All of Statistics: A Concise Course in Statistical Inference

Brief Contents

Chapter 1

Introduction

The goal of this book is to provide a broad background in probability and statistics for students in statistics, computer science (especially data mining and machine learning), mathematics, and related disciplines. This book covers a much wider range of topics than a typical introductory text on mathematical statistics. It includes modern topics like nonparametric curve estimation, bootstrapping and classification, topics that are usually relegated to follow-up courses. The reader is assumed to know calculus and a little linear algebra. No previous knowledge of probability and statistics is required. The text is suitable for advanced undergraduates and graduate students.

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 though possible. Formal statistical theory is more pervasive than computer scientists had realized. All agree students who deal with the analysis of data should be well grounded in basic probability and mathematical statistics. Using fancy tools like neu-

ral nets, boosting and support vector machines without understanding basic statistics is like doing brain surgery before knowing how to use a bandaid.

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 a understanding of modern statistics quickly?" The typical mathematical statistic 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 honors undergraduates in math, statistics and computer science as well as graduate students in computer science and other quantitative fields.

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. In general, I try to avoid belaboring tedious calculations in favor of emphasizing concepts.

5. I cover nonparametric inference before parametric inference. This is the opposite of most statistics books but I believe it is the right way to do it. Parametric models are unrealistic and pedagogically unnatural. (How would we know the everything about the distribution except for one or two parameters?) I introduce statistical functionals and bootstrapping very early and students find this quite natural.

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.

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. Anyway, slow courses are boring.

8. As Richard Feynman pointed out, 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.

9. On my website are files with R code which students can use for doing all the computing. However, the book is not tied to R and any computing language can be used.

The first part 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?

The second part of the book 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 1.1. 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. The second part of

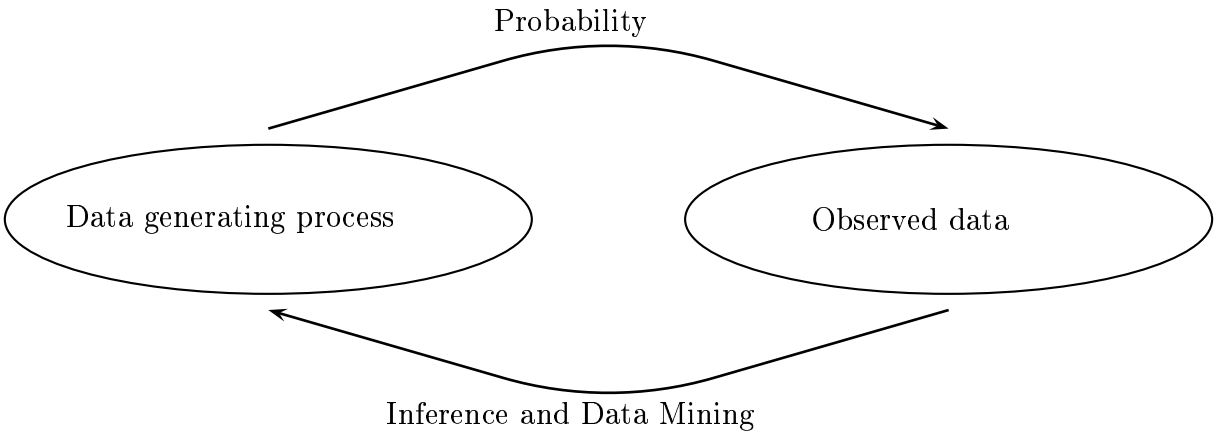


Figure 1.1: Probability and inference.

the book contains one more chapter on probability that covers stochastic processes including Markov chains.

I have drawn heavily on other books in many places. Most chapters contain a section called Bilbliographic 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 excercises.

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 $X \in \mathcal{X}$
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 unknown quantity with a prescribed frequency
directed acyclic graph	Bayes net	multivariate distribution with specified conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update subjective beliefs
frequentist inference	—	statistical methods for producing point estimates and confidence intervals with guarantees on frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

Notation

Symbol	Meaning
$\mathbb{R}$	real numbers
$\inf_{x\in A} f(x)$	infimum: the largest number y such that $y \leq f(x)$ for all $x \in A$
	think of this as the minimum of f
$\sup_{x\in A} f(x)$	supremum: the smallest number y such that $y \geq f(x)$ for all $x \in A$
	think of this as the maximum of f
$n!$	$n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$
$\binom{n}{k}$	$\frac{n!}{k!(n-k)!}$
$\Gamma(\alpha)$	Gamma function $\int_0^\infty y^{\alpha-1}e^{-y}dy$
ω	outcome
Ω	sample space (set of outcomes)
A	event (subset of Ω)
$I_A(\omega)$	indicator function; 1 if $\omega \in A$ and 0 otherwise
$\mathbb{P}(A)$	probability of event A
$\|A\|$	number of points in set A
F_X	cumulative distribution function
f_X	probability density (or mass) function
$X \sim F$	X has distribution F
$X \sim f$	X has density f
$X \stackrel{d}{=} Y$	X and Y have the same distribution
$X_1, \ldots, X_n \sim F$	sample of size n from F
ϕ	standard Normal probability density
Φ	standard Normal distribution function
z_α	upper α quantile of $N(0,1)$ i.e. $\Phi^{-1}(1-\alpha)$
$\mathbb{E}(X) = \int x dF(x)$	expected value (mean) of random variable X
$\mathbb{E}(r(X)) = \int r(x) dF(x)$	expected value (mean) of $r(X)$
$\mathbb{V}(X)$	variance of random variable X
$\mathsf{Cov}(X, Y)$	covariance between X and Y
$X_1, \ldots, X_n$	data
n	sample size
$\xrightarrow{\text{P}}$	convergence in probability
$\rightsquigarrow$	convergence in distribution
$\xrightarrow{\text{qm}}$	convergence in quadratic mean
$X_n \approx N(\mu, \sigma_n^2)$	$(X_n - \mu)/\sigma_n \rightsquigarrow N(0,1)$

Notation Continued

Symbol	Meaning
$\mathfrak{F}$	statistical model; a set of distribution functions, density functions or regression functions
θ	parameter
$\widehat{\theta}$	estimate of parameter
$T(F)$	statistical functional (the mean, for example)
$\mathcal{L}_n(\theta)$	likelihood function
$x_n = o(a_n)$	$x_n/a_n \to 0$
$x_n = O(a_n)$	$\lvert x_n/a_n \rvert$ is bounded for large n
$X_n = o_P(a_n)$	$X_n/a_n \xrightarrow{\mathrm{P}} 0$
$X_n = O_P(a_n)$	$\lvert X_n/a_n \rvert$ is bounded in probability for large n

Useful Math Facts

$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \cdots$

$\sum_{j=k}^{\infty} r^j = \frac{r^k}{1-r}$ for $0 < r < 1$

$\lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n = e^a$

The Gamma function is is defined by $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy$ for $\alpha \geq 0$. If $\alpha > 1$ then $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$. If n is an integer then $\Gamma(n) = (n-1)!$. Some special values are: $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$.

Part I

Probability

Chapter 2

Probability

2.1 Introduction

Probability is the mathematical language for quantifying uncertainty. We can apply probability theory to a diverse set of problems, from coin flipping to the analysis of computer algorithms. The starting point is to specify sample space, the set of possible outcomes.

2.2 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** or **realizations. Events** are subsets of Ω.

Example 2.1 *If we toss a coin twice then* $\Omega = \{HH, HT, TH, TT\}$. *The event that the first toss is heads is* $A = \{HH, HT\}$. ■

Example 2.2 *Let* ω *be the outcome of a measurement of some physical quantity, for example, temperature. Then* $\Omega = \mathbb{R} = (-\infty, \infty)$. *The event that the measurement is larger than 10 but less than or equal to 23 is* $A = (10, 23]$. ■

Example 2.3 *If we toss a coin forever then the sample space is the infinite set*

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

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

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

Given an event A, let $A^c = \{\omega \in \Omega;\ \omega \notin A\}$ denote the complement of A. Informally, A^c can be read as "not A." The complement of Ω is the empty set $\emptyset$. The union of events A and B is defined $A \bigcup B = \{\omega \in \Omega;\ \omega \in A \text{ or } \omega \in B \text{ or } \omega \in \text{both}\}$ which can be thought of as "A or B." If $A_1, A_2, \ldots$ is a sequence of sets then

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

The intersection of A and B is $A \bigcap B = \{\omega \in \Omega;\ \omega \in A \text{ and } \omega \in B\}$ read "A and B." Sometimes we write $A \bigcap B$ as AB. If $A_1, A_2, \ldots$ is a sequence of sets then

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

Let $A - B = \{\omega :\ \omega \in A, \omega \notin B\}$. 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 Table 1 for a summary.

Table 1. Sample space and events.

Ω	sample space
ω	outcome
A	event (subset of Ω)
$\|A\|$	number of points in A (if A is finite)
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 (points in A that are not in B)
$A \subset B$	set inclusion (A is a subset of or equal to B)
$\emptyset$	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 $A_n \to A$.

Example 2.4 *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 $\bigcap_{i=1}^{\infty} A_i = \emptyset$.* ■

2.3 Probability

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

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 technical appendix for details.

Definition 2.5 *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*

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

There are many interpretations of $\mathbb{P}(A)$. The two common interpretations are frequencies and degrees of beliefs. In 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 $1/2$, when 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 later.

One can derive many properties of $\mathbb{P}$ from the axioms. Here are a few:

$$\begin{aligned} \mathbb{P}(\emptyset) &= 0 \\ A \subset B &\implies \mathbb{P}(A) \leq \mathbb{P}(B) \\ 0 \leq \mathbb{P}(A) &\leq 1 \\ \mathbb{P}(A^c) &= 1 - \mathbb{P}(A) \\ A \bigcap B = \emptyset &\implies \mathbb{P}\left(A \bigcup B\right) = \mathbb{P}(A) + \mathbb{P}(B). \end{aligned} \tag{2.1}$$

A less obvious property is given in the following Lemma.

Lemma 2.6 *For any events A and B, $\mathbb{P}(A \bigcup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$.*

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

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

Example 2.7 *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, that is, $\mathbb{P}(\{H_1, H_2\}) = \mathbb{P}(\{H_1, T_2\}) = \mathbb{P}(\{T_1, H_2\}) = \mathbb{P}(\{T_1, T_2\}) = 1/4$, *then* $\mathbb{P}(H_1 \bigcup H_2) = \mathbb{P}(H_1) + \mathbb{P}(H_2) - \mathbb{P}(H_1 H_2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = 3/4$. ■

Theorem **2.8** (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 excercise 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,

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

2.4 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.

In general, 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

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \tag{2.2}$$

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:

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

2.5 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.

Definition 2.9 *Two events A and B are* **independent** *if*

$$\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B) \tag{2.3}$$

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.

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 a the sets in a Venn diagram.

Example 2.10 *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*

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

Example 2.11 *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 1 misses, 2 misses, 1 succeeds. This has probability $\mathbb{P}(A_2) = (2/3)(3/4)(1/3) = (1/2)(1/3)$. Fol-

lowing 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* $\sum_{j=k}^{\infty} r^j = r^k/(1-r)$. ■

Summary of Independence

1. A and B are independent 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.

2.6 Conditional Probability

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

Definition 2.12 *If* $\mathbb{P}(B) > 0$ *then the* **conditional probability** *of* A *given* B *is*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}. \tag{2.4}$$

Think of $\mathbb{P}(A|B)$ as the fraction of times A occurs among those in which B occurs. Here are some facts about conditional probabilities. 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.

Example 2.13 *A medical test for a disease D has outcomes $+$ and $-$. The probabilities are:*

	D	D^c
$+$	*.0081*	*.0900*
$-$	*.0090*	*.9010*

From the definition of conditional probability, $\mathbb{P}(+|D) = \mathbb{P}(+, D)/\mathbb{P}(D) = .0081/(.0081 + .0009) = .9$ *and* $\mathbb{P}(-|D^c) = \mathbb{P}(-, D^c)/\mathbb{P}(D^c) = .9010/(.9010 + .0900) \approx .9$. *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* $\mathbb{P}(D|+) = \mathbb{P}(+, D)/\mathbb{P}(+) = .0081/(.0081 + .0900) = .08$. *The lesson here is that you need to compute the answer numerically. Don't trust your intuition.* ■

If A and B are independent events then

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

So another interpretation of independence is that knowing B doesn't change the probability of A.

From the definition of conditional probability we can write $\mathbb{P}(AB) = \mathbb{P}(A|B)\mathbb{P}(B)$ and also $\mathbb{P}(AB) = \mathbb{P}(B|A)\mathbb{P}(A)$. Often, these formulae give us a convenient way to compute $\mathbb{P}(AB)$ when A and B are not independent.

Example 2.14 *Draw two cards from a deck, without replacement. Let A be the event that the first draw is Ace of Clubs and let B be the event that the second draw is Queen of Diamonds. Then* $\mathbb{P}(A,B) = \mathbb{P}(A)\mathbb{P}(B|A) = (1/52)\times(1/51)$.
■

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

2.7 Bayes' Theorem

Bayes' theorem is a useful result that is the basis of "expert systems" and "Bayes' nets." First, we need a preliminary result.

Theorem **2.15** (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. ■

Theorem **2.16** (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 $i = 1, \ldots, k$,*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}. \tag{2.5}$$

Remark 2.17 *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:

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

Example 2.18 *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,*

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

2.8 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), and at the advanced level, Billingsley (1979) and Breiman (1968). I adapted many examples and problems from DeGroot and Schervish (2002) and Grimmett and Stirzaker (1982).

2.9 Technical 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 P is a probability measure defined on $\mathcal{A}$ then $(\Omega, \mathcal{A}, P)$ 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**.

2.10 Excercises

1. Fill in the details of the proof of Theorem 2.8. Also, prove the monotone decreasing case.

2. Prove the statements in equation (2.1).

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 $C_1 \subset B_2 \subset \cdots$.

   (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 between 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 5 children.

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

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

16. Show that

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

17. 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 $i = 2, \ldots, k$.

18. Suppose that 30 percent of computer owners use a Macintosh, 50 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?

19. 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."

20. (Computer Experiment.) Suppose a coin has probability p of falling heads. 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 $n = 1000$ and simulate n coin flips. Plot the proportion of heads as a function of n. Repeat for $p = .03$.

21. (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 $n = 10, 100, 1000$.

22. (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 $\widehat{P}(AB) = \widehat{P}(A)\widehat{P}(B)$ where $\widehat{P}(A)$ is the proportion of times A occurred in the simulation and similarly for $\widehat{P}(AB)$ and $\widehat{P}(B)$. Now find two events A and B that are not independent. Compute $\widehat{P}(A), \widehat{P}(B)$ and $\widehat{P}(AB)$. Compare the calculated values to their theoretical values. Report your results and interpret.

Chapter 3

Random Variables

3.1 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.

Definition 3.1 *A* **random variable** *is a mapping* $X : \Omega \to \mathbb{R}$ *that assigns a real number* $X(\omega)$ *to each outcome* ω.

Technically, a random variable must be measurable. See the technical appendix for details.

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

Example 3.2 *Flip a coin ten times. Let* $X(\omega)$ *be the number of heads in the sequence* ω. *For example, if* $\omega = HHTHHTHHTT$ *then* $X(\omega) = 6$. ■

Example 3.3 *Let* $\Omega = \{(x, y);\ x^2 + y^2 \leq 1\}$ *be the unit disc. 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* $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$, $W(\omega) = \sqrt{x^2 + y^2}$. ■

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

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

X denotes the random variable and x denotes a possible value of X.

Example 3.4 *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. ■

3.2 Distribution Functions and Probability Functions

Given a random variable X, we define an important function called the cumulative distribution function (or distribution function) in the following way.

Definition 3.5 *The* **cumulative distribution function** CDF $F_X : \mathbb{R} \to [0, 1]$ *of a random variable X is defined by*

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

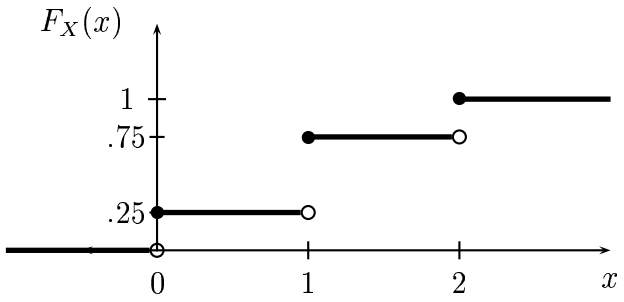


Figure 3.1: CDF for flipping a coin twice (Example 3.6.)

You might wonder why we bother to define the CDF . You will see later that the CDF is a useful function: it effectively contains all the information about the random variable.

Example 3.6 *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 3.1. 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 $F(1.4) = .75$? ■

The following result, which we do not prove, shows that the CDF completely determines the distribution of a random variable.

Theorem **3.7** *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.*

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

Theorem **3.8** *A function F mapping the real line to $[0, 1]$ is a CDF for some probability measure $\mathbb{P}$ if and only if it satisfies the following three conditions:*

(i) F is non-decreasing i.e. $x_1 < x_2$ implies that $F(x_1) \leq F(x_2)$.
(ii) F is normalized: $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.
(iii) F is right-continuous, i.e. $F(x) = F(x^+)$ 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. ■

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.

Definition 3.9 *X is* **discrete** *if it takes countably 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 $\sum_i f_X(x_i) = 1$. The CDF of X is related to f_X by

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

Sometimes we write f_X and F_X simply as f and F.

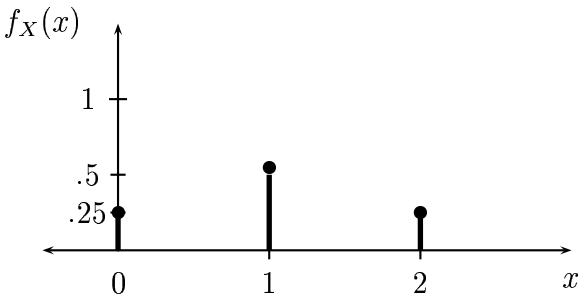


Figure 3.2: Probability function for flipping a coin twice (Example 3.6.)

Example 3.10 *The probability function for Example 3.6 is*

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

See Figure 3.2. ■

Definition 3.11 *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 $a \leq b$,*

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx. \tag{3.1}$$

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

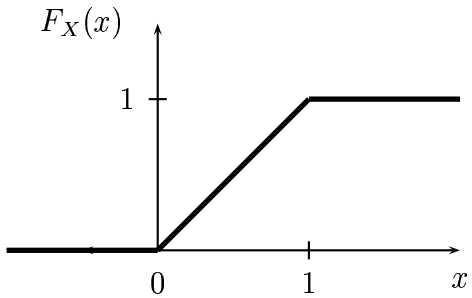


Figure 3.3: CDF for Uniform (0,1).

Example 3.12 *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. 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 3.3. ■

Example 3.13 *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.

Example 3.14 *Let*

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

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

Lemma 3.15 *Let F be the* CDF *for a random variable X. Then:*

(i) $\mathbb{P}(X = x) = F(x) - F(x^-)$ *where* $F(x^-) = \lim_{y \uparrow x} F(y)$,
(ii) $\mathbb{P}(x < X \leq y) = F(y) - F(x)$,
(iii) $\mathbb{P}(X > x) = 1 - F(x)$,
(iv) If X is continuous then

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b).$$

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

Definition 3.16 *Let X be a random variable with* CDF *F. The* **inverse** CDF *or* **quantile function** *is defined by*

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

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

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

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.

3.3 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.

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 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 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 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 the mass function is called a Binomial random variable and we write $X \sim \text{Binomial}(n, p)$. If $X_1 \sim \text{Binomial}(n, p_1)$ and $X_2 \sim \text{Binomial}(n, p_2)$ then $X_1 + X_2 \sim \text{Binomial}(n, p_1 + p_2)$.

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

$$\sum_{k=1}^{\infty} \mathbb{P}(X = k) = p \sum_{k=0}^{\infty} (1-p)^k = \frac{p}{1-(1-p)} = 1.$$

Think of X as the number of flips needed until the first heads 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 $X_1 \sim \text{Poisson}(n, \lambda_1)$ and $X_2 \sim \text{Poisson}(n, \lambda_2)$ then $X_1 + X_2 \sim \text{Poisson}(n, \lambda_1 + \lambda_2)$.

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.

3.4 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

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

where $\mu \in \mathbb{R}$ and $\sigma > 0$. Later we shall see that μ is the "center" (or mean) of the distribution and σ is the "spread" (or standard deviation) of the distribution. The Normal plays an important role in probability and statistics. Many phenomena in nature have approximately Normal distributions. Later, we shall see that the distribution of a sum of random variables can be approximated by a Normal distribution (the central limit theorem).

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 3.4. 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

$$\mathbb{P}(a < X < b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{3.2}$$

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 compute $\Phi(z)$ and $\Phi^{-1}(z)$. All statistics texts, including this one, have a table of values of $\Phi(z)$.

Example 3.17 *Suppose that* $X \sim N(3, 5)$. *Find* $\mathbb{P}(X > 1)$. *The solution is*

$$\mathbb{P}(X > 1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}\left(Z < \frac{1-3}{\sqrt{5}}\right) = 1 - \Phi(-0.8944) = .81.$$

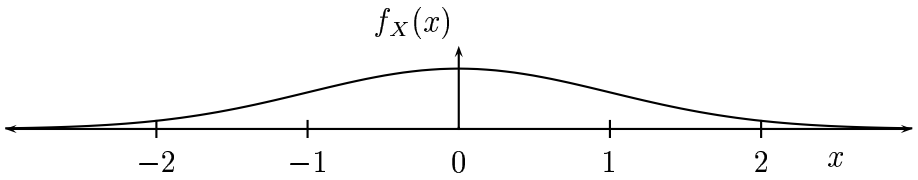


Figure 3.4: Density of a standard Normal.

Now find q such that $\mathbb{P}(X < q) = .2$. In other words, find $q = \Phi^{-1}(.2)$. We solve this by writing

$$.2 = \mathbb{P}(X < q) = \mathbb{P}\left(Z < \frac{q-\mu}{\sigma}\right) = \Phi\left(\frac{q-\mu}{\sigma}\right).$$

From the Normal table, $\Phi(-.8416) = .2$. Therefore,

$$-.8416 = \frac{q-\mu}{\sigma} = \frac{q-3}{\sqrt{5}}$$

and hence $q = 3 - .8416\sqrt{5} = 1.1181$. ■

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 an 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, let's do the integral:

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

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 $\sum_{i=1}^p Z_i^2 \sim \chi^2_p$.

3.5 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.

Example 3.18 *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*	*1/3*
	1/3	*1/3*	*1*

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

Definition 3.19 *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, 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 $F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$.*

Example 3.20 *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, $\mathbb{P}(X < 1/2, Y < 1/2) = 1/4$. ■

Example 3.21 *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. ■

Example 3.22 *If the distribution is defined over a non-rectangular region, then the calcuations are a bit more complicated. Here is an nice 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 3.5. 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,

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

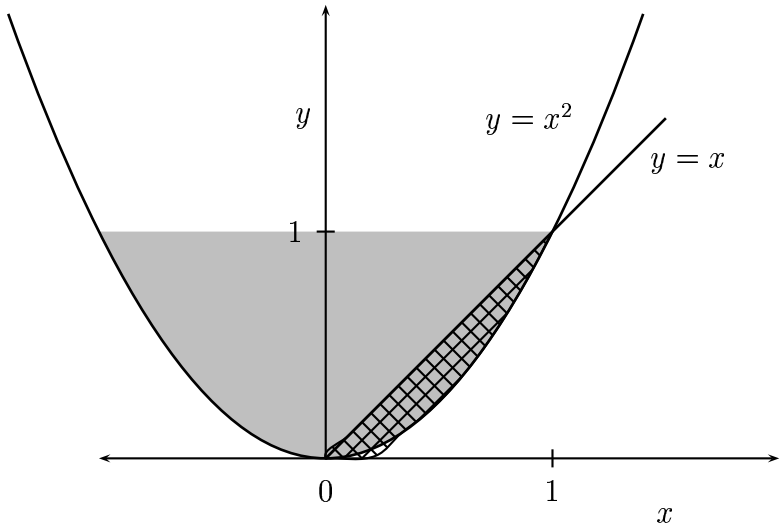


Figure 3.5: The light shaded region is $x^2 \leq y \leq 1$. The density is positive over this region. The hatched region is the event $X \geq Y$ intersected with $x^2 \leq y \leq 1$.

3.6 Marginal Distributions

Definition 3.23 *If (X, Y) have joint distribution with mass function $f_{X,Y}$, then the* **marginal mass function for** X *is defined by*

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f(x, y) \tag{3.3}$$

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

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y) = \sum_x f(x, y). \tag{3.4}$$

Example 3.24 *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* $f_X(1) = 7/10$. ■

Definition 3.25 *For continuous random variables, the marginal densities are*

$$f_X(x) = \int f(x,y)dy, \quad \text{and} \quad f_Y(y) = \int f(x,y)dx. \tag{3.5}$$

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

Example 3.26 *Suppose that*

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

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

Example 3.27 *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

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

Example 3.28 *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.* ■

3.7 Independent Random Variables

Definition 3.29 *Two random variables X and Y are* **independent** *if, for every A and B,*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B). \tag{3.6}$$

We write $X \amalg Y$.

In principle, to check whether X and Y are independent we need to check equation (3.6) 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.

Theorem **3.30** *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.*

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

Example 3.31 *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* $f(0,1) = 0$. ■

Example 3.32 *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*

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

Now,

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

The following result is helpful for verifying independence.

Theorem **3.33** *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.*

Example 3.34 *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,* $X \amalg Y$. ■

3.8 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 mass function as follows.

Definition 3.35 *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 $f_Y(y) > 0$.

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 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. We simply take it as a definition.

For continuous distributions we use the same definitions. 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.

Definition 3.36 *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,*

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

Example 3.37 *Let X and Y have a uniform distribution on the unit square. Verify that $f_{X|Y}(x|y) = 1$ for $0 \leq x \leq 1$ and 0 otherwise. Thus, given $Y = y$, X is* Uniform *(0,1). We can write this as $X|Y = y \sim \text{Unif}(0, 1)$.* ■

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 the next example.

Example 3.38 *Let*

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

Let us find $\mathbb{P}(X < 1/4|Y = 1/3)$. *In example 3.27 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,

$$\begin{aligned} P\left(X < \frac{1}{4} \,\middle|\, Y = \frac{1}{3}\right) &= \int_0^{1/4} f_{X|Y}\left(x \,\middle|\, \frac{1}{3}\right) dx \\ &= \int_0^{1/4} \frac{x+\frac{1}{3}}{\frac{1}{3}+\frac{1}{2}} dx = \frac{\frac{1}{32}+\frac{1}{3}}{\frac{1}{3}+\frac{1}{2}} = \frac{14}{32}. \quad \blacksquare \end{aligned}$$

Example 3.39 *Suppose that* $X \sim \text{Unif}(0,1)$. *After obtaining a value of* X *we generate* $Y|X = x \sim \text{Uniform}(x,1)$. *What is the marginal distribution of* Y? *First note that,*

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

and

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

So,

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

The marginal for Y *is*

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

for $0 < y < 1$. $\blacksquare$

Example 3.40 *Consider the density in Example 3.28. 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 computed by first noting that* $f_{Y|X}(y|1/2) = 32y/15$. *Thus,*

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

3.9 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 $A_1, \ldots, A_n$,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i). \tag{3.7}$$

It suffices to check that $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$. If $X_1, \ldots, X_n$ are independent and each has the same marginal distribution with density f, we say that $X_1, \ldots, X_n$ are IID (independent and identically distributed). We shall write this as $X_1, \ldots X_n \sim f$ or, in terms of the CDF , $X_1, \ldots X_n \sim F$. This means that $X_1, \ldots, X_n$ are independent draws from the same distribution. We also call $X_1, \ldots, X_n$ a *random sample* from F.

