\widehat{\theta} = X_1$. The same applies with n coin flips. Again define $\widehat{\theta} = X_1$ and $T = \sum_i X_i$. Then $\widetilde{\theta} = \mathbb{E}(X_1|T) = n^{-1}\sum_i X_i$ has improved MSE. ■

9.13.3 Exponential Families

Most of the parametric models we have studied so far are special cases of a general class of models called exponential families. We say that $\{f(x;\theta) : \theta \in \Theta\}$ is a **one-parameter exponential family** if there are functions $\eta(\theta)$, $B(\theta)$, $T(x)$ and $h(x)$ such that

$$f(x;\theta) = h(x)e^{\eta(\theta)T(x) - B(\theta)}.$$

It is easy to see that $T(X)$ is sufficient. We call T the **natural sufficient statistic**.

9.44 Example. Let $X \sim \text{Poisson}(\theta)$. Then

$$f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!} = \frac{1}{x!} e^{x \log \theta - \theta}$$

and hence, this is an exponential family with $\eta(\theta) = \log\theta$, $B(\theta) = \theta$, $T(x) = x$, $h(x) = 1/x!$. ■

9.45 Example. Let $X \sim \text{Binomial}(n, \theta)$. Then

$$f(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{n}{x}\exp\left\{x\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\}.$$

In this case,

$$\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right), B(\theta) = -n\log(\theta)$$

and

$$T(x) = x, h(x) = \binom{n}{x}.$$

■

We can rewrite an exponential family as

$$f(x;\eta) = h(x)e^{\eta T(x) - A(\eta)}$$

where $\eta = \eta(\theta)$ is called the **natural parameter** and

$$A(\eta) = \log \int h(x) e^{\eta T(x)} dx.$$

For example a Poisson can be written as $f(x;\eta) = e^{\eta x - e^\eta}/x!$ where the natural parameter is $\eta = \log\theta$.

Let $X_1, \ldots, X_n$ be IID from an exponential family. Then $f(x^n;\theta)$ is an exponential family:

$$f(x^n;\theta) = h_n(x^n)h_n(x^n)e^{\eta(\theta)T_n(x^n) - B_n(\theta)}$$

where $h_n(x^n) = \prod_i h(x_i)$, $T_n(x^n) = \sum_i T(x_i)$ and $B_n(\theta) = nB(\theta)$. This implies that $\sum_i T(X_i)$ is sufficient.

9.46 Example. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. Then

$$f(x^n;\theta) = \frac{1}{\theta^n} I(x_{(n)} \leq \theta)$$

where I is 1 if the term inside the brackets is true and 0 otherwise, and $x_{(n)} = max\{x_1, \ldots, x_n\}$. Thus $T(X^n) = max\{X_1, \ldots, X_n\}$ is sufficient. But since $T(X^n) \neq \sum_i T(X_i)$, this cannot be an exponential family. ■

9.47 Theorem. *Let X have density in an exponential family. Then,*

$$\mathbb{E}(T(X)) = A'(\eta),\ \mathbb{V}(T(X)) = A''(\eta).$$

If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector, then we say that $f(x;\theta)$ has exponential family form if

$$f(x;\theta) = h(x) \exp\left\{ \sum_{j=1}^{k} \eta_j(\theta) T_j(x) - B(\theta) \right\}.$$

Again, $T = (T_1, \ldots, T_k)$ is sufficient. An IID sample of size n also has exponential form with sufficient statistic $(\sum_i T_1(X_i), \ldots, \sum_i T_k(X_i))$.

9.48 Example. Consider the normal family with $\theta = (\mu, \sigma)$. Now,

$$f(x;\theta) = \exp\left\{ \frac{\mu}{\sigma^2} x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right\}.$$

This is exponential with

$$\eta_1(\theta) = \frac{\mu}{\sigma^2},\ T_1(x) = x$$

$$\eta_2(\theta) = -\frac{1}{2\sigma^2},\ T_2(x) = x^2$$

$$B(\theta) = \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right),\ h(x) = 1.$$

Hence, with n IID samples, $(\sum_i X_i, \sum_i X_i^2)$ is sufficient. ∎

As before we can write an exponential family as

$$f(x;\eta) = h(x) \exp\left\{ T^T(x)\eta - A(\eta) \right\},$$

where $A(\eta) = \log \int h(x) e^{T^T(x)\eta} dx$. It can be shown that

$$\mathbb{E}(T(X)) = \dot{A}(\eta)\ \ \mathbb{V}(T(X)) = \ddot{A}(\eta),$$

where the first expression is the vector of partial derivatives and the second is the matrix of second derivatives.

9.13.4 Computing Maximum Likelihood Estimates

In some cases we can find the MLE $\widehat{\theta}$ analytically. More often, we need to find the MLE by numerical methods. We will briefly discuss two commonly

used methods: (i) Newton-Raphson, and (ii) the EM algorithm. Both are iterative methods that produce a sequence of values $\theta^0, \theta^1, \ldots$ that, under ideal conditions, converge to the MLE $\widehat{\theta}$. In each case, it is helpful to use a good starting value θ^0. Often, the method of moments estimator is a good starting value.

NEWTON-RAPHSON. To motivate Newton-Raphson, let's expand the derivative of the log-likelihood around θ^j:

$$0 = \ell'(\widehat{\theta}) \approx \ell'(\theta^j) + (\widehat{\theta} - \theta^j)\ell''(\theta^j).$$

Solving for $\widehat{\theta}$ gives

$$\widehat{\theta} \approx \theta^j - \frac{\ell'(\theta^j)}{\ell''(\theta^j)}.$$

This suggests the following iterative scheme:

$$\widehat{\theta}^{j+1} = \theta^j - \frac{\ell'(\theta^j)}{\ell''(\theta^j)}.$$

In the multiparameter case, the mle $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ is a vector and the method becomes

$$\widehat{\theta}^{j+1} = \theta^j - H^{-1}\ell'(\theta^j)$$

where $\ell'(\theta^j)$ is the vector of first derivatives and H is the matrix of second derivatives of the log-likelihood.

THE EM ALGORITHM. The letters EM stand for Expectation-Maximization. The idea is to iterate between taking an expectation then maximizing. Suppose we have data Y whose density $f(y;\theta)$ leads to a log-likelihood that is hard to maximize. But suppose we can find another random variable Z such that $f(y;\theta) = \int f(y,z;\theta)\,dz$ and such that the likelihood based on $f(y,z;\theta)$ is easy to maximize. In other words, the model of interest is the marginal of a model with a simpler likelihood. In this case, we call Y the observed data and Z the hidden (or latent or missing) data. If we could just "fill in" the missing data, we would have an easy problem. Conceptually, the EM algorithm works by filling in the missing data, maximizing the log-likelihood, and iterating.

9.49 Example (Mixture of Normals)**.** Sometimes it is reasonable to assume that the distribution of the data is a mixture of two normals. Think of heights of people being a mixture of men and women's heights. Let $\phi(y;\mu,\sigma)$ denote a normal density with mean μ and standard deviation σ. The density of a mixture of two Normals is

$$f(y;\theta) = (1-p)\phi(y;\mu_0,\sigma_0) + p\phi(y;\mu_1,\sigma_1).$$

The idea is that an observation is drawn from the first normal with probability p and the second with probability $1-p$. However, we don't know which Normal it was drawn from. The parameters are $\theta = (\mu_0, \sigma_0, \mu_1, \sigma_1, p)$. The likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left[(1-p)\phi(y_i; \mu_0, \sigma_0) + p\phi(y_i; \mu_1, \sigma_1)\right].$$

Maximizing this function over the five parameters is hard. Imaging that we were given extra information telling us which of the two normals every observation came from. These "complete" data are of the form $(Y_1, Z_1), \ldots, (Y_n, Z_n)$, where $Z_i = 0$ represents the first normal and $Z_i = 1$ represents the second. Note that $\mathbb{P}(Z_i = 1) = p$. We shall soon see that the likelihood for the complete data $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ is much simpler than the likelihood for the observed data $Y_1, \ldots, Y_n$. ■

Now we describe the EM algorithm.

The EM Algorithm

(0) Pick a starting value θ^0. Now for $j = 1, 2, \ldots$, repeat steps 1 and 2 below:

(1) (The E-step): Calculate

$$J(\theta|\theta^j) = \mathbb{E}_{\theta^j}\left(\log \frac{f(Y^n, Z^n; \theta)}{f(Y^n, Z^n; \theta^j)} \;\middle|\; Y^n = y^n\right).$$

The expectation is over the missing data Z^n treating θ^i and the observed data Y^n as fixed.

(2) Find θ^{j+1} to maximize $J(\theta|\theta^j)$.

We now show that the EM algorithm always increases the likelihood, that is, $\mathcal{L}(\theta^{j+1}) \geq \mathcal{L}(\theta^j)$. Note that

$$\begin{aligned} J(\theta^{j+1}|\theta^j) &= \mathbb{E}_{\theta^j}\left(\log \frac{f(Y^n, Z^n; \theta^{j+1})}{f(Y^n, Z^n; \theta^j)} \;\middle|\; Y^n = y^n\right) \\ &= \log \frac{f(y^n; \theta^{j+1})}{f(y^n; \theta^j)} + \mathbb{E}_{\theta^j}\left(\log \frac{f(Z^n|Y^n; \theta^{j+1})}{f(Z^n|Y^n; \theta^j)} \;\middle|\; Y^n = y^n\right) \end{aligned}$$

and hence

$$\frac{\mathcal{L}(\theta^{j+1})}{\mathcal{L}(\theta^j)} = \log \frac{f(y^n; \theta^{j+1})}{f(y^n; \theta^j)}$$

$$\begin{aligned} &= J(\theta^{j+1}|\theta^j) - \mathbb{E}_{\theta^j}\left(\log \frac{f(Z^n|Y^n;\theta^{j+1})}{f(Z^n|Y^n;\theta^j)} \;\middle|\; Y^n = y^n\right) \\ &= J(\theta^{j+1}|\theta^j) + K(f_j, f_{j+1}) \end{aligned}$$

where $f_j = f(y^n;\theta^j)$ and $f_{j+1} = f(y^n;\theta^{j+1})$ and $K(f,g) = \int f(x)\log(f(x)/g(x))\,dx$ is the Kullback-Leibler distance. Now, θ^{j+1} was chosen to maximize $J(\theta|\theta^j)$. Hence, $J(\theta^{j+1}|\theta^j) \geq J(\theta^j|\theta^j) = 0$. Also, by the properties of Kullback-Leibler divergence, $K(f_j, f_{j+1}) \geq 0$. Hence, $\mathcal{L}(\theta^{j+1}) \geq \mathcal{L}(\theta^j)$ as claimed.

9.50 Example (Continuation of Example 9.49)**.** Consider again the mixture of two normals but, for simplicity assume that $p = 1/2$, $\sigma_1 = \sigma_2 = 1$. The density is

$$f(y;\mu_1,\mu_2) = \frac{1}{2}\phi(y;\mu_0,1) + \frac{1}{2}\phi(y;\mu_1,1).$$

Directly maximizing the likelihood is hard. Introduce latent variables $Z_1,\ldots,Z_n$ where $Z_i = 0$ if Y_i is from $\phi(y;\mu_0,1)$, and $Z_i = 1$ if Y_i is from $\phi(y;\mu_1,1)$, $\mathbb{P}(Z_i = 1) = P(Z_i = 0) = 1/2$, $f(y_i|Z_i = 0) = \phi(y;\mu_0,1)$ and $f(y_i|Z_i = 1) = \phi(y;\mu_1,1)$. So $f(y) = \sum_{z=0}^{1} f(y,z)$ where we have dropped the parameters from the density to avoid notational overload. We can write

$$f(z,y) = f(z)f(y|z) = \frac{1}{2}\phi(y;\mu_0,1)^{1-z}\phi(y;\mu_1,1)^z.$$

Hence, the complete likelihood is

$$\prod_{i=1}^{n} \phi(y_i;\mu_0,1)^{1-z_i}\phi(y_i;\mu_1,1)^{z_i}.$$

The complete log-likelihood is then

$$\widetilde{\ell} = -\frac{1}{2}\sum_{i=1}^{n}(1-z_i)(y_i-\mu_0) - \frac{1}{2}\sum_{i=1}^{n} z_i(y_i-\mu_1).$$

And so

$$J(\theta|\theta^j) = -\frac{1}{2}\sum_{i=1}^{n}(1-\mathbb{E}(Z_i|y^n,\theta^j))(y_i-\mu_0) - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}(Z_i|y^n,\theta^j))(y_i-\mu_1).$$

Since Z_i is binary, $\mathbb{E}(Z_i|y^n,\theta^j) = \mathbb{P}(Z_i = 1|y^n,\theta^j)$ and, by Bayes' theorem,

$$\begin{aligned} \mathbb{P}(Z_i = 1|y^n,\theta^i) &= \frac{f(y^n|Z_i = 1;\theta^j)\mathbb{P}(Z_i = 1)}{f(y^n|Z_i = 1;\theta^j)\mathbb{P}(Z_i = 1) + f(y^n|Z_i = 0;\theta^j)\mathbb{P}(Z_i = 0)} \\ &= \frac{\phi(y_i;\mu_1^j,1)\frac{1}{2}}{\phi(y_i;\mu_1^j,1)\frac{1}{2} + \phi(y_i;\mu_0^j,1)\frac{1}{2}} \\ &= \frac{\phi(y_i;\mu_1^j,1)}{\phi(y_i;\mu_1^j,1) + \phi(y_i;\mu_0^j,1)} \\ &= \tau(i). \end{aligned}$$

Take the derivative of $J(\theta|\theta^j)$ with respect to μ_1 and μ_2, set them equal to 0 to get

$$\widehat{\mu}_1^{j+1} = \frac{\sum_{i=1}^n \tau_i y_i}{\sum_{i=1}^n \tau_i}$$

and

$$\widehat{\mu}_0^{j+1} = \frac{\sum_{i=1}^n (1-\tau_i) y_i}{\sum_{i=1}^n (1-\tau_i)}.$$

We then recompute τ_i using $\widehat{\mu}_1^{j+1}$ and $\widehat{\mu}_0^{j+1}$ and iterate. ■

9.14 Exercises

1. Let $X_1, \ldots, X_n \sim \text{Gamma}(\alpha, \beta)$. Find the method of moments estimator for α and β.

2. Let $X_1, \ldots, X_n \sim \text{Uniform}(a, b)$ where a and b are unknown parameters and $a < b$.

   (a) Find the method of moments estimators for a and b.

   (b) Find the MLE $\widehat{a}$ and $\widehat{b}$.

   (c) Let $\tau = \int x \, dF(x)$. Find the MLE of τ.

   (d) Let $\widehat{\tau}$ be the MLE of τ. Let $\widetilde{\tau}$ be the nonparametric plug-in estimator of $\tau = \int x \, dF(x)$. Suppose that $a = 1$, $b = 3$, and $n = 10$. Find the MSE of $\widehat{\tau}$ by simulation. Find the MSE of $\widetilde{\tau}$ analytically. Compare.

3. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Let τ be the .95 percentile, i.e. $\mathbb{P}(X < \tau) = .95$.

   (a) Find the MLE of τ.

   (b) Find an expression for an approximate $1 - \alpha$ confidence interval for τ.

   (c) Suppose the data are:

```
3.23 -2.50  1.88 -0.68  4.43  0.17
1.03 -0.07 -0.01  0.76  1.76  3.18
0.33 -0.31  0.30 -0.61  1.52  5.43
1.54  2.28  0.42  2.33 -1.03  4.00
0.39
```

   Find the MLE $\widehat{\tau}$. Find the standard error using the delta method. Find the standard error using the parametric bootstrap.

4. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. Show that the MLE is consistent. Hint: Let $Y = \max\{X_1, ..., X_n\}$. For any c, $\mathbb{P}(Y < c) = \mathbb{P}(X_1 < c, X_2 < c, ..., X_n < c) = \mathbb{P}(X_1 < c)\mathbb{P}(X_2 < c)...\mathbb{P}(X_n < c)$.

5. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. Find the method of moments estimator, the maximum likelihood estimator and the Fisher information $I(\lambda)$.

6. Let $X_1, ..., X_n \sim N(\theta, 1)$. Define

$$Y_i = \begin{cases} 1 & \text{if } X_i > 0 \\ 0 & \text{if } X_i \leq 0. \end{cases}$$

   Let $\psi = \mathbb{P}(Y_1 = 1)$.

   (a) Find the maximum likelihood estimator $\widehat{\psi}$ of ψ.

   (b) Find an approximate 95 percent confidence interval for ψ.

   (c) Define $\widetilde{\psi} = (1/n)\sum_i Y_i$. Show that $\widetilde{\psi}$ is a consistent estimator of ψ.

   (d) Compute the asymptotic relative efficiency of $\widetilde{\psi}$ to $\widehat{\psi}$. Hint: Use the delta method to get the standard error of the MLE. Then compute the standard error (i.e. the standard deviation) of $\widetilde{\psi}$.

   (e) Suppose that the data are not really normal. Show that $\widehat{\psi}$ is not consistent. What, if anything, does $\widehat{\psi}$ converge to?

7. (Comparing two treatments.) n_1 people are given treatment 1 and n_2 people are given treatment 2. Let X_1 be the number of people on treatment 1 who respond favorably to the treatment and let X_2 be the number of people on treatment 2 who respond favorably. Assume that $X_1 \sim \text{Binomial}(n_1, p_1)$ $X_2 \sim \text{Binomial}(n_2, p_2)$. Let $\psi = p_1 - p_2$.

   (a) Find the MLE $\widehat{\psi}$ for ψ.

   (b) Find the Fisher information matrix $I(p_1, p_2)$.

   (c) Use the multiparameter delta method to find the asymptotic standard error of $\widehat{\psi}$.

   (d) Suppose that $n_1 = n_2 = 200$, $X_1 = 160$ and $X_2 = 148$. Find $\widehat{\psi}$. Find an approximate 90 percent confidence interval for ψ using (i) the delta method and (ii) the parametric bootstrap.

8. Find the Fisher information matrix for Example 9.29.

9. Let $X_1, ..., X_n \sim \text{Normal}(\mu, 1)$. Let $\theta = e^{\mu}$ and let $\widehat{\theta} = e^{\overline{X}}$ be the MLE. Create a data set (using $\mu = 5$) consisting of n=100 observations.

(a) Use the delta method to get $\widehat{\mathsf{se}}$ and a 95 percent confidence interval for θ. Use the parametric bootstrap to get $\widehat{\mathsf{se}}$ and 95 percent confidence interval for θ. Use the nonparametric bootstrap to get $\widehat{\mathsf{se}}$ and 95 percent confidence interval for θ. Compare your answers.

(b) Plot a histogram of the bootstrap replications for the parametric and nonparametric bootstraps. These are estimates of the distribution of $\widehat{\theta}$. The delta method also gives an approximation to this distribution namely, Normal$(\widehat{\theta}, \mathsf{se}^2)$. Compare these to the true sampling distribution of $\widehat{\theta}$ (which you can get by simulation). Which approximation — parametric bootstrap, bootstrap, or delta method — is closer to the true distribution?

10. Let $X_1, ..., X_n \sim \text{Uniform}(0, \theta)$. The MLE is $\widehat{\theta} = X_{(n)} = \max\{X_1, ..., X_n\}$. Generate a dataset of size 50 with $\theta = 1$.

(a) Find the distribution of $\widehat{\theta}$ analytically. Compare the true distribution of $\widehat{\theta}$ to the histograms from the parametric and nonparametric bootstraps.

(b) This is a case where the nonparametric bootstrap does very poorly. Show that for the parametric bootstrap $\mathbb{P}(\widehat{\theta}^* = \widehat{\theta}) = 0$, but for the nonparametric bootstrap $\mathbb{P}(\widehat{\theta}^* = \widehat{\theta}) \approx .632$. Hint: show that, $\mathbb{P}(\widehat{\theta}^* = \widehat{\theta}) = 1 - (1 - (1/n))^n$ then take the limit as n gets large. What is the implication of this?

10
Hypothesis Testing and p-values

Suppose we want to know if exposure to asbestos is associated with lung disease. We take some rats and randomly divide them into two groups. We expose one group to asbestos and leave the second group unexposed. Then we compare the disease rate in the two groups. Consider the following two hypotheses:

The Null Hypothesis: The disease rate is the same in the two groups.

The Alternative Hypothesis: The disease rate is not the same in the two groups.

If the exposed group has a much higher rate of disease than the unexposed group then we will reject the null hypothesis and conclude that the evidence favors the alternative hypothesis. This is an example of hypothesis testing.

More formally, suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1. \tag{10.1}$$

We call H_0 the **null hypothesis** and H_1 the **alternative hypothesis**.

Let X be a random variable and let $\mathcal{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $R \subset \mathcal{X}$ called the **rejection**

	Retain Null	Reject Null
H_0 true	$\surd$	type I error
H_1 true	type II error	$\surd$

TABLE 10.1. Summary of outcomes of hypothesis testing.

region. If $X \in R$ we reject the null hypothesis, otherwise, we do not reject the null hypothesis:

$$\begin{aligned} X \in R &\implies \text{reject } H_0 \\ X \notin R &\implies \text{retain (do not reject) } H_0 \end{aligned}$$

Usually, the rejection region R is of the form

$$R = \Big\{ x : \ T(x) > c \Big\} \tag{10.2}$$

where T is a **test statistic** and c is a **critical value**. The problem in hypothesis testing is to find an appropriate test statistic T and an appropriate critical value c.

Warning! There is a tendency to use hypothesis testing methods even when they are not appropriate. Often, estimation and confidence intervals are better tools. Use hypothesis testing only when you want to test a well-defined hypothesis.

Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that he is guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0. There are two types of errors we can make. Rejecting H_0 when H_0 is true is called a **type I error**. Retaining H_0 when H_1 is true is called a **type II error**. The possible outcomes for hypothesis testing are summarized in Tab. 10.1.

10.1 Definition. *The* **power function** *of a test with rejection region R is defined by*

$$\beta(\theta) = \mathbb{P}_\theta(X \in R). \tag{10.3}$$

The **size** *of a test is defined to be*

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta). \tag{10.4}$$

A test is said to have **level** α *if its size is less than or equal to α.*

A hypothesis of the form $\theta = \theta_0$ is called a **simple hypothesis**. A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a **composite hypothesis**. A test of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

is called a **two-sided test**. A test of the form

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$

or

$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

is called a **one-sided test**. The most common tests are two-sided.

10.2 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ where σ is known. We want to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. Hence, $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, \infty)$. Consider the test:

$$\text{reject } H_0 \text{ if } T > c$$

where $T = \overline{X}$. The rejection region is

$$R = \Big\{(x_1, \ldots, x_n) : \ T(x_1, \ldots, x_n) > c\Big\}.$$

Let Z denote a standard Normal random variable. The power function is

$$\begin{aligned}
\beta(\mu) &= \mathbb{P}_\mu\left(\overline{X} > c\right) \\
&= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} > \frac{\sqrt{n}(c - \mu)}{\sigma}\right) \\
&= \mathbb{P}\left(Z > \frac{\sqrt{n}(c - \mu)}{\sigma}\right) \\
&= 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right).
\end{aligned}$$

This function is increasing in μ. See Figure 10.1. Hence

$$\text{size} = \sup_{\mu \leq 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right).$$

For a size α test, we set this equal to α and solve for c to get

$$c = \frac{\sigma\, \Phi^{-1}(1 - \alpha)}{\sqrt{n}}.$$

We reject when $\overline{X} > \sigma\, \Phi^{-1}(1 - \alpha)/\sqrt{n}$. Equivalently, we reject when

$$\frac{\sqrt{n}\,(\overline{X} - 0)}{\sigma} > z_\alpha.$$

where $z_\alpha = \Phi^{-1}(1 - \alpha)$. ■

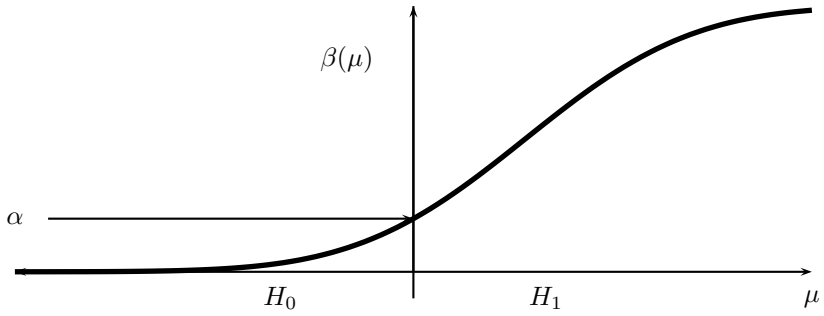


FIGURE 10.1. The power function for Example 10.2. The size of the test is the largest probability of rejecting H_0 when H_0 is true. This occurs at $\mu = 0$ hence the size is $\beta(0)$. We choose the critical value c so that $\beta(0) = \alpha$.

It would be desirable to find the test with highest power under H_1, among all size α tests. Such a test, if it exists, is called **most powerful**. Finding most powerful tests is hard and, in many cases, most powerful tests don't even exist. Instead of going into detail about when most powerful tests exist, we'll just consider four widely used tests: the Wald test,[1] the χ^2 test, the permutation test, and the likelihood ratio test.

10.1 The Wald Test

Let θ be a scalar parameter, let $\widehat{\theta}$ be an estimate of θ and let $\widehat{\mathsf{se}}$ be the estimated standard error of $\widehat{\theta}$.

[1]The test is named after Abraham Wald (1902–1950), who was a very influential mathematical statistician. Wald died in a plane crash in India in 1950.

10.3 Definition. The Wald Test

Consider testing

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Assume that $\widehat{\theta}$ *is asymptotically Normal:*

$$\frac{(\widehat{\theta} - \theta_0)}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1).$$

The size α **Wald test** *is: reject* H_0 *when* $|W| > z_{\alpha/2}$ *where*

$$W = \frac{\widehat{\theta} - \theta_0}{\widehat{\mathsf{se}}}. \tag{10.5}$$

10.4 Theorem. *Asymptotically, the Wald test has size* α*, that is,*

$$\mathbb{P}_{\theta_0}\left(|W| > z_{\alpha/2}\right) \to \alpha$$

as $n \to \infty$*.*

Proof. Under $\theta = \theta_0$, $(\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}} \rightsquigarrow N(0,1)$. Hence, the probability of rejecting when the null $\theta = \theta_0$ is true is

$$\begin{aligned} \mathbb{P}_{\theta_0}\left(|W| > z_{\alpha/2}\right) &= \mathbb{P}_{\theta_0}\left(\frac{|\widehat{\theta} - \theta_0|}{\widehat{\mathsf{se}}} > z_{\alpha/2}\right) \\ &\to \mathbb{P}\left(|Z| > z_{\alpha/2}\right) \\ &= \alpha \end{aligned}$$

where $Z \sim N(0,1)$. ∎

10.5 Remark. An alternative version of the Wald test statistic is $W = (\widehat{\theta} - \theta_0)/\mathsf{se}_0$ where se_0 is the standard error computed at $\theta = \theta_0$. Both versions of the test are valid.

Let us consider the power of the Wald test when the null hypothesis is false.

10.6 Theorem. *Suppose the true value of* θ *is* $\theta_\star \neq \theta_0$*. The power* $\beta(\theta_\star)$ *— the probability of correctly rejecting the null hypothesis — is given (approximately) by*

$$1 - \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} - z_{\alpha/2}\right). \tag{10.6}$$

Recall that $\widehat{\mathsf{se}}$ tends to 0 as the sample size increases. Inspecting (10.6) closely we note that: (i) the power is large if $\theta_\star$ is far from θ_0, and (ii) the power is large if the sample size is large.

10.7 Example (Comparing Two Prediction Algorithms)**.** We test a prediction algorithm on a test set of size m and we test a second prediction algorithm on a second test set of size n. Let X be the number of incorrect predictions for algorithm 1 and let Y be the number of incorrect predictions for algorithm 2. Then $X \sim \text{Binomial}(m, p_1)$ and $Y \sim \text{Binomial}(n, p_2)$. To test the null hypothesis that $p_1 = p_2$ write

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta \neq 0$$

where $\delta = p_1 - p_2$. The MLE is $\widehat{\delta} = \widehat{p}_1 - \widehat{p}_2$ with estimated standard error

$$\widehat{\mathsf{se}} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}.$$

The size α Wald test is to reject H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}}.$$

The power of this test will be largest when p_1 is far from p_2 and when the sample sizes are large.

What if we used the same test set to test both algorithms? The two samples are no longer independent. Instead we use the following strategy. Let $X_i = 1$ if algorithm 1 is correct on test case i and $X_i = 0$ otherwise. Let $Y_i = 1$ if algorithm 2 is correct on test case i, and $Y_i = 0$ otherwise. Define $D_i = X_i - Y_i$. A typical dataset will look something like this:

Test Case	X_i	Y_i	$D_i = X_i - Y_i$
1	1	0	1
2	1	1	0
3	1	1	0
4	0	1	-1
5	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	0	1	-1

Let

$$\delta = \mathbb{E}(D_i) = \mathbb{E}(X_i) - \mathbb{E}(Y_i) = \mathbb{P}(X_i = 1) - \mathbb{P}(Y_i = 1).$$

The nonparametric plug-in estimate of δ is $\widehat{\delta} = \overline{D} = n^{-1}\sum_{i=1}^n D_i$ and $\widehat{\mathsf{se}}(\widehat{\delta}) = S/\sqrt{n}$, where $S^2 = n^{-1}\sum_{i=1}^n (D_i - \overline{D})^2$. To test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$

we use $W = \widehat{\delta}/\widehat{\mathsf{se}}$ and reject H_0 if $|W| > z_{\alpha/2}$. This is called a **paired comparison.** ■

10.8 Example (Comparing Two Means)**.** Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent samples from populations with means μ_1 and μ_2, respectively. Let's test the null hypothesis that $\mu_1 = \mu_2$. Write this as $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ where $\delta = \mu_1 - \mu_2$. Recall that the nonparametric plug-in estimate of δ is $\widehat{\delta} = \overline{X} - \overline{Y}$ with estimated standard error

$$\widehat{\mathsf{se}} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where s_1^2 and s_2^2 are the sample variances. The size α Wald test rejects H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}. \quad ■$$

10.9 Example (Comparing Two Medians)**.** Consider the previous example again but let us test whether the medians of the two distributions are the same. Thus, $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ where $\delta = \nu_1 - \nu_2$ where ν_1 and ν_2 are the medians. The nonparametric plug-in estimate of δ is $\widehat{\delta} = \widehat{\nu}_1 - \widehat{\nu}_2$ where $\widehat{\nu}_1$ and $\widehat{\nu}_2$ are the sample medians. The estimated standard error $\widehat{\mathsf{se}}$ of $\widehat{\delta}$ can be obtained from the bootstrap. The Wald test statistic is $W = \widehat{\delta}/\widehat{\mathsf{se}}$. ■

There is a relationship between the Wald test and the $1 - \alpha$ asymptotic confidence interval $\widehat{\theta} \pm \widehat{\mathsf{se}}\, z_{\alpha/2}$.

10.10 Theorem. *The size α Wald test rejects $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$ where*

$$C = (\widehat{\theta} - \widehat{\mathsf{se}}\, z_{\alpha/2},\ \widehat{\theta} + \widehat{\mathsf{se}}\, z_{\alpha/2}).$$

Thus, testing the hypothesis is equivalent to checking whether the null value is in the confidence interval.

Warning! When we reject H_0 we often say that the result is **statistically significant.** A result might be statistically significant and yet the size of the effect might be small. In such a case we have a result that is statistically significant but not scientifically or practically significant. The difference between statistical significance and scientific significance is easy to understand in light of Theorem 10.10. Any confidence interval that excludes θ_0 corresponds to rejecting H_0. But the values in the interval could be close to θ_0 (not scientifically significant) or far from θ_0 (scientifically significant). See Figure 10.2.

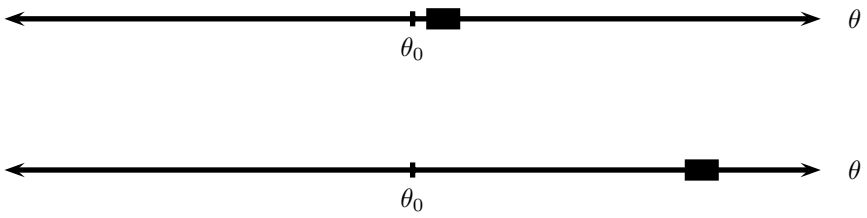


FIGURE 10.2. Scientific significance versus statistical significance. A level α test rejects $H_0 : \theta = \theta_0$ if and only if the $1 - \alpha$ confidence interval does not include θ_0. Here are two different confidence intervals. Both exclude θ_0 so in both cases the test would reject H_0. But in the first case, the estimated value of θ is close to θ_0 so the finding is probably of little scientific or practical value. In the second case, the estimated value of θ is far from θ_0 so the finding is of scientific value. This shows two things. First, statistical significance does not imply that a finding is of scientific importance. Second, confidence intervals are often more informative than tests.

10.2 p-values

Reporting "reject H_0" or "retain H_0" is not very informative. Instead, we could ask, for every α, whether the test rejects at that level. Generally, if the test rejects at level α it will also reject at level $\alpha' > \alpha$. Hence, there is a smallest α at which the test rejects and we call this number the p-value. See Figure 10.3.