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

3.10 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

$$f(x) = \binom{n}{x_1 \ldots x_k} p_1^{x_1} \cdots p_k^{x_k} \tag{3.8}$$

where

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

Lemma 3.41 *Suppose that $X \sim$ 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 had 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

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

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

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

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

$$f(x;\, \mu, \Sigma) = \frac{1}{(2\pi)^{k/2} \det(\Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\} \tag{3.9}$$

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

A matrix Σ is positive definite if, for all non-zero vectors x, $x^T \Sigma X > 0$.

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 $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$.

Theorem **3.42** *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

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

Theorem **3.43** *Let $X \sim N(\mu, \Sigma)$. Then:*

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

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

3.11 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

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

Example 3.44 *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

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(r(X) \leq y) \\ &= \mathbb{P}(\{x;\ r(x) \leq y\}) = \int_{A_y} f_X(x)dx \end{aligned} \tag{3.10}$$

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

Example 3.45 *Let* $f_X(x) = e^{-x}$ *for* $x > 0$. *Then* $F_X(x) = \int_0^x f_X(s)ds = 1 - e^{-x}$. *Let* $Y = r(X) = \log X$. *Then* $A_y = \{x :\ x \leq e^y\}$ *and*

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

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

Example 3.46 *Let* $X \sim \text{Unif}(-1, 3)$. *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 case: (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

$$f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right|. \tag{3.11}$$

3.12 Transformations of Several Random Variables

In some cases we are interested in transformation 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:

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 $f_Z(z) = F_Z'(z)$.

Example 3.47 *Let $X_1, X_2 \sim$ Unif$(0,1)$ 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,

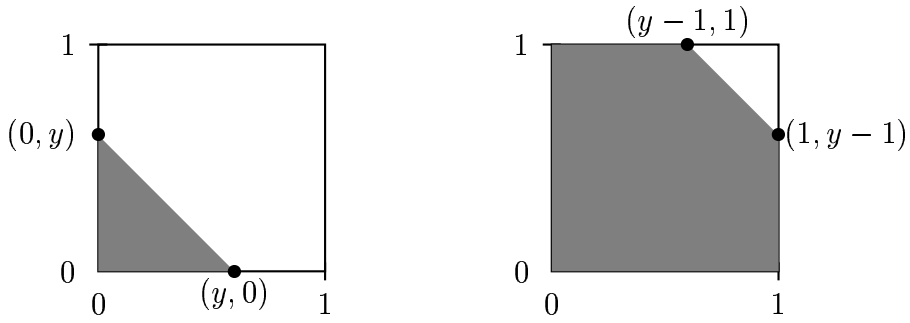
$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(r(X_1,X_2) \leq y) \\ &= \mathbb{P}(\{(x_1,x_2) : \ \ r(x_1,x_2) \leq y\}) = \int\int_{A_y} f(x_1,x_2)dx_1dx_2. \end{aligned}$$

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 3.6. In this case, $\int\int_{A_y} f(x_1,x_2)dx_1dx_2$ is the area of this triangle which $y^2/2$. 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 $1 - y^2/2$. Therefore,

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

By differentiation, the PDF *is*

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



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

Figure 3.6: The set A_y for example 3.47.

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

3.14 Excercises

1. Show that
$$\mathbb{P}(X = x) = F(x^+) - F(x^-)$$
and
$$F(x_2) - F(x_1) = \mathbb{P}(X \leq x_2) - \mathbb{P}(X \leq x_1).$$

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 3.15.

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 3.33.

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 disc $\{(x, y) : x^2 + y^2 \leq 1\}$. 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
$$f_{X,Y}(x,y) = \begin{cases} c(x + y^2) & 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$
    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 $\mathbb{P}(0 \leq X < 4)$.

(e) Find x such that $\mathbb{P}(|X| > |x|) = .05$.

19. Prove formula (3.11).

20. Let $X, Y \sim \text{Unif}(0, 1)$ 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: $Y \leq y$ if and only if $X_i \leq y$ for $i = 1, \ldots, n$.

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

23. 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) (10 points) Find $\mathbb{E}(Y|Z)$. Hint: What values z can Z take? Now find $\mathbb{E}(Y|Z = z)$.

Chapter 4

Expectation

4.1 Expectation of a Random Variable

The expectation (or mean) of a random variable X is the average value of X. The formal definition is as follows.

Definition 4.1 *The* **expected value**, *or* **mean**, *or* **first moment**, *of* X *is defined to be*

$$\mathbb{E}(X) = \int x\, dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) dx & \text{if } X \text{ is continuous} \end{cases} \tag{4.1}$$

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

$$\mathbb{E}(X) = \mathbb{E}X = \int x\, dF(x) = \mu = \mu_X. \tag{4.2}$$

The expectation is a one-number summary of the distribution. Think of $\mathbb{E}(X)$ as the average value you would obtain if you computed the numerical average $n^{-1}\sum_{i=1}^n X_i$ of a large number of IID draws $X_1, \ldots, X_n$. The fact that $\mathbb{E}(X) \approx n^{-1}\sum_{i=1}^n X_i$ is actually more than a heuristic: it is a theorem called the law of large numbers that we will discuss later. 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 $\sum_x xf(x)$ 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 anlysis 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.

Example 4.2 *Let* $X \sim \text{Bernoulli}(p)$. *Then* $\mathbb{E}(X) = \sum_{x=0}^{1} xf(x) = (0 \times (1-p)) + (1 \times p) = p$. ■

Example 4.3 *Flip a fair coin two times. Let* X *be the number of heads. Then,* $\mathbb{E}(X) = \int x dF_X(x) = \sum_x x f_X(x) = (0 \times f(0)) + (1 \times f(1)) + (2 \times f(2)) = (0 \times (1/4)) + (1 \times (1/2)) + (2 \times (1/4)) = 1$. ■

Example 4.4 *Let* $X \sim \text{Unif}(-1, 3)$. *Then,* $\mathbb{E}(X) = \int x dF_X(x) = \int x f_X(x) dx = \frac{1}{4}\int_{-1}^{3} x\, dx = 1$. ■

Example 4.5 *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 $\mathbb{E}(Y) = \int y f_Y(y) dy$. But there is an easier way.

Theorem **4.6** (The rule of the lazy statistician.) *Let $Y = r(X)$. Then*

$$\mathbb{E}(Y) = \mathbb{E}(r(X)) = \int r(x)dF_X(x). \tag{4.3}$$

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.

Example 4.7 *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, $\mathbb{E}(Y) = \int_1^e y\, f(y)dy = e - 1$. ■

Example 4.8 *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,*

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

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

$$\mathbb{E}(Z) = \mathbb{E}(r(X,Y)) = \int\int r(x,y)dF(x,y). \tag{4.4}$$

Example 4.9 *Let (X, Y) have a jointly uniform distribution on the unit square. Let $Z = r(X, Y) = X^2 + Y^2$. Then,*

$$\mathbb{E}(Z) = \int\int r(x,y)dF(x,y) = \int_0^1\int_0^1 (x^2+y^2)\,dxdy = \int_0^1 x^2\,dx + \int_0^1 y^2\,dy = \frac{2}{3}. \quad \blacksquare$$

The k^{th} **moment** of X is defined to be $\mathbb{E}(X^k)$ assuming that $\mathbb{E}(|X|^k) < \infty$. We shall rarely make much use of moments beyond $k = 2$.

4.2 Properties of Expectations

Theorem **4.10** *If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then*

$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i \mathbb{E}(X_i). \tag{4.5}$$

Example 4.11 *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 $\mathbb{E}(X) = \mathbb{E}(\sum_i X_i) = \sum_i \mathbb{E}(X_i) = np$.

Theorem **4.12** *Let $X_1, \ldots, X_n$ be independent random variables. Then,*

$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_i \mathbb{E}(X_i). \tag{4.6}$$

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

4.3 Variance and Covariance

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

The variance measures the "spread" of a distribution.

Definition 4.13 *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 $\mathbb{V}(X)$ or $\mathbb{V}X$ — is defined by*

$$\sigma^2 = \mathbb{E}(X-\mu)^2 = \int (x-\mu)^2 dF(x) \tag{4.7}$$

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

Theorem **4.14** *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*

$$\mathbb{V}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i). \tag{4.8}$$

Example 4.15 *Let $X \sim \mathrm{Binomial}(n,p)$. 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*

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

Now,

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

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

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4.9}$$

and the **sample variance** to be

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2. \tag{4.10}$$

Theorem **4.16** *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.

Definition 4.17 *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*

$$\mathsf{Cov}(X, Y) = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right] \tag{4.11}$$

and the **correlation** *by*

$$\rho = \rho_{X,Y} = \rho(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sigma_x \sigma_Y}. \tag{4.12}$$

Theorem **4.18** *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 $Y = a + bX$ for some constants a and b then $\rho(X,Y) = 1$ if $b > 0$ and $\rho(X,Y) = -1$ if $b < 0$. If X and Y are independent, then $\mathsf{Cov}(X,Y) = \rho = 0$. The converse is not true in general.

Theorem **4.19** *$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathsf{Cov}(X,Y)$ and $\mathbb{V}(X-Y) = \mathbb{V}(X)+\mathbb{V}(Y)-2\mathsf{Cov}(X,Y)$. More generally, for random variables $X_1, \ldots, X_n$,*

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

4.4 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)	p	n p (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 excercises.

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 is binomial, that is $X_i \sim \text{Binomial}(n, p_i)$. Thus, $\mathbb{E}(X_i) = np_i$ and $\mathbb{V}(X_i) = np_i(1-p_i)$. Note 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 $\mathbb{V}(X_i + X_j) = \mathbb{V}(X_i) + \mathbb{V}(X_i) + 2\mathsf{Cov}(X_i, X_j) = np_i(1-p_i) + np_j(1-p_j) + 2\mathsf{Cov}(X_i, X_j)$. If you equate this formula with $n(p_i + p_j)(p_i + p_j)$ and solve, one gets $\mathsf{Cov}(X_i, X_j) = -np_ip_j$.

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

Lemma 4.20 *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 $\mathbb{V}(AX) = A\Sigma A^T$.*

4.5 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.

Definition 4.21 *The conditional expectation of X given $Y = y$ is*

$$\mathbb{E}(X|Y=y) = \begin{cases} \sum x\, f_{X|Y}(x|y)\, dx & \text{discrete case} \\ \int x\, f_{X|Y}(x|y)\, dx & \text{continuous case.} \end{cases} \tag{4.13}$$

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

$$\mathbb{E}(r(X,Y)|Y=y) = \begin{cases} \sum r(x,y)\, f_{X|Y}(x|y)\, dx & \text{discrete case} \\ \int r(x,y)\, f_{X|Y}(x|y)\, dx & \text{continuous case.} \end{cases} \tag{4.14}$$

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.

Example 4.22 *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. ■

Theorem **4.23** (The rule of iterated expectations.) *For random variables X and Y, assuming the expectations exist, we have that*

$$\mathbb{E}\left[\mathbb{E}(Y|X)\right] = \mathbb{E}(Y) \quad \text{and} \quad \mathbb{E}\left[\mathbb{E}(X|Y)\right] = \mathbb{E}(X). \tag{4.15}$$

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

$$\mathbb{E}\left[\mathbb{E}(r(X,Y)|X)\right] = \mathbb{E}(r(X,Y)) \quad \text{and} \quad \mathbb{E}\left[\mathbb{E}(r(X,Y)|X)\right] = \mathbb{E}(r(X,Y)). \tag{4.16}$$

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

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

Example 4.24 *Consider example 4.22. 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,*

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

Definition 4.25 *The conditional variance is defined as*

$$\mathbb{V}(Y|X=x) = \int (y-\mu(x))^2 f(y|x)dy \tag{4.17}$$

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

Theorem **4.26** *For random variables X and Y,*

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

Example 4.27 *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 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 P has a Uniform (0,1) distribution. Then, $\mathbb{E}(X) = \mathbb{E}\mathbb{E}(X|Q) = \mathbb{E}(nQ) = n\mathbb{E}(Q) = n/2$. 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, $\mathbb{V}(X) = (n/6) + (n^2/12)$.* ■

4.6 Technical Appendix

4.6.1 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 $\int r(x)dF(x) = \sum_{i=1}^k a_i \mathbb{P}(r(X) \in A_i)$. 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 $r^-(x) = -\min\{r(x), 0\}$.

4.6.2 Moment Generating Functions

Definition 4.28 *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 small neighborhood of 0. 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. The mgf is useful for several reasons. First, it helps us compute the moments of a distribution. Second, it helps us find the distribution of sums of random variables. Third, it is used to prove the central limit theorem which we discuss later.

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

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

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

Example 4.29 *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 $\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2 = 2 - 1 = 1$.

Lemma 4.30 *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.

Example 4.31 *Let $X \sim \text{Binomial}(n,p)$. As before we know that $X = \sum_i X_i$ 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, $\psi_X(t) = \prod_i \psi_i(t) = (pe^t + q)^n$.*

Theorem **4.32** *Let X and Y be random variables. If $\psi_X(t) = \psi_Y(t)$ for all t in an open interval around 0, then $X \stackrel{d}{=} Y$.*

Example 4.33 *Let $X \sim \text{Binomial}(n_1,p)$ and $X \sim \text{Binomial}(n_2,p)$ be independent. Let $Y = X_1 + X_2$. Now*

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

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 $Y \sim Bin(n_1+n_2,p)$.*

Moment Generating Function for Some Common Distributions

Distribution	mgf
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)	$\left(\frac{\beta}{\beta-t}\right)^\alpha$ for $t < \beta$

4.7 Excercises

1. Suppose we play a game where we start with c dollars. On each play of the game you either double or half 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 4.6 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 4.16.

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 4.4 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: i.e. $\mathbb{E}(\text{stuff}) = \mathbb{E}(\mathbb{E}(\text{stuff}|X))$.

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 4.16 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. Comment. 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 4.20.

Chapter 5

Inequalities

5.1 Markov and Chebychev 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.

Theorem 5.1 (Markov's Inequality.) *Let X be a non-negative random variable and suppose that $\mathbb{E}(X)$ exists. For any $t > 0$,*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}. \tag{5.1}$$

PROOF. $\mathbb{E}(X) = \int_0^\infty x f(x)dx = \int_0^t x f(x)dx + \int_t^\infty x f(x)dx \geq \int_t^\infty x f(x)dx \geq t\int_t^\infty f(x)dx = t\mathbb{P}(X > t)$. ■

Theorem 5.2 (Chebyshev's inequality.) *Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathbb{V}(X)$. Then,*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \text{and} \quad \mathbb{P}(|Z| \geq k) \leq \frac{1}{k^2} \tag{5.2}$$

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 $t = k\sigma$. ■

Example 5.3 *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 $\mathbb{V}(\overline{X}_n) = \mathbb{V}(X_1)/n^2 = p(1-p)/n$ and*

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

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

5.2 Hoeffding's Inequality

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. The proofs are in the technical appendix.

Theorem 5.4 *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 $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n Y_i \geq \epsilon\right) \leq e^{-t\epsilon} \prod_{i=1}^n e^{t^2(b_i-a_i)^2/8}. \tag{5.3}$$

Theorem 5.5 *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(|\overline{X}_n - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2} \tag{5.4}$$

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

Example 5.6 *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Let $n = 100$ and $\epsilon = .2$. We saw that Chebyshev's yielded*

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

According to Hoeffding's inequality,

$$\mathbb{P}(|\overline{X} - p| > .2) \leq 2e^{-2100(.2)^2} = .00067$$

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 later but here is the basic idea. Fix $\alpha > 0$ and let

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

By Hoeffding's inequality,

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

Let $C = (\overline{X}_n - \epsilon, \overline{X}_n + \epsilon)$. Then, $\mathbb{P}(C \not\ni p) = \mathbb{P}(|\overline{X}_n - p| > \epsilon) \leq \alpha$. 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.

5.3 Cauchy-Schwarz and Jensen Inequalities

This section contains two inequalities on expected values that are often useful.

Theorem **5.7** (Cauchy-Schwarz inequality.) *If X and Y have finite variances then*

$$\mathbb{E}\,|XY| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}. \tag{5.5}$$

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, then convexity reduces to checking that $g''(x) \geq 0$ for all x. It can be shown that if g is convex then it 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 $g(x) = \log x$.

Theorem **5.8** (Jensen's Inequality.) *If g is convex then*

$$\mathbb{E}g(X) \geq g(\mathbb{E}X). \tag{5.6}$$

If g is concave then

$$\mathbb{E}g(X) \leq g(\mathbb{E}X). \tag{5.7}$$

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 $\mathbb{E}(1/X) \geq 1/\mathbb{E}(X)$. Since log is concave, $\mathbb{E}(\log X) \leq \log \mathbb{E}(X)$. For example, suppose that $X \sim N(3,1)$. Then $\mathbb{E}(1/X) \geq 1/3$.

5.4 Technical 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 5.4. For any $t > 0$, we have, from Markov's inequality, that

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^n Y_i \geq \epsilon\right) &= \mathbb{P}\left(t\sum_{i=1}^n Y_i \geq t\epsilon\right) = \mathbb{P}\left(e^{t\sum_{i=1}^n Y_i} \geq e^{t\epsilon}\right) \\ &\leq e^{-t\epsilon}\mathbb{E}\left(e^{t\sum_{i=1}^n Y_i}\right) = e^{-t\epsilon}\prod_i \mathbb{E}(e^{tY_i}). \end{aligned} \tag{5.8}$$

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 $E(Y_i) = 0$ to get

$$Ee^{tY_i} \leq -\frac{a_i}{b_i - a_i}e^{tb_i} + \frac{b_i}{b_i - a_i}e^{ta_i} = e^{g(u)} \tag{5.9}$$

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 (5.8). ■

PROOF of Theorem 5.5. 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 the last Theorem 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 $\mathbb{P}\left(|\overline{X}_n - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$. ■

5.5 Bibliographic Remarks

An excellent reference on probability inequalities and their use in statistics and pattern recognition is Devroye, Györfi and Lugosi (1996). The proof of Hoeffding's inequality is from that text.

5.6 Excercises

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

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

   We call C_n a $1 - \alpha$ *confidence interval for* p. 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?

6
Convergence of Random Variables

6.1 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**. This material is extremely important for statistical inference. 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, $\lim_n x_n = x$. 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 $\mathbb{P}(X_n = Z) = 0$ 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. But $P(X_n = 0) = 0$ for all n. This chapter develops appropriate methods of discussing convergence of random variables.

There are two main ideas in this chapter:

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

2. The **central limit theorem** says that sample average has approximately a Normal distribution for large n. More precisely, $\sqrt{n}(\overline{X}_n - \mu)$ **converges in distribution** to a Normal$(0, \sigma^2)$ distribution, where $\sigma^2 = \mathbb{V}(X)$.

6.2 Types of Convergence

The two main types of convergence are defined as follows.

Definition 6.1 *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 $\epsilon > 0$,*

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \tag{6.1}$$

as $n \to \infty$.

2. X_n **converges to X in distribution**, *written* $X_n \rightsquigarrow X$, *if,*

$$\lim_{n\to\infty} F_n(t) = F(t) \tag{6.2}$$

at all t for which F is continuous.

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

Definition 6.2 X_n **converges to X in quadratic mean** *(also called convergence in L_2), written* $X_n \xrightarrow{\text{qm}} X$, *if,*

$$\mathbb{E}(X_n - X)^2 \to 0 \tag{6.3}$$

as $n \to \infty$.

If X is a point mass at c – that is $\mathbb{P}(X = c) = 1$ – we write $X_n \xrightarrow{\text{qm}} c$ instead of $X \xrightarrow{\text{qm}} X$. Similarly, we write $X_n \xrightarrow{\text{P}} c$ and $X_n \rightsquigarrow c$.

Example 6.3 *Let $X_n \sim N(0, 1/n)$. Intuitively, X_n is concentrating at 0 so we would like to say that $X_n \rightsquigarrow 0$. 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* $X_n \rightsquigarrow 0$. *But notice that* $F_n(0) = 1/2 \neq F(1/2) = 1$ *so convergence fails at* $t = 0$. *But 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.* ■

The next theorem gives the relationship between the types of convergence. The results are summarized in Figure 6.1.

Theorem **6.4** *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 Chebyshev'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 if it 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,

$$F(x - \epsilon) = \mathbb{P}(X \leq x - \epsilon) = \mathbb{P}(X \leq x - \epsilon, X_n \leq x) + \mathbb{P}(X \leq x + \epsilon, X_n > x)$$

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

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,

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

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 $\mathbb{P}(|X_n| > \epsilon) = \mathbb{P}(\sqrt{n} I_{(0,1/n)}(U) > \epsilon) = \mathbb{P}(0 \leq U < 1/n) = 1/n \to 0$. Hence, Then $X_n \xrightarrow{\text{P}} 0$. 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, $X_n \xrightarrow{d} X$. But $\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|2X| > \epsilon) = \mathbb{P}(|X| > \epsilon/2) \neq 0$. So X_n does not tend to X in probability. ■

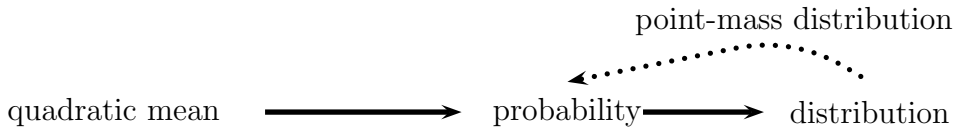


FIGURE 6.1. Relationship between types of convergence.

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

Warning! One might conjecture that if $X_n \xrightarrow{\text{P}} b$ then $\mathbb{E}(X_n) \to b$. This is not 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, $Z \xrightarrow{\text{P}} 0$. 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 6.1.

Some convergence properties are preserved under transformations.

Theorem **6.5** *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 $g(X_n) \rightsquigarrow g(X)$.

6.3 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 and let $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \mathbb{V}(X_1)$. 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 $\mathbb{V}(\overline{X}_n) = \sigma^2/n$.

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 $\mu = \mathbb{E}(X_1)$.

Theorem 6.6 (The Weak Law of Large Numbers (WLLN).) *If $X_1, \ldots, X_n$ are IID , then $\overline{X}_n \xrightarrow{\text{P}} \mu$.*

Interpretation of 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 $n \to \infty$. ■

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

Example 6.7 *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 $n = 84$. ■

6.4 The Central Limit Theorem

Suppose that $X_1, \ldots, X_n$ are iid with mean μ and variance σ. 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.

Theorem 6.8 (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*

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

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:

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

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

Example 6.9 *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 $\mathbb{P}(\overline{X} < 5.5)$. Let $\mu = \mathbb{E}(X_1) = \lambda = 5$ and $\sigma^2 = \mathbb{V}(X_1) = \lambda = 5$. Then,*

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

The central limit theorem tells us that $Z_n = \sqrt{n}(\overline{X} - \mu)/\sigma$ is approximately N(0,1). However, we rarely know σ. 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.

Theorem **6.10** *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.

Theorem **6.11** (Berry-Essèen.) *Suppose that $\mathbb{E}|X_1|^3 < \infty$. Then*

$$\sup_z |\mathbb{P}(Z_n \leq z) - \Phi(z)| \leq \frac{33}{4}\frac{\mathbb{E}|X_1 - \mu|^3}{\sqrt{n}\sigma^3}. \tag{6.4}$$

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

Theorem **6.12 (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 $\overline{X}_i = n^{-1} \sum_{i=1}^n X_{1i}$. *Then,*

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

6.5 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.

Theorem **6.13 (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,

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

Example 6.14 *Let $X_1, \ldots, X_n$ be* IID *with finite mean μ and finite variance σ^2. By the central limit theorem, $\sqrt{n}(\overline{X}_n)/\sigma \rightsquigarrow N(0,1)$. 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.

Theorem **6.15 (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 $g : \mathbb{R}^k \to \mathbb{R}$ ane 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 non-zero. Then

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

Example 6.16 *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 $\mu = (\mu_1, \mu_2)^T$ and variance Σ. Let*

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

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

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

Therefore,

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

6.6 Bibliographic Remarks

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

6.7 Technical Appendix

6.7.1 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 $n \to \infty$.

Theorem **6.17** *Let X_n and X be random vaiables. 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.

Theorem **6.18** (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

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

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

6.7.2 Proof of the Central Limit Theorem

If X is a random variable, define its moment generating function (mgf) by $\psi_X(t) = Ee^{tX}$. Assume in what follows that the mgf is finite in a neighborhood around $t = 0$.

Lemma 6.19 *Let $Z_1, Z_2, \ldots$ be a sequence of random variables. Let ψ_n 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 $\psi'(0) = E(Y_1) = 0$, $\psi''(0) = E(Y_1^2) = Var(Y_1) = 1$. 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

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

6.8 Excercises

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 6.5.

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

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

   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 4 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 $t > 0$.

(a) Show that, for any $k > 0$,

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

(b) (Mill's inequality.) Show that

$$\mathbb{P}(|Z| > t) \leq \left\{\frac{2}{\pi}\right\}^{1/2} \frac{e^{-t^2/2}}{t}.$$

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 $x > t$.

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 i.i.d., 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.

Part II

Statistical Inference

7

Models, Statistical Inference and Learning

7.1 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. The basic statistical inference problem is this:

> We observe $X_1, \ldots, X_n \sim F$. We want to infer (or estimate or learn) F or some feature of F such as its mean.

7.2 Parametric and Nonparametric Models

A **statistical model** is a set of distributions (or a set of densities) $\mathfrak{F}$. 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

$$\mathfrak{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \quad \mu \in \mathbb{R},\ \sigma > 0 \right\}. \tag{7.1}$$

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

$$\mathfrak{F} = \left\{ f(x;\theta) :\ \theta \in \Theta \right\} \tag{7.2}$$

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.

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

Example 7.1 (One-dimensional Parametric Estimation.) *Let $X_1, \ldots, X_n$ be independent* Bernoulli(p) *observations. The problem is to estimate the parameter p.* ■

Example 7.2 (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 (7.1). 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.* ■

Example 7.3 (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 $F \in \mathfrak{F}_{\mathsf{ALL}} = \{\text{all CDF } 's\}$.* ■

Example 7.4 (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."* ■

Example 7.5 (Nonparametric estimation of functionals.) *Let* $X_1, \ldots, X_n \sim F$. *Suppose we want to estimate* $\mu = E(X_1) = \int x\, dF(x)$ *assuming only that* 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 functions are the variance* $T(F) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$ *and the median* $T(F) = F^{-1/2}$. ■

Example 7.6 (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* $r(x) = E(Y|X = x)$ *the* **regression function**. *If we assume that* $f \in \mathfrak{F}$ *where* $\mathfrak{F}$ *is finite dimensional – the set of straight lines for example – then we have a* **parametric regression model**. *If we assume that* $f \in \mathfrak{F}$ *where* $\mathfrak{F}$ *is not finite dimensional then we have a* **parametric 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 functin* f, *then we call this* **regression** *or* **curve estimation**. *Regression models*

are sometimes written as

$$Y = f(X) + \epsilon \tag{7.3}$$

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

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 pro's and con's 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.