10.11 Definition. *Suppose that for every $\alpha \in (0, 1)$ we have a size α test with rejection region R_α. Then,*

$$\text{p-value} = \inf\left\{\alpha : \ T(X^n) \in R_\alpha\right\}.$$

That is, the p-value is the smallest level at which we can reject H_0.

Informally, the p-value is a measure of the evidence against H_0: the smaller the p-value, the stronger the evidence against H_0. Typically, researchers use the following evidence scale:

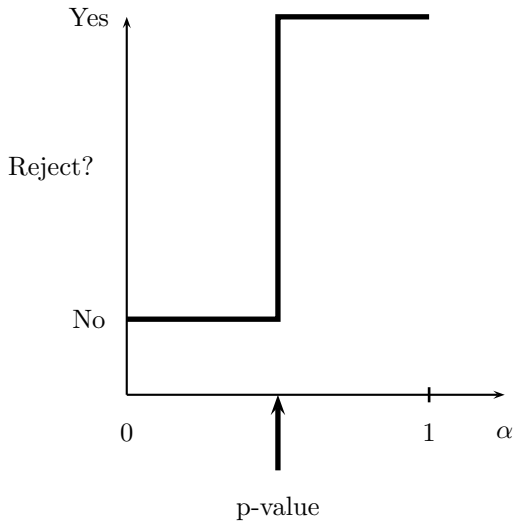


FIGURE 10.3. p-values explained. For each α we can ask: does our test reject H_0 at level α? The p-value is the smallest α at which we do reject H_0. If the evidence against H_0 is strong, the p-value will be small.

p-value	evidence
$< .01$	very strong evidence against H_0
$.01 - .05$	strong evidence against H_0
$.05 - .10$	weak evidence against H_0
$> .1$	little or no evidence against H_0

Warning! A large p-value is not strong evidence in favor of H_0. A large p-value can occur for two reasons: (i) H_0 is true or (ii) H_0 is false but the test has low power.

Warning! Do not confuse the p-value with $\mathbb{P}(H_0|\text{Data})$. [2] **The p-value is not the probability that the null hypothesis is true.**

The following result explains how to compute the p-value.

[2] We discuss quantities like $\mathbb{P}(H_0|\text{Data})$ in the chapter on Bayesian inference.

10.12 Theorem. *Suppose that the size α test is of the form*

$$\text{reject } H_0 \text{ if and only if } T(X^n) \geq c_\alpha.$$

Then,

$$\text{p-value} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X^n) \geq T(x^n))$$

where x^n is the observed value of X^n. If $\Theta_0 = \{\theta_0\}$ then

$$\text{p-value} = \mathbb{P}_{\theta_0}(T(X^n) \geq T(x^n)).$$

We can express Theorem 10.12 as follows:

> The p-value is the probability (under H_0) of observing a value of the test statistic the same as or more extreme than what was actually observed.

10.13 Theorem. *Let $w = (\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}}$ denote the observed value of the Wald statistic W. The p-value is given by*

$$\text{p} - \text{value} = \mathbb{P}_{\theta_0}(|W| > |w|) \approx \mathbb{P}(|Z| > |w|) = 2\Phi(-|w|) \tag{10.7}$$

where $Z \sim N(0,1)$.

To understand this last theorem, look at Figure 10.4.
Here is an important property of p-values.

10.14 Theorem. *If the test statistic has a continuous distribution, then under $H_0 : \theta = \theta_0$, the p-value has a Uniform (0,1) distribution. Therefore, if we reject H_0 when the p-value is less than α, the probability of a type I error is α.*

In other words, if H_0 is true, the p-value is like a random draw from a $\text{Unif}(0,1)$ distribution. If H_1 is true, the distribution of the p-value will tend to concentrate closer to 0.

10.15 Example. Recall the cholesterol data from Example 7.15. To test if the means are different we compute

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{216.2 - 195.3}{\sqrt{5^2 + 2.4^2}} = 3.78.$$

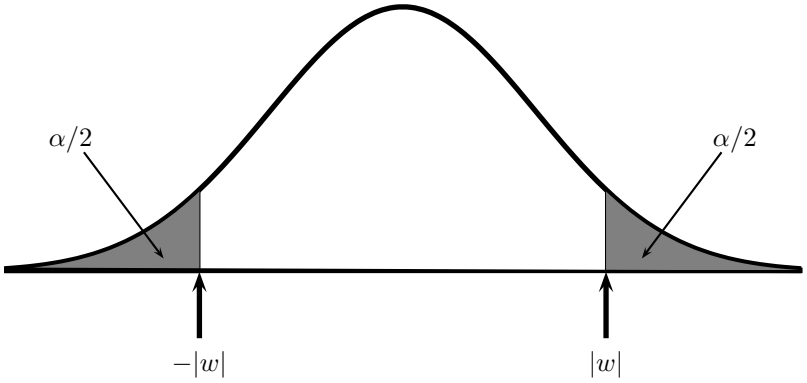


FIGURE 10.4. The p-value is the smallest α at which you would reject H_0. To find the p-value for the Wald test, we find α such that $|w|$ and $-|w|$ are just at the boundary of the rejection region. Here, w is the observed value of the Wald statistic: $w = (\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}}$. This implies that the p-value is the tail area $\mathbb{P}(|Z| > |w|)$ where $Z \sim N(0,1)$.

To compute the p-value, let $Z \sim N(0,1)$ denote a standard Normal random variable. Then,

$$\text{p-value} = \mathbb{P}(|Z| > 3.78) = 2\mathbb{P}(Z < -3.78) = .0002$$

which is very strong evidence against the null hypothesis. To test if the medians are different, let $\widehat{\nu}_1$ and $\widehat{\nu}_2$ denote the sample medians. Then,

$$W = \frac{\widehat{\nu}_1 - \widehat{\nu}_2}{\widehat{\mathsf{se}}} = \frac{212.5 - 194}{7.7} = 2.4$$

where the standard error 7.7 was found using the bootstrap. The p-value is

$$\text{p-value} = \mathbb{P}(|Z| > 2.4) = 2\mathbb{P}(Z < -2.4) = .02$$

which is strong evidence against the null hypothesis. ■

10.3 The χ^2 Distribution

Before proceeding we need to discuss the χ^2 distribution. Let $Z_1, \ldots, Z_k$ be independent, standard Normals. Let $V = \sum_{i=1}^k Z_i^2$. Then we say that V has a χ^2 distribution with k degrees of freedom, written $V \sim \chi^2_k$. The probability density of V is

$$f(v) = \frac{v^{(k/2)-1} e^{-v/2}}{2^{k/2}\Gamma(k/2)}$$

for $v > 0$. It can be shown that $\mathbb{E}(V) = k$ and $\mathbb{V}(V) = 2k$. We define the upper α quantile $\chi^2_{k,\alpha} = F^{-1}(1-\alpha)$ where F is the CDF. That is, $\mathbb{P}(\chi^2_k > \chi^2_{k,\alpha}) = \alpha$.

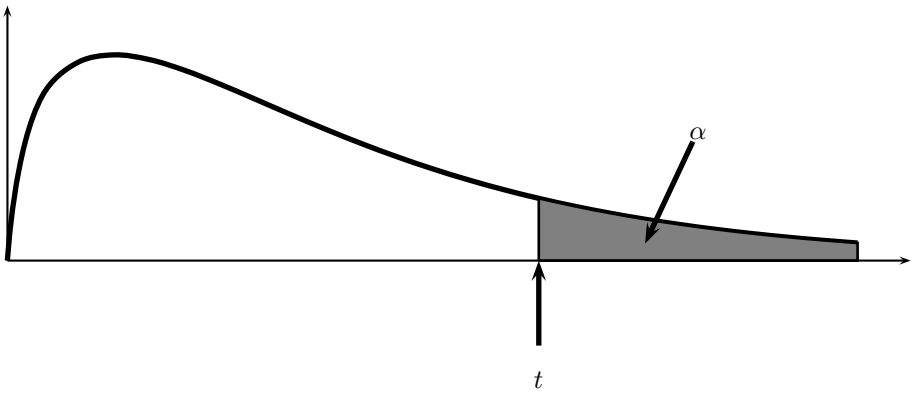


FIGURE 10.5. The p-value is the smallest α at which we would reject H_0. To find the p-value for the χ^2_{k-1} test, we find α such that the observed value t of the test statistic is just at the boundary of the rejection region. This implies that the p-value is the tail area $\mathbb{P}(\chi^2_{k-1} > t)$.

10.4 Pearson's χ^2 Test For Multinomial Data

Pearson's χ^2 test is used for multinomial data. Recall that if $X = (X_1, \ldots, X_k)$ has a multinomial (n, p) distribution, then the MLE of p is $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_k) = (X_1/n, \ldots, X_k/n)$.

Let $p_0 = (p_{01}, \ldots, p_{0k})$ be some fixed vector and suppose we want to test

$$H_0 : p = p_0 \text{ versus } H_1 : p \neq p_0.$$

10.16 Definition. Pearson's χ^2 statistic *is*

$$T = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}} = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}$$

where $E_j = \mathbb{E}(X_j) = np_{0j}$ is the expected value of X_j under H_0.

10.17 Theorem. *Under H_0, $T \rightsquigarrow \chi^2_{k-1}$. Hence the test: reject H_0 if $T > \chi^2_{k-1,\alpha}$ has asymptotic level α. The p-value is $\mathbb{P}(\chi^2_{k-1} > t)$ where t is the observed value of the test statistic.*

Theorem 10.17 is illustrated in Figure 10.5.

10.18 Example (Mendel's peas)**.** Mendel bred peas with round yellow seeds and wrinkled green seeds. There are four types of progeny: round yellow, wrinkled yellow, round green, and wrinkled green. The number of each type is multinomial with probability $p = (p_1, p_2, p_3, p_4)$. His theory of inheritance predicts that p is equal to

$$p_0 \equiv \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right).$$

In $n = 556$ trials he observed $X = (315, 101, 108, 32)$. We will test $H_0 : p = p_0$ versus $H_1 : p \neq p_0$. Since, $np_{01} = 312.75, np_{02} = np_{03} = 104.25$, and $np_{04} = 34.75$, the test statistic is

$$\begin{aligned}\chi^2 &= \frac{(315 - 312.75)^2}{312.75} + \frac{(101 - 104.25)^2}{104.25} \\ &\quad + \frac{(108 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} = 0.47.\end{aligned}$$

The $\alpha = .05$ value for a χ^2_3 is 7.815. Since 0.47 is not larger than 7.815 we do not reject the null. The p-value is

$$\text{p-value} = \mathbb{P}(\chi^2_3 > .47) = .93$$

which is not evidence against H_0. Hence, the data do not contradict Mendel's theory.[3] ■

In the previous example, one could argue that hypothesis testing is not the right tool. Hypothesis testing is useful to see if there is evidence to reject H_0. This is appropriate when H_0 corresponds to the status quo. It is not useful for proving that H_0 is true. Failure to reject H_0 might occur because H_0 is true, but it might occur just because the test has low power. Perhaps a confidence set for the distance between p and p_0 might be more useful in this example.

10.5 The Permutation Test

The permutation test is a nonparametric method for testing whether two distributions are the same. This test is "exact," meaning that it is not based on large sample theory approximations. Suppose that $X_1, \ldots, X_m \sim F_X$ and $Y_1, \ldots, Y_n \sim F_Y$ are two independent samples and H_0 is the hypothesis that

[3] There is some controversy about whether Mendel's results are "too good."

the two samples are identically distributed. This is the type of hypothesis we would consider when testing whether a treatment differs from a placebo. More precisely we are testing

$$H_0 : F_X = F_Y \quad \text{versus} \quad H_1 : F_X \neq F_Y.$$

Let $T(x_1, \ldots, x_m, y_1, \ldots, y_n)$ be some test statistic, for example,

$$T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = |\overline{X}_m - \overline{Y}_n|.$$

Let $N = m + n$ and consider forming all $N!$ permutations of the data X_1, ..., X_m, Y_1, ..., Y_n. For each permutation, compute the test statistic T. Denote these values by $T_1, \ldots, T_{N!}$. Under the null hypothesis, each of these values is equally likely. [4] The distribution $\mathbb{P}_0$ that puts mass $1/N!$ on each T_j is called the **permutation distribution** of T. Let t_{obs} be the observed value of the test statistic. Assuming we reject when T is large, the p-value is

$$\text{p-value} = \mathbb{P}_0(T > t_{obs}) = \frac{1}{N!} \sum_{j=1}^{N!} I(T_j > t_{obs}).$$

10.19 Example. Here is a toy example to make the idea clear. Suppose the data are: $(X_1, X_2, Y_1) = (1, 9, 3)$. Let $T(X_1, X_2, Y_1) = |\overline{X} - \overline{Y}| = 2$. The permutations are:

permutation	value of T	probability
(1,9,3)	2	1/6
(9,1,3)	2	1/6
(1,3,9)	7	1/6
(3,1,9)	7	1/6
(3,9,1)	5	1/6
(9,3,1)	5	1/6

The p-value is $\mathbb{P}(T > 2) = 4/6$. ■

Usually, it is not practical to evaluate all $N!$ permutations. We can approximate the p-value by sampling randomly from the set of permutations. The fraction of times $T_j > t_{obs}$ among these samples approximates the p-value.

[4] More precisely, under the null hypothesis, given the ordered data values, $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ is uniformly distributed over the $N!$ permutations of the data.

Algorithm for Permutation Test

1. Compute the observed value of the test statistic $t_{\text{obs}} = T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$.
2. Randomly permute the data. Compute the statistic again using the permuted data.
3. Repeat the previous step B times and let $T_1, \ldots, T_B$ denote the resulting values.
4. The approximate p-value is

$$\frac{1}{B}\sum_{j=1}^{B} I(T_j > t_{\text{obs}}).$$

10.20 Example. DNA microarrays allow researchers to measure the expression levels of thousands of genes. The data are the levels of messenger RNA (mRNA) of each gene, which is thought to provide a measure of how much protein that gene produces. Roughly, the larger the number, the more active the gene. The table below, reproduced from Efron et al. (2001) shows the expression levels for genes from ten patients with two types of liver cancer cells. There are 2,638 genes in this experiment but here we show just the first two. The data are log-ratios of the intensity levels of two different color dyes used on the arrays.

	Type I					Type II				
Patient	1	2	3	4	5	6	7	8	9	10
Gene 1	230	-1,350	-1,580	-400	-760	970	110	-50	-190	-200
Gene 2	470	-850	-.8	-280	120	390	-1730	-1360	-1	-330
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Let's test whether the median level of gene 1 is different between the two groups. Let ν_1 denote the median level of gene 1 of Type I and let ν_2 denote the median level of gene 1 of Type II. The absolute difference of sample medians is $T = |\widehat{\nu}_1 - \widehat{\nu}_2| = 710$. Now we estimate the permutation distribution by simulation and we find that the estimated p-value is .045. Thus, if we use a $\alpha = .05$ level of significance, we would say that there is evidence to reject the null hypothesis of no difference. ■

In large samples, the permutation test usually gives similar results to a test that is based on large sample theory. The permutation test is thus most useful for small samples.

10.6 The Likelihood Ratio Test

The Wald test is useful for testing a scalar parameter. The likelihood ratio test is more general and can be used for testing a vector-valued parameter.

10.21 Definition. *Consider testing*

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \notin \Theta_0.$$

The **likelihood ratio statistic** *is*

$$\lambda = 2 \log \left(\frac{\sup_{\theta \in \Theta} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)} \right) = 2 \log \left(\frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(\widehat{\theta}_0)} \right)$$

where $\widehat{\theta}$ is the MLE *and $\widehat{\theta}_0$ is the* MLE *when θ is restricted to lie in Θ_0.*

You might have expected to see the maximum of the likelihood over Θ_0^c instead of Θ in the numerator. In practice, replacing Θ_0^c with Θ has little effect on the test statistic. Moreover, the theoretical properties of λ are much simpler if the test statistic is defined this way.

The likelihood ratio test is most useful when Θ_0 consists of all parameter values θ such that some coordinates of θ are fixed at particular values.

10.22 Theorem. *Suppose that $\theta = (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_r)$. Let*

$$\Theta_0 = \{\theta : \ (\theta_{q+1}, \ldots, \theta_r) = (\theta_{0,q+1}, \ldots, \theta_{0,r})\}.$$

Let λ be the likelihood ratio test statistic. Under $H_0 : \theta \in \Theta_0$,

$$\lambda(x^n) \rightsquigarrow \chi^2_{r-q,\alpha}$$

where $r - q$ is the dimension of Θ minus the dimension of Θ_0. The p-value for the test is $\mathbb{P}(\chi^2_{r-q} > \lambda)$.