7.3 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.

7.3.1 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 $\widehat{\theta}$. 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 point from some distribution F. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

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

We define

$$\mathsf{bias}\,(\widehat{\theta}_n) = \mathbb{E}_\theta(\widehat{\theta}_n) - \theta \tag{7.4}$$

to be the bias of $\widehat{\theta}_n$. We say that $\widehat{\theta}_n$ is **unbiased** if $\mathbb{E}(\widehat{\theta}_n) = \theta$. Unbiasedness used to receive much attention but these days it is not considered very important; many of the estimators we will use are biased. A point estimator $\widehat{\theta}_n$ of a parameter θ is **consistent** if $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$. Consistency is a reasonable requirement for estimators. 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:

$$\mathsf{se} = \mathsf{se}\,(\widehat{\theta}_n) = \sqrt{\mathbb{V}(\widehat{\theta}_n)}. \tag{7.5}$$

Often, it is not possible to compute the standard error but usually we can estimate the standard error. The estimated standard error is denoted by $\widehat{\mathsf{se}}$.

Example 7.7 *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

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

Recall 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 density for θ.

Theorem **7.8** *The* MSE *can be written as*

$$\mathsf{MSE} = \mathsf{bias}\,(\widehat{\theta}_n)^2 + \mathbb{V}_\theta(\widehat{\theta}_n). \tag{7.6}$$

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

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

Theorem **7.9** *If* bias $\to 0$ *and* se $\to 0$ *as* $n \to \infty$ *then* $\widehat{\theta}_n$ *is consistent, that is,* $\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta$.

PROOF. If bias $\to 0$ and se $\to 0$ then, by Theorem 7.8, MSE $\to 0$. It follows that $\widehat{\theta}_n \xrightarrow{\mathrm{qm}} \theta$. (Recall definition 6.3.) The result follows from part (b) of Theorem 6.4. ■

Example 7.10 *Returning to the coin flipping example, we have that* $\mathbb{E}_p(\widehat{p}_n) = p$ *so that* bias $= p - p = 0$ *and* se $= \sqrt{p(1-p)/n} \to 0$. *Hence,* $\widehat{p}_n \xrightarrow{\mathrm{P}} p$, *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.

Definition 7.11 *An estimator is* **asymptotically Normal** *if*

$$\frac{\widehat{\theta}_n - \theta}{\mathsf{se}} \rightsquigarrow N(0,1). \tag{7.7}$$

7.3.2 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

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha, \quad \text{for all } \theta \in \Theta. \tag{7.8}$$

In words, (a,b) traps θ with probability $1-\alpha$. We call $1-\alpha$ the **coverage** of the confidence interval. Commonly, people use 95 per cent confidence intervals which corresponds to choosing $\alpha = 0.05$. Note: C_n **is random and** θ **is fixed!** If θ is a vector then we use a confidence set (such 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 per cent 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 per cent confidence interval for a parameter θ_1. On day

> 2, you collect new data and construct a 95 per cent confidence interval for an unrelated parameter θ_2. On day 3, you collect new data and construct a 95 per cent 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 per cent your intervals will trap the true parameter value. There is no need to introduce the idea of repeating the same experiment over and over.

Example 7.12 *Every day, newspapers report opinion polls. For example, they might say that "83 per cent of the population favor arming pilots with guns." Usually, you will see a statement like "this poll is accurate to within 4 points 95 per cent of the time." They are saying that* 83 ± 4 *is a 95 per cent 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 everyday for the rest of your life, 95 per cent 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.*

Later, 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 C_n, given the data, is 95 per cent." However, these Bayesian intervals refer to degree-of-belief probabilities. These Bayesian intervals will not, in general, trap the parameter 95 per cent of the time.

Example 7.13 *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 (5.4) 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 (7.7) holds, that is, $\widehat{\theta}_n \approx N(\theta, \widehat{\mathsf{se}}^2)$. In this case we can construct (approximate) confidence intervals as follows.

Theorem **7.14** (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*

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}}\,,\ \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}\,). \tag{7.9}$$

Then

$$\mathbb{P}_\theta(\theta \in C_n) \to 1-\alpha. \tag{7.10}$$

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 per cent confidence intervals, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96 \approx 2$ leading to the approximate 95 per cent confidence interval $\widehat{\theta}_n \pm 2\widehat{\mathsf{se}}$. We will discuss the construction of confidence intervals in more generality in the rest of the book.

Example 7.15 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ *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,* $\widehat{p}_n \approx N(p, \widehat{se}^2)$. *Therefore, an approximate* $1-\alpha$ *confidence interval is*

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

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

7.3.3 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.

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

Example 7.16 (Testing if a Coin is Fair) *Suppose* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ *denote* 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. ■

7.4 Bibliographic Remarks

Statistical inference is covered in many texts. Elementary texts include DeGroot and Schervish (2001) and Larsen and Marx (1986). At the intermediate level I recommend Casella and Berger (2002) and Bickel and Doksum (2001). At the advanced level, Lehmann and Casella (1998), Lehmann (1986) and van der Vaart (1998).

7.5 Technical Appendix

Our definition of confidence interval requires that $\mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha$ for all $\theta \in \Theta$. An **pointwise asymptotic** confidence interval requires that $\liminf_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha$ for all $\theta \in \Theta$. An **uniform asymptotic** confidence interval requres 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. In general, it might not be a uniform asymptotic confidence interval.

8

Estimating the CDF and Statistical Functionals

The first inference problem we will consider is nonparametric estimation of the CDF F and functions of the CDF.

8.1 The Empirical Distribution Function

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

Definition 8.1 *The* **empirical distribution function** $\widehat{F}_n$ *is the* CDF *that puts mass* $1/n$ *at each data point* X_i. *Formally,*

$$\begin{aligned} \widehat{F}_n(x) &= \frac{\sum_{i=1}^n I(X_i \leq x)}{n} \\ &= \frac{\text{number of observations less than or equal to } x}{n} \end{aligned} \tag{8.1}$$

where

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

Example 8.2 (Nerve Data.) *Cox and Lewis (1966) reported 799 waiting times between successive pulses along a nerve fibre. The first plot in Figure 8.1 shows the a "toothpick plot" where each toothpick shows the location of one data point. The second plot shows that empirical* CDF $\widehat{F}_n$. *Suppose we want to estimate the fraction of waiting times between .4 and .6 seconds. The estimate is* $\widehat{F}_n(.6) - \widehat{F}_n(.4) = .93 - .84 = .09$. ■

The following theorems give some properties of $\widehat{F}_n(x)$.

Theorem **8.3** *At any fixed value of* x,

$$\mathbb{E}\left(\widehat{F}_n(x)\right) = F(x) \quad \text{and} \quad \mathbb{V}\left(\widehat{F}_n(x)\right) = \frac{F(x)(1-F(x))}{n}.$$

Thus,

$$\mathsf{MSE} = \frac{F(x)(1-F(x))}{n} \to 0$$

and hence, $\widehat{F}_n(x) \xrightarrow{\text{P}} F(x)$.

Actually, $\sup_x |\widehat{F}_n(x) - F(x)|$ converges to 0 almost surely.

Theorem **8.4** (Glivenko-Cantelli Theorem.) *Let* $X_1, \ldots, X_n \sim F$. *Then*

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

8.2 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 $m = F^{-1}(1/2)$.

Definition 8.5 *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.

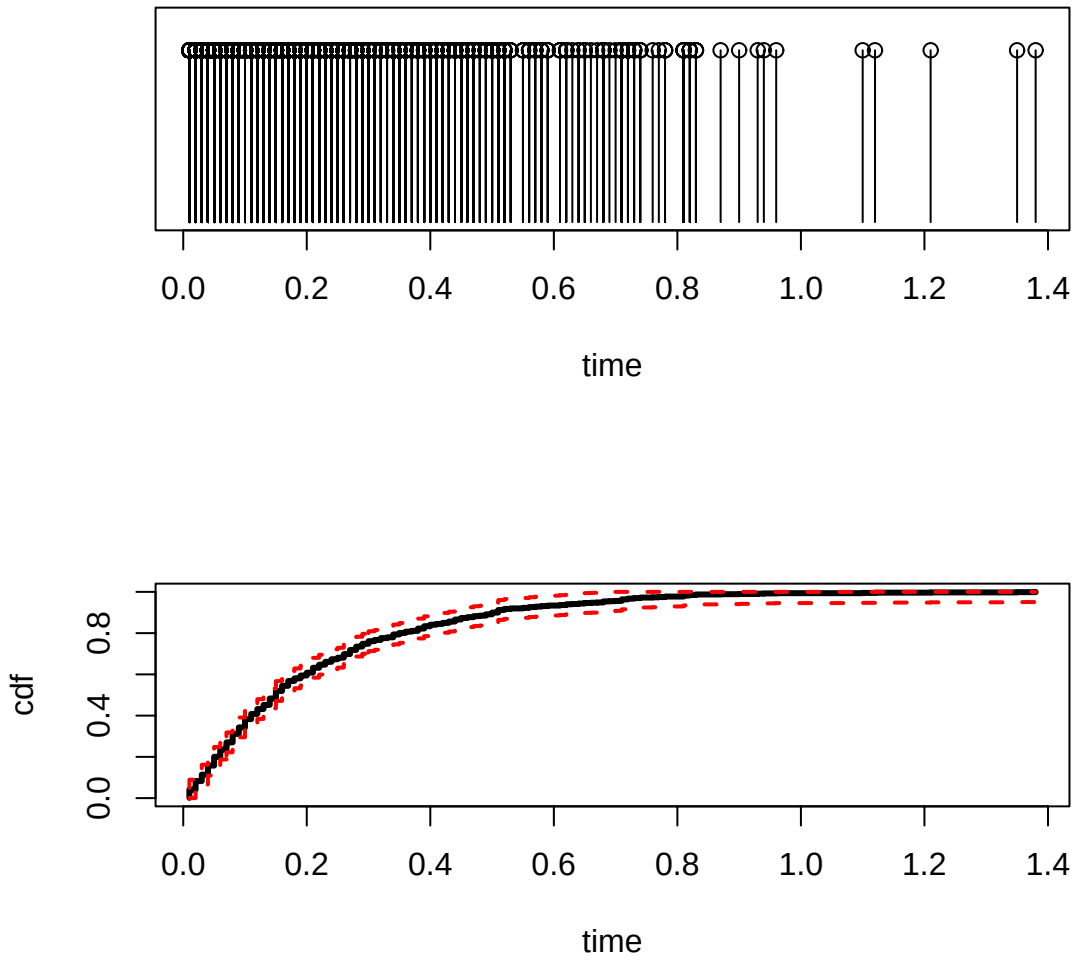


FIGURE 8.1. Nerve data. The solid line in the middle is the empirical distribution function. The lines above and below the middle line are a 95 per cent confidence band. The confidence band is explained in the appendix.

A functional of the form $\int r(x)dF(x)$ is called a **linear functional.** 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:

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

$$T(\widehat{F}_n) = \int r(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} r(X_i). \tag{8.2}$$

Sometimes we can find the estimated standard error $\widehat{se}$ 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 $\widehat{se}$. For now, let us just assume that somehow we can find $\widehat{se}$. In many cases, it turns out that

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

By equation (7.10), an approximate $1-\alpha$ confidence interval for $T(F)$ is then

$$T(\widehat{F}_n) \pm z_{\alpha/2}\, \widehat{se}. \tag{8.3}$$

We will call this the **Normal-based interval.** For a 95 per cent confidence interval, $z_{\alpha/2} = z_{.05/2} = 1.96 \approx 2$ so the interval is $T(\widehat{F}_n) \pm 2\,\widehat{se}$.

Example 8.6 (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* $\text{se} = \sqrt{\mathbb{V}(\overline{X}_n)} = \sigma/\sqrt{n}$. *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* $\overline{X}_n \pm z_{\alpha/2}\,\mathrm{se}^2$. ■

Example 8.7 (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*

$$\begin{aligned}\widehat{\sigma}^2 &= \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.\end{aligned}$$

Another reasoable 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 the our last example, we now see that the estimated standard error of the estimate of the mean is* $\widehat{se} = \widehat{\sigma}/\sqrt{n}$. ■

Example 8.8 (The Skewness) *Let* μ *and* σ^2 *denote the mean and variance of a random variable* X. *The skewness is defined to be*

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

The skewness measure 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*

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

Example 8.9 (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 $T_1(F)$, ..., $T_5(F)$, 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**. ■

Example 8.10 (Quantiles.) *Let F be strictly increasing with density f. The $T(F) = F^{-1}(p)$ be the p^{th} quantile. 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 $\widehat{F}_n^{-1}(p)$ 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 formulae for standard errors and confidence intervals. But in our nonparametric setting we need something else. In the next chapter, we will introduce two methods – the jackknife and the bootstrap – for getting standard errors and confidence intervals.

Example 8.11 (Plasma Cholesterol.) *Figure 8.2 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*

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

which we estimate by

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

where

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

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

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*

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

and we estimate this by

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

An approximate 95 per cent confidence interval for θ *is* $\widehat{\theta} \pm 2\widehat{\mathsf{se}} = (9.8, 32.0)$. *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 later in this text.* ■

8.3 Technical Appendix

In this appendix we explain how to construct a confidence band for the CDF .

Theorem **8.12** (Dvoretzky-Kiefer-Wolfowitz (DKW) inequality.) *Let* $X_1, \ldots, X_n$ *be iid from* F. *Then, for any* $\epsilon > 0$,

$$\mathbb{P}\left(\sup_x |F(x) - \widehat{F}_n(x)| > \epsilon \right) \leq 2e^{-2n\epsilon^2}. \tag{8.4}$$

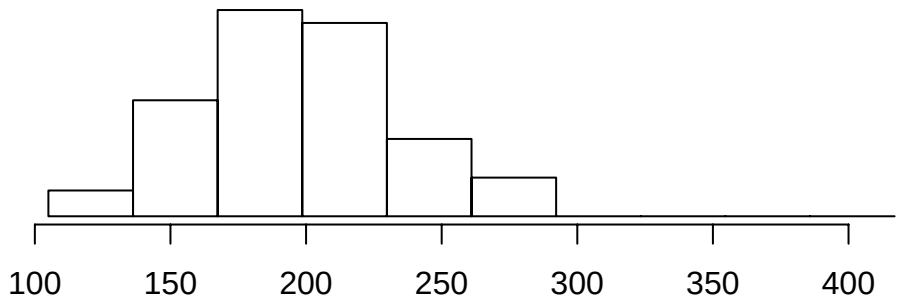
From the DKW inequality, we can construct a confidence set. Let $\epsilon_n^2 = \log(2/\alpha)/(2n)$, $L(x) = \max\{\widehat{F}_n(x) - \epsilon_n, 0\}$ and $U(x) = \min\{\widehat{F}_n(x) + \epsilon_n, 1\}$. It follows from (8.4) that for any F,

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

Thus, C_n is a nonparametric $1-\alpha$ confidence set for F. A better name for C_n is a **confidence band**. To summarize:

A $1-\alpha$ nonparametric confidence band for F is $(L(x),\ U(x))$ where

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



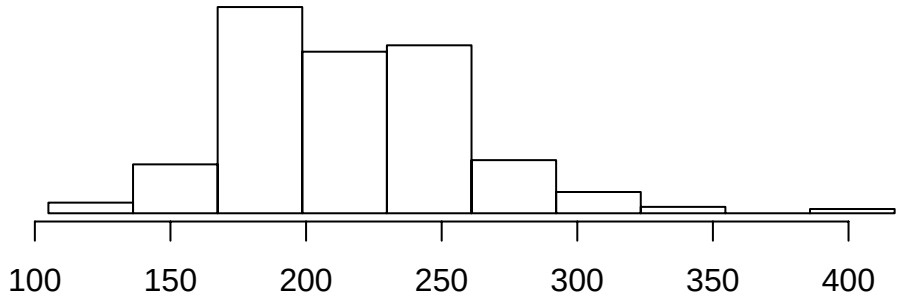


FIGURE 8.2. Plasma cholesterol for 51 patients with no heart disease and 320 patients with narrowing of the arteries.

Example 8.13 *The dashed lines in Figure 8.1 give a 95 per cent confidence band using* $\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{.05}\right)} = .048$. ■

8.4 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).

8.5 Exercises

1. Prove Theorem 8.3.

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 per cent confidence interval for p. Find the plug-in estimator and estimated standard error for $p-q$. Find an approximate 90 per cent confidence interval for $p - q$.

3. (Computer Experiment.) Generate 100 observations from a N(0,1) distribution. Compute a 95 per cent confidence band for the CDF F. 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 course website. Estimate the cdf $F(x)$. Compute and plot a 95 per cent confidence envelope for F. Find an approximate 95 per cent 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 course website. Estimate the mean waiting time and give a standard error for the estimate. Also, give a 90 per cent 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 per cent confidence interval and a 95 per cent 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 median precipitation from the two groups. Estimate θ. Estimate the standard error of the estimate and produce a 95 per cent confidence interval.

9
The Bootstrap

The bootstrap is a nonparametric method for estimating standard errors and computing confidence intervals. Let

$$T_n = g(X_1, \ldots, X_n)$$

be a **statistic**, that 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 $T_n = n^{-1}\sum_{i=1}^n X_i$ then $\mathbb{V}_F(T_n) = \sigma^2/n$ where $\sigma^2 = \int (x-\mu)^2 dF(x)$ and $\mu = \int x dF(x)$. The bootstrap idea has two parts. First we estimate $\mathbb{V}_F(T_n)$ with $\mathbb{V}_{\widehat{F}_n}(T_n)$. Thus, for $T_n = n^{-1}\sum_{i=1}^n X_i$ we have $\mathbb{V}_{\widehat{F}_n}(T_n) = \widehat{\sigma}^2/n$ where $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)$. However, in more complicated cases we cannot write down a simple formula for $\mathbb{V}_{\widehat{F}_n}(T_n)$. This leads us to the second step which is to approximate $\mathbb{V}_{\widehat{F}_n}(T_n)$ using simulation. Before proceeding, let us discuss the idea of simulation.

9.1 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,

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

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

9.2 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.** Thu, 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:

Boostrap 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

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,b}^*\right)^2. \tag{9.1}$$

Example 9.1 *The following pseudo-code shows how to use the bootstrap to estimate the standard 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:N){
    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:

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

Example 9.2 *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 medin example except we compute the skewness from each bootstrap sample. When applied to the nerve data, the bootstrap, based on $B = 1000$ *replications, yields a standard error for the estimated skewness of .16.*

9.3 Bootstrap Confidence Intervals

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

Normal Interval. The simplest is the Normal interval

$$T_n \pm z_{\alpha/2}\ \widehat{\mathsf{se}}_{\text{boot}} \tag{9.2}$$

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

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:

$$H(r) = \mathbb{P}_F(R_n \leq r). \tag{9.3}$$

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

$$a = \widehat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \quad \text{and} \quad b = \widehat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right). \tag{9.4}$$

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, $C^\star_n$ 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:

$$\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R^*_{n,b} \leq r) \tag{9.5}$$

where $R^*_{n,b} = \widehat{\theta}^*_{n,b} - \widehat{\theta}_n$. Let r^*_β denote the β sample qauntile of $(R^*_{n,1}, \ldots, R^*_{n,B})$ and let θ^*_β denote the β sample qauntile of $(\theta^*_{n,1}, \ldots, \theta^*_{n,B})$. 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

$$C_n = \left(2\widehat{\theta}_n - \widehat{\theta}^*_{1-\alpha/2},\ 2\widehat{\theta}_n - \widehat{\theta}^*_{\alpha/2}\right). \tag{9.6}$$

Typically, this is a pointwise, asymptotic confidence interval.

Theorem **9.3** *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 (9.6).*

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.

Example 9.4 *For estimating the skewness of the nerve data, here are the various confidence intervals.*

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

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.

Example 9.5 (The Plasma Cholesterol Data.) *Let us return to the cholesterol data. Suppose we are interested in the difference of the medians. Pseudo-code 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 per cent confidence intervals are as follows:

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

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.

Example 9.6 *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 9.1 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 per cent 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.*

Example 9.7 *This example is borrowed from* An Introduction to the Bootstrap *by B. Efron and R. Tibshirani. When drug companies introduce new medications, they are sometimes requires 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

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*

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

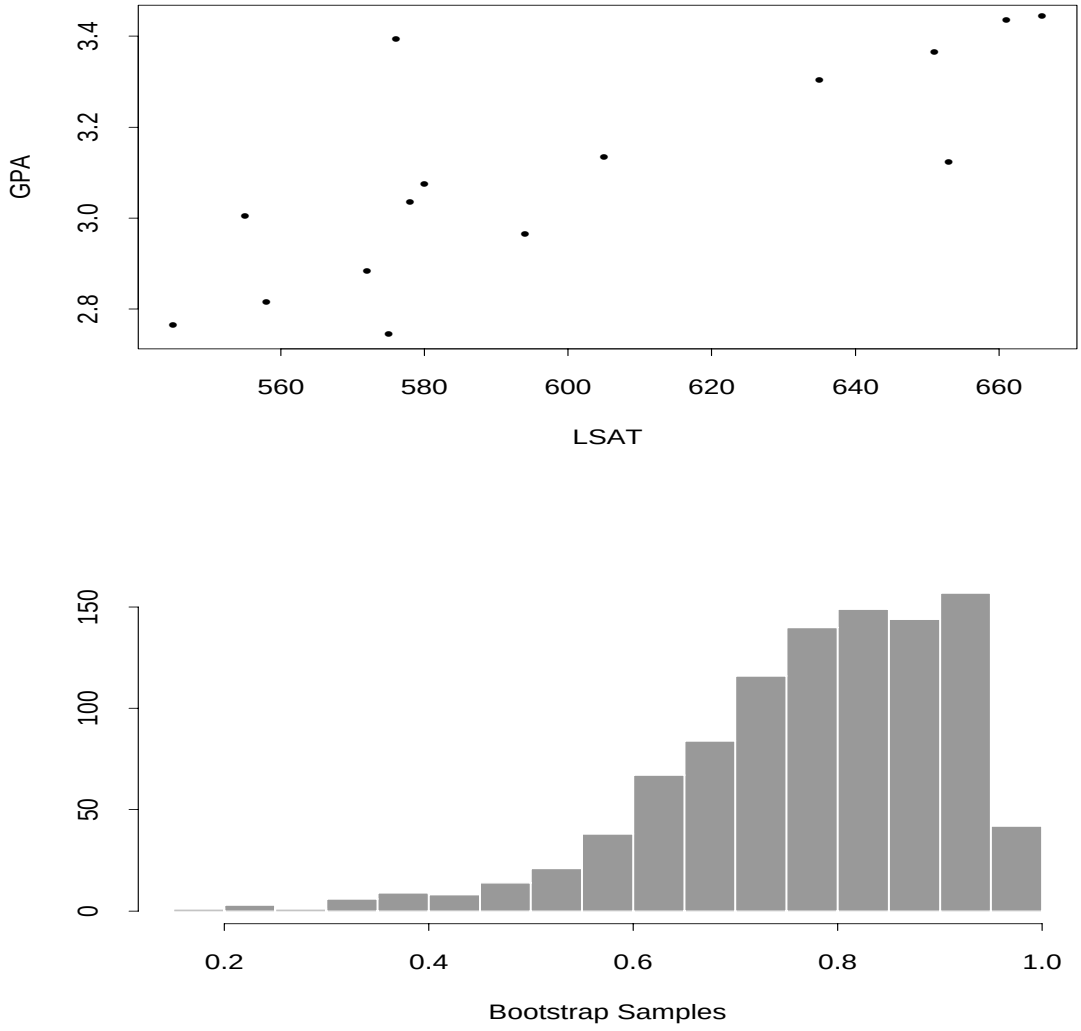


FIGURE 9.1. Law school data.

The estimate of θ *is*

$$\widehat{\theta} = \frac{\overline{Y}}{\overline{Z}} = \frac{-452.3}{6342} = -.0713.$$

The bootstrap standard error is $\widehat{se} = .105$. *To answer the bioequivalence question, we compute a confidence interval. From* $B = 1000$ *bootstrap replications we get the 95 per cent interval is (-.24,.15). This is not quite contained in [-.20,.20] so at the 95 per cent level we have not demonstrated bioequivalence. Figure 9.2 shows the histogram of the bootstrap values.*

9.4 Bibliographic Remarks

The boostrap 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).

9.5 Technical Appendix

9.5.1 The Jackknife

There is another method for computing standard errors called the jackknife, due to Quenouille (1949). It is less computationally expensive than the bootstrap 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 $\widehat{se}_{jack} = \sqrt{v_{jack}}$. Under suitable conditions on T, it can be shown that v_{jack} consistently estimates $\text{var}(T_n)$ in the sense that $v_{\text{jack}}/\text{var}(\text{T}_\text{n}) \xrightarrow{p}$

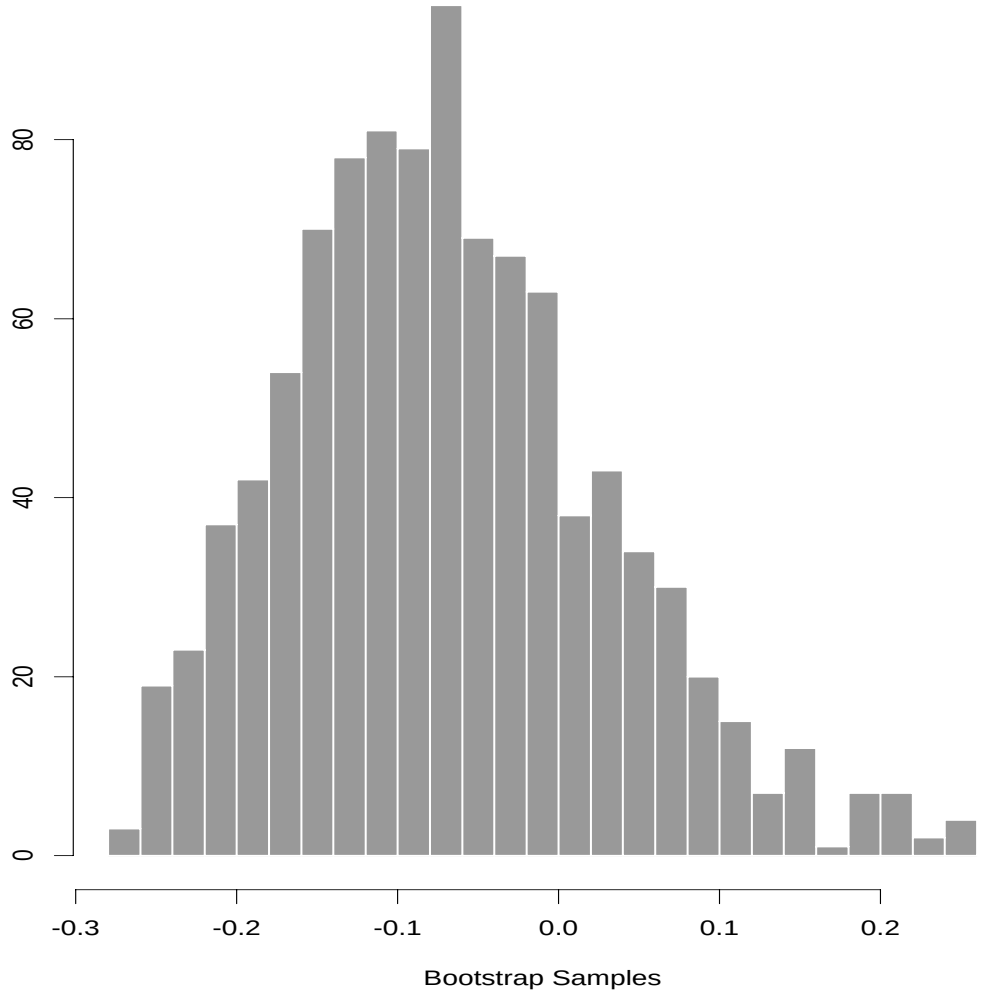


FIGURE 9.2. Patch data.

1. However, unlike the bootstrap, the jackknife does not produce consistent estimates of the standard error of sample quantiles.

9.5.2 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 exist. 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.

9.6 Excercises

1. Consider the data in Example 9.6. Find the plug-in estimate of the correlation coefficient. Estimate the standard error using the bootstrap. Find a 95 per cent confidence inerval using all three methods.

2. (Computer Experiment.) Conduct a simulation to compare the four bootstrap confidence interval methods. Let $n = 50$ and let $T(F) = \int (x - \mu)^3 dF(x)/\sigma^3$ be the skewness. Draw $Y_1, \ldots, X_n \sim N(0, 1)$ and set $X_i = e^{Y_i}$, $i = 1, \ldots, n$. Construct the four types of bootstrap 95 per cent 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 four 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.

   Remark: The jackknife does not give a consistent estimator of the variance of a quantile.

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 $\widehat{\theta} = e^{\overline{X}}$ be the mle. 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 for the parametric and nonparametric bootstraps. These are estimates of the distribution of $\widehat{\theta}$. Compare this to the true sampling distribution of $\widehat{\theta}$.

7. Let $X_1, ..., X_n$ Unif$(0, \theta)$. The mle is $\widehat{\theta} = X_{max} = \max\{X_1, ..., X_n\}$. 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 parametric and nonparametric bootstraps.

   (b) This is a case where the nonparametric bootstrap does very poorly. In fact, we can prove that this is the case. Show that, for the parametric bootstrap $P(\widehat{\theta}^* = \widehat{\theta}) = 0$ but for the nonparametric bootstrap $P(\widehat{\theta}^* = \widehat{\theta}) \approx .632$. 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

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

10
Parametric Inference

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

$$\mathfrak{F} = \Big\{ f(x;\theta) : \ \theta \in \Theta \Big\} \tag{10.1}$$

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 will begin with a brief dicussion 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.

10.1 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 can be a complicated function of θ as in the following example.

Example 10.1 *Let $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$. The parameter is $\theta = (\mu, \sigma)$ 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*

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

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

Example 10.2 *Recall that X has a $\text{Gamma}(\alpha, \beta)$ 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* $T(\alpha, \beta) = \mathbb{E}_\theta(X_1) = \alpha\beta$. ■

10.2 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**

$$\alpha_j \equiv \alpha_j(\theta) = \mathbb{E}_\theta(X^j) = \int x^j dF_\theta(x) \tag{10.2}$$

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

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j. \tag{10.3}$$

Definition 10.3 *The* **method of moments estimator** $\widehat{\theta}_n$ *is defined to be the value of* θ *such that*

$$\begin{aligned} \alpha_1(\widehat{\theta}_n) &= \widehat{\alpha}_1 \\ \alpha_2(\widehat{\theta}_n) &= \widehat{\alpha}_2 \\ &\vdots \quad \vdots \quad \vdots \\ \alpha_k(\widehat{\theta}_n) &= \widehat{\alpha}_k. \end{aligned} \tag{10.4}$$

Formula (10.4) defines a system of k equations with k unknowns.

Example 10.4 *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*

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

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

Example 10.5 *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*

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

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

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

Theorem **10.6** *Let $\widehat{\theta}_n$ denote the method of moments estimator. Under the the conditions given in the appendix, 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.

10.3 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 $f(x;\theta)$.

Definition 10.7 *The* **likelihood function** *is defined by*

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta). \tag{10.5}$$

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.

Definition 10.8 *The* **maximum likelihood estimator** MLE , *denoted by* $\widehat{\theta}_n$, *is the value of* θ *that maximizes* $\mathcal{L}_n(\theta)$.

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.

Remark 10.9 *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*

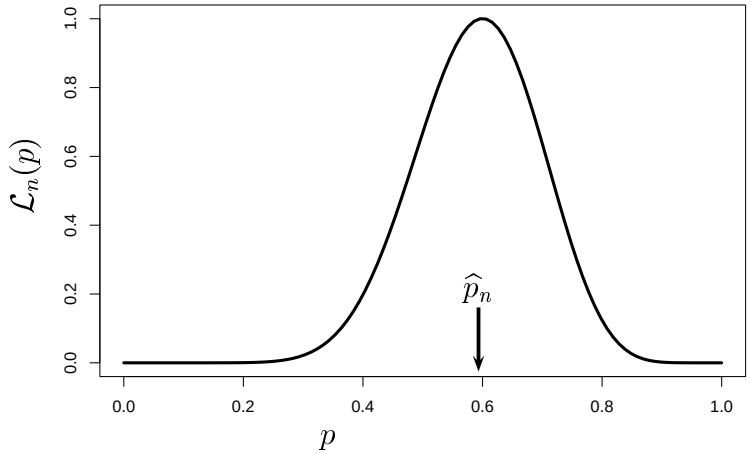


FIGURE 10.1. Likelihood function for Bernoulli with $n = 20$ and $\sum_{i=1}^n X_i = 12$. The MLE is $\widehat{p}_n = 12/20 = 0.6$.

shall often be sloppy about dropping constants in the likelihood function.

Example 10.10 *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 10.1.* ■

Example 10.11 *Let* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. *The parameter is* $\theta = (\mu, \sigma)$ *and the likelihood function is (ignoring some constants)*

$$\mathcal{L}_n(\mu, \sigma) = \prod_i \frac{1}{\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(X_i - \mu)^2 \right\}$$

$$= \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (X_i - \mu)^2\right\}$$
$$= \sigma^{-n} \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{X} - \mu)^2}{2\sigma^2}\right\}$$

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.* ■

Example 10.12 (A Hard Example.) *Here is the example that confuses everyone. Let* $X_1, \ldots, X_n \sim Unif(0, \theta)$. *Recall that*

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise.} \end{cases}$$

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 10.2. Now $\mathcal{L}_n(\theta)$ *is strictly decreasing over the interval* $[X_{(n)}, \infty)$. *Hence,* $\widehat{\theta}_n = X_{(n)}$. ■

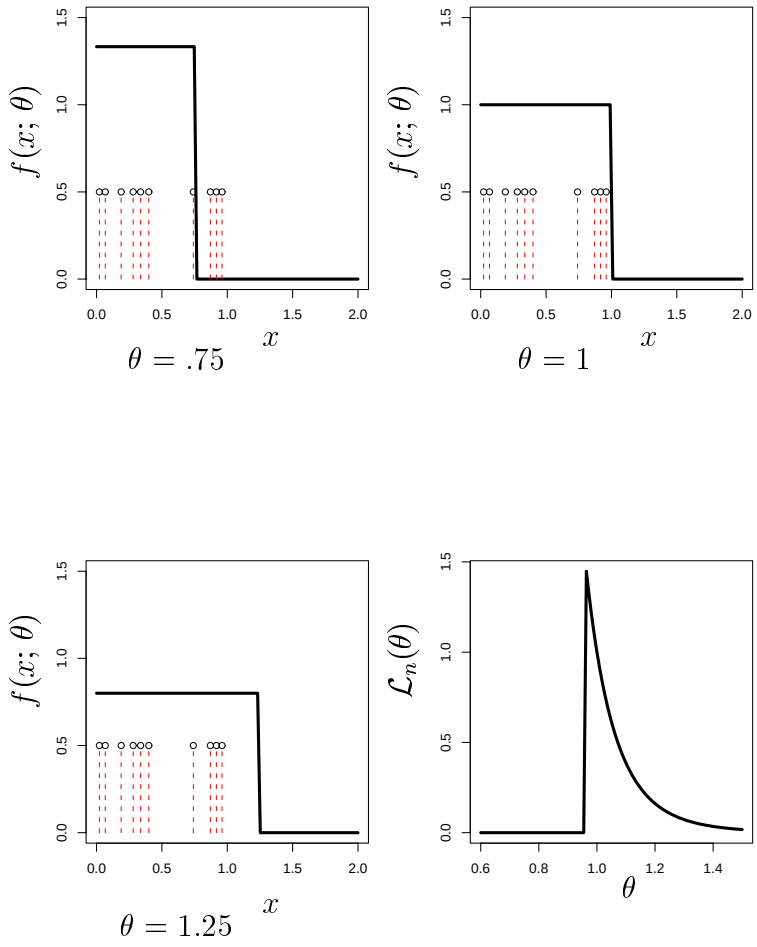


FIGURE 10.2. Likelihood function for Uniform $(0, \theta)$. The vertical lines show the observed data. The first three plots show $f(x;\theta)$ for three different values of θ. When $\theta < X_{(n)} = \max\{X_1, \ldots, X_n\}$, as in the first plot, $f(X_{(n)};\theta) = 0$ and hence $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i;\theta) = 0$. 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.

10.4 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:

(1) It is **consistent**: $\widehat{\theta}_n \xrightarrow{\text{P}} \theta_\star$ where $\theta_\star$ denotes the true value of the parameter θ;

(2) It is **equivariant**: if $\widehat{\theta}_n$ is the MLE of θ then $g(\widehat{\theta}_n)$ is the MLE of $g(\theta)$;

(3) It is **asymptotically Normal**: $\sqrt{n}(\widehat{\theta} - \theta_\star)/\widehat{\mathsf{se}} \rightsquigarrow N(0,1)$ where $\widehat{\mathsf{se}}$ can be computed analytically;

(4) It 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. (To 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.**

10.5 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** between f and g to be

$$D(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx. \tag{10.6}$$

This is not a distance in the formal sense.

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 assume that $\theta \neq \psi$ implies that $D(\theta,\psi) > 0$.

Let $\theta_\star$ denote the true value of θ. Maximizing $\ell_n(\theta)$ is equivalent to maximizing

$$M_n(\theta) = \frac{1}{n}\sum_i \log \frac{f(X_i;\,\theta)}{f(X_i;\,\theta_\star)}.$$

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 $-D(\theta_\star,\theta_\star) < 0$ for $\theta \neq \theta_\star$. Hence, 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.

Theorem **10.13** *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

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\text{P}} 0 \tag{10.7}$$

and that, for every $\epsilon > 0$,

$$\sup_{\theta:|\theta-\theta_\star|\geq\epsilon} M(\theta) < M(\theta_\star). \tag{10.8}$$

Let $\widehat{\theta}_n$ *denote the mle. Then* $\widehat{\theta}_n \xrightarrow{\text{P}} \theta_\star$.

PROOF. See appendix. ■

10.6 Equivariance of the MLE

Theorem **10.14** *Let* $\tau = g(\theta)$ *be a one-to-one 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, $L(\tau) = \prod_i f(x_i; h(\tau)) = \prod_i f(x_i; \theta) = \mathcal{L}(\theta)$ where $\theta = h(\tau)$. Hence, for ana τ, $\mathcal{L}_n(\tau) = \mathcal{L}(\theta) \leq \mathcal{L}(\widehat{\theta}) = \mathcal{L}_n(\widehat{\tau})$. ■

Example 10.15 *Let* $X_1, \ldots, X_n \sim N(\theta, 1)$. *The mle for* θ *is* $\widehat{\theta}_n = \overline{X}_n$. *Let* $\tau = e^\theta$. *Then, the mle for* τ *is* $\widehat{\tau} = e^{\widehat{\theta}} = e^{\overline{X}}$. ■

10.7 Asymptotic Normality

It turns out that $\widehat{\theta}_n$ is approximately Normal and we can compute its variance analytically. To explore this, we first need a few definitions.

Definition 10.16 *The* **score function** *is defined to be*

$$s(X;\theta) = \frac{\partial \log f(X;\theta)}{\partial \theta}. \tag{10.9}$$

The **Fisher information** *is defined to be*

$$\begin{aligned} I_n(\theta) &= \mathbb{V}_\theta\left(\sum_{i=1}^n s(X_i;\theta)\right) \\ &= \sum_{i=1}^n \mathbb{V}_\theta\left(s(X_i;\theta)\right). \end{aligned} \tag{10.10}$$

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.

Theorem **10.17** *$I_n(\theta) = nI(\theta)$ and*

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta\left(\frac{\partial^2 \log f(X;\theta)}{\partial^2 \theta^2}\right) \\ &= -\int \left(\frac{\partial^2 \log f(x;\theta)}{\partial^2 \theta^2}\right) f(x;\,\theta)dx. \end{aligned} \tag{10.11}$$

Theorem **10.18 (Asymptotic Normality of the MLE .)** *Under appropriate regularity conditions, the following hold:*

1. *Let* $\mathsf{se} = \sqrt{1/I_n(\theta)}$. *Then,*

$$\frac{(\widehat{\theta}_n - \theta)}{\mathsf{se}} \rightsquigarrow N(0,1). \tag{10.12}$$

2. *Let* $\widehat{\mathsf{se}} = \sqrt{1/I_n(\widehat{\theta}_n)}$. *Then,*

$$\frac{(\widehat{\theta}_n - \theta)}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1). \tag{10.13}$$

The proof is in the appendix. The first statement says that $\widehat{\theta}_n \approx N(\theta, \mathsf{se})$ where the 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 $N(\theta, \widehat{se}^2)$. From this fact we can construct an (asymptotic) confidence interval.

Theorem **10.19** *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,

$$\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}_\theta\left(\widehat{\theta}_n - z_{\alpha/2}\widehat{\mathsf{se}} \le \theta \le \widehat{\theta}_n + z_{\alpha/2}\widehat{\mathsf{se}}\right) \\
&= \mathbb{P}_\theta\left(-z_{\alpha/2} \le \frac{\widehat{\theta}_n - \theta}{\widehat{\mathsf{se}}} \le z_{\alpha/2}\right)
\end{aligned}$$

$$\rightarrow \quad \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha. \quad \blacksquare$$

For $\alpha = .05$, $z_{\alpha/2} = 1.96 \approx 2$, so:

$$\widehat{\theta}_n \pm 2\,\widehat{\mathsf{se}} \tag{10.14}$$

is an approximate 95 per cent 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 per cent of the time. They are simply giving a 95 per cent confidence interval of the form $\widehat{\theta}_n \pm 2\widehat{\mathsf{se}}$.

Example 10.20 *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. The* MLE *is $\widehat{p}_n = \sum_i X_i/n$ and $f(x;p) = p^x(1-p)^{1-x}$, $\log f(x;p) = x\log p + (1-x)\log(1-p)$, $s(X;p) = (x/p) - (1-x)/(1-p)$, and $-s'(X;p) = (x/p^2) + (1-x)/(1-p)^2$. 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 per cent confidence interval is

$$\widehat{p}_n \pm 2\left\{\frac{\widehat{p}(1-\widehat{p})}{n}\right\}^{1/2}.$$

Compare this with the Hoeffding interval. $\blacksquare$

Example 10.21 *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 10.18, $\overline{X}_n \approx N(\theta, \sigma^2/n)$. In fact, in this case, the distribution is exact.* $\blacksquare$

Example 10.22 *Let* $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. *Then* $\widehat{\lambda}_n = \overline{X}_n$ *and some calculations show that* $I_1(\lambda) = 1/\lambda$, *so*

$$\widehat{\text{se}} = \frac{1}{\sqrt{nI_1(\widehat{\lambda}_n)}} = \sqrt{\frac{\widehat{\lambda}_n}{n}}.$$

Therefore, an approximate $1-\alpha$ *confidence interval for* λ *is* $\widehat{\lambda}_n \pm z_{\alpha/2}\sqrt{\widehat{\lambda}_n/n}$. ■

10.8 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 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 $ARE(U,T) = t^2/u^2$. In the Normal example, $ARE(\widetilde{\theta}_n, \widehat{\theta}_n) = 2/\pi = .63$. The interpretation is that if you use the median, you are only using about 63 per cent of the available data.

Theorem **10.23** *If* $\widehat{\theta}_n$ *is the* MLE *and* $\widetilde{\theta}_n$ *is any other estimator then*

The result is actually more subtle than this but we needn't worry about the details.

$$ARE(\widetilde{\theta}_n, \widehat{\theta}_n) \leq 1.$$

Thus, the MLE *has the smallest (asymptotic) variance and we say that* 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.

10.9 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 $\widehat{\tau}$?

Theorem **10.24 (The Delta Method)** *If $\tau = g(\theta)$ where g is differentiable and $g'(\theta) \neq 0$ then*

$$\frac{\sqrt{n}(\widehat{\tau}_n - \tau)}{\widehat{\mathsf{se}}\,(\widehat{\tau})} \rightsquigarrow N(0,1) \tag{10.15}$$

where $\widehat{\tau}_n = g(\widehat{\theta}_n)$ and

$$\widehat{\mathsf{se}}\,(\widehat{\tau}_n) = |g'(\widehat{\theta})|\,\widehat{\mathsf{se}}\,(\widehat{\theta}_n) \tag{10.16}$$

Hence, if

$$C_n = \Big(\widehat{\tau}_n - z_{\alpha/2}\widehat{\mathsf{se}}\,(\widehat{\tau}_n),\ \widehat{\tau}_n + z_{\alpha/2}\widehat{\mathsf{se}}\,(\widehat{\tau}_n)\Big) \tag{10.17}$$

then $\mathbb{P}_\theta(\tau \in C_n) \to 1 - \alpha$ as $n \to \infty$.

Example 10.25 *Let $X_1, \ldots, X_n \sim$ Bernoulli(p) and let $\psi = g(p) = \log(p/(1-p))$. The Fisher information function is $I(p) = 1/(p(1-p))$ so the standard error of the* MLE *$\widehat{p}_n$ is $\mathsf{se} = \{\widehat{p}_n(1-\widehat{p}_n)/n\}^{1/2}$.*

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 per cent confidence interval is

$$\widehat{\psi}_n \pm \frac{2}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}. \quad \blacksquare$$

Example 10.26 *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*

$$\widehat{\sigma} = \left\{ \frac{\sum_i (X_i - \mu)^2}{n} \right\}^{1/2}.$$

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

$$I(\sigma) = -\frac{1}{\sigma^2} + \frac{3\sigma^2}{\sigma^4} = \frac{2}{\sigma^2}.$$

Hence, $\widehat{\mathsf{se}} = \widehat{\sigma}_n/\sqrt{2n}$. *Let* $\psi = g(\sigma) = \log(\sigma)$. *Then,* $\widehat{\psi}_n = \log \widehat{\sigma}_n$. *Since,* $g' = 1/\sigma$,

$$\widehat{\mathsf{se}}\,(\widehat{\psi}_n) = \frac{1}{\widehat{\sigma}} \frac{\widehat{\sigma}}{\sqrt{2n}} = \frac{1}{\sqrt{2n}}$$

and an approximate 95 per cent confidence interval is $\widehat{\psi}_n \pm 2/\sqrt{2n}$. $\blacksquare$

10.10 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

$$I_n(\theta) = - \begin{bmatrix} \mathbb{E}_\theta(H_{11}) & \mathbb{E}_\theta(H_{12}) & \cdots & \mathbb{E}_\theta(H_{1k}) \\ \mathbb{E}_\theta(H_{21}) & \mathbb{E}_\theta(H_{22}) & \cdots & \mathbb{E}_\theta(H_{2k}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{E}_\theta(H_{k1}) & \mathbb{E}_\theta(H_{k2}) & \cdots & \mathbb{E}_\theta(H_{kk}) \end{bmatrix}. \tag{10.18}$$

Let $J_n(\theta) = I_n^{-1}(\theta)$ be the inverse of I_n.

Theorem **10.27** *Under appropriate regularity conditions,*

$$\sqrt{n}(\widehat{\theta} - \theta) \approx N(0, J_n).$$

Also, if $\widehat{\theta}_j$ is the j^{th} component of $\widehat{\theta}$, then

$$\frac{\sqrt{n}(\widehat{\theta}_j - \theta_j)}{\widehat{\mathsf{se}}_j} \rightsquigarrow N(0,1)$$

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

$$\nabla g = \begin{pmatrix} \dfrac{\partial g}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial g}{\partial \theta_k} \end{pmatrix}$$

be the gradient of g.

Theorem **10.28 (Multiparameter delta method)** *Suppose that* ∇g *evaluated at* $\widehat{\theta}$ *is not 0. Let* $\widehat{\tau} = g(\widehat{\theta})$. *Then*

$$\frac{\sqrt{n}(\widehat{\tau} - \tau)}{\widehat{\mathsf{se}}\,(\widehat{\tau})} \rightsquigarrow N(0,1)$$

where

$$\widehat{\mathsf{se}}\,(\widehat{\tau}) = \sqrt{(\widehat{\nabla} g)^T \widehat{J}_n (\widehat{\nabla} g)}, \tag{10.19}$$

$\widehat{J}_n = J_n(\widehat{\theta}_n)$ *and* $\widehat{\nabla} g$ *is* ∇g *evaluated at* $\theta = \widehat{\theta}$.

Example 10.29 *Let* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. *Let* $\tau = g(\mu, \sigma) = \sigma/\mu$. *In homework question 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,

$$\widehat{\mathsf{se}}\,(\widehat{\tau}) = \sqrt{(\widehat{\nabla} g)^T \widehat{J}_n (\widehat{\nabla} g)} = \frac{1}{\sqrt{n}} \left\{ \frac{1}{\widehat{\mu}^4} + \frac{\widehat{\sigma}^2}{2\widehat{\mu}^2} \right\}^{1/2}. \quad \blacksquare$$

10.11 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.

Example 10.30 *Consider example 10.29. To get to 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.

10.12 Technical Appendix

10.12.1 Proofs

Proof of Theorem 10.13. Since $\widehat{\theta}_n$ maximizes $M_n(\theta)$, we have $M_n(\widehat{\theta}_n) \geq M_n(\theta_\star)$. Hence,

$$\begin{aligned} M(\theta_\star) - M(\widehat{\theta}_n) &= M_n(\theta_\star) - M(\widehat{\theta}_n) + M(\theta_\star) - M_n(\theta_\star) \\ &\leq M_n(\widehat{\theta}) - M(\widehat{\theta}_n) + M(\theta_\star) - M_n(\theta_\star) \\ &\leq \sup_\theta |M_n(\theta) - M(\theta)| + M(\theta_\star) - M_n(\theta_\star) \\ &\xrightarrow{\text{P}} 0 \end{aligned}$$

where the last line follows from (10.7). 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 (10.8), 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 10.18. First we need a Lemma.

Lemma 10.31 *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 10.18. 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,

$$\text{TOP} = n^{-1/2}\sum_i Y_i = \sqrt{n}\overline{Y} = \sqrt{n}(\overline{Y} - 0) \rightsquigarrow W \sim N(0, I)$$

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 6.5 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

$$\begin{aligned} \frac{\widehat{\theta}_n - \theta}{\widehat{se}} &= \sqrt{n} I^{1/2}(\widehat{\theta}_n)(\widehat{\theta}_n - \theta) \\ &= \left\{\sqrt{n} I^{1/2}(\theta)(\widehat{\theta}_n - \theta)\right\} \left\{\frac{I(\widehat{\theta}_n)}{I(\theta)}\right\}^{1/2}. \end{aligned}$$

The first terms tends in distribution to N(0,1). The second term tends in probability to 1. The result follows from Theorem 6.5 part (e). ■

Outline of Proof of Theorem 10.24. Write,

$$\widehat{\tau} = g(\widehat{\theta}) \approx g(\theta) + (\widehat{\theta} - \theta) g'(\theta) = \tau + (\widehat{\theta} - \theta) g'(\theta).$$

Thus,

$$\sqrt{n}(\widehat{\tau} - \tau) \approx \sqrt{n}(\widehat{\theta} - \theta) g'(\theta)$$

and hence

$$\frac{\sqrt{nI(\theta)}(\widehat{\tau} - \tau)}{g'(\theta)} \approx \sqrt{nI(\theta)}(\widehat{\theta} - \theta).$$

Theorem 10.18 tells us that the right hand side tends in distribution to a N(0,1). Hence,

$$\frac{\sqrt{nI(\theta)}(\widehat{\tau} - \tau)}{g'(\theta)} \rightsquigarrow N(0,1)$$

or, in other words,

$$\widehat{\tau} \approx N\left(\tau, \mathsf{se}^{\,2}(\widehat{\tau}_n)\right)$$

where

$$\mathsf{se}^{\,2}(\widehat{\tau}_n) = \frac{(g'(\theta))^2}{nI(\theta)}.$$

The result remains true if we substitute $\widehat{\theta}$ for θ by Theorem 6.5 part (e). ■

10.12.2 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 defintions.

Definition 10.32 *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 $T(X^n)$.

Example 10.33 *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Then $\mathcal{L}(p) = p^S(1-p)^{n-S}$ where $S = \sum_i X_i$ so S is sufficient.* ■

Example 10.34 *Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ and let $T = (\overline{X}, S)$. Then,*

$$\begin{aligned} f(X^n;\mu,\sigma) &= \prod_i f(X_i;\mu,\sigma) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (X_i-\mu)^2\right\} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

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 $g(a_1, a_2, a_3) = (a_1, a_2)$. ■

Definition 10.35 *A statistic T is* **minimally sufficient** *if (i) it is sufficient and (ii) it is a function of every other sufficient statistic.*

Theorem **10.36** *T is minimal sufficient if $T(x^n) = T(y^n)$ if and only if $x^n \leftrightarrow y^n$.*

A statistic induces a partition on the set if outcomes. We can think of sufficiency in terms of these partitions.

Example 10.37 *Let $X_1, X_2, X_3 \sim$ Bernoulli(θ). 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 $x^n = (1,0,1)$ and $y^n = (0,1,1)$, then $x^n \leftrightarrow y^n$ yet $U(x^n) \neq U(y^n)$. The statistic $W = 17\,T$ generates the same partition as T. It is also minimal sufficient. ■

Example 10.38 *For a $N(\mu, \sigma^2)$ model, $T = (\overline{X}, S)$ is a minimally sufficient statistic. For the Bernoulli model, $T = \sum_i X_i$ is a minimally sufficient statistic. For the Poisson model, $T = \sum_i X_i$ is a minimally 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.

Example 10.39 *Two coin flips. Let $X = (X_1, X_2) \sim$ 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* $T = 1$*:*

$$P(X_1 = 0, X_2 = 0|t = 1) = 0, P(X_1 = 0, X_2 = 1|t = 1) = \frac{1}{2},$$

$$P(X_1 = 1, X_2 = 0|t = 1) = \frac{1}{2}, P(X_1 = 1, X_2 = 1|t = 1) = 0$$

(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.* ■

Theorem **10.40 (Factorization Theorem)** *T is sufficient if and only if there are functions $g(t, \theta)$ and $h(x)$ such that $f(x^n; \theta) = g(t(x^n), \theta)h(x^n)$.*

Example 10.41 *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.

Theorem **10.42 (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, $R(\theta, \widetilde{\theta}) \leq R(\theta, \widehat{\theta})$.

Example 10.43 *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 . ■

10.12.3 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**.

Example 10.44 *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 $\eta(\theta) = \log\theta$, $B(\theta) = \theta$, $T(x) = x$, $h(x) = 1/x!$. ■

Example 10.45 *Let $X \sim Bin(n, \theta)$. 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 a 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.