For example, if $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ and we want to test the null hypothesis that $\theta_4 = \theta_5 = 0$ then the limiting distribution has $5 - 3 = 2$ degrees of freedom.

10.23 Example (Mendel's Peas Revisited)**.** Consider example 10.18 again. The likelihood ratio test statistic for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ is

$$\begin{aligned}
\lambda &= 2\log\left(\frac{\mathcal{L}(\widehat{p})}{\mathcal{L}(p_0)}\right) \\
&= 2\sum_{j=1}^{4} X_j \log\left(\frac{\widehat{p}_j}{p_{0j}}\right) \\
&= 2\left(315\log\left(\frac{\frac{315}{556}}{\frac{9}{16}}\right) + 101\log\left(\frac{\frac{101}{556}}{\frac{3}{16}}\right)\right. \\
&\quad \left. +108\log\left(\frac{\frac{108}{556}}{\frac{3}{16}}\right) + 32\log\left(\frac{\frac{32}{556}}{\frac{1}{16}}\right)\right) \\
&= 0.48.
\end{aligned}$$

Under H_1 there are four parameters. However, the parameters must sum to one so the dimension of the parameter space is three. Under H_0 there are no free parameters so the dimension of the restricted parameter space is zero. The difference of these two dimensions is three. Therefore, the limiting distribution of λ under H_0 is χ^2_3 and the p-value is

$$\text{p-value} = \mathbb{P}(\chi^2_3 > .48) = .92.$$

The conclusion is the same as with the χ^2 test. ■

When the likelihood ratio test and the χ^2 test are both applicable, as in the last example, they usually lead to similar results as long as the sample size is large.

10.7 Multiple Testing

In some situations we may conduct many hypothesis tests. In example 10.20, there were actually 2,638 genes. If we tested for a difference for each gene, we would be conducting 2,638 separate hypothesis tests. Suppose each test is conducted at level α. For any one test, the chance of a false rejection of the null is α. But the chance of at least one false rejection is much higher. This is the **multiple testing problem.** The problem comes up in many data mining situations where one may end up testing thousands or even millions of hypotheses. There are many ways to deal with this problem. Here we discuss two methods.

Consider m hypothesis tests:

$$H_{0i} \quad \text{versus} \quad H_{1i}, \quad i = 1, \ldots, m$$

and let $P_1, \ldots, P_m$ denote the m p-values for these tests.

The Bonferroni Method

Given p-values $P_1, \ldots, P_m$, reject null hypothesis H_{0i} if

$$P_i < \frac{\alpha}{m}.$$

10.24 Theorem. *Using the Bonferroni method, the probability of falsely rejecting any null hypotheses is less than or equal to α.*

PROOF. Let R be the event that at least one null hypothesis is falsely rejected. Let R_i be the event that the i^{th} null hypothesis is falsely rejected. Recall that if $A_1, \ldots, A_k$ are events then $\mathbb{P}(\bigcup_{i=1}^k A_i) \leq \sum_{i=1}^k \mathbb{P}(A_i)$. Hence,

$$\mathbb{P}(R) = \mathbb{P}\left(\bigcup_{i=1}^m R_i\right) \leq \sum_{i=1}^m \mathbb{P}(R_i) = \sum_{i=1}^m \frac{\alpha}{m} = \alpha$$

from Theorem 10.14. ∎

10.25 Example. In the gene example, using $\alpha = .05$, we have that $.05/2,638 = .00001895375$. Hence, for any gene with p-value less than .00001895375, we declare that there is a significant difference. ∎

The Bonferroni method is very conservative because it is trying to make it unlikely that you would make even one false rejection. Sometimes, a more reasonable idea is to control the **false discovery rate** (FDR) which is defined as the mean of the number of false rejections divided by the number of rejections.

Suppose we reject all null hypotheses whose p-values fall below some threshold. Let m_0 be the number of null hypotheses that are true and let $m_1 = m - m_0$. The tests can be categorized in a 2×2 as in Table 10.2.

Define the **False Discovery Proportion** (FDP)

$$\text{FDP} = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{if } R = 0. \end{cases}$$

The FDP is the proportion of rejections that are incorrect. Next define FDR $= \mathbb{E}(\text{FDP})$.

	H_0 Not Rejected	H_0 Rejected	Total
H_0 True	U	V	m_0
H_0 False	T	S	m_1
Total	$m-R$	R	m

TABLE 10.2. Types of outcomes in multiple testing.

The Benjamini-Hochberg (BH) Method

1. Let $P_{(1)} < \cdots < P_{(m)}$ denote the ordered p-values.
2. Define

$$\ell_i = \frac{i\alpha}{C_m m}, \quad \text{and} \quad R = \max\left\{ i : P_{(i)} < \ell_i \right\} \tag{10.8}$$

   where C_m is defined to be 1 if the p-values are independent and $C_m = \sum_{i=1}^{m}(1/i)$ otherwise.
3. Let $T = P_{(R)}$; we call T the **BH rejection threshold**.
4. Reject all null hypotheses H_{0i} for which $P_i \leq T$.

10.26 Theorem (Benjamini and Hochberg). *If the procedure above is applied, then regardless of how many nulls are true and regardless of the distribution of the p-values when the null hypothesis is false,*

$$\text{FDR} = \mathbb{E}(\text{FDP}) \leq \frac{m_0}{m}\alpha \leq \alpha.$$

10.27 Example. Figure 10.6 shows six ordered p-values plotted as vertical lines. If we tested at level α without doing any correction for multiple testing, we would reject all hypotheses whose p-values are less than α. In this case, the four hypotheses corresponding to the four smallest p-values are rejected. The Bonferroni method rejects all hypotheses whose p-values are less than α/m. In this case, this leads to no rejections. The BH threshold corresponds to the last p-value that falls under the line with slope α. This leads to two hypotheses being rejected in this case. ■

10.28 Example. Suppose that 10 independent hypothesis tests are carried leading to the following ordered p-values:

```
0.00017 0.00448 0.00671 0.00907 0.01220
0.33626 0.39341 0.53882 0.58125 0.98617
```

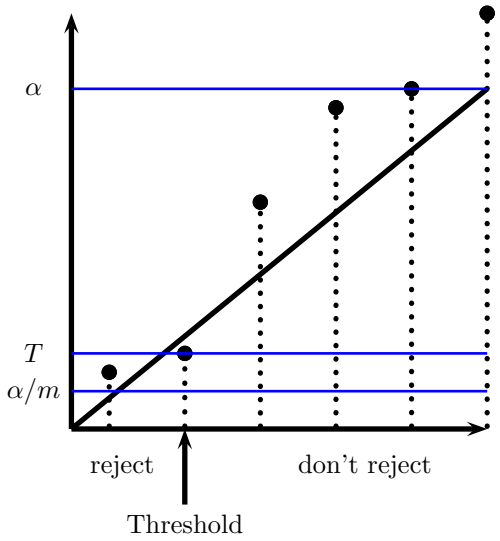


FIGURE 10.6. The Benjamini-Hochberg (BH) procedure. For uncorrected testing we reject when $P_i < \alpha$. For Bonferroni testing we reject when $P_i < \alpha/m$. The BH procedure rejects when $P_i \leq T$. The BH threshold T corresponds to the rightmost undercrossing of the upward sloping line.

With $\alpha = 0.05$, the Bonferroni test rejects any hypothesis whose p-value is less than $\alpha/10 = 0.005$. Thus, only the first two hypotheses are rejected. For the BH test, we find the largest i such that $P_{(i)} < i\alpha/m$, which in this case is $i = 5$. Thus we reject the first five hypotheses. ■

10.8 Goodness-of-fit Tests

There is another situation where testing arises, namely, when we want to check whether the data come from an assumed parametric model. There are many such tests; here is one.

Let $\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\}$ be a parametric model. Suppose the data take values on the real line. Divide the line into k disjoint intervals $I_1, \ldots, I_k$. For $j = 1, \ldots, k$, let

$$p_j(\theta) = \int_{I_j} f(x;\theta)\, dx$$

be the probability that an observation falls into interval I_j under the assumed model. Here, $\theta = (\theta_1, \ldots, \theta_s)$ are the parameters in the assumed model. Let N_j be the number of observations that fall into I_j. The likelihood for θ based

on the counts $N_1, \ldots, N_k$ is the multinomial likelihood

$$Q(\theta) = \prod_{j=1}^{k} p_i(\theta)^{N_j}.$$

Maximizing $Q(\theta)$ yields estimates $\widetilde{\theta} = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_s)$ of θ. Now define the test statistic

$$Q = \sum_{j=1}^{k} \frac{(N_j - np_j(\widetilde{\theta}))^2}{np_j(\widetilde{\theta})}. \tag{10.9}$$

10.29 Theorem. *Let H_0 be the null hypothesis that the data are* IID *draws from the model $\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\}$. Under $H - 0$, the statistic Q defined in equation (10.9) converges in distribution to a χ^2_{k-1-s} random variable. Thus, the (approximate) p-value for the test is $\mathbb{P}(\chi^2_{k-1-s} > q)$ where q denotes the observed value of Q.*

It is tempting to replace $\widetilde{\theta}$ in (10.9) with the MLE $\widehat{\theta}$. However, this will not result in a statistic whose limiting distribution is a χ^2_{k-1-s}. However, it can be shown — due to a theorem of Herman Chernoff and Erich Lehmann from 1954 — that the p-value is bounded approximately by the p-values obtained using a χ^2_{k-1-s} and a χ^2_{k-1}.

Goodness-of-fit testing has some serious limitations. If reject H_0 then we conclude we should not use the model. But if we do not reject H_0 we cannot conclude that the model is correct. We may have failed to reject simply because the test did not have enough power. This is why it is better to use nonparametric methods whenever possible rather than relying on parametric assumptions.

10.9 Bibliographic Remarks

The most complete book on testing is Lehmann (1986). See also Chapter 8 of Casella and Berger (2002) and Chapter 9 of Rice (1995). The FDR method is due to Benjamini and Hochberg (1995). Some of the exercises are from Rice (1995).

10.10 Appendix

10.10.1 The Neyman-Pearson Lemma

In the special case of a simple null $H_0 : \theta = \theta_0$ and a simple alternative $H_1 : \theta = \theta_1$ we can say precisely what the most powerful test is.

10.30 Theorem (Neyman-Pearson). *Suppose we test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Let*

$$T = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{\prod_{i=1}^n f(x_i;\theta_1)}{\prod_{i=1}^n f(x_i;\theta_0)}.$$

Suppose we reject H_0 when $T > k$. If we choose k so that $\mathbb{P}_{\theta_0}(T > k) = \alpha$ then this test is the most powerful, size α test. That is, among all tests with size α, this test maximizes the power $\beta(\theta_1)$.

10.10.2 The t-test

To test $H_0 : \mu = \mu_0$ where $\mu = \mathbb{E}(X_i)$ is the mean, we can use the Wald test. When the data are assumed to be Normal and the sample size is small, it is common instead to use the **t-test**. A random variable T has a *t-distribution with k degrees of freedom* if it has density

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)\left(1+\frac{t^2}{k}\right)^{(k+1)/2}}.$$

When the degrees of freedom $k \to \infty$, this tends to a Normal distribution. When $k = 1$ it reduces to a Cauchy.

Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$ are both unknown. Suppose we want to test $\mu = \mu_0$ versus $\mu \neq \mu_0$. Let

$$T = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S_n}$$

where S_n^2 is the sample variance. For large samples $T \approx N(0,1)$ under H_0. The exact distribution of T under H_0 is t_{n-1}. Hence if we reject when $|T| > t_{n-1,\alpha/2}$ then we get a size α test. However, when n is moderately large, the t-test is essentially identical to the Wald test.

10.11 Exercises

1. Prove Theorem 10.6.

2. Prove Theorem 10.14.

3. Prove Theorem 10.10.

4. Prove Theorem 10.12.

5. Let $X_1, ..., X_n \sim \text{Uniform}(0, \theta)$ and let $Y = \max\{X_1, ..., X_n\}$. We want to test

   $H_0 : \theta = 1/2$ versus $H_1 : \theta > 1/2$.

   The Wald test is not appropriate since Y does not converge to a Normal. Suppose we decide to test this hypothesis by rejecting H_0 when $Y > c$.

   (a) Find the power function.

   (b) What choice of c will make the size of the test .05?

   (c) In a sample of size $n = 20$ with Y=0.48 what is the p-value? What conclusion about H_0 would you make?

   (d) In a sample of size $n = 20$ with Y=0.52 what is the p-value? What conclusion about H_0 would you make?

6. There is a theory that people can postpone their death until after an important event. To test the theory, Phillips and King (1988) collected data on deaths around the Jewish holiday Passover. Of 1919 deaths, 922 died the week before the holiday and 997 died the week after. Think of this as a binomial and test the null hypothesis that $\theta = 1/2$. Report and interpret the p-value. Also construct a confidence interval for θ.

7. In 1861, 10 essays appeared in the *New Orleans Daily Crescent*. They were signed "Quintus Curtius Snodgrass" and some people suspected they were actually written by Mark Twain. To investigate this, we will consider the proportion of three letter words found in an author's work. From eight Twain essays we have:

   .225 .262 .217 .240 .230 .229 .235 .217

   From 10 Snodgrass essays we have:

   .209 .205 .196 .210 .202 .207 .224 .223 .220 .201

   (a) Perform a Wald test for equality of the means. Use the nonparametric plug-in estimator. Report the p-value and a 95 per cent confidence interval for the difference of means. What do you conclude?

   (b) Now use a permutation test to avoid the use of large sample methods. What is your conclusion? (Brinegar (1963)).

8. Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Consider testing

$$H_0 : \theta = 0 \text{ versus } \theta = 1.$$

Let the rejection region be $R = \{x^n : \ T(x^n) > c\}$ where $T(x^n) = n^{-1} \sum_{i=1}^n X_i$.

(a) Find c so that the test has size α.

(b) Find the power under H_1, that is, find $\beta(1)$.

(c) Show that $\beta(1) \to 1$ as $n \to \infty$.

9. Let $\widehat{\theta}$ be the MLE of a parameter θ and let $\widehat{\mathsf{se}} = \{nI(\widehat{\theta})\}^{-1/2}$ where $I(\theta)$ is the Fisher information. Consider testing

$$H_0 : \theta = \theta_0 \text{ versus } \theta \neq \theta_0.$$

Consider the Wald test with rejection region $R = \{x^n : \ |Z| > z_{\alpha/2}\}$ where $Z = (\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}}$. Let $\theta_1 > \theta_0$ be some alternative. Show that $\beta(\theta_1) \to 1$.

10. Here are the number of elderly Jewish and Chinese women who died just before and after the Chinese Harvest Moon Festival.

Week	Chinese	Jewish
-2	55	141
-1	33	145
1	70	139
2	49	161

Compare the two mortality patterns. (Phillips and Smith (1990)).

11. A randomized, double-blind experiment was conducted to assess the effectiveness of several drugs for reducing postoperative nausea. The data are as follows.

	Number of Patients	Incidence of Nausea
Placebo	80	45
Chlorpromazine	75	26
Dimenhydrinate	85	52
Pentobarbital (100 mg)	67	35
Pentobarbital (150 mg)	85	37

(a) Test each drug versus the placebo at the 5 per cent level. Also, report the estimated odds–ratios. Summarize your findings.

(b) Use the Bonferroni and the FDR method to adjust for multiple testing. (Beecher (1959)).

12. Let $X_1, ..., X_n \sim \text{Poisson}(\lambda)$.

(a) Let $\lambda_0 > 0$. Find the size α Wald test for

$$H_0 : \lambda = \lambda_0 \quad \text{versus} \quad H_1 : \lambda \neq \lambda_0.$$

(b) (Computer Experiment.) Let $\lambda_0 = 1$, $n = 20$ and $\alpha = .05$. Simulate $X_1, \ldots, X_n \sim \text{Poisson}(\lambda_0)$ and perform the Wald test. Repeat many times and count how often you reject the null. How close is the type I error rate to .05?

13. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Construct the likelihood ratio test for

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

Compare to the Wald test.

14. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Construct the likelihood ratio test for

$$H_0 : \sigma = \sigma_0 \quad \text{versus} \quad H_1 : \sigma \neq \sigma_0.$$

Compare to the Wald test.

15. Let $X \sim \text{Binomial}(n, p)$. Construct the likelihood ratio test for

$$H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0.$$

Compare to the Wald test.

16. Let θ be a scalar parameter and suppose we test

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Let W be the Wald test statistic and let λ be the likelihood ratio test statistic. Show that these tests are equivalent in the sense that

$$\frac{W^2}{\lambda} \xrightarrow{\text{P}} 1$$

as $n \to \infty$. Hint: Use a Taylor expansion of the log-likelihood $\ell(\theta)$ to show that

$$\lambda \approx \left(\sqrt{n}(\widehat{\theta} - \theta_0)\right)^2 \left(-\frac{1}{n}\ell''(\widehat{\theta})\right).$$

11
Bayesian Inference

11.1 The Bayesian Philosophy

The statistical methods that we have discussed so far are known as **frequentist (or classical)** methods. The frequentist point of view is based on the following postulates:

F1 Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.