Example 10.46 *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. ■

Theorem **10.47** *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 and n iid samples also has exponential form with sufficient statistic $(\sum_i T_1(X_i), \ldots, \sum_i T_k(X_i))$.

Example 10.48 *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 $A(\eta) = \int h(x) e^{T^T(x)\eta} dx$. 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.

10.13 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 from (1bc). 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 estimate $\widehat{\psi}$ of ψ.

(b) Find an approximate 95 per cent 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 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 10.29.

9. Let $X_1, ..., X_n$ 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 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$(\widehat{\theta}, \widehat{\mathsf{se}}^2)$. Compare these to the true sampling distribution of $\widehat{\theta}$ (whihc you can get by simulation). Which approximation, parametric bootstrap, bootstrap, or delta method is closer to the true distribution?

10. Let $X_1, ..., X_n$ Unif$(0, \theta)$. The MLE is $\widehat{\theta} = X_{(n)} = \max\{X_1, ..., X_n\}$. Generate a data set 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?

11
Hypothesis Testing and p-values

Suppose we want to know if a certain chemical causes cancer. We take some rats and randomly divide them into two groups. We expose one group to the chemical and then we compare the cancer rate in the two groups. Consider the following two hypotheses:

> **The Null Hypothesis**: The cancer rate is the same in the two groups
> **The Alternative Hypothesis**: The cancer rate is not the same in the two groups.

If the exposed group has a much higher rate of cancer than the unexposed group then we will reject the null hypothesis and conclude that the evidence favors the alternative hypothesis; in other words we will conclude that there is evidence that the chemical causes cancer. 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

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1. \tag{11.1}$$

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 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

$$R = \Big\{ x : \ T(x) > c \Big\}$$

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 cutoff value c.

Warning! There is a tendency to use hypthesis 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 asume 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 when can make. Rejecting H_0 when H_0 is true is called a **type I error**. Rejecting H_1 when H_1 is true is called a **type II error**. The possible outcomes for hypothesis testing are summarized in the Table below:

	Retain Null	Reject Null
H_0 true	$\surd$	type I error
H_1 true	type II error	$\surd$

Summary of outcomes of hypothesis testing.

Definition 11.1 *The* **power function** *of a test with rejection region R is defined by*

$$\beta(\theta) = P_\theta(X \in R). \tag{11.2}$$

The **size** *of a test is defined to be*

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta). \tag{11.3}$$

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.

Example 11.2 *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 $R = \left\{x^n : \ T(x^n) > c\right\}$. Let Z denote a standard normal random variable. The power function is

$$\begin{aligned}
\beta(\mu) &= P_\mu\left(\overline{X} > c\right) \\
&= P_\mu\left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} > \frac{\sqrt{n}(c - \mu)}{\sigma}\right) \\
&= P\left(Z > \frac{\sqrt{n}(c - \mu)}{\sigma}\right) \\
&= 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right).
\end{aligned}$$

This function is increasing in μ. Hence

$$\text{size} = \sup_{\mu < 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right).$$

To get a size α test, set this equal to α and solve for c to get

$$c = \frac{\sigma \Phi^{-1}(1 - \alpha)}{\sqrt{n}}.$$

So we reject when $\overline{X} > \sigma\Phi^{-1}(1-\alpha)/\sqrt{n}$. Equivalently, we reject when

$$\frac{\sqrt{n}\left(\overline{X} - 0\right)}{\sigma} > z_\alpha. \quad \blacksquare$$

Finding most powerful tests is hard and, in many cases, most powerful tests don't even exist. Instead of searching for most powerful tests, we'll just consider three widely used tests: the Wald test, the χ^2 test and the permutation test. A fourth test, the likelihood ratio test, is dicussed in the appendix.

11.1 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 $\widehat{\theta}$.

The test is named after Abraham Wald (1902-1950), who was a very influential mathematical statistician.

Definition 11.3 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:

$$\frac{\sqrt{n}(\widehat{\theta} - \theta_0)}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1).$$

The size α **Wald test** *is: reject H_0 when $|W| > z_{\alpha/2}$ where*

$$W = \frac{\widehat{\theta} - \theta_0}{\widehat{se}}.$$

Theorem **11.4** *Asymptotically, the Wald test has size α, that is,*

$$\mathbb{P}_{\theta_0}\left(|Z| > z_{\alpha/2}\right) \to \alpha$$

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

$$\begin{aligned} \mathbb{P}_{\theta_0}\left(|W| > z_{\alpha/2}\right) &= \mathbb{P}_{\theta_0}\left(\frac{|\widehat{\theta} - \theta_0|}{\widehat{\mathsf{se}}} > z_{\alpha/2}\right) \\ &\to \mathbb{P}\left(|N(0,1)| > z_{\alpha/2}\right) \\ &= \alpha. \quad \blacksquare \end{aligned}$$

Remark 11.5 *Most texts define the Wald test slightly differently. They use the standard error computed at $\theta = \theta_0$ rather then at the estimated value $\widehat{\theta}$. Both versions are valid.*

Let us consider the power of the Wald test when the null hypothesis is false.

Theorem **11.6** *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*

$$1 - \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} - z_{\alpha/2}\right). \tag{11.4}$$

Recall that $\widehat{\mathsf{se}}$ tends to 0 as the sample size increases. Inspecting (11.4) 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.

Example 11.7 (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 $Y_i = 0$ otherwise. A typical data set 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

$$\delta = \mathbb{E}(D_i) = \mathbb{E}(X_i) - \mathbb{E}(Y_i) = \mathbb{P}(X_i = 1) - \mathbb{P}(Y_i = 1).$$

Then $\widehat{\delta} = \overline{D} = n^{-1}\sum_{i=1}^n D_i$ 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.** ■

Example 11.8 (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

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}. \quad \blacksquare$$

Example 11.9 (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 $W = \widehat{\delta}/\widehat{\mathsf{se}}$.* $\blacksquare$

There is a relationship between the Wald test and the $1-\alpha$ asymptotic confidence interval $\widehat{\theta} \pm \widehat{\mathsf{se}}\, z_{\alpha/2}$.

Theorem **11.10** *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.

11.2 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.

Definition 11.11 *Suppose that for every $\alpha \in (0,1)$ we have a size α test with rejection region R_α. Then,*

$$\text{p-value} = \inf\Big\{\alpha : \ T(X^n) \in R_\alpha\Big\}.$$

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:

p-value	evidence
$< .01$	very strong evidence againt H_0
.01 - .05	strong evidence againt H_0
.05 - .10	weak evidence againt H_0
$> .1$	little or no evidence againt 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.

But do not confuse the p-value with $\mathbb{P}(H_0|\text{Data})$. **The p-value is not the probability that the null hypothesis is true.**

We discuss quantities like $\mathbb{P}(H_0|\text{Data})$ in the chapter on Bayesian inference.

The following result explains how to compute the p-value.

Theorem 11.12 *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,

$$\text{p-value} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X^n) \geq T(x^n)).$$

In words, the p-value is the probability (under H_0) of observing a value of the test statistic as or more extreme than what was actually observed.

For a Wald test, W has an aproximate $N(0,1)$ distribution under H_0. Hence, the p-value is

$$\text{p-value} \approx \mathbb{P}\Big(|Z| > |w|\Big) = 2\mathbb{P}\Big(Z < -|w|\Big) = 2\Phi\Big(Z < -|w|\Big) \tag{11.5}$$

where $Z \sim N(0,1)$ and $w = (\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}}$ is the observed value of the test statistic.

Here is an important property of p-values.

Theorem **11.13** *If the test statistic has a continuous distribution, then under $H_0 : \theta = \theta_0$, the p-value has a Uniform (0,1) distribution.*

If the p-value is less than .05 then people often report that "the result is statistically significant at the 5 per cent level." This just means that the null would be rejected if we used a size $\alpha = 0.05$ test. In general, the size α test rejects if and only if p-value $< \alpha$.

Example 11.14 *Recall the cholesterol data from Example 8.11. To test of the means ard different we compute*

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{216.2 - 195.3}{5^2 + 2.4^2} = 3.78.$$

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 boostrap. 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. ■

Warning! 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 practically significant. It is wise to report a confidence interval as well.

11.3 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 $\mathbb{V}(k) = 2k$. We define the upper α quantile $\chi^2_{k,\alpha} = F^{-1}(1-\alpha)$ where F is the CDF . That is, $\mathbb{P}(\chi^2_k > \chi^2_{k,\alpha}) = \alpha$.

11.4 Pearson's χ^2 Test For Multinomial Data

Pearson's χ^2 test is used for multinomial data. Recall that $X = (X_1, \ldots, X_k)$ has a multinomial distribution if

$$f(x_1, \ldots, x_k;\ p) = \binom{n}{x_1 \ldots x_k} p_1^{x_1} \cdots p_k^{x_k}$$

where

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

The MLE of p is $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_k) = (X_1/n, \ldots, X_k/n)$.

Let $(p_{01}, \ldots, p_{0k})$ be some fixed set of probabilities and suppose we want to test

$$H_0 : (p_1, \ldots, p_k) = (p_{01}, \ldots, p_{0k}) \text{ versus } H_1 : (p_1, \ldots, p_k) \neq (p_{01}, \ldots, p_{0k}).$$

Pearson's χ^2 statistic is

$$T = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}} = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

where $O_j = X_j$ is the observed data and $E_j = \mathbb{E}(X_j) = np_{0j}$ is the expected value of X_j under H_0.

Theorem **11.15** *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 $\mathbb{P}(\chi^2_k > t)$ where t is the observed value of the test statistic.*

Example 11.16 (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 (4) wrinkled green. The number of each type is multinomial with probability (p_1, p_2, p_3, p_4). His theory of inheritance predicts that*

$$p = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right) \equiv p_0.$$

In $n = 556$ trials he observed $X = (315, 101, 108, 32)$. Since, $np_{10} = 312.75, np_{20} = np_{30} = 104.25$ and $np_{40} = 34.75$, the test statistic is

$$\chi^2 = \frac{(315 - 312.75)^2}{312.75} + \frac{(101 - 104.25)^2}{104.25} + \frac{(108 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} = 0.47.$$

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

$$\text{p-value} = \mathbb{P}(\chi^2_3 > .47) = .07$$

which is only moderate evidence againt H_0. Hence, the data do not contradict Mendel's theory. Interestingly, there is some controversey about whether Mendel's results are "too good." ■

11.5 The Permutation Test

The permutation test is a nonparametric method for testing whether two distribution 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 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 $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. 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. The distribution 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

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.

$$\text{p-value} = \mathbb{P}_0(T > t_{obs}) = \frac{1}{N!} \sum_{j=1}^{N!} I(T_j > t_{obs}).$$

Example 11.17 *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.

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

$$\frac{1}{B}\sum_{j=1}^{B} I(T_j > t_{\text{obs}}).$$

Example 11.18 *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 how much protein that gene produces. Roughly, the larger the number, the more active the gene. The table below, reproduced from Efron, et. al. (JASA, 2001, p. 1160) shows the expression levels for genes from 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	*230.0*	*-1,350*	*-1,580.0*	*-400*	*-760*	*970*	*110*	*-50*	*-190*	*-200*
Gene 2	*470.0*	*-850*	*-.8*	*-280*	*120*	*390*	*-1730*	*-1360*	*-.8*	*-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.

11.6 Multiple Testing

In some situations we may conduct many hypothesis tests. In example 11.18, 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 m p-values for these tests.

The Bonferroni Method

Given p-values $P_1, \ldots, P_m$, reject null hypothesis H_{0i} if $P_i < \alpha/m$.

Theorem **11.19** *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 hypotheses 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,

$$\mathbb{P}(R) = \mathbb{P}\left(\bigcup_{i=1}^m R_i\right) \leq \sum_{i=1}^m P(R_i) = \sum_{i=1}^m \frac{\alpha}{m} = \alpha$$

from Theorem 11.13. ■

Example 11.20 *In the gene example, using $\alpha = .05$, we have that $.05/2638 = .00001895375$. 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 are true and $m_1 = m - m_0$ null hypotheses are false. The tests can be categorize in a 2×2 as in the following table.

	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

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 $\text{FDR} = \mathbb{E}(\text{FDP})$.

The Benjamini-Hochberg (BH) Method

1. Let $P_{(1)} < \cdots < P_{(m)}$ denote the ordered p-values.
2. Define

$$\ell_i = \frac{i\alpha}{C_m m}, \quad \text{and} \quad R = \max\left\{i : P_{(i)} < \ell_i\right\} \tag{11.6}$$

   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. Rejects all null hypotheses H_{0i} for which $P_i \leq t$.

Theorem **11.21 (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 hypthesis is false,*

$$\text{FDR} = \mathbb{E}(\text{FDP}) \leq \frac{m_0}{m}\alpha \leq \alpha.$$

Example 11.22 *Figure 11.1 shows 7 ordered p-values plotted as vertical lines. If we tested at level α without doing any corection for mutiple testing, we would reject all hypotheses whose p-values are less than α. In this case, the 5 hypotheses corresponding to the 5 smallest p-values are rejected. The Bonferroni method rejects all hypotheses whose p-values are less than α/m. In this example, 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 three hypotheses being rejected in this case.* ■

SUMMARY. The Bonferonni method controls the probability of a single false rejection. This is very strict and leads to low power when there are many tests. The FDR method controls the fraction of false discoveries which is a more reasonable criterion when there are many tests.

11.7 Technical Appendix

11.7.1 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.

Theorem **11.23 (Neyman-Pearson.)** *Suppose we test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Let*

$$T = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{\prod_{i=1}^n f(x_i;\theta_1)}{\prod_{i=1}^n f(x_i;\theta_0)}.$$

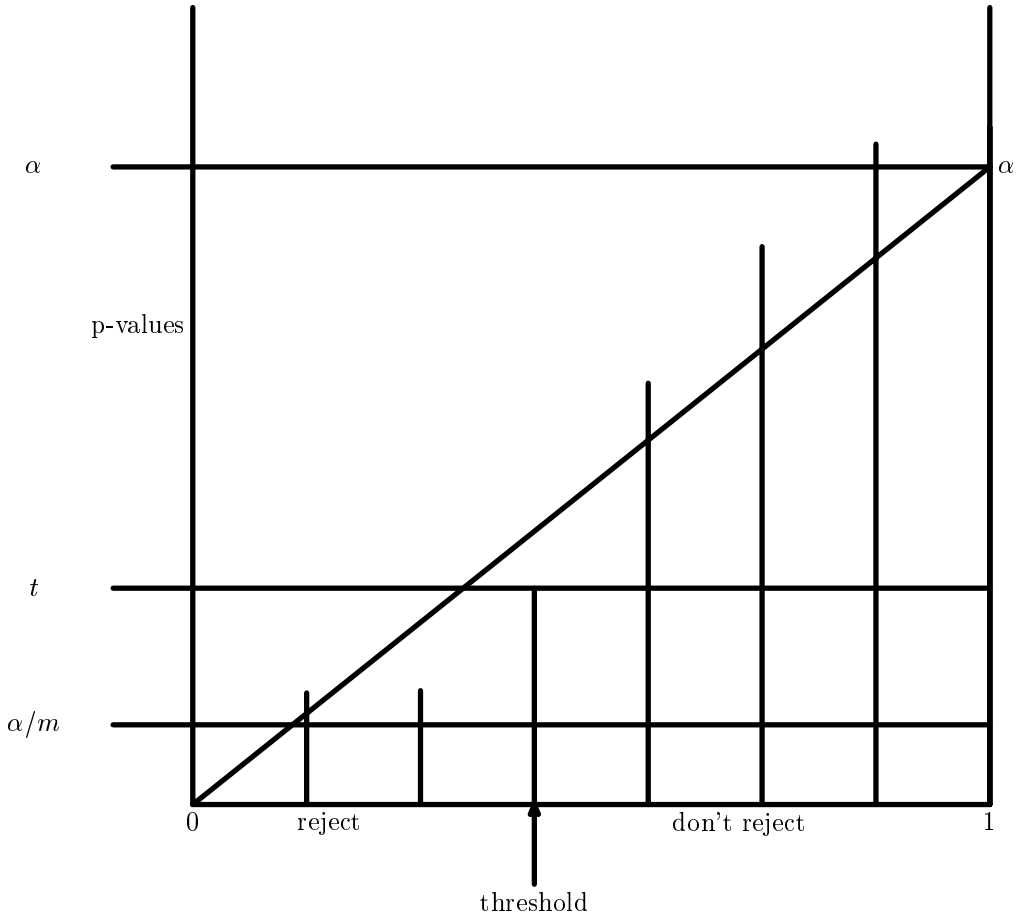


FIGURE 11.1. Schematic illustration of Benjamini-Hochberg procedure. All hypotheses corresponding to the last undercrossing are rejected.

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 $\beta(\theta_1)$.

11.7.2 Power of the Wald Test

PROOF OF THEOREM 11.6.
Let $Z \sim N(0,1)$. Then,

$$\begin{aligned}
\text{Power} &= \beta(\theta_\star) \\
&= \mathbb{P}_{\theta_\star}(\text{Reject } H_0) \\
&= \mathbb{P}_{\theta_\star}(|W| > z_{\alpha/2}) \\
&= \mathbb{P}_{\theta_\star}\left(\frac{|\widehat{\theta} - \theta_0|}{\widehat{\mathsf{se}}} > z_{\alpha/2}\right) \\
&= \mathbb{P}_{\theta_\star}\left(\frac{\widehat{\theta} - \theta_0}{\widehat{\mathsf{se}}} > z_{\alpha/2}\right) + \mathbb{P}_{\theta_\star}\left(\frac{\widehat{\theta} - \theta_0}{\widehat{\mathsf{se}}} < -z_{\alpha/2}\right) \\
&= \mathbb{P}_{\theta_\star}\left(\widehat{\theta} > \theta_0 + \widehat{\mathsf{se}}\; z_{\alpha/2}\right) + \mathbb{P}_{\theta_\star}\left(\widehat{\theta} < \theta_0 - \widehat{\mathsf{se}}\; z_{\alpha/2}\right) \\
&= \mathbb{P}_{\theta_\star}\left(\frac{\widehat{\theta} - \theta_\star}{\widehat{\mathsf{se}}} > \frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} + z_{\alpha/2}\right) + \mathbb{P}_{\theta_\star}\left(\frac{\widehat{\theta} - \theta_\star}{\widehat{\mathsf{se}}} < \frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} - z_{\alpha/2}\right) \\
&\approx \mathbb{P}\left(Z > \frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} + z_{\alpha/2}\right) + \mathbb{P}\left(Z < \frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} - z_{\alpha/2}\right) \\
&= 1 - \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} - z_{\alpha/2}\right). \quad \blacksquare
\end{aligned}$$

11.7.3 The t-test

To test $H_0 : \mu = \mu_0$ where μ 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 sizeα test.

11.7.4 The Likelihood Ratio Test

Let $\theta = (\theta_1, \ldots, \theta_q, \theta_{q+1}, \ldots, \theta_r)$ and suppose that Θ_0 consists of all parameter values θ such that $(\theta_{q+1}, \ldots, \theta_r) = (\theta_{0,q+1}, \ldots, \theta_{0,r})$.

Definition 11.24 *Define the* **likelihood ratio statistic** *by*

$$\lambda = 2\log\left(\frac{\sup_{\theta\in\Theta}\mathcal{L}(\theta)}{\sup_{\theta\in\Theta_0}\mathcal{L}(\theta)}\right) = 2\log\left(\frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}(\widehat{\theta}_0)}\right)$$

where $\widehat{\theta}$ *is the* MLE *and* $\widehat{\theta}_0$ *is the* MLE *when* θ *is restricted to lie in* Θ_0. *The* **likelihood ratio test** *is: reject* H_0 *when* $\lambda(x^n) > \chi^2_{r-q,\alpha}$.

For example, if $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ and we want to test the null hypothesis that $\theta_3 = \theta_4 = 0$ then the limiting distribution has $4 - 2 = 2$ degrees of freedom.

Theorem **11.25** *Under H_0,*

$$2 \log \lambda(x^n) \xrightarrow{d} \chi^2_{r-q}.$$

Hence, asymptotically, the LR test is level α.

11.8 Bibliographic Remarks

The most complete book on testing is (1986). See also Casella and Berger (1990, Chapter 8). The FDR method is due to Benjamini and Hochberg (1995).

11.9 Exercises

1. Prove Theorem 11.13.

2. Prove Theorem 11.10.

3. 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?

4. 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.

Reference:
Phillips, D.P. and King, E.W. (1988).
Death takes a holiday: Mortality surrounding major social occasions.
The Lancet, 2, 728-732.

5. In 1861, 10 essays appeared in the New Orleans Daily Crescent. They were signed "Quintus Curtuis 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

   (source: Rice xxxx)

   (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?

6. 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, i.e. find $\beta(1)$.

(c) Show that $\beta(1) \to 1$ as $n \to \infty$.

7. 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 $Z = (\widehat{\theta} - \theta_0)/\widehat{se}$. Let $\theta_1 > \theta_0$ be some alternative. Show that $\beta(\theta_1) \to 1$.

8. 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.

9. A randomized, double-blind experiment was conducted to assess the effectiveness of several drugs for reducing post-operative 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

(source:)

(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 Bonferoni and the FDR method to adjust for multiple testing.

10. 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?

12
Bayesian Inference

12.1 The Bayesian Philosophy

The statistical theory and methods that we have discussed so far are known as **frequentist (or classical)** inference. 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, (usually unknown) constants. Because they are not fluctuating, no 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 per cent confidence interval should trap the true value of the parameter with limiting frequency at least 95per cent.

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 communuties 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 each approach.

12.2 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 degrees of 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 form 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,

$$P(\Theta = \theta|X = x) = \frac{P(X = x, \Theta = \theta)}{P(X = x)} = \frac{P(X = x|\Theta = \theta)P(\Theta = \theta)}{\sum_\theta P(X = x|\Theta = \theta)P(\Theta = \theta)}$$

which you may recognize from earlier in the course as **Bayes' theorem.** The version for continuous variables is obtained by using density functions:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}. \tag{12.1}$$

If we have n IID observations $X_1, \ldots, X_n$, we replace $f(x|\theta)$ with $f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$. Let us write X^n to mean $(X_1, \ldots, X_n)$ and x^n to mean $(x_1, \ldots, x_n)$. Then

$$f(\theta|x^n) = \frac{f(x^n|\theta)f(\theta)}{\int f(x^n|\theta)f(\theta)d\theta} = \frac{\mathcal{L}_n(\theta)f(\theta)}{\int \mathcal{L}_n(\theta)f(\theta)d\theta} \propto \mathcal{L}_n(\theta)f(\theta). \tag{12.2}$$

In the right hand side of the last equation, we threw away the denominator $\int \mathcal{L}_n(\theta)f(\theta)d\theta$ which is a constant that does not depend on θ; we call this quantity the **normalizing constant.** We can summarize all this by writing:

"posterior is proportional to likelihood times prior." (12.3)

You might wonder, doesn't it cause a problem to throw away the constant $\int \mathcal{L}_n(\theta)f(\theta)d\theta$? The answer is that we can always

recover the constant is since we know that $\int f(\theta|x^n)d\theta = 1$. Hence, we often omit the constant until we really need it.

What do we do with the posterior? 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

$$\overline{\theta}_n = \int \theta f(\theta|x^n)d\theta = \frac{\int \theta \mathcal{L}_n(\theta)f(\theta)}{\int \mathcal{L}_n(\theta)f(\theta)d\theta}. \tag{12.4}$$

We can also obtain a Bayesian interval estimate. Define a and b by $\int_{-\infty}^{a} f(\theta|x^n)d\theta = \int_b^{\infty} f(\theta|x^n)d\theta = \alpha/2$. 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.**

Example 12.1 *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 $s = \sum_i x_i$ *is the number of heads. Recall that 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 Beta* (α, β) *is* $\alpha/(\alpha+\beta)$ *so the Bayes estimator is*

$$\overline{p} = \frac{s+1}{n+2}.$$

It is instructive to rewrite the estimator as

$$\overline{p} = \lambda_n \widehat{p} + (1-\lambda_n)\widetilde{p}$$

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 per cent 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**.

Example 12.2 *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 homework, it is shown that the posterior for* θ *is*

$$\theta|X^n \sim N(a, b^2) \tag{12.5}$$

where

$$\overline{\theta} = w\overline{X} + (1-w)a$$

where

$$w = \frac{\frac{1}{se^2}}{\frac{1}{se^2} + \frac{1}{b^2}} \quad \text{and} \quad \frac{1}{\tau^2} = \frac{1}{se^2} + \frac{1}{b^2}$$

and $se = \sigma/\sqrt{n}$ *is the standard error of the mle* $\overline{X}$*. This is another example of a conjugate prior. Note that* $w \to 1$ *and* $\tau/se \to 1$ *as* $n \to \infty$*. So, for large* n*, the posterior is approximately* $N(\widehat{\theta}, se^2)$*. 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 is find $C = (c, d)$ *such that* $Pr(\theta \in C|X^n) = .95$*. We can do this by choosing* c *such that* $Pr(\theta < c|X^n) = .025$ *and* $Pr(\theta > d|X^n) = .025$*. So, we want to find* c *such that*

$$\begin{aligned} P(\theta < c|X^n) &= P\left(\frac{\theta - \overline{\theta}}{\tau} < \frac{c - \overline{\theta}}{\tau} \;\middle|\; X^n\right) \\ &= P\left(Z < \frac{c - \overline{\theta}}{\tau}\right) = .025. \end{aligned}$$

Now, we know that $P(Z < -1.96) = .025$*. 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 per cent Bayesian interval is* $\overline{\theta} \pm 1.96\,\tau$*. Since* $\overline{\theta} \approx \widehat{\theta}$ *and* $\tau \approx se$*, the 95 per cent Bayesian interval is approximated by* $\widehat{\theta} \pm 1.96\,se$ *which is the frequentist confidence interval.* ■

12.3 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

$$H(\tau|x^n) = \mathbb{P}(g(\theta) \leq \tau) = \int_A f(\theta|x^n)d\theta$$

where $A = \{\theta : \ g(\theta) \leq \tau\}$. The posterior density is $h(\tau|x^n) = H'(\tau|x^n)$.

Example 12.3 *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 $\psi \in \mathbb{R}$. ■

12.4 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 later in the book.

Example 12.4 *Consider again Example 12.3. 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 $h(\psi|x^n)$.* ■

12.5 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.

Theorem **12.5** *Under appropriate regularity conditions, we have that the posterior is approximately $N(\widehat{\theta}, \widehat{\mathsf{se}}^{\,2})$ where $\widehat{\theta}_n$ is the* MLE *and $\widehat{\mathsf{se}} = 1/\sqrt{nI(\widehat{\theta}_n)}$. Hence, $\overline{\theta}_n \approx \widehat{\theta}_n$. Also, if $C = (\widehat{\theta}_n - z_{\alpha/2}\widehat{\mathsf{se}}\,, \widehat{\theta}_n + z_{\alpha/2}\widehat{\mathsf{se}}\,)$ is the asymptotic frequentist $1-\alpha$ confidence interval, then C_n is also an approximate $1-\alpha$ Bayesian posterior interval:*

$$\mathbb{P}(\theta \in C|X^n) \to 1-\alpha.$$

There is also a Bayesian delta method. Let $\tau = g(\theta)$. Then

$$\tau|X^n \approx N(\widehat{\tau}, \widetilde{se}^2)$$

where $\widehat{\tau} = g(\widehat{\theta})$ and $\widetilde{se} = se\,|g'(\widehat{\theta})|$.