F2 Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.

F3 Statistical procedures should be designed to have well-defined long run frequency properties. For example, a 95 percent confidence interval should trap the true value of the parameter with limiting frequency at least 95 percent.

There is another approach to inference called **Bayesian inference.** The Bayesian approach is based on the following postulates:

B1 Probability describes degree of belief, not limiting frequency. As such, we can make probability statements about lots of things, not just data which are subject to random variation. For example, I might say that "the probability that Albert Einstein drank a cup of tea on August 1, 1948" is .35. This does not refer to any limiting frequency. It reflects my strength of belief that the proposition is true.

B2 We can make probability statements about parameters, even though they are fixed constants.

B3 We make inferences about a parameter θ by producing a probability distribution for θ. Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

Bayesian inference is a controversial approach because it inherently embraces a subjective notion of probability. In general, Bayesian methods provide no guarantees on long run performance. The field of statistics puts more emphasis on frequentist methods although Bayesian methods certainly have a presence. Certain data mining and machine learning communities seem to embrace Bayesian methods very strongly. Let's put aside philosophical arguments for now and see how Bayesian inference is done. We'll conclude this chapter with some discussion on the strengths and weaknesses of the Bayesian approach.

11.2 The Bayesian Method

Bayesian inference is usually carried out in the following way.

1. We choose a probability density $f(\theta)$ — called the **prior distribution** — that expresses our beliefs about a parameter θ before we see any data.

2. We choose a statistical model $f(x|\theta)$ that reflects our beliefs about x given θ. Notice that we now write this as $f(x|\theta)$ instead of $f(x;\theta)$.

3. After observing data $X_1, \ldots, X_n$, we update our beliefs and calculate the **posterior** distribution $f(\theta|X_1, \ldots, X_n)$.

To see how the third step is carried out, first suppose that θ is discrete and that there is a single, discrete observation X. We should use a capital letter

now to denote the parameter since we are treating it like a random variable, so let Θ denote the parameter. Now, in this discrete setting,

$$\begin{aligned}\mathbb{P}(\Theta = \theta | X = x) &= \frac{\mathbb{P}(X = x, \Theta = \theta)}{\mathbb{P}(X = x)} \\ &= \frac{\mathbb{P}(X = x|\Theta = \theta)\mathbb{P}(\Theta = \theta)}{\sum_\theta \mathbb{P}(X = x|\Theta = \theta)\mathbb{P}(\Theta = \theta)}\end{aligned}$$

which you may recognize from Chapter 1 as **Bayes' theorem.** The version for continuous variables is obtained by using density functions:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}. \tag{11.1}$$

If we have n IID observations $X_1, \ldots, X_n$, we replace $f(x|\theta)$ with

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta) = \mathcal{L}_n(\theta).$$

NOTATION. We will write X^n to mean $(X_1, \ldots, X_n)$ and x^n to mean $(x_1, \ldots, x_n)$. Now,

$$f(\theta|x^n) = \frac{f(x^n|\theta)f(\theta)}{\int f(x^n|\theta)f(\theta)d\theta} = \frac{\mathcal{L}_n(\theta)f(\theta)}{c_n} \propto \mathcal{L}_n(\theta)f(\theta) \tag{11.2}$$

where

$$c_n = \int \mathcal{L}_n(\theta)f(\theta)d\theta \tag{11.3}$$

is called the **normalizing constant**. Note that c_n does not depend on θ. We can summarize by writing:

> Posterior is proportional to Likelihood times Prior
>
> or, in symbols,
>
> $$f(\theta|x^n) \propto \mathcal{L}(\theta)f(\theta).$$

You might wonder, doesn't it cause a problem to throw away the constant c_n? The answer is that we can always recover the constant later if we need to.

What do we do with the posterior distribution? First, we can get a point estimate by summarizing the center of the posterior. Typically, we use the mean or mode of the posterior. The posterior mean is

$$\overline{\theta}_n = \int \theta f(\theta|x^n)d\theta = \frac{\int \theta\mathcal{L}_n(\theta)f(\theta)}{\int \mathcal{L}_n(\theta)f(\theta)d\theta}. \tag{11.4}$$

We can also obtain a Bayesian interval estimate. We find a and b such that $\int_{-\infty}^{a} f(\theta|x^n)d\theta = \int_b^{\infty} f(\theta|x^n)d\theta = \alpha/2$. Let $C = (a, b)$. Then

$$\mathbb{P}(\theta \in C|x^n) = \int_a^b f(\theta|x^n)\, d\theta = 1 - \alpha$$

so C **is a** $1-\alpha$ **posterior interval.**

11.1 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Suppose we take the uniform distribution $f(p) = 1$ as a prior. By Bayes' theorem, the posterior has the form

$$f(p|x^n) \propto f(p)\mathcal{L}_n(p) = p^s(1-p)^{n-s} = p^{s+1-1}(1-p)^{n-s+1-1}$$

where $s = \sum_{i=1}^n x_i$ is the number of successes. Recall that a random variable has a Beta distribution with parameters α and β if its density is

$$f(p;\, \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

We see that the posterior for p is a Beta distribution with parameters $s+1$ and $n-s+1$. That is,

$$f(p|x^n) = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} p^{(s+1)-1}(1-p)^{(n-s+1)-1}.$$

We write this as

$$p|x^n \sim \text{Beta}(s+1, n-s+1).$$

Notice that we have figured out the normalizing constant without actually doing the integral $\int \mathcal{L}_n(p)f(p)dp$. The mean of a $\text{Beta}(\alpha, \beta)$ distribution is $\alpha/(\alpha+\beta)$ so the Bayes estimator is

$$\overline{p} = \frac{s+1}{n+2}. \tag{11.5}$$

It is instructive to rewrite the estimator as

$$\overline{p} = \lambda_n \widehat{p} + (1-\lambda_n)\widetilde{p} \tag{11.6}$$

where $\widehat{p} = s/n$ is the MLE, $\widetilde{p} = 1/2$ is the prior mean and $\lambda_n = n/(n+2) \approx 1$. A 95 percent posterior interval can be obtained by numerically finding a and b such that $\int_a^b f(p|x^n)\, dp = .95$.

Suppose that instead of a uniform prior, we use the prior $p \sim \text{Beta}(\alpha, \beta)$. If you repeat the calculations above, you will see that $p|x^n \sim \text{Beta}(\alpha+s, \beta+$

$n-s$). The flat prior is just the special case with $\alpha=\beta=1$. The posterior mean is

$$\overline{p}=\frac{\alpha+s}{\alpha+\beta+n}=\left(\frac{n}{\alpha+\beta+n}\right)\widehat{p}+\left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)p_0$$

where $p_0=\alpha/(\alpha+\beta)$ is the prior mean. ■

In the previous example, the prior was a Beta distribution and the posterior was a Beta distribution. When the prior and the posterior are in the same family, we say that the prior is **conjugate** with respect to the model.

11.2 Example. Let $X_1,\ldots,X_n\sim N(\theta,\sigma^2)$. For simplicity, let us assume that σ is known. Suppose we take as a prior $\theta\sim N(a,b^2)$. In problem 1 in the exercises it is shown that the posterior for θ is

$$\theta|X^n\sim N(\overline{\theta},\tau^2) \tag{11.7}$$

where

$$\overline{\theta}=w\overline{X}+(1-w)a,$$

$$w=\frac{\frac{1}{\mathsf{se}^2}}{\frac{1}{\mathsf{se}^2}+\frac{1}{b^2}},\quad \frac{1}{\tau^2}=\frac{1}{\mathsf{se}^2}+\frac{1}{b^2},$$

and $\mathsf{se}=\sigma/\sqrt{n}$ is the standard error of the MLE $\overline{X}$. This is another example of a conjugate prior. Note that $w\to 1$ and $\tau/\mathsf{se}\to 1$ as $n\to\infty$. So, for large n, the posterior is approximately $N(\widehat{\theta},\mathsf{se}^2)$. The same is true if n is fixed but $b\to\infty$, which corresponds to letting the prior become very flat.

Continuing with this example, let us find $C=(c,d)$ such that $\mathbb{P}(\theta\in C|X^n)=.95$. We can do this by choosing c and d such that $\mathbb{P}(\theta<c|X^n)=.025$ and $\mathbb{P}(\theta>d|X^n)=.025$. So, we want to find c such that

$$\begin{aligned}\mathbb{P}(\theta<c|X^n) &= \mathbb{P}\left(\frac{\theta-\overline{\theta}}{\tau}<\frac{c-\overline{\theta}}{\tau}\;\middle|\;X^n\right)\\ &= \mathbb{P}\left(Z<\frac{c-\overline{\theta}}{\tau}\right)=.025.\end{aligned}$$

We know that $\mathbb{P}(Z<-1.96)=.025$. So,

$$\frac{c-\overline{\theta}}{\tau}=-1.96$$

implying that $c=\overline{\theta}-1.96\tau$. By similar arguments, $d=\overline{\theta}+1.96$. So a 95 percent Bayesian interval is $\overline{\theta}\pm 1.96\,\tau$. Since $\overline{\theta}\approx\widehat{\theta}$ and $\tau\approx\mathsf{se}$, the 95 percent Bayesian interval is approximated by $\widehat{\theta}\pm 1.96\,\mathsf{se}$ which is the frequentist confidence interval. ■

11.3 Functions of Parameters

How do we make inferences about a function $\tau = g(\theta)$? Remember in Chapter 3 we solved the following problem: given the density f_X for X, find the density for $Y = g(X)$. We now simply apply the same reasoning. The posterior CDF for τ is

$$H(\tau|x^n) = \mathbb{P}(g(\theta) \leq \tau|x^n) = \int_A f(\theta|x^n)d\theta$$

where $A = \{\theta : \ g(\theta) \leq \tau\}$. The posterior density is $h(\tau|x^n) = H'(\tau|x^n)$.

11.3 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and $f(p) = 1$ so that $p|X^n \sim \text{Beta}(s+1, n-s+1)$ with $s = \sum_{i=1}^n x_i$. Let $\psi = \log(p/(1-p))$. Then

$$\begin{aligned} H(\psi|x^n) &= \mathbb{P}(\Psi \leq \psi|x^n) = \mathbb{P}\left(\log\left(\frac{P}{1-P}\right) \leq \psi \ \middle| \ x^n\right) \\ &= \mathbb{P}\left(P \leq \frac{e^\psi}{1+e^\psi} \ \middle| \ x^n\right) \\ &= \int_0^{e^\psi/(1+e^\psi)} f(p|x^n)\, dp \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \int_0^{e^\psi/(1+e^\psi)} p^s(1-p)^{n-s}\, dp \end{aligned}$$

and

$$\begin{aligned} h(\psi|x^n) &= H'(\psi|x^n) \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^s \left(\frac{1}{1+e^\psi}\right)^{n-s} \left(\frac{\partial\left(\frac{e^\psi}{1+e^\psi}\right)}{\partial\psi}\right) \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^s \left(\frac{1}{1+e^\psi}\right)^{n-s} \left(\frac{1}{1+e^\psi}\right)^2 \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^s \left(\frac{1}{1+e^\psi}\right)^{n-s+2} \end{aligned}$$

for $\psi \in \mathbb{R}$. ■

11.4 Simulation

The posterior can often be approximated by simulation. Suppose we draw $\theta_1, \ldots, \theta_B \sim p(\theta|x^n)$. Then a histogram of $\theta_1, \ldots, \theta_B$ approximates the posterior density $p(\theta|x^n)$. An approximation to the posterior mean $\overline{\theta}_n = \mathbb{E}(\theta|x^n)$ is

$B^{-1}\sum_{j=1}^{B}\theta_j$. The posterior $1-\alpha$ interval can be approximated by $(\theta_{\alpha/2}, \theta_{1-\alpha/2})$ where $\theta_{\alpha/2}$ is the $\alpha/2$ sample quantile of $\theta_1, \ldots, \theta_B$.

Once we have a sample $\theta_1, \ldots, \theta_B$ from $f(\theta|x^n)$, let $\tau_i = g(\theta_i)$. Then $\tau_1, \ldots, \tau_B$ is a sample from $f(\tau|x^n)$. This avoids the need to do any analytical calculations. Simulation is discussed in more detail in Chapter 24.

11.4 Example. Consider again Example 11.3. We can approximate the posterior for ψ without doing any calculus. Here are the steps:

1. Draw $P_1, \ldots, P_B \sim \text{Beta}(s+1, n-s+1)$.
2. Let $\psi_i = \log(P_i/(1-P_i))$ for $i = 1, \ldots, B$.

Now $\psi_1, \ldots, \psi_B$ are IID draws from $h(\psi|x^n)$. A histogram of these values provides an estimate of $h(\psi|x^n)$. ■

11.5 Large Sample Properties of Bayes' Procedures

In the Bernoulli and Normal examples we saw that the posterior mean was close to the MLE. This is true in greater generality.

11.5 Theorem. *Let $\widehat{\theta}_n$ be the* MLE *and let $\widehat{\mathsf{se}} = 1/\sqrt{nI(\widehat{\theta}_n)}$. Under appropriate regularity conditions, the posterior is approximately Normal with mean $\widehat{\theta}_n$ and standard deviation $\widehat{\mathsf{se}}$. Hence, $\overline{\theta}_n \approx \widehat{\theta}_n$. Also, if $C_n = (\widehat{\theta}_n - z_{\alpha/2}\widehat{\mathsf{se}}, \widehat{\theta}_n + z_{\alpha/2}\widehat{\mathsf{se}})$ is the asymptotic frequentist $1-\alpha$ confidence interval, then C_n is also an approximate $1-\alpha$ Bayesian posterior interval:*

$$\mathbb{P}(\theta \in C_n|X^n) \to 1-\alpha.$$

There is also a Bayesian delta method. Let $\tau = g(\theta)$. Then

$$\tau|X^n \approx N(\widehat{\tau}, \widetilde{\mathsf{se}}^2)$$

where $\widehat{\tau} = g(\widehat{\theta})$ and $\widetilde{\mathsf{se}} = \widehat{\mathsf{se}}\,|g'(\widehat{\theta})|$.

11.6 Flat Priors, Improper Priors, and "Noninformative" Priors

An important question in Bayesian inference is: where does one get the prior $f(\theta)$? One school of thought, called **subjectivism** says that the prior should

reflect our subjective opinion about θ (before the data are collected). This may be possible in some cases but is impractical in complicated problems especially if there are many parameters. Moreover, injecting subjective opinion into the analysis is contrary to the goal of making scientific inference as objective as possible. An alternative is to try to define some sort of "noninformative prior." An obvious candidate for a noninformative prior is to use a flat prior $f(\theta) \propto$ constant.

In the Bernoulli example, taking $f(p) = 1$ leads to $p|X^n \sim \text{Beta}(s+1, n-s+1)$ as we saw earlier, which seemed very reasonable. But unfettered use of flat priors raises some questions.

Improper Priors. Let $X \sim N(\theta, \sigma^2)$ with σ known. Suppose we adopt a flat prior $f(\theta) \propto c$ where $c > 0$ is a constant. Note that $\int f(\theta)d\theta = \infty$ so this is not a probability density in the usual sense. We call such a prior an **improper prior.** Nonetheless, we can still formally carry out Bayes' theorem and compute the posterior density by multiplying the prior and the likelihood: $f(\theta) \propto \mathcal{L}_n(\theta) f(\theta) \propto \mathcal{L}_n(\theta)$. This gives $\theta|X^n \sim N(\overline{X}, \sigma^2/n)$ and the resulting point and interval estimators agree exactly with their frequentist counterparts. In general, improper priors are not a problem as long as the resulting posterior is a well-defined probability distribution.

Flat Priors are Not Invariant. Let $X \sim \text{Bernoulli}(p)$ and suppose we use the flat prior $f(p) = 1$. This flat prior presumably represents our lack of information about p before the experiment. Now let $\psi = \log(p/(1-p))$. This is a transformation of p and we can compute the resulting distribution for ψ, namely,

$$f_\Psi(\psi) = \frac{e^\psi}{(1+e^\psi)^2}$$

which is not flat. But if we are ignorant about p then we are also ignorant about ψ so we should use a flat prior for ψ. This is a contradiction. In short, the notion of a flat prior is not well defined because a flat prior on a parameter does not imply a flat prior on a transformed version of the parameter. Flat priors are not **transformation invariant.**

Jeffreys' Prior. Jeffreys came up with a rule for creating priors. The rule is: take

$$f(\theta) \propto I(\theta)^{1/2}$$

where $I(\theta)$ is the Fisher information function. This rule turns out to be transformation invariant. There are various reasons for thinking that this prior might be a useful prior but we will not go into details here.

11.6 Example. Consider the Bernoulli (p) model. Recall that

$$I(p) = \frac{1}{p(1-p)}.$$

Jeffreys' rule says to use the prior

$$f(p) \propto \sqrt{I(p)} = p^{-1/2}(1-p)^{-1/2}.$$

This is a Beta (1/2,1/2) density. This is very close to a uniform density. ∎

In a multiparameter problem, the Jeffreys' prior is defined to be $f(\theta) \propto \sqrt{|I(\theta)|}$ where $|A|$ denotes the determinant of a matrix A and $I(\theta)$ is the Fisher information matrix.

11.7 Multiparameter Problems

Suppose that $\theta = (\theta_1, \ldots, \theta_p)$. The posterior density is still given by

$$f(\theta|x^n) \propto \mathcal{L}_n(\theta) f(\theta). \tag{11.8}$$

The question now arises of how to extract inferences about one parameter. The key is to find the marginal posterior density for the parameter of interest. Suppose we want to make inferences about θ_1. The marginal posterior for θ_1 is

$$f(\theta_1|x^n) = \int \cdots \int f(\theta_1, \cdots, \theta_p|x^n) d\theta_2 \ldots d\theta_p. \tag{11.9}$$

In practice, it might not be feasible to do this integral. Simulation can help. Draw randomly from the posterior:

$$\theta^1, \ldots, \theta^B \sim f(\theta|x^n)$$

where the superscripts index the different draws. Each θ^j is a vector $\theta^j = (\theta_1^j, \ldots, \theta_p^j)$. Now collect together the first component of each draw:

$$\theta_1^1, \ldots, \theta_1^B.$$

These are a sample from $f(\theta_1|x^n)$ and we have avoided doing any integrals.

11.7 Example (Comparing Two Binomials)**.** Suppose we have n_1 control patients and n_2 treatment patients and that X_1 control patients survive while X_2 treatment patients survive. We want to estimate $\tau = g(p_1, p_2) = p_2 - p_1$. Then,

$$X_1 \sim \text{Binomial}(n_1, p_1) \quad \text{and} \quad X_2 \sim \text{Binomial}(n_2, p_2).$$

If $f(p_1, p_2) = 1$, the posterior is

$$f(p_1, p_2|x_1, x_2) \propto p_1^{x_1}(1-p_1)^{n_1-x_1} p_2^{x_2}(1-p_2)^{n_2-x_2}.$$

Notice that (p_1, p_2) live on a rectangle (a square, actually) and that

$$f(p_1, p_2|x_1, x_2) = f(p_1|x_1) f(p_2|x_2)$$

where

$$f(p_1|x_1) \propto p_1^{x_1}(1-p_1)^{n_1-x_1} \quad \text{and} \quad f(p_2|x_2) \propto p_2^{x_2}(1-p_2)^{n_2-x_2}$$

which implies that p_1 and p_2 are independent under the posterior. Also, $p_1|x_1 \sim \text{Beta}(x_1+1, n_1-x_1+1)$ and $p_2|x_2 \sim \text{Beta}(x_2+1, n_2-x_2+1)$. If we simulate $P_{1,1}, \ldots, P_{1,B} \sim \text{Beta}(x_1+1, n_1-x_1+1)$ and $P_{2,1}, \ldots, P_{2,B} \sim \text{Beta}(x_2+1, n_2-x_2+1)$, then $\tau_b = P_{2,b} - P_{1,b}$, $b = 1, \ldots, B$, is a sample from $f(\tau|x_1, x_2)$. ■

11.8 Bayesian Testing

Hypothesis testing from a Bayesian point of view is a complex topic. We will only give a brief sketch of the main idea here. The Bayesian approach to testing involves putting a prior on H_0 and on the parameter θ and then computing $\mathbb{P}(H_0|X^n)$. Consider the case where θ is scalar and we are testing

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

It is usually reasonable to use the prior $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$ (although this is not essential in what follows). Under H_1 we need a prior for θ. Denote this prior density by $f(\theta)$. From Bayes' theorem

$$\begin{aligned}
\mathbb{P}(H_0|X^n = x^n) &= \frac{f(x^n|H_0)\mathbb{P}(H_0)}{f(x^n|H_0)\mathbb{P}(H_0) + f(x^n|H_1)\mathbb{P}(H_1)} \\
&= \frac{\frac{1}{2}f(x^n \mid \theta_0)}{\frac{1}{2}f(x^n \mid \theta_0) + \frac{1}{2}f(x^n \mid H_1)} \\
&= \frac{f(x^n \mid \theta_0)}{f(x^n \mid \theta_0) + \int f(x^n \mid \theta) f(\theta) d\theta} \\
&= \frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_0) + \int \mathcal{L}(\theta) f(\theta) d\theta}.
\end{aligned}$$

We saw that, in estimation problems, the prior was not very influential and that the frequentist and Bayesian methods gave similar answers. This is not

the case in hypothesis testing. Also, one can't use improper priors in testing because this leads to an undefined constant in the denominator of the expression above. Thus, if you use Bayesian testing you must choose the prior $f(\theta)$ very carefully. It is possible to get a prior-free bound on $\mathbb{P}(H_0|X^n = x^n)$. Notice that $0 \leq \int \mathcal{L}(\theta)f(\theta)d\theta \leq \mathcal{L}(\widehat{\theta})$. Hence,

$$\frac{\mathcal{L}(\theta_0)}{\mathcal{L}(\theta_0) + \mathcal{L}(\widehat{\theta})} \ \leq \ \mathbb{P}(H_0|X^n = x^n) \ \leq \ 1.$$

The upper bound is not very interesting, but the lower bound is non-trivial.

11.9 Strengths and Weaknesses of Bayesian Inference

Bayesian inference is appealing when prior information is available since Bayes' theorem is a natural way to combine prior information with data. Some people find Bayesian inference psychologically appealing because it allows us to make probability statements about parameters. In contrast, frequentist inference provides confidence sets C_n which trap the parameter 95 percent of the time, but we cannot say that $\mathbb{P}(\theta \in C_n|X^n)$ is .95. In the frequentist approach we can make probability statements about C_n, not θ. However, psychological appeal is not a compelling scientific argument for using one type of inference over another.

In parametric models, with large samples, Bayesian and frequentist methods give approximately the same inferences. In general, they need not agree.

Here are three examples that illustrate the strengths and weakness of Bayesian inference. The first example is Example 6.14 revisited. This example shows the psychological appeal of Bayesian inference. The second and third show that Bayesian methods can fail.

11.8 Example (Example 6.14 revisited)**.** We begin by reviewing the example. Let θ be a fixed, known real number and let X_1, X_2 be independent random variables such that $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. Now define $Y_i = \theta + X_i$ and suppose that you only observe Y_1 and Y_2. Let

$$C = \begin{cases} \{Y_1 - 1\} & \text{if } Y_1 = Y_2 \\ \{(Y_1 + Y_2)/2\} & \text{if } Y_1 \neq Y_2. \end{cases}$$

This is a 75 percent confidence set since, no matter what θ is, $\mathbb{P}_\theta(\theta \in C) = 3/4$.

Suppose we observe $Y_1 = 15$ and $Y_2 = 17$. Then our 75 percent confidence interval is $\{16\}$. However, we are certain, in this case, that $\theta = 16$. So calling

this a 75 percent confidence set, bothers many people. Nonetheless, C is a valid 75 percent confidence set. It will trap the true value 75 percent of the time.

The Bayesian solution is more satisfying to many. For simplicity, assume that θ is an integer. Let $f(\theta)$ be a prior mass function such that $f(\theta) > 0$ for every integer θ. When $Y = (Y_1, Y_2) = (15, 17)$, the likelihood function is

$$\mathcal{L}(\theta) = \begin{cases} 1/4 & \theta = 16 \\ 0 & \text{otherwise.} \end{cases}$$

Applying Bayes' theorem we see that

$$\mathbb{P}(\Theta = \theta | Y = (15, 17)) = \begin{cases} 1 & \theta = 16 \\ 0 & \text{otherwise.} \end{cases}$$

Hence, $\mathbb{P}(\theta \in C | Y = (15, 17)) = 1$. There is nothing wrong with saying that $\{16\}$ is a 75 percent confidence interval. But is it not a probability statement about θ. ■

11.9 Example. This is a simplified version of the example in Robins and Ritov (1997). The data consist of n IID triples

$$(X_1, R_1, Y_1), \ldots, (X_n, Y_n, R_n).$$

Let B be a finite but very large number, like $B = 100^{100}$. Any realistic sample size n will be small compared to B. Let

$$\theta = (\theta_1, \ldots, \theta_B)$$

be a vector of unknown parameters such that $0 \leq \theta_j \leq 1$ for $1 \leq j \leq B$. Let

$$\xi = (\xi_1, \ldots, \xi_B)$$

be a vector of **known** numbers such that

$$0 < \delta \leq \xi_j \leq 1 - \delta < 1, \quad 1 \leq j \leq B,$$

where δ is some, small, positive number. Each data point (X_i, R_i, Y_i) is drawn in the following way:

1. Draw X_i uniformly from $\{1, \ldots, B\}$.
2. Draw $R_i \sim \text{Bernoulli}(\xi_{X_i})$.
3. If $R_i = 1$, then draw $Y_i \sim \text{Bernoulli}(\theta_{X_i})$. If $R_i = 0$, do not draw Y_i.

The model may seem a little artificial but, in fact, it is caricature of some real **missing data** problems in which some data points are not observed. In this example, $R_i = 0$ can be thought of as meaning "missing." Our goal is to estimate

$$\psi = \mathbb{P}(Y_i = 1).$$

Note that

$$\begin{aligned} \psi &= \mathbb{P}(Y_i = 1) = \sum_{j=1}^{B} \mathbb{P}(Y_i = 1 | X = j)\mathbb{P}(X = j) \\ &= \frac{1}{B}\sum_{j=1}^{B} \theta_j \equiv g(\theta) \end{aligned}$$

so $\psi = g(\theta)$ is a function of θ.

Let us consider a Bayesian analysis first. The likelihood of a single observation is

$$f(X_i, R_i, Y_i) = f(X_i) f(R_i|X_i) f(Y_i|X_i)^{R_i}.$$

The last term is raised to the power R_i since, if $R_i = 0$, then Y_i is not observed and hence that term drops out of the likelihood. Since $f(X_i) = 1/B$ and that Y_i and R_i are Bernoulli,

$$f(X_i) f(R_i|X_i) f(Y_i|X_i)^{R_i} = \frac{1}{B}\, \xi_{X_i}^{R_i}\, (1 - \xi_{X_i})^{1-R_i}\, \theta_{X_i}^{Y_i R_i}\, (1 - \theta_{X_i})^{(1-Y_i)R_i}.$$

Thus, the likelihood function is

$$\begin{aligned} \mathcal{L}(\theta) &= \prod_{i=1}^{n} f(X_i) f(R_i|X_i) f(Y_i|X_i)^{R_i} \\ &= \prod_{i=1}^{n} \frac{1}{B}\, \xi_{X_i}^{R_i}\, (1 - \xi_{X_i})^{1-R_i}\, \theta_{X_i}^{Y_i R_i}\, (1 - \theta_{X_i})^{(1-Y_i)R_i} \\ &\propto \theta_{X_i}^{Y_i R_i} (1 - \theta_{X_i})^{(1-Y_i)R_i}. \end{aligned}$$

We have dropped all the terms involving B and the ξ_j's since these are known constants, not parameters. The log-likelihood is

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} Y_i\, R_i \log \theta_{X_i} + (1 - Y_i)\, R_i\, \log(1 - \theta_{X_i}) \\ &= \sum_{j=1}^{B} n_j \log \theta_j + \sum_{j=1}^{B} m_j \log(1 - \theta_j) \end{aligned}$$

where

$$\begin{aligned} n_j &= \#\{i:\ Y_i = 1, R_i = 1, X_i = j\} \\ m_j &= \#\{i:\ Y_i = 0, R_i = 1, X_i = j\}. \end{aligned}$$

Now, $n_j = m_j = 0$ for most j since B is so much larger than n. This has several implications. First, the MLE for most θ_j is not defined. Second, for most θ_j, the posterior distribution is equal to the prior distribution, since those θ_j do not appear in the likelihood. Hence, $f(\theta|\text{Data}) \approx f(\theta)$. It follows that $f(\psi|\text{Data}) \approx f(\psi)$. In other words, the data provide little information about ψ in a Bayesian analysis.

Now we consider a frequentist solution. Define

$$\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\xi_{X_i}}. \tag{11.10}$$

We will now show that this estimator is unbiased and has small mean-squared error. It can be shown (see Exercise 7) that

$$\mathbb{E}(\widehat{\psi}) = \psi \quad \text{and} \quad \mathbb{V}(\widehat{\psi}) \leq \frac{1}{n\delta^2}. \tag{11.11}$$

Therefore, the MSE is of order $1/n$ which goes to 0 fairly quickly as we collect more data, no matter how large B is. The estimator defined in (11.10) is called the **Horwitz-Thompson** estimator. It cannot be derived from a Bayesian or likelihood point of view since it involves the terms ξ_{X_i}. These terms drop out of the log-likelihood and hence will not show up in any likelihood-based method including Bayesian estimators.

The moral of the story is this. Bayesian methods are tied to the likelihood function. But in high dimensional (and nonparametric) problems, the likelihood may not yield accurate inferences. ■

11.10 Example. Suppose that f is a probability density function and that

$$f(x) = cg(x)$$

where $g(x) > 0$ is a known function and c is unknown. In principle we can compute c since $\int f(x)\,dx = 1$ implies that $c = 1/\int g(x)\,dx$. But in many cases we can't do the integral $\int g(x)\,dx$ since g might be a complicated function and x could be high dimensional. Despite the fact that c is not known, it is often possible to draw a sample $X_1, \ldots, X_n$ from f; see Chapter 24. Can we use the sample to estimate the normalizing constant c? Here is a frequentist solution:

Let $\widehat{f}_n(x)$ be a consistent estimate of the density f. Chapter 20 explains how to construct such an estimate. Choose any point x and note that $c = f(x)/g(x)$. Hence, $\widehat{c} = \widehat{f}(x)/g(x)$ is a consistent estimate of c. Now let us try to solve this problem from a Bayesian approach. Let $\pi(c)$ be a prior such that $\pi(c) > 0$ for all $c > 0$. The likelihood function is

$$\mathcal{L}_n(c) = \prod_{i=1}^n f(X_i) = \prod_{i=1}^n cg(X_i) = c^n \prod_{i=1}^n g(X_i) \ \propto \ c^n.$$

Hence the posterior is proportional to $c^n\pi(c)$. The posterior does not depend on $X_1, \ldots, X_n$, so we come to the startling conclusion that, from the Bayesian point of view, there is no information in the data about c. Moreover, the posterior mean is

$$\frac{\int_0^\infty c^{n+1}\pi(c)\,dc}{\int_0^\infty c^n\pi(c)\,dc}$$

which tends to infinity as n increases. ■

These last two examples illustrate an important point. Bayesians are slaves to the likelihood function. When the likelihood goes awry, so will Bayesian inference.

What should we conclude from all this? The important thing is to understand that frequentist and Bayesian methods are answering different questions. To combine prior beliefs with data in a principled way, use Bayesian inference. To construct procedures with guaranteed long run performance, such as confidence intervals, use frequentist methods. Generally, Bayesian methods run into problems when the parameter space is high dimensional. In particular, 95 percent posterior intervals need not contain the true value 95 percent of the time (in the frequency sense).

11.10 Bibliographic Remarks

Some references on Bayesian inference include Carlin and Louis (1996), Gelman et al. (1995), Lee (1997), Robert (1994), and Schervish (1995). See Cox (1993), Diaconis and Freedman (1986), Freedman (1999), Barron et al. (1999), Ghosal et al. (2000), Shen and Wasserman (2001), and Zhao (2000) for discussions of some of the technicalities of nonparametric Bayesian inference. The Robins-Ritov example is discussed in detail in Robins and Ritov (1997) where it is cast more properly as a nonparametric problem. Example 11.10 is due to Edward George (personal communication). See Berger and Delampady (1987)

and Kass and Raftery (1995) for a discussion of Bayesian testing. See Kass and Wasserman (1996) for a discussion of noninformative priors.

11.11 Appendix

Proof of Theorem 11.5.