12.6 Flat Priors, Improper Priors and "Noninformative" Priors.

A big question in Bayesian inference is: where do you 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 seems impractical in complicated problems especially if there are many parameters. 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 $p|X^n \sim Beta(s+1, n-s+1)$ as we saw earlier which seemed very reasonable. But unfettered use of flat priors raises some questions.

Improper Priors. Consider the $N(\theta, 1)$ example. 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 real probability density in the usual sense. We call such a prior an **improper prior.** Nonetheless, we can still carry out Bayes' theorem and compute the posterior density $f(\theta) \propto \mathcal{L}_n(\theta)f(\theta) \propto \mathcal{L}_n(\theta)$. In the normal example, 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. Go back to the Bernoulli example and consider using the flat prior $f(p) = 1$. Recall that a flat prior presumably represents our lack of information about p before the experiment. Now let $\psi = \log(p/(1-p))$. This is a transformation and we can compute the resulting distribution

for ψ. It turns out that

$$f_\Psi(\psi) = \frac{e^\psi}{(1+e^\psi)^2}.$$

But one could argue that if we are ignorant about p then we are also ignorant about ψ so shouldn't we use a flat prior for ψ? This contradicts the prior $f_\Psi(\psi)$ for ψ that is implied by using a flat prior for p. 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.

Example 12.6 *Consider the Bernoulli (p). Recall that*

$$I(p) = \frac{1}{p(1-p)}.$$

Jeffrey's 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 $f(\theta) \propto \sqrt{\det I(\theta)}$ where $\det(A)$ denotes the determinant of a matrix A.

12.7 Multiparameter Problems

In principle, multiparameter problems are handled the same way. Suppose that $\theta = (\theta_1, \ldots, \theta_p)$. The posterior density is still given by

$$p(\theta|x^n) \propto \mathcal{L}_n(\theta) f(\theta).$$

The question now arises of how to extract inferences about one parameter. The key is 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

$$f(\theta_1|x^n) = \int \cdots \int f(\theta_1, \cdots, \theta_p|x^n) d\theta_2 \ldots d\theta_p.$$

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.

Example 12.7 (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,*

$$X_1 \sim \text{Binomial}(n_1, p_1) \quad \textit{and} \quad X_2 \sim \text{Binomial}(n_2, p_2).$$

Suppose we take $f(p_1, p_2) = 1$. 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

$$f(p_1|x_1) \propto p_1^{x_1}(1-p_1)^{n_1-x_1} \quad \textit{and} \quad f(p_2|x_2) \propto p_2^{x_2}(1-p_2)^{n_2-x_2}$$

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* $f(\tau|x_1, x_2)$. ■

12.8 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 per cent 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 really an 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. Consider the following example.

Example 12.8 *Let* $X \sim N(\theta, 1)$ *and suppose we use the prior* $\theta \sim N(0, \tau^2)$. *From (12.5), the posterior is*

$$\theta|x \sim N\left(\frac{x}{1+\frac{1}{\tau^2}}, \frac{1}{1+\frac{1}{\tau^2}}\right) = N(cx, c)$$

where $c = \tau^2/(\tau^2+1)$. *A* $1-\alpha$ *per cent posterior interval is* $C = (a, b)$ *where*

$$a = cx - \sqrt{c}z_{\alpha/2} \quad \text{and} \quad b = cx + \sqrt{c}z_{\alpha/2}.$$

Thus, $\mathbb{P}(\theta \in C|X) = 1-\alpha$. *We can now ask, from a frequntist perspective, what is the coverage of* C, *that is, how often will this interval contain the true value? The answer is*

$$\begin{aligned}
\mathbb{P}_\theta(a < \theta < b) &= \mathbb{P}_\theta(cX - z_{\alpha/2}\sqrt{c} < \theta < cX + z_{\alpha/2}\sqrt{c}) \\
&= \mathbb{P}_\theta\left(\frac{\theta - z_{\alpha/2}\sqrt{c} - c\theta}{c} < X - \theta < \frac{\theta + z_{\alpha/2}\sqrt{c} - c\theta}{c}\right) \\
&= \mathbb{P}_\theta\left(\frac{\theta(1-c) - z_{\alpha/2}\sqrt{c}}{c} < Z < \frac{\theta(1-c) + z_{\alpha/2}\sqrt{c}}{c}\right) \\
&= \Phi\left(\frac{\theta(1-c) + z_{\alpha/2}\sqrt{c}}{c}\right) - \Phi\left(\frac{\theta(1-c) - z_{\alpha/2}\sqrt{c}}{c}\right)
\end{aligned}$$

where $Z \sim N(0,1)$. *Figure 12.1 shows the coverage as a function of* θ *for* $\tau = 1$ *and* $\alpha = .05$. *Unless the true value of* θ *is close to 0, the coverage can be very small. Thus, upon repeated use, the Bayesian 95 per cent interval might contain the true value with frequency near 0! In contrast, a confidence interval has coverage 95 per cent coverage no matter what the true value of* θ *is.* ■

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. It is worth remarking that it is possible to develop nonparametric Bayesian methods similar to plug-in estimation and the bootstrap. be forewarned, however, that the frequency properties of nonparametric Bayesian methods can sometimes be quite poor.

12.9 Appendix

Proof of Theorem 12.5.

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,

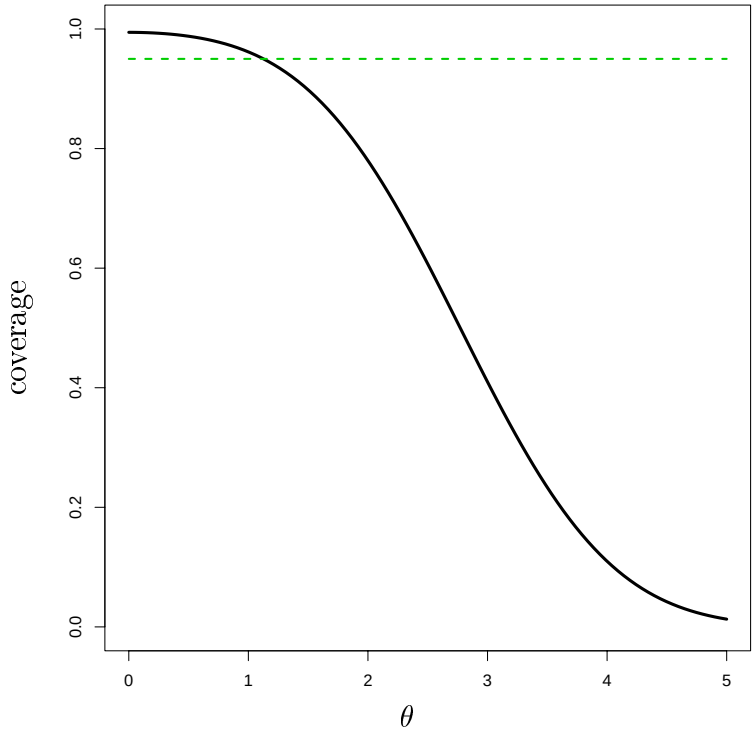


FIGURE 12.1. Frequentist coverage of 95 per cent Bayesian posterior interval as a function of the true value θ. The dotted line marks the 95 per cent level.

$\log f(\theta|X^n) \approx \ell(\theta)$. 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

$$\begin{aligned}
\sigma_n^{-2} &= -\ell''(\widehat{\theta}_n) = \sum_i -\ell_i''(\widehat{\theta}_n) \\
&= n\left(\frac{1}{n}\right)\sum_i -\ell_i''(\widehat{\theta}_n) \approx n\mathbb{E}_\theta\left[-\ell_i''(\widehat{\theta}_n)\right] \\
&= nI(\widehat{\theta}_n)
\end{aligned}$$

and hence $\sigma_n \approx se(\widehat{\theta})$. ■

12.10 Bibliographic Remarks

Some references on Bayesian inference include Carlin and Louis (1996), Gelman, Carlin, Stern and Rubin (1995), Lee (1997), Robert (1994) and Schervish (1995). See Cox (1997), Diaconis and Freedman (1996), Freedman (2001), Barron, Schervish and Wasserman (1999), Ghosal, Ghosh and van der Vaart (2001), Shen and Wasserman (2001) and Zhao (2001) for discussions of some of the technicalities of nonparametric Bayesian inference.

12.11 Exercises

1. Verify (12.5).

2. Let $X_1, ..., X_n$ 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 1000 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 per cent posterior interval for θ.

(f) Find a 95 per cent confidence interval for θ.

3. Let $X_1, ..., X_n$ 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 per cent confidence interval using the delta method.

   (b) Find the standard error and 90 per cent 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 per cent 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 per cent confidence interval for ψ.

   (e) Use simulation to find the posterior mean and posterior 90 per cent 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.

13
Statistical Decision Theory

13.1 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, a 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 $\theta = \widehat{\theta}$ and 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.

Definition 13.1 *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):

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \text{MSE} \ = \mathbb{V}_\theta(\widehat{\theta}) + \mathsf{bias}_\theta^2(\widehat{\theta}).$$

In the rest of the chapter, **if we do not state what loss function we are using, assume the loss function is squared error.**

13.2 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.

Example 13.2 *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*

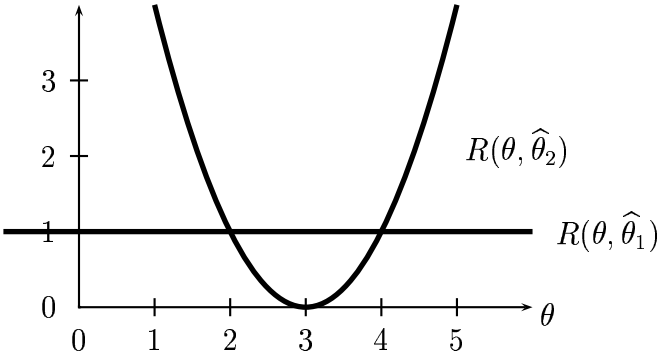


FIGURE 13.1. Comparing two risk functions. Neither dominates the other at all values of θ.

risk functions are $R(\theta, \widehat{\theta}_1) = \mathbb{E}_\theta(X - \theta)^2 = 1$ *and* $R(\theta, \widehat{\theta}_2) = \mathbb{E}_\theta(3 - \theta)^2 = (3 - \theta)^2$. *Notice that, 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 13.1. Obviously* $\widehat{\theta}_2$ *is a ridiculous estimator but it serves to illustrate the point that it is not obvious how to compare two risk functions.* ■

Example 13.3 *Let* $X_1, \ldots, X_n \sim$ 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,*

$$\begin{aligned} R(p, \widehat{p}_2) &= \mathbb{V}_p(\widehat{p}_2) + (\mathsf{bias}_p(\widehat{p}_2))^2 \\ &= \mathbb{V}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathbb{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2. \end{aligned}$$

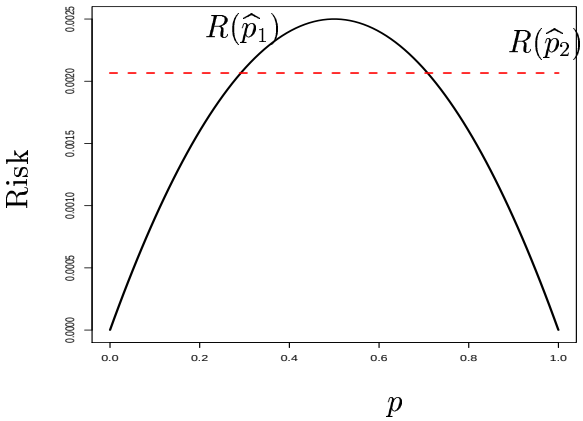


FIGURE 13.2. Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 13.3.

Now let $\alpha = \beta = \sqrt{n/4}$. *(In Example 13.13 we will explain this choice.) The resulting estimator is*

$$\widehat{p}_2 = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

and risk function is

$$R(p, \widehat{p}_2) = \frac{n}{4(n + \sqrt{n})^2}.$$

The risk functions are plotted in figure 13.2. 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.

Definition 13.4 *The* **maximum risk** *is*

$$\overline{R}(\widehat{\theta}) = \sup_{\theta} R(\theta, \widehat{\theta}) \tag{13.1}$$

and **the Bayes risk** *is*

$$r(\pi, \widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta \tag{13.2}$$

where $\pi(\theta)$ is a prior for θ.

Example 13.5 *Consider again the two estimators in Example 13.3. We have*

$$\overline{R}(\widehat{p}_1) = \max_{0 \le p \le 1} \frac{p(1-p)}{n} = \frac{1}{4n}$$

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 $\pi(p) = 1$. Then

$$r(\pi, \widehat{p}_1) = \int R(p, \widehat{p}_1)dp = \int \frac{p(1-p)}{n}dp = \frac{1}{6n}$$

and

$$r(\pi, \widehat{p}_2) = \int R(p, \widehat{p}_2)dp = \frac{n}{4(n+\sqrt{n})^2}.$$

For $n \ge 20$, $r(\pi, \widehat{p}_2) > r(\pi, \widehat{p}_1)$ 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 advtantage of using maximum risk, despite its problems, is that it does not require

one to choose a prior. In high-dimensional, complex problems, choosing a defensible prior can be extremely difficult. ■

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.

Definition 13.6 *A decision rule that minimizes the Bayes risk is called a* **Bayes rule**. *Formally, $\widehat{\theta}$ is a Byes rule for prior π if*

$$R(\theta, \widehat{\theta}) = \inf_{\widetilde{\theta}} r(\pi, \widetilde{\theta}) \tag{13.3}$$

where the infimum is over all estimators $\widetilde{\theta}$.

Definition 13.7 *An estimator that minimizes the maximum risk is called a* **minimax rule**. *Formally, $\widehat{\theta}$ is minimax if*

$$R(\theta, \widehat{\theta}) = \inf_{\widetilde{\theta}} \sup_{\theta} R(\theta, \widetilde{\theta}) \tag{13.4}$$

where the infimum is over all estimators $\widetilde{\theta}$.

13.3 Bayes Estimators

Let π be a prior. From Bayes' theorem, the posterior density is

$$f(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \tag{13.5}$$

where $m(x) = \int f(x, \theta)d\theta = \int f(x|\theta)\pi(\theta)d\theta$ is the **marginal distribution** of X. Define the **posterior risk** of an estimator

$\widehat{\theta}(x)$ by

$$r(\widehat{\theta}|x) = \int L(\theta, \widehat{\theta}(x)) f(\theta|x) d\theta. \tag{13.6}$$

Theorem 13.8 *The Bayes risk $r(\pi, \widehat{\theta})$ satisfies*

$$r(\pi, \widehat{\theta}) = \int r(\widehat{\theta}|x) m(x)\, dx.$$

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:

$$\begin{aligned} r(\pi, \widehat{\theta}) &= \int R(\theta, \widehat{\theta}) \pi(\theta) d\theta = \int \left(\int L(\theta, \widehat{\theta}(x)) f(x|\theta) dx \right) \pi(\theta) d\theta \\ &= \int \int L(\theta, \widehat{\theta}(x)) f(x, \theta) dx d\theta = \int \int L(\theta, \widehat{\theta}(x)) f(\theta|x) m(x) dx d\theta \\ &= \int \left(\int L(\theta, \widehat{\theta}(x)) f(\theta|x) d\theta \right) m(x)\, dx = \int r(\widehat{\theta}|x) m(x)\, dx. \end{aligned}$$

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.

Theorem 13.9 *If $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$ then the Bayes estimator is*

$$\widehat{\theta}(x) = \int \theta f(\theta|x) d\theta = \mathbb{E}(\theta|X = x). \tag{13.7}$$

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 13.7. ■

Example 13.10 *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*

$$\widehat{\theta}(X_1, \ldots, X_n) = \frac{b^2}{b^2 + \frac{\sigma^2}{n}}\overline{X} + \frac{\frac{\sigma^2}{n}}{b^2 + \frac{\sigma^2}{n}}a. \quad \blacksquare$$

13.4 Minimax Rules

The problem of 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.

Theorem **13.11** *Let $\widehat{\theta}^\pi$ be the Bayes rule for some prior π:*

$$r(\pi, \widehat{\theta}^\pi) = \inf_{\widehat{\theta}} r(\pi, \widehat{\theta}). \tag{13.8}$$

Suppose that

$$R(\theta, \widehat{\theta}^\pi) \leq r(\pi, \widehat{\theta}^\pi) \quad \text{for all } \theta. \tag{13.9}$$

Then $\widehat{\theta}^\pi$ is minimax and π is called a **least favorable prior**.

Proof. Suppose that $\widehat{\theta}^\pi$ is not minimax. Then there is another rule $\widehat{\theta}_0$ such that $\sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}^\pi)$. Since the average of a function is always less than or equal to its maximum, we have that $r(\pi, \widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0)$. Hence,

$$r(\pi, \widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}^\pi) \leq r(\pi, \widehat{\theta}^\pi)$$

which contradicts (13.8). ■

Theorem **13.12** *Suppose that $\widehat{\theta}$ is the Bayes rule with respect to some prior π. 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 $r(\pi, \widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta = c$ and hence $R(\theta, \widehat{\theta}) \leq r(\pi, \widehat{\theta})$ for all θ. Now apply the previous theorem. ■

Example 13.13 *Consider the Bernoulli model with squared error loss. In example 13.3 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 $\text{Beta}(\alpha, \beta)$ *with* $\alpha = \beta = \sqrt{n/4}$. *Hence, by the previous theorem, this estimator is minimax.* ■

Example 13.14 *Consider again the Bernoulli but with loss function*

$$L(p, \widehat{p}) = \frac{(p - \widehat{p})^2}{p(1-p)}.$$

Let

$$\widehat{p}(X^n) = \widehat{p} = \frac{Y}{n}.$$

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 $\pi(p) = 1$. Hence, $\widehat{p}$ is minimax. ■

A natural question to ask is: what is the minimax estimator for a Normal model?

Theorem 13.15 *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. It is the only estimator with this property.*

"Well-behaved" means that the level sets must be convex and symmetric about the origin. The result holds up to sets of measure 0.

If the parameter space is restricted, the theorem above does not apply as the next example shows.

Example 13.16 *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 $R(\theta, \widehat{\theta}) \leq r(\pi, \widehat{\theta})$ for all θ; see Figure 13.3. Hence, Theorem 13.11 implies that $\widehat{\theta}$ is minimax.* ■

13.5 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:

Typically, the squared bias is order $O(n^{-2})$ while the variance is of order $O(n^{-1})$.

$$R(\theta, \widehat{\theta}) = \mathbb{V}_\theta(\widehat{\theta}) + \text{bias}^2 \approx \mathbb{V}_\theta(\widehat{\theta}).$$

As we saw in the Chapter on parametric models, the variance of the MLE is approximately

$$\mathbb{V}(\widehat{\theta}) \approx \frac{1}{nI(\theta)}$$

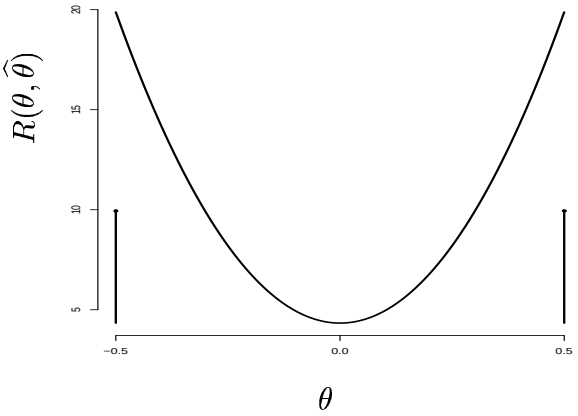


FIGURE 13.3. Risk functions for constrained Normal with m=.5. The two short lines show the least favorable prior which puts its mass at two points.

where $I(\theta)$ is the Fisher information. Hence,

$$nR(\theta, \widehat{\theta}) \approx \frac{1}{I(\theta)}. \tag{13.10}$$

For any other estimator θ', it can be shown that for large n, $R(\theta, \theta') \geq R(\theta, \widehat{\theta})$. More precisely,

$$\lim_{\epsilon \to 0} \limsup_{n \to \infty} \sup_{|\theta - \theta'| < \epsilon} nR(\theta', \widehat{\theta}) \geq \frac{1}{I(\theta)}. \tag{13.11}$$

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 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.

Example 13.17 (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

$$\theta \in \Theta_n \equiv \left\{ (\theta_1, \ldots, \theta_n) : \sum_{i=1}^{n} \theta_i^2 \leq c^2 \right\}$$

for some $c > 0$. In this model, there are as many parameters as observations. 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.*

The many Normal means problem is more general than it looks. Many nonparametric estimation problems are mathematically equivalent to this model.

13.6 Admissibility

Minimax estimator and Bayes estimator are "good estimator" in the sense that they have small risk. Sometimes it is also useful to characterize bad estimator.

Definition 13.18 *An estimator $\widehat{\theta}$ is* **inadmissible** *if there exists another rule $\widehat{\theta}'$ such that*

$$\begin{aligned} R(\theta, \widehat{\theta}') &\leq R(\theta, \widehat{\theta}) \quad \text{for all } \theta \text{ and} \\ R(\theta, \widehat{\theta}') &< R(\theta, \widehat{\theta}) \quad \text{for at least one } \theta. \end{aligned}$$

Example 13.19 *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.* ■

A prior density has **full support** if for every θ and every $\epsilon > 0$, $\int_{\theta-\epsilon}^{\theta+\epsilon} \pi(\theta)d\theta > 0$.

Theorem **13.20 (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 π be a prior density with full support and let $\widehat{\theta}^\pi$ be the Bayes' rule. If the Bayes risk is finite then $\widehat{\theta}^\pi$ is admissible.*

PROOF. Suppose $\widehat{\theta}^\pi$ is inadmissible. Then there exists a better rule $\widehat{\theta}$ such that $R(\theta, \widehat{\theta}) \leq R(\theta, \widehat{\theta}^\pi)$ for all θ and $R(\theta_0, \widehat{\theta}) < R(\theta_0, \widehat{\theta}^\pi)$ for some θ_0. Let $\nu = R(\theta_0, \widehat{\theta}^\pi) - R(\theta_0, \widehat{\theta}) > 0$. Since R is continuous, there is an $\epsilon > 0$ such that $R(\theta, \widehat{\theta}^\pi) - R(\theta, \widehat{\theta}) > \nu/2$ for all $\theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$. Now,

$$\begin{aligned} r(\pi, \widehat{\theta}^\pi) - r(\pi, \widehat{\theta}) &= \int R(\theta, \widehat{\theta}^\pi)\pi(\theta)d\theta - \int R(\theta, \widehat{\theta})\pi(\theta)d\theta \\ &= \int \left[R(\theta, \widehat{\theta}^\pi) - R(\theta, \widehat{\theta})\right] \pi(\theta)d\theta \\ &\geq \int_{\theta_0-\epsilon}^{\theta_0+\epsilon} \left[R(\theta, \widehat{\theta}^\pi) - R(\theta, \widehat{\theta})\right] \pi(\theta)d\theta \\ &\geq \frac{\nu}{2} \int_{\theta_0-\epsilon}^{\theta_0+\epsilon} \pi(\theta)d\theta \\ &> 0. \end{aligned}$$

Hence, $r(\pi, \widehat{\theta}^\pi) > r(\pi, \widehat{\theta})$. This implies that $\widehat{\theta}^\pi$ does not minimize $r(\pi, \widehat{\theta})$ which contradicts the fact that $\widehat{\theta}^\pi$ is the Bayes rule. ■

Theorem **13.21** *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.

Theorem **13.22** *Suppose that $\widehat{\theta}$ has constant risk and is admissible. Then it is minimax.*

PROOF. The risk is $R(\widehat{\theta}, \widehat{\theta}) = c$ for some c. If $\widehat{\theta}$ were not minimax then there exists a rule $\widehat{\theta}'$ such that

$$R(\theta, \widehat{\theta}') \leq \sup_\theta R(\theta, \widehat{\theta}') < \sup_\theta R(\widehat{\theta}, \widehat{\theta}) = c.$$

This would imply that $\widehat{\theta}$ is inadmissible. ■

Now we can prove a restricted version of Theorem 13.15 for squared error loss.

Theorem **13.23** *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 13.21, $\widehat{\theta}$ is admissible. The risk of $\widehat{\theta}$ is $1/n$ which is constant. The result follows from Theorem 13.22. ■

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.

Theorem **13.24** *If $\widehat{\theta}$ is minimax then it is not strongly inadmissible.*

13.7 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 following estimator, known as the James-Stein estimator, has smaller risk:

$$\widehat{\theta}_S(X) = \left(1 - \frac{k-2}{\sum_i X_i^2}\right)^+ X_i \tag{13.12}$$

where $(z)^+ = \max\{z, 0\}$. 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.

13.8 Bibliographic Remarks

It is difficult to find books that cover modern decision theory in great detail. Aspects of decision theory can be found in Casella and Berger (2002), Berger (1985), Ferguson (1967) and Lehmann and Casella (1998).

13.9 Exercises

1. In each of the following models, find (i) 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.) 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 (13.12) by simulation. Try various values of n and various vectors θ. Summarize your results.

Part III

Statistical Models and Methods

14
Linear Regression

Regression is a method for studying the relationship between a **response variable** Y and a **covariates** X. The covariate is also called a **predictor variable** or a **feature.** Later, we will generalize and allow for more than one covariate. The data are of the form

$$(Y_1, X_1), \ldots, (Y_n, X_n).$$

One way to summarize the relationship between X and Y is through the **regression function**

$$r(x) = \mathbb{E}(Y|X = x) = \int y\, f(y|x)dy. \tag{14.1}$$

Most of this chapter is concerned with estimating the regression function.

The term "regression" is due to Sir Francis Galton (1822-1911) who noticed that tall and short men tend to have sons with heights closer to the mean. He called this "regression towards the mean."

14.1 Simple Linear Regression

The simplest version of regression is when X_i is simple (a scalar not a vector) and $r(x)$ is assumed to be linear:

$$r(x) = \beta_0 + \beta_1 x.$$

This model is called the **the simple linear regression model**. Let $\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$. Then,

$$\begin{aligned}\mathbb{E}(\epsilon_i|X_i) &= \mathbb{E}(Y_i - (\beta_0 + \beta_1 X_i)|X_i) = \mathbb{E}(Y_i|X_i) - (\beta_0 + \beta_1 X_i) \\ &= r(X_i) - (\beta_0 + \beta_1 X_i) \\ &= (\beta_0 + \beta_1 X_i) - (\beta_0 + \beta_1 X_i) \\ &= 0.\end{aligned}$$

Let $\sigma^2(x) = \mathbb{V}(\epsilon_i|X = x)$. We will make the further simplifying assumption that $\sigma^2(x) = \sigma^2$ does not depend on x. We can thus write the linear regression model as follows.

Definition 14.1 The Linear Regresion Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{14.2}$$

where $\mathbb{E}(\epsilon_i|X_i) = 0$ *and* $\mathbb{V}(\epsilon_i|X_i) = \sigma^2$.