It can be shown that the effect of the prior diminishes as n increases so that $f(\theta|X^n) \propto \mathcal{L}_n(\theta)f(\theta) \approx \mathcal{L}_n(\theta)$. Hence, $\log f(\theta|X^n) \approx \ell(\theta)$. Now, $\ell(\theta) \approx \ell(\widehat{\theta}) + (\theta - \widehat{\theta})\ell'(\widehat{\theta}) + [(\theta - \widehat{\theta})^2/2]\ell''(\widehat{\theta}) = \ell(\widehat{\theta}) + [(\theta - \widehat{\theta})^2/2]\ell''(\widehat{\theta})$ since $\ell'(\widehat{\theta}) = 0$. Exponentiating, we get approximately that

$$f(\theta|X^n) \propto \exp\left\{-\frac{1}{2}\frac{(\theta - \widehat{\theta})^2}{\sigma_n^2}\right\}$$

where $\sigma_n^2 = -1/\ell''(\widehat{\theta}_n)$. So the posterior of θ is approximately Normal with mean $\widehat{\theta}$ and variance σ_n^2. Let $\ell_i = \log f(X_i|\theta)$, then

$$\begin{aligned}
\frac{1}{\sigma_n^2} &= -\ell''(\widehat{\theta}_n) = \sum_i -\ell_i''(\widehat{\theta}_n) \\
&= n\left(\frac{1}{n}\right)\sum_i -\ell_i''(\widehat{\theta}_n) \approx n\mathbb{E}_\theta\left[-\ell_i''(\widehat{\theta}_n)\right] \\
&= nI(\widehat{\theta}_n)
\end{aligned}$$

and hence $\sigma_n \approx \mathsf{se}(\widehat{\theta})$. ∎

11.12 Exercises

1. Verify (11.7).
2. Let $X_1, ..., X_n \sim \text{Normal}(\mu, 1)$.
   (a) Simulate a data set (using $\mu = 5$) consisting of n=100 observations.
   (b) Take $f(\mu) = 1$ and find the posterior density. Plot the density.
   (c) Simulate 1,000 draws from the posterior. Plot a histogram of the simulated values and compare the histogram to the answer in (b).
   (d) Let $\theta = e^\mu$. Find the posterior density for θ analytically and by simulation.
   (e) Find a 95 percent posterior interval for μ.
   (f) Find a 95 percent confidence interval for θ.

3. Let $X_1, ..., X_n \sim \text{Uniform}(0, \theta)$. Let $f(\theta) \propto 1/\theta$. Find the posterior density.

4. Suppose that 50 people are given a placebo and 50 are given a new treatment. 30 placebo patients show improvement while 40 treated patients show improvement. Let $\tau = p_2 - p_1$ where p_2 is the probability of improving under treatment and p_1 is the probability of improving under placebo.

   (a) Find the MLE of τ. Find the standard error and 90 percent confidence interval using the delta method.

   (b) Find the standard error and 90 percent confidence interval using the parametric bootstrap.

   (c) Use the prior $f(p_1, p_2) = 1$. Use simulation to find the posterior mean and posterior 90 percent interval for τ.

   (d) Let
   $$\psi = \log\left(\left(\frac{p_1}{1-p_1}\right) \div \left(\frac{p_2}{1-p_2}\right)\right)$$
   be the log-odds ratio. Note that $\psi = 0$ if $p_1 = p_2$. Find the MLE of ψ. Use the delta method to find a 90 percent confidence interval for ψ.

   (e) Use simulation to find the posterior mean and posterior 90 percent interval for ψ.

5. Consider the Bernoulli(p) observations

   0 1 0 1 0 0 0 0 0 0

   Plot the posterior for p using these priors: Beta(1/2,1/2), Beta(1,1), Beta(10,10), Beta(100,100).

6. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$.

   (a) Let $\lambda \sim \text{Gamma}(\alpha, \beta)$ be the prior. Show that the posterior is also a Gamma. Find the posterior mean.

   (b) Find the Jeffreys' prior. Find the posterior.

7. In Example 11.9, verify (11.11).

8. Let $X \sim N(\mu, 1)$. Consider testing
   $$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0.$$

Take $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$. Let the prior for μ under H_1 be $\mu \sim N(0, b^2)$. Find an expression for $\mathbb{P}(H_0|X = x)$. Compare $\mathbb{P}(H_0|X = x)$ to the p-value of the Wald test. Do the comparison numerically for a variety of values of x and b. Now repeat the problem using a sample of size n. You will see that the posterior probability of H_0 can be large even when the p-value is small, especially when n is large. This disagreement between Bayesian and frequentist testing is called the Jeffreys-Lindley paradox.

12
Statistical Decision Theory

12.1 Preliminaries

We have considered several point estimators such as the maximum likelihood estimator, the method of moments estimator, and the posterior mean. In fact, there are many other ways to generate estimators. How do we choose among them? The answer is found in **decision theory** which is a formal theory for comparing statistical procedures.

Consider a parameter θ which lives in a parameter space Θ. Let $\widehat{\theta}$ be an estimator of θ. In the language of decision theory, an estimator is sometimes called a **decision rule** and the possible values of the decision rule are called **actions**.

We shall measure the discrepancy between θ and $\widehat{\theta}$ using a **loss function** $L(\theta, \widehat{\theta})$. Formally, L maps $\Theta \times \Theta$ into $\mathbb{R}$. Here are some examples of loss functions:

$L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$	squared error loss,
$L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert$	absolute error loss,
$L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert^p$	L_p loss,
$L(\theta, \widehat{\theta}) = 0$ if $\theta = \widehat{\theta}$ or 1 if $\theta \neq \widehat{\theta}$	zero–one loss,
$L(\theta, \widehat{\theta}) = \int \log\left(\frac{f(x;\theta)}{f(x;\widehat{\theta})}\right) f(x;\, \theta)dx$	Kullback–Leibler loss.

Bear in mind in what follows that an estimator $\widehat{\theta}$ is a function of the data. To emphasize this point, sometimes we will write $\widehat{\theta}$ as $\widehat{\theta}(X)$. To assess an estimator, we evaluate the average loss or risk.

12.1 Definition. *The* **risk** *of an estimator* $\widehat{\theta}$ *is*

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta\Big(L(\theta, \widehat{\theta})\Big) = \int L(\theta, \widehat{\theta}(x)) f(x;\theta) dx.$$

When the loss function is squared error, the risk is just the MSE (mean squared error):

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \text{MSE} = \mathbb{V}_\theta(\widehat{\theta}) + \mathsf{bias}^2_\theta(\widehat{\theta}).$$

In the rest of the chapter, if we do not state what loss function we are using, assume the loss function is squared error.

12.2 Comparing Risk Functions

To compare two estimators we can compare their risk functions. However, this does not provide a clear answer as to which estimator is better. Consider the following examples.

12.2 Example. Let $X \sim N(\theta, 1)$ and assume we are using squared error loss. Consider two estimators: $\widehat{\theta}_1 = X$ and $\widehat{\theta}_2 = 3$. The risk functions are $R(\theta, \widehat{\theta}_1) = \mathbb{E}_\theta(X - \theta)^2 = 1$ and $R(\theta, \widehat{\theta}_2) = \mathbb{E}_\theta(3 - \theta)^2 = (3 - \theta)^2$. If $2 < \theta < 4$ then $R(\theta, \widehat{\theta}_2) < R(\theta, \widehat{\theta}_1)$, otherwise, $R(\theta, \widehat{\theta}_1) < R(\theta, \widehat{\theta}_2)$. Neither estimator uniformly dominates the other; see Figure 12.1. ■

12.3 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Consider squared error loss and let $\widehat{p}_1 = \overline{X}$. Since this has 0 bias, we have that

$$R(p, \widehat{p}_1) = \mathbb{V}(\overline{X}) = \frac{p(1-p)}{n}.$$

Another estimator is

$$\widehat{p}_2 = \frac{Y + \alpha}{\alpha + \beta + n}$$

where $Y = \sum_{i=1}^n X_i$ and α and β are positive constants. This is the posterior mean using a Beta (α, β) prior. Now,

$$R(p, \widehat{p}_2) \quad = \quad \mathbb{V}_p(\widehat{p}_2) + (\mathsf{bias}_p(\widehat{p}_2))^2$$

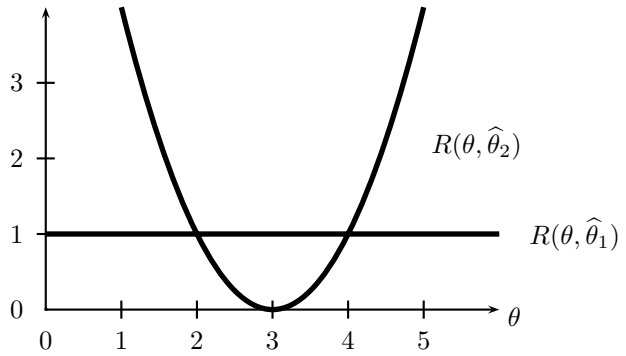


FIGURE 12.1. Comparing two risk functions. Neither risk function dominates the other at all values of θ.

$$= \mathbb{V}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathbb{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2$$
$$= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2.$$

Let $\alpha = \beta = \sqrt{n/4}$. (In Example 12.12 we will explain this choice.) The resulting estimator is

$$\widehat{p}_2 = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

and the risk function is

$$R(p, \widehat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}.$$

The risk functions are plotted in figure 12.2. As we can see, neither estimator uniformly dominates the other.

These examples highlight the need to be able to compare risk functions. To do so, we need a one-number summary of the risk function. Two such summaries are the maximum risk and the Bayes risk.

12.4 Definition. *The* **maximum risk** *is*

$$\overline{R}(\widehat{\theta}) = \sup_\theta R(\theta, \widehat{\theta}) \tag{12.1}$$

and **the Bayes risk** *is*

$$r(f, \widehat{\theta}) = \int R(\theta, \widehat{\theta}) f(\theta) d\theta \tag{12.2}$$

where $f(\theta)$ is a prior for θ.

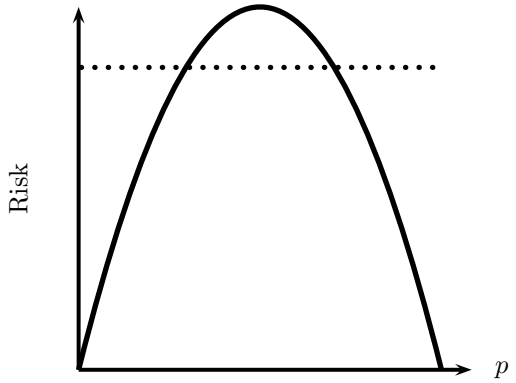


FIGURE 12.2. Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 12.3. The solid curve is $R(\widehat{p}_1)$. The dotted line is $R(\widehat{p}_2)$.

12.5 Example. Consider again the two estimators in Example 12.3. We have

$$\overline{R}(\widehat{p}_1) = \max_{0\leq p\leq 1} \frac{p(1-p)}{n} = \frac{1}{4n}$$

and

$$\overline{R}(\widehat{p}_2) = \max_{p} \frac{n}{4(n+\sqrt{n})^2} = \frac{n}{4(n+\sqrt{n})^2}.$$

Based on maximum risk, $\widehat{p}_2$ is a better estimator since $\overline{R}(\widehat{p}_2) < \overline{R}(\widehat{p}_1)$. However, when n is large, $\overline{R}(\widehat{p}_1)$ has smaller risk except for a small region in the parameter space near $p = 1/2$. Thus, many people prefer $\widehat{p}_1$ to $\widehat{p}_2$. This illustrates that one-number summaries like maximum risk are imperfect. Now consider the Bayes risk. For illustration, let us take $f(p) = 1$. Then

$$r(f, \widehat{p}_1) = \int R(p, \widehat{p}_1)dp = \int \frac{p(1-p)}{n} dp = \frac{1}{6n}$$

and

$$r(f, \widehat{p}_2) = \int R(p, \widehat{p}_2)dp = \frac{n}{4(n+\sqrt{n})^2}.$$

For $n \geq 20$, $r(f, \widehat{p}_2) > r(f, \widehat{p}_1)$ which suggests that $\widehat{p}_1$ is a better estimator. This might seem intuitively reasonable but this answer depends on the choice of prior. The advantage of using maximum risk, despite its problems, is that it does not require one to choose a prior. ■

These two summaries of the risk function suggest two different methods for devising estimators: choosing $\widehat{\theta}$ to minimize the maximum risk leads to

minimax estimators; choosing $\widehat{\theta}$ to minimize the Bayes risk leads to Bayes estimators.

12.6 Definition. *A decision rule that minimizes the Bayes risk is called a* **Bayes rule**. *Formally, $\widehat{\theta}$ is a Bayes rule with respect to the prior f if*

$$r(f,\widehat{\theta}) = \inf_{\widetilde{\theta}} r(f,\widetilde{\theta}) \tag{12.3}$$

where the infimum is over all estimators $\widetilde{\theta}$. An estimator that minimizes the maximum risk is called a **minimax rule**. *Formally, $\widehat{\theta}$ is minimax if*

$$\sup_{\theta} R(\theta,\widehat{\theta}) = \inf_{\widetilde{\theta}} \sup_{\theta} R(\theta,\widetilde{\theta}) \tag{12.4}$$

where the infimum is over all estimators $\widetilde{\theta}$.

12.3 Bayes Estimators

Let f be a prior. From Bayes' theorem, the posterior density is

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{m(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \tag{12.5}$$

where $m(x) = \int f(x,\theta)d\theta = \int f(x|\theta)f(\theta)d\theta$ is the **marginal distribution** of X. Define the **posterior risk** of an estimator $\widehat{\theta}(x)$ by

$$r(\widehat{\theta}|x) = \int L(\theta,\widehat{\theta}(x))f(\theta|x)d\theta. \tag{12.6}$$

12.7 Theorem. *The Bayes risk $r(f,\widehat{\theta})$ satisfies*

$$r(f,\widehat{\theta}) = \int r(\widehat{\theta}|x)m(x)\,dx.$$

Let $\widehat{\theta}(x)$ be the value of θ that minimizes $r(\widehat{\theta}|x)$. Then $\widehat{\theta}$ is the Bayes estimator.

PROOF. We can rewrite the Bayes risk as follows:

$$\begin{aligned} r(f,\widehat{\theta}) &= \int R(\theta,\widehat{\theta})f(\theta)d\theta = \int\left(\int L(\theta,\widehat{\theta}(x))f(x|\theta)dx\right)f(\theta)d\theta \\ &= \int\int L(\theta,\widehat{\theta}(x))f(x,\theta)dxd\theta = \int\int L(\theta,\widehat{\theta}(x))f(\theta|x)m(x)dxd\theta \\ &= \int\left(\int L(\theta,\widehat{\theta}(x))f(\theta|x)d\theta\right)m(x)\,dx = \int r(\widehat{\theta}|x)m(x)\,dx. \end{aligned}$$

If we choose $\widehat{\theta}(x)$ to be the value of θ that minimizes $r(\widehat{\theta}|x)$ then we will minimize the integrand at every x and thus minimize the integral $\int r(\widehat{\theta}|x)m(x)dx$. ■

Now we can find an explicit formula for the Bayes estimator for some specific loss functions.

12.8 Theorem. *If $L(\theta,\widehat{\theta}) = (\theta-\widehat{\theta})^2$ then the Bayes estimator is*

$$\widehat{\theta}(x) = \int \theta f(\theta|x)d\theta = \mathbb{E}(\theta|X=x). \tag{12.7}$$

If $L(\theta,\widehat{\theta}) = |\theta-\widehat{\theta}|$ then the Bayes estimator is the median of the posterior $f(\theta|x)$. If $L(\theta,\widehat{\theta})$ is zero–one loss, then the Bayes estimator is the mode of the posterior $f(\theta|x)$.

PROOF. We will prove the theorem for squared error loss. The Bayes rule $\widehat{\theta}(x)$ minimizes $r(\widehat{\theta}|x) = \int(\theta-\widehat{\theta}(x))^2 f(\theta|x)d\theta$. Taking the derivative of $r(\widehat{\theta}|x)$ with respect to $\widehat{\theta}(x)$ and setting it equal to 0 yields the equation $2\int(\theta-\widehat{\theta}(x))f(\theta|x)d\theta = 0$. Solving for $\widehat{\theta}(x)$ we get 12.7. ■

12.9 Example. Let $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$ where σ^2 is known. Suppose we use a $N(a,b^2)$ prior for μ. The Bayes estimator with respect to squared error loss is the posterior mean, which is

$$\widehat{\theta}(X_1,\ldots,X_n) = \frac{b^2}{b^2+\frac{\sigma^2}{n}}\overline{X} + \frac{\frac{\sigma^2}{n}}{b^2+\frac{\sigma^2}{n}}a. \quad ■$$

12.4 Minimax Rules

Finding minimax rules is complicated and we cannot attempt a complete coverage of that theory here but we will mention a few key results. The main message to take away from this section is: Bayes estimators with a constant risk function are minimax.