Example 14.2 *Figure 14.1 shows a plot of Log surface temperature (Y) versus Log light intensity (X) for some nearby stars. Also on the plot is an estimated linear regression line which will be explained shortly.*

The unknown parameters in the model are the intercept β_0 and the slope β_1 and the variance σ^2. Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ denote estimates of β_0 and β_1. The **fitted line** is defined to be

$$\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x. \tag{14.3}$$

The **predicted values** or **fitted values** are $\widehat{Y}_i = \widehat{r}(X_i)$ and the **residuals** are defined to be

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right). \tag{14.4}$$

The **residual sums of squares** or RSS is defined by

$$\text{RSS} = \sum_{i=1}^{n} \widehat{\epsilon}_i^2.$$

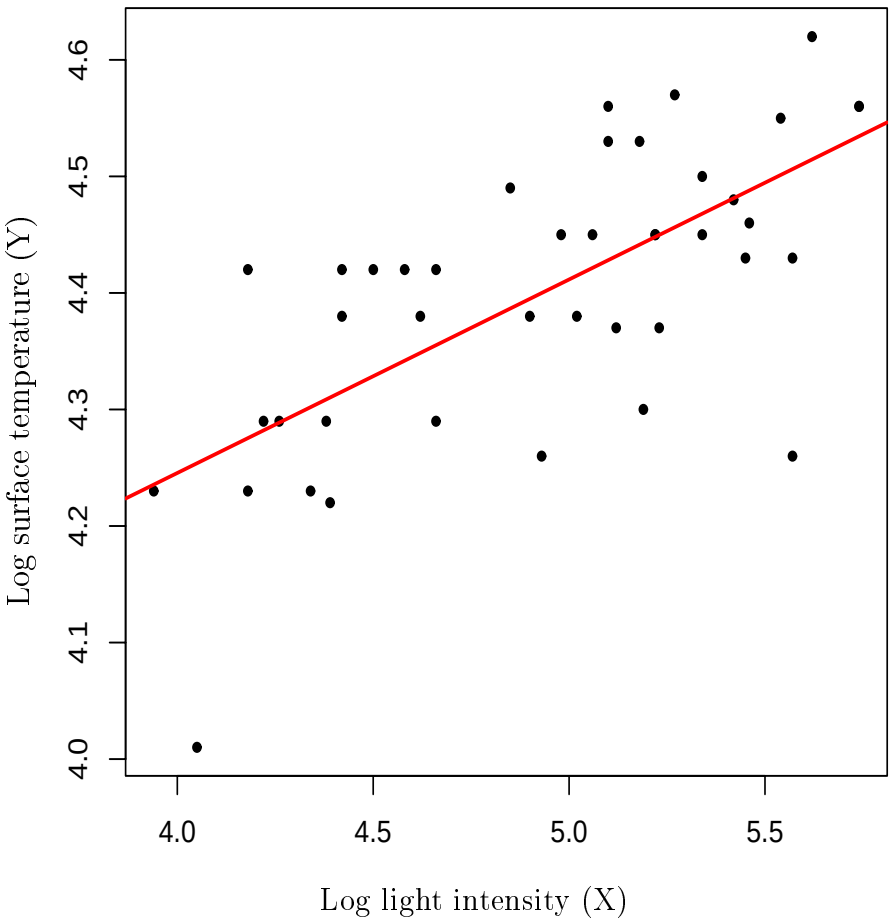


FIGURE 14.1. Data on stars in nearby stars.

The quantity RSS measures how well the fitted line fits the data.

Definition 14.3 *The* **least squares estimates** *are the values* $\widehat{\beta}_0$ *and* $\widehat{\beta}_1$ *that minimize* RSS $= \sum_{i=1}^n \widehat{\epsilon}_i^2$.

Theorem **14.4** *The least squares estimates are given by*

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^n (X_i - \overline{X}_n)^2}, \tag{14.5}$$

$$\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n. \tag{14.6}$$

An unbiased estimate of σ^2 *is*

$$\widehat{\sigma}^2 = \left(\frac{1}{n-2}\right) \sum_{i=1}^n \widehat{\epsilon}_i^2. \tag{14.7}$$

Example 14.5 *Consider the star data from Example 14.2. The least squares estimates are* $\widehat{\beta}_0 = 3.58$ *and* $\widehat{\beta}_1 = 0.166$. *The fitted line* $\widehat{r}(x) = 3.58 + 0.166x$ *is shown in Figure 14.1.* ■

Example 14.6 (The 2001 Presidential Election.) *Figure 14.2 shows the plot of votes for Buchanan (Y) versus votes for Bush (X) in Florida. The least squares estimates (omitting Palm Beach County) and the standard errors are*

$$\begin{aligned} \widehat{\beta}_0 &= 66.0991 \quad \widehat{\mathsf{se}}\,(\widehat{\beta}_0) = 17.2926 \\ \widehat{\beta}_1 &= 00.0035 \quad \widehat{\mathsf{se}}\,(\widehat{\beta}_1) = 0.0002. \end{aligned}$$

The fitted line is

$$\text{Buchanon} = 66.0991 + .0035\,\text{Bush}.$$

(We will see later how the standard errors were computed.) Figure 14.2 also shows the residuals. The inferences from linear

regression are most accurate when the residuals behave like random normal numbers. Based on the residual plot, this is not the case in this example. If we repeat the analysis replacing votes with log(votes) we get

$$\begin{aligned} \widehat{\beta}_0 &= -2.3298 \quad \widehat{\mathsf{se}}\,(\widehat{\beta}_0) = .3529 \\ \widehat{\beta}_1 &= 0.730300 \quad \widehat{\mathsf{se}}\,(\widehat{\beta}_1) = 0.0358. \end{aligned}$$

This gives the fit

$$\log(\text{Buchanon}) = -2.3298 + .7303\ \log(\text{Bush}).$$

The residuals look much healthier. Later, we shall address two interesting questions: (1) how do we see if Palm Beach County has a statistically plausible outcome? (2) how do we do this problem nonparametrically? ■

14.2 Least Squares and Maximum Likelihood

Suppose we add the assumption that $\epsilon_i | X_i \sim N(0, \sigma^2)$, that is,

$$Y_i | X_i \sim N(\mu_i, \sigma_i^2)$$

where $\mu_i = \beta_0 + \beta_1 X_i$. The likelihood function is

$$\begin{aligned} \prod_{i=1}^n f(X_i, Y_i) &= \prod_{i=1}^n f_X(X_i) f_{Y|X}(Y_i|X_i) \\ &= \prod_{i=1}^n f_X(X_i) \times \prod_{i=1}^n f_{Y|X}(Y_i|X_i) \\ &= \mathcal{L}_1 \times \mathcal{L}_2 \end{aligned}$$

where $\mathcal{L}_1 = \prod_{i=1}^n f_X(X_i)$ and

$$\mathcal{L}_2 = \prod_{i=1}^n f_{Y|X}(Y_i|X_i). \tag{14.8}$$

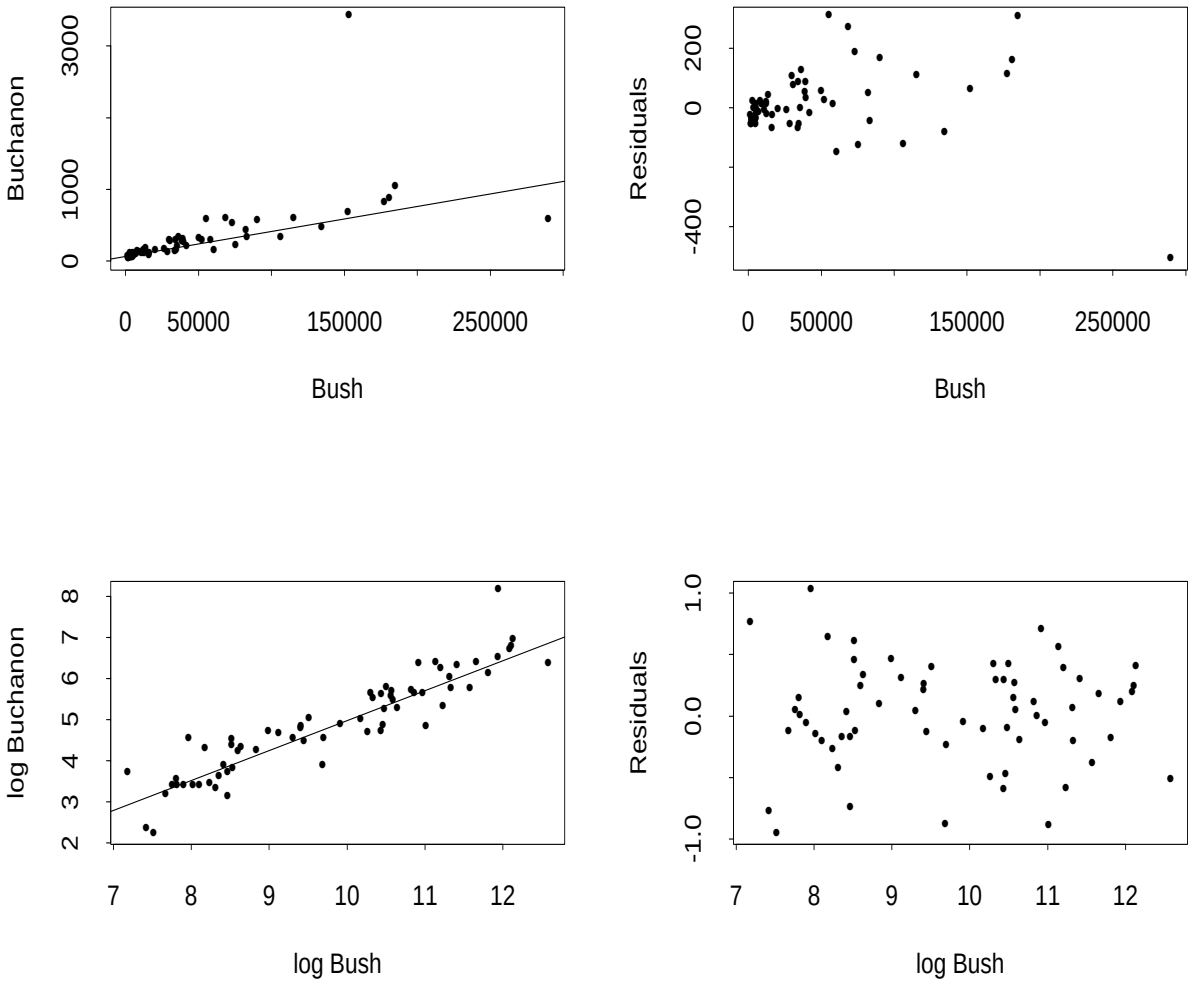


FIGURE 14.2. Voting Data for Election 2000.

The term $\mathcal{L}_1$ does not involve the parameters β_0 and β_1. We shall focus on the second term $\mathcal{L}_2$ which is called the **conditional likelihood**, given by

$$\mathcal{L}_2 \equiv \mathcal{L}(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (Y_i - \mu_i)^2\right\}.$$

The conditional log-likelihood is

$$\ell(\beta_0, \beta_1, \sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(Y_i - (\beta_0 + \beta_1 X_i)\Big)^2. \quad (14.9)$$

To find the MLE of (β_0, β_1) we maximize $\ell(\beta_0, \beta_1, \sigma)$. From (14.9) we see that maximizing the likelihood is the same as minimizing the RSS $\sum_{i=1}^{n}\Big(Y_i - (\beta_0 + \beta_1 X_i)\Big)^2$. Therefore, we have shown the following.

Theorem **14.7** *Under the assumption of Normality, the least squares estimator is also the maximum likelihood estimator.*

We can also maximize $\ell(\beta_0, \beta_1, \sigma)$ over σ yielding the MLE

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i \widehat{\epsilon}_i^2. \quad (14.10)$$

This estimator is similar to, but not identical to, the unbiased estimator. Common practice is to use the unbiased estimator (14.7).

14.3 Properties of the Least Squares Estimators

We now record the standard errors and limiting distribution of the least squares estimator. In regression problems, we usually focus on the properties of the estimators conditional on $X^n = (X_1, \ldots, X_n)$. Thus, we state the means and variances as conditional means and variances.

Theorem 14.8 *Let* $\widehat{\beta}^T = (\widehat{\beta}_0, \widehat{\beta}_1)^T$ *denote the least squares estimators. Then,*

$$\begin{aligned} \mathbb{E}(\widehat{\beta}|X^n) &= \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \\ \mathbb{V}(\widehat{\beta}|X^n) &= \frac{\sigma^2}{n\, s_X^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix} \end{aligned} \tag{14.11}$$

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

The estimated standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are obtained by taking the square roots of the corresponding diagonal terms of $\mathbb{V}(\widehat{\beta}|X^n)$ and inserting the estimate $\widehat{\sigma}$ for σ. Thus,

$$\widehat{\mathsf{se}}\,(\widehat{\beta}_0) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}} \tag{14.12}$$

$$\widehat{\mathsf{se}}\,(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}. \tag{14.13}$$

We should really write these as $\widehat{\mathsf{se}}\,(\widehat{\beta}_0|X^n)$ and $\widehat{\mathsf{se}}\,(\widehat{\beta}_1|X^n)$ but we will use the shorter notation $\widehat{\mathsf{se}}\,(\widehat{\beta}_0)$ and $\widehat{\mathsf{se}}\,(\widehat{\beta}_1)$.

Theorem 14.9 *Under appropriate conditions we have:*

1. *(Consistency):* $\widehat{\beta}_0 \xrightarrow{\text{P}} \beta_0$ *and* $\widehat{\beta}_1 \xrightarrow{\text{P}} \beta_1$.
2. *(Asymptotic Normality):*

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\mathsf{se}}\,(\widehat{\beta}_0)} \rightsquigarrow N(0,1) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathsf{se}}\,(\widehat{\beta}_1)} \rightsquigarrow N(0,1).$$

3. *Approximate* $1-\alpha$ *confidence intervals for* β_0 *and* β_1 *are*

$$\widehat{\beta}_0 \pm z_{\alpha/2}\,\widehat{\mathsf{se}}\,(\widehat{\beta}_0) \quad \text{and} \quad \widehat{\beta}_1 \pm z_{\alpha/2}\,\widehat{\mathsf{se}}\,(\widehat{\beta}_1). \tag{14.14}$$

The Wald test statistic for testing $H_0 : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$ *is: reject* H_0 *if* $W > z_{\alpha/2}$ *where* $W = \widehat{\beta}_1/\widehat{\mathsf{se}}\,(\widehat{\beta}_1)$.

Example 14.10 *For the election data, on the log scale, a 95 per cent confidence interval is* $.7303 \pm 2(.0358) = (.66, .80)$. *The fact that the interval excludes 0 The Wald statistics for testing* $H_0 : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$ *is* $W = |.7303 - 0|/.0358 = 20.40$ *with a p-value of* $\mathbb{P}(|Z| > 20.40) \approx 0$. *This is strong evidence that that the true slope is not 0.* ■

14.4 Prediction

Suppose we have estimated a regression model $\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1$ from data $(X_1, Y_1), \ldots, (X_n, Y_n)$. We observe the value $X = x_*$ of the covariate for a new subject and we want to predict their outcome Y_*. An estimate of Y_* is

$$\widehat{Y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 X_*. \tag{14.15}$$

Using the formula for the variance of the sum of two random variables,

$$\mathbb{V}(\widehat{Y}_*) = \mathbb{V}(\widehat{\beta}_0 + \widehat{\beta}_1 x_*) = \mathbb{V}(\widehat{\beta}_0) + x_*^2 \mathbb{V}(\widehat{\beta}_1) + 2x_* \mathsf{Cov}(\widehat{\beta}_0, \widehat{\beta}_1).$$

Theorem 14.8 gives the formulas for all the terms in this equation. The estimated standard error $\widehat{\mathsf{se}}(\widehat{Y}_*)$ is the square root of this variance, with $\widehat{\sigma}^2$ in place of σ^2. However, the confidence interval for Y_* is **not** of the usual form $\widehat{Y}_* \pm z_{\alpha/2}$. The appendix explains why. The correct form of the confidence interval is given in the following Theorem. We call the interval a **prediction interval**.

Theorem **14.11 (Prediction Interval)** *Let*

$$\begin{aligned}\widehat{\xi}_n^2 &= \widehat{\mathsf{se}}^{\,2}(\widehat{Y}_*) + \widehat{\sigma}^2 \\ &= \widehat{\sigma}^2\left(\frac{\sum_{i=1}^n (X_i - X_*)^2}{n\sum_i (X_i - \overline{X})^2} + 1\right). \qquad (14.16)\end{aligned}$$

An approximate $1-\alpha$ *predicition interval for* Y_* *is*

$$\widehat{Y}_* \pm z_{\alpha/2}\xi_n. \qquad (14.17)$$

Example 14.12 (Election Data Revisited.) *On the log-scale, our linear regression gives the following prediction equation:* $log(Buchanon) = -2.3298 + .7303 log(Bush)$. *In Palm Beach, Bush had 152954 votes and Buchanan had 3467 votes. On the log scale this is 11.93789 and 8.151045. How likely is this outcome, assuming our regression model is appropriate? Our prediction for log Buchanan votes -2.3298 + .7303 (11.93789)=6.388441. Now 8.151045 is bigger than 6.388441 but is is "significantly" bigger? Let us compute a confidence interval. We find that* $\widehat{\xi}_n = .093775$ *and the approximate 95 per cent confidence interval is (6.200,6.578) which clearly excludes 8.151. Indeed, 8.151 is nearly 20 standard errors from* $\widehat{Y}_*$. *Going back to the vote scale by exponentiating, the confidence interval is (493,717) compared to the actual number of votes which is 3467.* ■

14.5 Multiple Regression

Now suppose we have k covariates $X_1, \ldots, X_k$. The data are of the form

$$(Y_1, X_1), \ldots, (Y_i, X_i), \ldots, (Y_n, X_n)$$

where

$$X_i = (X_{i1}, \ldots, X_{ik}).$$

Here, X_i is the vector of k covariate values for the i^{th} observation. The linear regression model is

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \tag{14.18}$$

for $i = 1, \ldots, n$, where $\mathbb{E}(\epsilon_i | X_{1i}, \ldots, X_{ki}) = 0$. Usually we want to include an intercept in the model which we can do by setting $X_{i1} = 1$ for $i = 1, \ldots, n$. At this point it will be more convenient to express the model in matrix notation. The outcomes will be denoted by

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

and the covariates will be denoted by

$$X = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ X_{21} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{nk} \end{pmatrix}.$$

Each row is one observation; the columns correspond to the k covariates. Thus, X is a $(n \times k)$ matrix. Let

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \text{and} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Then we can write (14.18) as

$$Y = X\beta + \epsilon. \tag{14.19}$$

Theorem **14.13** *Assuming that the $(k \times k)$ matrix $X^T X$ is invertible, the least squares estimate is*

$$\widehat{\beta} = (X^T X)^{-1} X^T Y. \tag{14.20}$$

The estimated regression function is

$$\widehat{r}(x) = \sum_{j=1}^{k} \widehat{\beta}_j x_j. \tag{14.21}$$

The variance-covariance matrix of $\widehat{\beta}$ *is*

$$\mathbb{V}(\widehat{\beta}|X^n) = \sigma^2 (X^T X)^{-1}.$$

Under appropriate conditions,

$$\widehat{\beta} \approx N(\beta, \sigma^2 (X^T X)^{-1}).$$

An unbiased estimate of σ^2 *is*

$$\widehat{\sigma}^2 = \left(\frac{1}{n-k}\right) \sum_{i=1}^{n} \widehat{\epsilon}_i^2$$

where $\widehat{\epsilon} = X\widehat{\beta} - Y$ *is the vector of residuals. An approximate* $1-\alpha$ *confidence interval for* β_j *is*

$$\widehat{\beta}_j \pm z_{\alpha/2} \widehat{\mathsf{se}}\,(\widehat{\beta}_j) \tag{14.22}$$

where $\widehat{\mathsf{se}}^{\,2}(\widehat{\beta}_j)$ *is the* j^{th} *diagonal element of the matrix* $\widehat{\sigma}^2 (X^T X)^{-1}$.

Example 14.14 *Crime data on 47 states in 1960 can be obtained at http://lib.stat.cmu.edu/DASL/Stories/USCrime.html. If we fit a linear regression of crime rate on 10 variables we get the following:*

Covariate	Least Squares Estimate	Estimated Standard Error	t value	p-value	
(Intercept)	*-589.39*	*167.59*	*-3.51*	*0.001*	**
Age	*1.04*	*0.45*	*2.33*	*0.025*	*
Southern State	*11.29*	*13.24*	*0.85*	*0.399*	
Education	*1.18*	*0.68*	*1.7*	*0.093*	
Expenditures	*0.96*	*0.25*	*3.86*	*0.000*	***
Labor	*0.11*	*0.15*	*0.69*	*0.493*	
Number of Males	*0.30*	*0.22*	*1.36*	*0.181*	
Population	*0.09*	*0.14*	*0.65*	*0.518*	
Unemployment (14-24)	*-0.68*	*0.48*	*-1.4*	*0.165*	
Unemployment (25-39)	*2.15*	*0.95*	*2.26*	*0.030*	*
Wealth	*-0.08*	*0.09*	*-0.91*	*0.367*	

This table is typical of the output of a multiple regression program. The "t-value" is the Wald test statistic for testing $H_0 : \beta_j = 0$ *versus* $H_1 : \beta_j \neq 0$. *The asterisks denote "degree of significance" with more asterisks being significant at a smaller level. The example raises several important questions. In particular: (1) should we eliminate some variables from this model? (2) should we interpret this relationships as causal? For example, should we conclude that low crime prevention expenditures cause high crime rates? We will address question (1) in the next section. We will not address question (2) until a later Chapter.*

14.6 Model Selection

Example 14.14 illustrates a problem that often arises in multiple regression. We may have data on many covariates but we may not want to include all of them in the model. A smaller model with fewer covariates has two advantages: it might give better predictions than a big model and it is more parsimonious (simpler). Generally, as you add more variables to a regression,

the bias of the predictions decreases and the variance increases. Too few covariates yields high bias; too many covariates yields high variance. Good predictions result from achieving a good balance between bias and variance.

In model selection there are two problems: (i) assigning a "score" to each model which measures, in some sense, how good the model is and (ii) searching through all the models to find the model with the best score.

Let us first discuss the problem of scoring models. Let $S \subset \{1, \ldots, k\}$ and let $\mathcal{X}_S = \{X_j : \ j \in S\}$ denote a subset of the covariates. Let β_S denote the coefficients of the corresponding set of covariates and let $\widehat{\beta}_S$ denote the least squares estimate of β_S. Also, let X_S denote the X matrix for this subset of covariates and define $\widehat{r}_S(x)$ to be the estimated regression function from (14.21). The predicted valus from model S are denoted by $\widehat{Y}_i(S) = \widehat{r}_S(X_i)$. The **prediction risk** is defined to be

$$R(S) = \sum_{i=1}^{n} \mathbb{E}(\widehat{Y}_i(S) - Y_i^*)^2 \tag{14.23}$$

where Y_i^* denotes the value of a future observation of Y_i at covariate value X_i. Our goal is to choose S to make $R(S)$ small.

The **training error** is defined to be

$$\widehat{R}_{\text{tr}}(S) = \sum_{i=1}^{n} (\widehat{Y}_i(S) - Y_i)^2.$$

This estimate is very biased and under-estimates $R(S)$.

Theorem **14.15** *The training error is a downward biased estimate of the prediction risk:*

$$\mathbb{E}(\widehat{R}_{\text{tr}}(S)) < R(S).$$

In fact,

$$\mathsf{bias}\,(\widehat{R}_{\text{tr}}(S)) = \mathbb{E}(\widehat{R}_{\text{tr}}(S)) - R(S) = -2\sum_{i=1} \mathsf{Cov}(\widehat{Y}_i, Y_i). \tag{14.24}$$

The reason for the bias is that the data are being used twice: to estimate the parameters and to estimate the risk. When we fit a complex model with many parameters, the covariance $\mathsf{Cov}(\widehat{Y}_i, Y_i)$ will be large and the bias of the training error gets worse. In summary, the training error is a poor estimate of risk. Here are some better estimates.

Mallow's C_p statistic is defined by

$$\widehat{R}(S) = \widehat{R}_{\text{tr}}(S) + 2|S|\widehat{\sigma}^2 \tag{14.25}$$

where $|S|$ denotes the number of terms in S and $\widehat{\sigma}^2$ is the estimate of σ^2 obtained from the full model (with all covariates in the model). This is simply the training error plus a bias correction. This estimate is named in honor of Colin Mallows who invented it. The first term in (14.25) measures the fit of the model while the second measure the complexity of the model. Think of the C_p statistic as:

$$\text{lack of fit } + \text{ complexity penalty.}$$

Thus, **finding a good model involves trading off fit and complexity.**

A related method for estimating risk is **AIC (Akaike Information Criterion).** The idea is to choose S to maximize

$$\ell_S - |S| \tag{14.26}$$

where ℓ_S is the log-likelihood of the model evaluated at the MLE . This can be thought of "goodness of fit" minus "complexity." In linear regression with Normal errors, maximizing AIC is equivalent to minimizing Mallow's C_p; see exercise 8.

Some texts use a slightly different definition of AIC which involves multiplying the definition here by 2 or -2. This has no effect on which model is selected.

Yet another method for estimating risk is **leave-one-out cross-validation**. In this case, the risk estimator is

$$\widehat{R}_{CV}(S) = \sum_{i=1}^{n}(Y_i - \widehat{Y}_{(i)})^2 \tag{14.27}$$

where $\widehat{Y}_{(i)}$) is the prediction for Y_i obtained by fitting the model with Y_i omitted. It can be shown that

$$\widehat{R}_{CV}(S) = \sum_{i=1}^{n} \left(\frac{Y_i - \widehat{Y}_i(S)}{1 - U_{ii}(S)} \right)^2 \tag{14.28}$$

where $U_{ii}(S)$ is the i^{th} diagonal element of the matrix

$$U(S) = X_S(X_S^T X_S)^{-1} X_S^T. \tag{14.29}$$

Thus, one need not actually drop each observation and re-fit the model. A generalization is **k-fold cross-validation.** Here we divide the data into k groups; often people take $k = 10$. We omit one group of data and fit the models to the remaining data. We use the fitted model to predict the data in the group that was omitted. We then estimate the risk by $\sum_i (Y_i - \widehat{Y}_i)^2$ where the sum is over the the data points in the omitted group. This process is repeated for each of the k groups and the resulting risk estimates are averaged.

For linear regression, Mallows C_p and cross-validation often yield essentially the same results so one might as well use Mallows' method. In some of the more complex problems we will discuss later, cross-validation will be more useful.

Another scoring method is BIC (Bayesian information criterion). Here we choose a model to maximize

$$\text{BIC}(S) = \text{RSS}\ (S) + 2|S|\widehat{\sigma}^2. \tag{14.30}$$

The BIC score has a Bayesian interpretation. Let $\mathcal{S} = \{S_1, \ldots, S_m\}$ denote a set of models. Suppose we assign the prior $\mathbb{P}(S_j) = 1/m$ over the models. Also, assume we put a smooth prior on the parameters within each model. It can be shown that the posterior probability for a model is approximately,

$$\mathbb{P}(S_j|\text{data}) \approx \frac{e^{BIC(S_j)}}{\sum_r e^{BIC(S_r)}}.$$

Hence, choosing the model with highest BIC is like choosing the model with highest posterior probability. The BIC score also has an information-theoretic interpretation in terms of something called minimum description length. The BIC score is identical to Mallows C_p except that it puts a more severe penalty for complexity. It thus leads one to choose a smaller model than the other methods.