12.10 Theorem. *Let $\widehat{\theta}^f$ be the Bayes rule for some prior f:*

$$r(f,\widehat{\theta}^f) = \inf_{\widehat{\theta}} r(f,\widehat{\theta}). \tag{12.8}$$

Suppose that

$$R(\theta,\widehat{\theta}^f) \leq r(f,\widehat{\theta}^f) \text{ for all } \theta. \tag{12.9}$$

Then $\widehat{\theta}^f$ is minimax and f is called a **least favorable prior**.

PROOF. Suppose that $\widehat{\theta}^f$ is not minimax. Then there is another rule $\widehat{\theta}_0$ such that $\sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}^f)$. Since the average of a function is always less than or equal to its maximum, we have that $r(f, \widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0)$. Hence,

$$r(f, \widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}^f) \leq r(f, \widehat{\theta}^f)$$

which contradicts (12.8). ■

12.11 Theorem. *Suppose that $\widehat{\theta}$ is the Bayes rule with respect to some prior f. Suppose further that $\widehat{\theta}$ has constant risk: $R(\theta, \widehat{\theta}) = c$ for some c. Then $\widehat{\theta}$ is minimax.*

PROOF. The Bayes risk is $r(f, \widehat{\theta}) = \int R(\theta, \widehat{\theta}) f(\theta) d\theta = c$ and hence $R(\theta, \widehat{\theta}) \leq r(f, \widehat{\theta})$ for all θ. Now apply the previous theorem. ■

12.12 Example. Consider the Bernoulli model with squared error loss. In example 12.3 we showed that the estimator

$$\widehat{p}(X^n) = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}}$$

has a constant risk function. This estimator is the posterior mean, and hence the Bayes rule, for the prior Beta(α, β) with $\alpha = \beta = \sqrt{n/4}$. Hence, by the previous theorem, this estimator is minimax. ■

12.13 Example. Consider again the Bernoulli but with loss function

$$L(p, \widehat{p}) = \frac{(p - \widehat{p})^2}{p(1-p)}.$$

Let

$$\widehat{p}(X^n) = \widehat{p} = \frac{\sum_{i=1}^n X_i}{n}.$$

The risk is

$$R(p, \widehat{p}) = E\left(\frac{(\widehat{p} - p)^2}{p(1-p)}\right) = \frac{1}{p(1-p)}\left(\frac{p(1-p)}{n}\right) = \frac{1}{n}$$

which, as a function of p, is constant. It can be shown that, for this loss function, $\widehat{p}(X^n)$ is the Bayes estimator under the prior $f(p) = 1$. Hence, $\widehat{p}$ is minimax. ■

A natural question to ask is: what is the minimax estimator for a Normal model?

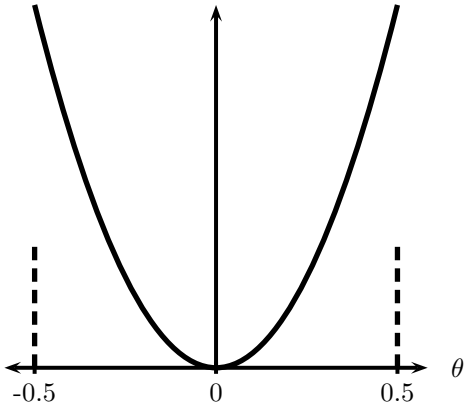


FIGURE 12.3. Risk function for constrained Normal with m=.5. The two short dashed lines show the least favorable prior which puts its mass at two points.

12.14 Theorem. *Let $X_1, \ldots, X_n \sim N(\theta, 1)$ and let $\widehat{\theta} = \overline{X}$. Then $\widehat{\theta}$ is minimax with respect to any well-behaved loss function.* [1] *It is the only estimator with this property.*

If the parameter space is restricted, then the theorem above does not apply as the next example shows.

12.15 Example. Suppose that $X \sim N(\theta, 1)$ and that θ is known to lie in the interval $[-m, m]$ where $0 < m < 1$. The unique, minimax estimator under squared error loss is

$$\widehat{\theta}(X) = m \tanh(mX)$$

where $\tanh(z) = (e^z - e^{-z})/(e^z + e^{-z})$. It can be shown that this is the Bayes rule with respect to the prior that puts mass $1/2$ at m and mass $1/2$ at $-m$. Moreover, it can be shown that the risk is not constant but it does satisfy $R(\theta, \widehat{\theta}) \leq r(f, \widehat{\theta})$ for all θ; see Figure 12.3. Hence, Theorem 12.10 implies that $\widehat{\theta}$ is minimax. ■

[1] "Well-behaved" means that the level sets must be convex and symmetric about the origin. The result holds up to sets of measure 0.

12.5 Maximum Likelihood, Minimax, and Bayes

For parametric models that satisfy weak regularity conditions, the maximum likelihood estimator is approximately minimax. Consider squared error loss which is squared bias plus variance. In parametric models with large samples, it can be shown that the variance term dominates the bias so the risk of the MLE $\widehat{\theta}$ roughly equals the variance:[2]

$$R(\theta, \widehat{\theta}) = \mathbb{V}_\theta(\widehat{\theta}) + \mathsf{bias}^2 \approx \mathbb{V}_\theta(\widehat{\theta}).$$

As we saw in Chapter 9, the variance of the MLE is approximately

$$\mathbb{V}(\widehat{\theta}) \approx \frac{1}{nI(\theta)}$$

where $I(\theta)$ is the Fisher information. Hence,

$$nR(\theta, \widehat{\theta}) \approx \frac{1}{I(\theta)}. \tag{12.10}$$

For any other estimator θ', it can be shown that for large n, $R(\theta, \theta') \geq R(\theta, \widehat{\theta})$. More precisely,

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} \sup_{|\theta - \theta'| < \epsilon} n\, R(\theta', \widehat{\theta}) \geq \frac{1}{I(\theta)}. \tag{12.11}$$

This says that, in a local, large sample sense, the MLE is minimax. It can also be shown that the MLE is approximately the Bayes rule.

In summary:

> In most parametric models, with large samples, the MLE is approximately minimax and Bayes.

There is a caveat: these results break down when the number of parameters is large as the next example shows.

12.16 Example (Many Normal means)**.** Let $Y_i \sim N(\theta_i, \sigma^2/n)$, $i = 1, \ldots, n$. Let $Y = (Y_1, \ldots, Y_n)$ denote the data and let $\theta = (\theta_1, \ldots, \theta_n)$ denote the unknown parameters. Assume that

$$\theta \in \Theta_n \equiv \left\{ (\theta_1, \ldots, \theta_n) : \sum_{i=1}^n \theta_i^2 \leq c^2 \right\}$$

[2] Typically, the squared bias is order $O(n^{-2})$ while the variance is of order $O(n^{-1})$.

for some $c > 0$. In this model, there are as many parameters as observations. [3] The MLE is $\widehat{\theta} = Y = (Y_1, \ldots, Y_n)$. Under the loss function $L(\theta, \widehat{\theta}) = \sum_{i=1}^n (\widehat{\theta}_i - \theta_i)^2$, the risk of the MLE is $R(\theta, \widehat{\theta}) = \sigma^2$. It can be shown that the minimax risk is approximately $\sigma^2/(\sigma^2 + c^2)$ and one can find an estimator $\widetilde{\theta}$ that achieves this risk. Since $\sigma^2/(\sigma^2 + c^2) < \sigma^2$, we see that $\widetilde{\theta}$ has smaller risk than the MLE. In practice, the difference between the risks can be substantial. This shows that maximum likelihood is not an optimal estimator in high dimensional problems. ■

12.6 Admissibility

Minimax estimators and Bayes estimators are "good estimators" in the sense that they have small risk. It is also useful to characterize bad estimators.

12.17 Definition. *An estimator $\widehat{\theta}$ is* **inadmissible** *if there exists another rule $\widehat{\theta}'$ such that*

$$\begin{aligned} R(\theta, \widehat{\theta}') &\leq R(\theta, \widehat{\theta}) \text{ for all } \theta \text{ and} \\ R(\theta, \widehat{\theta}') &< R(\theta, \widehat{\theta}) \text{ for at least one } \theta. \end{aligned}$$

Otherwise, $\widehat{\theta}$ is **admissible**.

12.18 Example. Let $X \sim N(\theta, 1)$ and consider estimating θ with squared error loss. Let $\widehat{\theta}(X) = 3$. We will show that $\widehat{\theta}$ is admissible. Suppose not. Then there exists a different rule $\widehat{\theta}'$ with smaller risk. In particular, $R(3, \widehat{\theta}') \leq R(3, \widehat{\theta}) = 0$. Hence, $0 = R(3, \widehat{\theta}') = \int (\widehat{\theta}'(x) - 3)^2 f(x; 3) dx$. Thus, $\widehat{\theta}'(x) = 3$. So there is no rule that beats $\widehat{\theta}$. Even though $\widehat{\theta}$ is admissible it is clearly a bad decision rule. ■

12.19 Theorem (Bayes Rules Are Admissible)**.** *Suppose that $\Theta \subset \mathbb{R}$ and that $R(\theta, \widehat{\theta})$ is a continuous function of θ for every $\widehat{\theta}$. Let f be a prior density with full support, meaning that, for every θ and every $\epsilon > 0$, $\int_{\theta-\epsilon}^{\theta+\epsilon} f(\theta) d\theta > 0$. Let $\widehat{\theta}^f$ be the Bayes' rule. If the Bayes risk is finite then $\widehat{\theta}^f$ is admissible.*

PROOF. Suppose $\widehat{\theta}^f$ is inadmissible. Then there exists a better rule $\widehat{\theta}$ such that $R(\theta, \widehat{\theta}) \leq R(\theta, \widehat{\theta}^f)$ for all θ and $R(\theta_0, \widehat{\theta}) < R(\theta_0, \widehat{\theta}^f)$ for some θ_0. Let

[3] The many Normal means problem is more general than it looks. Many nonparametric estimation problems are mathematically equivalent to this model.

$\nu = R(\theta_0, \widehat{\theta}^f) - R(\theta_0, \widehat{\theta}) > 0$. Since R is continuous, there is an $\epsilon > 0$ such that $R(\theta, \widehat{\theta}^f) - R(\theta, \widehat{\theta}) > \nu/2$ for all $\theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$. Now,

$$\begin{aligned} r(f, \widehat{\theta}^f) - r(f, \widehat{\theta}) &= \int R(\theta, \widehat{\theta}^f) f(\theta) d\theta - \int R(\theta, \widehat{\theta}) f(\theta) d\theta \\ &= \int \left[R(\theta, \widehat{\theta}^f) - R(\theta, \widehat{\theta}) \right] f(\theta) d\theta \\ &\geq \int_{\theta_0 - \epsilon}^{\theta_0 + \epsilon} \left[R(\theta, \widehat{\theta}^f) - R(\theta, \widehat{\theta}) \right] f(\theta) d\theta \\ &\geq \frac{\nu}{2} \int_{\theta_0 - \epsilon}^{\theta_0 + \epsilon} f(\theta) d\theta \\ &> 0. \end{aligned}$$

Hence, $r(f, \widehat{\theta}^f) > r(f, \widehat{\theta})$. This implies that $\widehat{\theta}^f$ does not minimize $r(f, \widehat{\theta})$ which contradicts the fact that $\widehat{\theta}^f$ is the Bayes rule. ∎

12.20 Theorem. *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Under squared error loss, $\overline{X}$ is admissible.*

The proof of the last theorem is quite technical and is omitted but the idea is as follows: The posterior mean is admissible for any strictly positive prior. Take the prior to be $N(a, b^2)$. When b^2 is very large, the posterior mean is approximately equal to $\overline{X}$.

How are minimaxity and admissibility linked? In general, a rule may be one, both, or neither. But here are some facts linking admissibility and minimaxity.

12.21 Theorem. *Suppose that $\widehat{\theta}$ has constant risk and is admissible. Then it is minimax.*

PROOF. The risk is $R(\theta, \widehat{\theta}) = c$ for some c. If $\widehat{\theta}$ were not minimax then there exists a rule $\widehat{\theta}'$ such that

$$R(\theta, \widehat{\theta}') \leq \sup_\theta R(\theta, \widehat{\theta}') < \sup_\theta R(\theta, \widehat{\theta}) = c.$$

This would imply that $\widehat{\theta}$ is inadmissible. ∎

Now we can prove a restricted version of Theorem 12.14 for squared error loss.

12.22 Theorem. *Let $X_1, \ldots, X_n \sim N(\theta, 1)$. Then, under squared error loss, $\widehat{\theta} = \overline{X}$ is minimax.*

PROOF. According to Theorem 12.20, $\widehat{\theta}$ is admissible. The risk of $\widehat{\theta}$ is $1/n$ which is constant. The result follows from Theorem 12.21. ∎

Although minimax rules are not guaranteed to be admissible they are "close to admissible." Say that $\widehat{\theta}$ is **strongly inadmissible** if there exists a rule $\widehat{\theta}'$ and an $\epsilon > 0$ such that $R(\theta, \widehat{\theta}') < R(\theta, \widehat{\theta}) - \epsilon$ for all θ.

12.23 Theorem. *If $\widehat{\theta}$ is minimax, then it is not strongly inadmissible.*

12.7 Stein's Paradox

Suppose that $X \sim N(\theta, 1)$ and consider estimating θ with squared error loss. From the previous section we know that $\widehat{\theta}(X) = X$ is admissible. Now consider estimating two, unrelated quantities $\theta = (\theta_1, \theta_2)$ and suppose that $X_1 \sim N(\theta_1, 1)$ and $X_2 \sim N(\theta_2, 1)$ independently, with loss $L(\theta, \widehat{\theta}) = \sum_{j=1}^{2}(\theta_j - \widehat{\theta}_j)^2$. Not surprisingly, $\widehat{\theta}(X) = X$ is again admissible where $X = (X_1, X_2)$. Now consider the generalization to k normal means. Let $\theta = (\theta_1, \ldots, \theta_k)$, $X = (X_1, \ldots, X_k)$ with $X_i \sim N(\theta_i, 1)$ (independent) and loss $L(\theta, \widehat{\theta}) = \sum_{j=1}^{k}(\theta_j - \widehat{\theta}_j)^2$. Stein astounded everyone when he proved that, if $k \geq 3$, then $\widehat{\theta}(X) = X$ is inadmissible. It can be shown that the **James-Stein estimator** $\widehat{\theta}^S$ has smaller risk, where $\widehat{\theta}^S = (\widehat{\theta}_1^S, \ldots, \widehat{\theta}_k^S)$,

$$\widehat{\theta}_i^S(X) = \left(1 - \frac{k-2}{\sum_i X_i^2}\right)^+ X_i \tag{12.12}$$

and $(z)^+ = \max\{z, 0\}$. This estimator shrinks the X_i's towards 0. The message is that, when estimating many parameters, there is great value in shrinking the estimates. This observation plays an important role in modern nonparametric function estimation.

12.8 Bibliographic Remarks

Aspects of decision theory can be found in Casella and Berger (2002), Berger (1985), Ferguson (1967), and Lehmann and Casella (1998).

12.9 Exercises

1. In each of the following models, find the Bayes risk and the Bayes estimator, using squared error loss.

   (a) $X \sim \text{Binomial}(n, p)$, $p \sim \text{Beta}(\alpha, \beta)$.

(b) $X \sim \text{Poisson}(\lambda)$, $\lambda \sim \text{Gamma}(\alpha, \beta)$.

(c) $X \sim N(\theta, \sigma^2)$ where σ^2 is known and $\theta \sim N(a, b^2)$.

2. Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ and suppose we estimate θ with loss function $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2/\sigma^2$. Show that $\overline{X}$ is admissible and minimax.

3. Let $\Theta = \{\theta_1, \ldots, \theta_k\}$ be a finite parameter space. Prove that the posterior mode is the Bayes estimator under zero–one loss.

4. (Casella and Berger (2002).) Let $X_1, \ldots, X_n$ be a sample from a distribution with variance σ^2. Consider estimators of the form bS^2 where S^2 is the sample variance. Let the loss function for estimating σ^2 be

$$L(\sigma^2, \widehat{\sigma}^2) = \frac{\widehat{\sigma}^2}{\sigma^2} - 1 - \log\left(\frac{\widehat{\sigma}^2}{\sigma^2}\right).$$

Find the optimal value of b that minimizes the risk for all σ^2.

5. (Berliner (1983).) Let $X \sim \text{Binomial}(n, p)$ and suppose the loss function is

$$L(p, \widehat{p}) = \left(1 - \frac{\widehat{p}}{p}\right)^2$$

where $0 < p < 1$. Consider the estimator $\widehat{p}(X) = 0$. This estimator falls outside the parameter space $(0, 1)$ but we will allow this. Show that $\widehat{p}(X) = 0$ is the unique, minimax rule.

6. (Computer Experiment.) Compare the risk of the MLE and the James-Stein estimator (12.12) by simulation. Try various values of n and various vectors θ. Summarize your results.