Now let us turn to the problem of model search. If there are k covariates then there are 2^k possible models. We need to search through all these models, assign a score to each one, and choose the model with the best score. If k is not too large we can do a complete search over all the models. When k is large, this is infeasible. In that case we need to search over a subset of all the models. Two common methods are **forward and backward stepwise regression.** In forward stepwise regression, we start with no covariates in the model. We then add the one variable that leads to the best score. We continue adding variables one at a time until the score does not improve. Backwards stepwise regression is the same except that we start with the biggest model and drop one variable at a time. Both are greedy searches; nether is guaranteed to find the model with the best score. Another popular method is to do random searching through the set of all models. However, there is no reason to expect this to be superior to a deterministic search.

Example 14.16 *We apply backwards stepwise regression to the crime data using AIC. The following was obtained from the program R. This program uses minus our version of AIC. Hence, we are seeking the smallest possible AIC. This is the same is minimizing Mallows* C_p.

The full model (which includes all covariates) has AIC= 310.37. The AIC scores, in ascending order, for deleting one variable are as follows:

variable	*Pop*	*Labor*	*South*	*Wealth*	*Males*	*U1*	*Educ.*	*U2*	*Age*	*Expend*
AIC	*308*	*309*	*309*	*309*	*310*	*310*	*312*	*314*	*315*	*324*

For example, if we dropped Pop from the model and kept the other terms, then the AIC score would be 308. Based on this information we drop "population" from the model and the current AIC score is 308. Now we consider dropping a variable from the current model. The AIC scores are:

variable	*South*	*Labor*	*Wealth*	*Males*	*U1*	*Education*	*U2*	*Age*	*Expend*
AIC	*308*	*308*	*308*	*309*	*309*	*310*	*313*	*313*	*329*

We then drop "Southern" from the model. This process is continued until there is no gain in AIC by dropping any variables. In the end, we are left with the following model:

$$\text{Crime} = 1.2\ \text{Age} + .75\ \text{Education} + .87\ \text{Expenditure} + .34\ \text{Males} - .86\ \text{U1} + 2.31\ \text{U2}.$$

Warning! *This does not yet address the question of which variables are* **causes** *of crime.*

14.7 The Lasso

There is an easier model search method although it addresses a slightly different question. The method, due to Tibshirani, is called the **Lasso**. In this section we assume that the covariates have all been rescaled to have the same variance. This puts each covariate on the same scale. Consider estimating $\beta = (\beta_1, \ldots, \beta_k)$ by minimizing the loss function

$$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \tag{14.31}$$

where $\lambda > 0$. The idea is to minimize the sums of squares but we include a penalty that gets large if any of the $\beta_j's$ are large. The solution $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_k)$ can be found numerically and will depend on the choice of λ. It can be shown that some of the $\widehat{\beta}_j$'s will be 0. We interpret this is meaning that the j^{th} is omitted from the model. Hence, we are doing estimation and model selection simultaneously. We need to choose a value of λ. We can do this by estimating the prediction risk $R(\lambda)$ as a function of λ and choosing λ to minimze the estimated risk. For example, we can estimate the risk using leave-one-out cross-validation.

Example 14.17 *Returning to the crime data, Figure 14.3 shows the results of the lasso. The first plot shows the leave-one-out cross-validation score as a function of λ. The minimum occurs at $\lambda = .7$. The second plot shows the estimated coefficients as a function of λ. You can see how some estimated parameters are zero until λ gets larger. At $\lambda = .7$, all the $\widehat{\beta}_j$ are non-zero so the Lasso chooses the full model.*

14.8 Technical Appendix

Why is the predicition interval of a different form than the other confidence intervals we have seen? The reason is that the quantity we want to estimate, Y_*, is not a fixed parameter, it is a random variable. To understand this point better, let $\theta = \beta_0 + \beta_1 X_*$ and let $\widehat{\theta} = \widehat{\beta}_0 + \widehat{\beta}_1 X_*$. Thus, $\widehat{Y}_* = \widehat{\theta}$ while $Y_* = \theta + \epsilon$. Now, $\widehat{\theta} \approx N(\theta, \mathsf{se}^2)$ where

$$\mathsf{se}^2 = \mathbb{V}(\widehat{\theta}) = \mathbb{V}(\widehat{\beta}_0 + \widehat{\beta}_1 x_*).$$

Note that $\mathbb{V}(\widehat{\theta})$ is the same as $\mathbb{V}(\widehat{Y}_*)$. Now, $\widehat{\theta} \pm 2\sqrt{Var(\widehat{\theta})}$ is a 95 per cent confidence interval for $\theta = \beta_0 + \beta_1 x_*$ using the usual

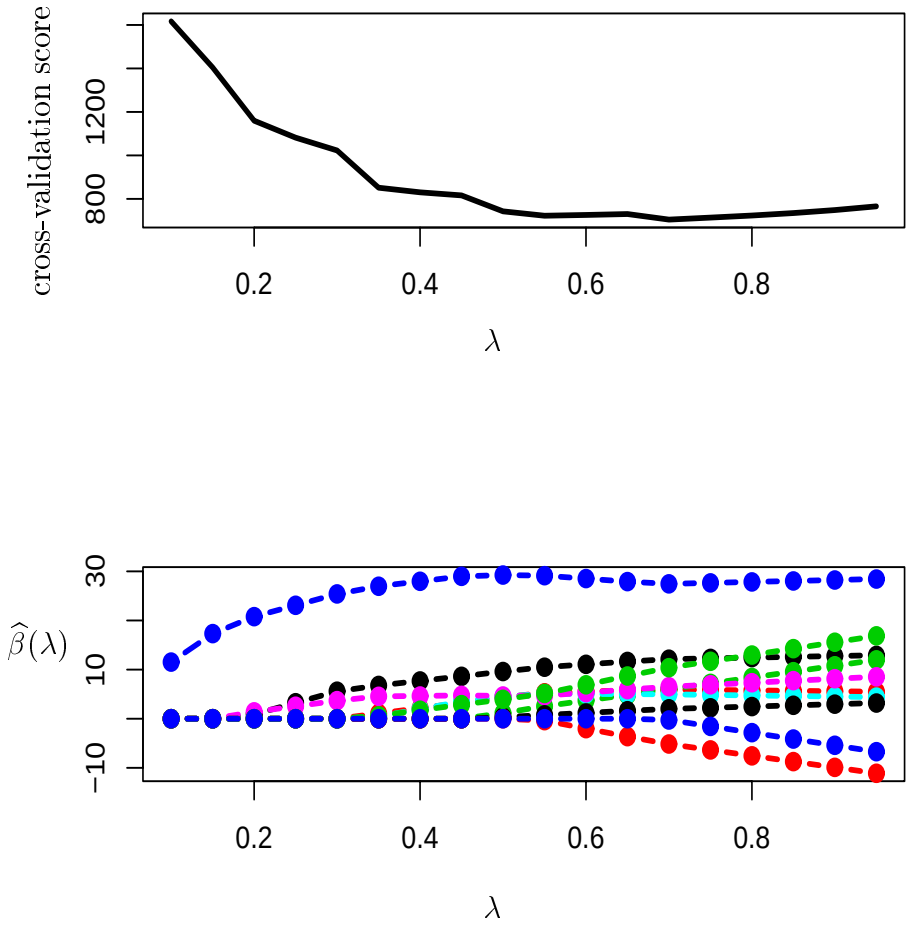


FIGURE 14.3. The Lasso applied to the crime data.

argument for a confidence interval. It is not a valid confidence interval for Y_*. To see why, let's compute the probability that $\widehat{Y}_* \pm 2\sqrt{\mathbb{V}(\widehat{Y}_*)}$ contains Y_*. Let $s = \sqrt{Var(\widehat{Y}_*)}$. Then,

$$\begin{aligned}
\mathbb{P}(\widehat{Y}_* - 2s < Y_* < \widehat{Y}_* + 2s) &= \mathbb{P}\left(-2 < \frac{\widehat{Y}_* - Y_*}{s} < 2\right) \\
&= \mathbb{P}\left(-2 < \frac{\widehat{\theta} - \theta - \epsilon}{s} < 2\right) \\
&= \mathbb{P}\left(-2 < \frac{\widehat{\theta} - \theta}{s} - \frac{\epsilon}{s} < 2\right) \\
&\approx \mathbb{P}\left(-2 < N(0,1) - N\left(0, \frac{\sigma^2}{s^2}\right) < 2\right) \\
&= \mathbb{P}\left(-2 < N\left(0, 1 + \frac{\sigma^2}{s^2}\right) < 2\right) \\
&\neq .95.
\end{aligned}$$

The problem is that the quantity of interest Y_* is equal to a parameter θ plus a random variable. We can fix this by defining

$$\xi_n^2 = \mathbb{V}(\widehat{Y}_*) + \sigma^2 = \left[\frac{\sum_i (x_i - x_*)^2}{n\sum_i (x_i - \overline{x})^2} + 1\right]\sigma^2.$$

In practice, we substitute $\widehat{\sigma}$ for σ and we denote the resulting quantity by $\widehat{\xi}_n$. Now consider the interval $\widehat{Y}_* \pm 2\widehat{\xi}_n$. Then,

$$\begin{aligned}
\mathbb{P}(\widehat{Y}_* - 2\widehat{\xi}_n < Y_* < \widehat{Y}_* + 2\widehat{\xi}_n) &= \mathbb{P}\left(-2 < \frac{\widehat{Y}_* - Y_*}{\widehat{\xi}_n} < 2\right) \\
&= \mathbb{P}\left(-2 < \frac{\widehat{\theta} - \theta - \epsilon}{\widehat{\xi}_n} < 2\right) \\
&\approx \mathbb{P}\left(-2 < \frac{N(0, s^2 + \sigma^2)}{\widehat{\xi}_n} < 2\right) \\
&\approx \mathbb{P}\left(-2 < \frac{N(0, s^2 + \sigma^2)}{\xi_n} < 2\right)
\end{aligned}$$

$$= \mathbb{P}(-2 < N(0,1) < 2) = .95.$$

Of course, a $1-\alpha$ interval is given by $\widehat{Y}_* \pm z_{\alpha/2}\widehat{\xi}_n$.

14.9 Exercises

1. Prove Theorem 14.4.

2. Prove the formulas for the standard errors in Theorem 14.8. You should regard the X_i's as fixed constants.

3. Consider the **regression through the origin** model:

$$Y_i = \beta X_i + \epsilon.$$

   Find the least squares estimate for β. Find the standard error of the estimate. Find conditions that guarantee that the estimate is consistent.

4. Prove equation (14.24).

5. In the simple linear regression model, construct a Wald test for $H_0 : \beta_1 = 17\beta_0$ versus $H_1 : \beta_1 \neq 17\beta_0$.

6. Get the passenger car mileage data from

   http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html

   (a) Fit a simple linear regression model to predict MPG (miles per gallon) from HP (horsepower). Summarize your analysis including a plot of the data with the fitted line.

   (b) Repeat the analysis but use log(MPG) as the response. Compare the analyses.

7. Get the passenger car mileage data from

   http://lib.stat.cmu.edu/DASL/Datafiles/carmpgdat.html

   (a) Fit a multiple linear regression model to predict MPG (miles per gallon) from HP (horsepower). Summarize your analysis.

(b) Use Mallow C_p to select a best sub-model. To search trhough the models try (i) all possible models, (ii) forward stepwise, (iii) backward stepwise. Summarize your findings.

(c) Repeat (b) but use BIC. Compare the results.

(d) Now use the Lasso and compare the results.

8. Assume that the errors are Normal. Show that the model with highest AIC (equation (14.26)) is the model with the lowest Mallows C_p statistic.

9. In this question we will take a closer look at the AIC method. Let $X_1, \ldots, X_n$ be iid observations. Consider two models $\mathcal{M}_0$ and $\mathcal{M}_1$. Under $\mathcal{M}_0$ the data are assumed to be $N(0,1)$ while under $\mathcal{M}_1$ the data are assumed to be $N(\theta, 1)$ for some unknown $\theta \in \mathbb{R}$:

$$\begin{aligned} \mathcal{M}_0 : \ X_1, \ldots, X_n &\sim N(0,1) \\ \mathcal{M}_1 : \ X_1, \ldots, X_n &\sim N(\theta, 1), \quad \theta \in \mathbb{R}. \end{aligned}$$

This is just another way to view the hypothesis testing problem: $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Let $\ell_n(\theta)$ be the log-likelihood function. The AIC score for a model is the log-likelihood at the mle minus the number of parameters. (Some people multiply this score by 2 but that is irrelevant.) Thus, the AIC score for $\mathcal{M}_0$ is $AIC_0 = \ell_n(0)$ and the AIC score for $\mathcal{M}_1$ is $AIC_1 = \ell_n(\widehat{\theta}) - 1$. Suppose we choose the model with the highest AIC score. Let J_n denote the selected model:

$$J_n = \begin{cases} 0 & \text{if } AIC_0 > AIC_1 \\ 1 & \text{if } AIC_1 > AIC_0. \end{cases}$$

(a) Suppose that $\mathcal{M}_0$ is the true model, i.e. $\theta = 0$. Find

$$\lim_{n \to \infty} \mathbb{P}(J_n = 0).$$

Now compute $\lim_{n\to\infty} \mathbb{P}(J_n = 0)$ when $\theta \neq 0$.

(b) The fact that $\lim_{n\to\infty} \mathbb{P}(J_n = 0) \neq 1$ when $\theta = 0$ is why some people say that AIC "overfits." But this is not quite true as we shall now see. Let $\phi_\theta(x)$ denote a Normal density function with mean θ and variance 1. Define

$$\widehat{f}_n(x) = \begin{cases} \phi_0(x) & \text{if } J_n = 0 \\ \phi_{\widehat{\theta}}(x) & \text{if } J_n = 1. \end{cases}$$

If $\theta = 0$, show that $D(\phi_0, \widehat{f}_n) \xrightarrow{p} 0$ as $n \to \infty$ where

$$D(f, g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

is the Kullback-Leibler distance. Show also that $D(\phi_\theta, \widehat{f}_n) \xrightarrow{p} 0$ if $\theta \neq 0$. Hence, AIC consistently estimates the true density even if it "overshoots" the correct model.

REMARK: If you are feeling ambitious, repeat this analysis for BIC which is the log-likelihood minus $(p/2)\log n$ where p is the number of parameters and n is sample size.

15
Multivariate Models

In this chapter we revisit the Multinomial model and the multivariate Normal, as we will need them in future chapters.

Let us first review some notation from linear algebra. Recall that if x and y are vectors then $x^T y = \sum_j x_j y_j$. If A is a matrix then $\det(A)$ denotes the determinant of A, A^T denotes the transpose of A and A^- denotes the the inverse of A (if the inverse exists). The trace of a square matrix A – denoted by $\text{tr}(A)$ – is the sum of its diagonal elements. The trace satisfies $\text{tr}(AB) = \text{tr}(BA)$ and $\text{tr}(A) + \text{tr}(B)$. Also, $\text{tr}(a) = a$ if a is a scalar (i.e. a real number). A matrix is positive definite if $x^T \Sigma x > 0$ for all non-zero vectors x. If a matrix Σ is symmetric and positive definite, there exists a matrix $\Sigma^{1/2}$ – called the square root of Σ – with the following properties:
1. $\Sigma^{1/2}$ is symmetric;
2. $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$;
3. $\Sigma^{1/2}\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma^{1/2} = I$ where $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$.

15.1 Random Vectors

Multivariate models 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

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_k) \end{pmatrix}. \tag{15.1}$$

The covariance matrix Σ is defined to be

$$\Sigma = \mathbb{V}(X) = \begin{bmatrix} \mathbb{V}(X_1) & \mathsf{Cov}(X_1, X_2) & \cdots & \mathsf{Cov}(X_1, X_k) \\ \mathsf{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathsf{Cov}(X_2, X_k) \\ \vdots & \vdots & \vdots & \vdots \\ \mathsf{Cov}(X_k, X_1) & \mathsf{Cov}(X_k, X_2) & \cdots & \mathbb{V}(X_k) \end{bmatrix}. \tag{15.2}$$

This is also called the variance matrix or the variance-covariance matrix.

Theorem **15.1** *Let a be a vector of length k and let X be a random vector of the same length 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 with k columns then $\mathbb{E}(AX) = A\mu$ and $\mathbb{V}(AX) = A\Sigma A^T$.*

Now suppose we have a random sample of n vectors:

$$\begin{pmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{k1} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \\ \vdots \\ X_{k2} \end{pmatrix}, \ \ldots, \begin{pmatrix} X_{1n} \\ X_{2n} \\ \vdots \\ X_{kn} \end{pmatrix}. \tag{15.3}$$

The sample mean $\overline{X}$ is a vector defined by

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

where $\overline{X}_i = n^{-1}\sum_{j=1}^n X_{ij}$. The sample variance matrix, also called the covariance matrix or the variance-covariance matrix, is

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{12} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ s_{1k} & s_{2k} & \cdots & s_{kk} \end{bmatrix} \tag{15.4}$$

where

$$s_{ab} = \frac{1}{n-1}\sum_{j=1}^n (X_{aj} - \overline{X}_a)(X_{bj} - \overline{X}_b).$$

It follows that $\mathbb{E}(\overline{X}) = \mu$. and $\mathbb{E}(S) = \Sigma$.

15.2 Estimating the Correlation

Consider n data points from a bivariate distribution:

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

Recall that the correlation between X_1 and X_2 is

$$\rho = \frac{\mathbb{E}((X_1 - \mu_1)(X_2 - \mu_2))}{\sigma_1 \sigma_2}. \tag{15.5}$$

The sample correlation (the plug-in estimator) is

$$\widehat{\rho} = \frac{\sum_{i=1}^n (X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{s_1 s_2}. \tag{15.6}$$

We can construct a confidence interval for ρ by applying the delta method as usual. However, it turns out that we get a more accurate confidence interval by first constructing a confidence interval for a function $\theta = f(\rho)$ and then applying the inverse function f^{-1}. The method, due to Fisher, is as follows. Define

$$f(r) = \frac{1}{2}\Big(\log(1+r) - \log(1-r)\Big)$$

akd let $\theta = f(\rho)$. The inverse of r is

$$g(z) \equiv f^{-1}(z) = \frac{e^{2z} - 1}{e^{2z} + 1}.$$

Now do the following steps:

Approximate Confidence Interval for The Correlation

1. Compute
$$\widehat{\theta} = f(\widehat{\rho}) = \frac{1}{2}\Big(\log(1 + \widehat{\rho}) - \log(1 - \widehat{\rho})\Big).$$
2. Compute the approximate standard error of $\widehat{\theta}$ which can be shown to be
$$\widehat{\mathsf{se}}\,(\widehat{\theta}) = \frac{1}{\sqrt{n-3}}.$$
3. An approximate $1 - \alpha$ confidence interval for $\theta = f(\rho)$ is
$$(a, b) \equiv \left(\widehat{\theta} - \frac{z_{\alpha/2}}{\sqrt{n-3}}, \widehat{\theta} + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right).$$
4. Apply the inverse transformation $f^{-1}(z)$ to get a confidence interval for ρ:
$$\left(\frac{e^{2a} - 1}{e^{2a} + 1}, \frac{e^{2b} - 1}{e^{2b} + 1}\right).$$

15.3 Multinomial

Let us now review the Multinomial distribution. 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 appeared. Hence, $n = \sum_{j=1}^k X_j$. We say that X has a Multinomial (n,p) distribution. The probability function is

$$f(x;\, p) = \binom{n}{x_1 \ldots x_k} p_1^{x_1} \cdots p_k^{x_k}$$

where

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

Theorem **15.2** *Let $X \sim$ Multinomial(n, p). Then the marginal distribution of X_j is $X_j \sim$ Binomial(n, p_j). The mean and variance of X are*

$$\mathbb{E}(X) = \begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix}$$

and

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

Proof. That $X_j \sim$ Binomial(n, p_j) follows easily. Hence, $\mathbb{E}(X_j) = np_j$ and $\mathbb{V}(X_j) = np_j(1-p_j)$. To compute $\mathsf{Cov}(X_i, X_j)$ we proceed as follows. Notice that $X_i + X_j \sim$ Binomial$(n, p_i + p_j)$ and so $\mathbb{V}(X_i + X_j) = n(p_i + p_j)(1 - p_i - p_j)$. On the other hand,

$$\begin{aligned} \mathbb{V}(X_i + X_j) &= \mathbb{V}(X_i) + \mathbb{V}(X_j) + 2\mathsf{Cov}(X_i, X_j) \\ &= np_i(1-p_i) + np_j(1-p_j) + 2\mathsf{Cov}(X_i, X_j). \end{aligned}$$

Equating this last expression with $n(p_i + p_j)(1 - p_i - p_j)$ implies that $\mathsf{Cov}(X_i, X_j) = -np_ip_j$. ■

Theorem **15.3** *The maximum likelihood estimator of p is*

$$\widehat{p} = \begin{pmatrix} \widehat{p}_1 \\ \vdots \\ \widehat{p}_k \end{pmatrix} = \begin{pmatrix} \frac{X_1}{n} \\ \vdots \\ \frac{X_k}{n} \end{pmatrix} = \frac{X}{n}.$$

PROOF. The log-likelihood (ignoring the constant) is

$$\ell(p) = \sum_{j=1}^{k} X_j \log p_j.$$

When we maximize ℓ we have to be careful since we must enforce the constraint that $\sum_j p_j = 1$. We use the method of Lagrange multipliers and instead maximize

$$A(p) = \sum_{j=1}^{k} X_j \log p_j + \lambda\Big(\sum_j p_j - 1\Big).$$

Now

$$\frac{\partial A(p)}{\partial p_j} = \frac{X_j}{p_j} + \lambda.$$

Setting $\frac{\partial A(p)}{\partial p_j} = 0$ yields $\widehat{p}_j = -X_j/\lambda$. Since $\sum_j \widehat{p}_j = 1$ we see that $\lambda = -n$ and hence $\widehat{p}_j = X_j/n$ as claimed. ■

Next we would like to know the variability of the MLE . We can either compute the variance matrix of $\widehat{p}$ directly or we can approximate the variability of the mle by computing the Fisher information matrix. These two approaches give the same answer in this case. The direct approach is easy: $\mathbb{V}(\widehat{p}) = \mathbb{V}(X/n) = n^{-2}\mathbb{V}(X)$ and so

$$\mathbb{V}(\widehat{p}) = \frac{1}{n}\begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \vdots & \vdots & \vdots \\ -p_1p_k & -p_2p_k & \cdots & p_k(1-p_k) \end{pmatrix}.$$

15.4 Multivariate Normal

Let us recall how the multivariate Normal distribution is defined. 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

$$f(z) = \frac{1}{(2\pi)^{k/2}} \exp\left\{ -\frac{1}{2} \sum_{j=1}^{k} z_j^2 \right\} = \frac{1}{(2\pi)^{k/2}} \exp\left\{ -\frac{1}{2} z^T z \right\}. \tag{15.7}$$

The variance matrix of Z is the identity matrix I. We write $Z \sim N(0, I)$ where it is understood that 0 denotes a vector of k zeroes. We say that Z has a standard multivariate Normal distribution.

More generally, a vector X has a multivariate Normal distribution, denoted by $X \sim N(\mu, \Sigma)$, if its density is

$$f(x;\, \mu, \Sigma) = \frac{1}{(2\pi)^{k/2} \det(\Sigma)^{1/2}} \exp\left\{ -\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right\} \tag{15.8}$$

where $\det(\cdot)$ denotes the determinant of a matrix, μ is a vector of length k and Σ is a $k \times k$ symmetric, positive definite matrix. Then $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = \Sigma$. Setting $\mu = 0$ and $\Sigma = I$ gives back the standard Normal.

Theorem **15.4** *The following properties hold:*

1. *If $Z \sim N(0,1)$ and $X = \mu + \Sigma^{1/2} Z$ then $X \sim N(\mu, \Sigma)$.*
2. *If $X \sim N(\mu, \Sigma)$, then $\Sigma^{-1/2}(X - \mu) \sim N(0,1)$.*
3. *If $X \sim N(\mu, \Sigma)$ a is a vector of the same length as X, then $a^T X \sim N(a^T \mu, a^T \Sigma a)$.*

4. Let

$$V = (X - \mu)^T \Sigma^{-1} (X - \mu).$$

Then $V \sim \chi^2_k$.

Suppose we partition a random Normal vector X into two parts $X = (X_a, X_b)$ We can similarly partition the the mean $\mu = (\mu_a, \mu_b)$ and the variance

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

Theorem **15.5** *Let* $X \sim N(\mu, \Sigma)$. *Then:*
(1) The marginal distribition of X_a *is* $X_a \sim N(\mu_a, \Sigma_{aa})$.
(2) The conditional distribition of X_b *given* $X_a = x_a$ *is*

$$X_b | X_a = x_a \sim N(\mu(x_a), \Sigma(x_a))$$

where

$$\mu(x_a) = \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a) \tag{15.9}$$
$$\Sigma(x_a) = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}. \tag{15.10}$$

Theorem **15.6** *Given a random sample of size* n *from a* $N(\mu, \Sigma)$, *the log-likelihood is (up to a constant not depending on* μ *or* Σ*) is given by*

$$\ell(\mu, \Sigma) = -\frac{n}{2}(\overline{X} - \mu)^T \Sigma^{-1} (\overline{X} - \mu) - \frac{n}{2} tr(\Sigma^{-1} S) - \frac{n}{2} \log \det(\Sigma).$$

The MLE *is*

$$\widehat{\mu} = \overline{X} \quad \text{and} \quad \widehat{\Sigma} = \left(\frac{n-1}{n}\right) S. \tag{15.11}$$

15.5 Appendix

Proof of Theorem 15.6. Denote the i^{th} random vector by X^i.
The log-likelihood is

$$\ell(\mu, \Sigma) = \sum_i f(X^i;\, \mu, \Sigma) = -\frac{kn}{2}\log(2\pi) - \frac{n}{2}\log\det(\Sigma) - \frac{1}{2}\sum_i (X^i - \mu)^T \Sigma^{-1}(X^i - \mu).$$

Now,

$$\begin{aligned}
\sum_i (X^i - \mu)^T \Sigma^{-1}(X^i - \mu) &= \sum_i [(X^i - \overline{X}) + (\overline{X} - \mu)]^T \Sigma^{-1}[(X^i - \overline{X}) + (\overline{X} - \mu)] \\
&= \sum_i \left[(X^i - \overline{X})^T \Sigma^{-1}(X^i - \overline{X})\right] + n(\overline{X} - \mu)^T \Sigma^{-1}(\overline{X} - \mu)
\end{aligned}$$

since $\sum_i (X^i - \overline{X})\Sigma^{-1}(\overline{X} - \mu) = 0$. Also, notice that $(X^i - \mu)^T \Sigma^{-1}(X^i - \mu)$ is a scalar, so

$$\begin{aligned}
\sum_i (X^i - \mu)^T \Sigma^{-1}(X^i - \mu) &= \sum_i tr\left[(X^i - \mu)^T \Sigma^{-1}(X^i - \mu)\right] \\
&= \sum_i tr\left[\Sigma^{-1}(X^i - \mu)(X^i - \mu)^T\right] \\
&= tr\left[\Sigma^{-1}\sum_i (X^i - \mu)(X^i - \mu)^T\right] \\
&= n\ tr\left[\Sigma^{-1} S\right]
\end{aligned}$$

and the conclusion follows.

15.6 Exercises

1. Prove Theorem 15.1.
2. Find the Fisher information matrix for the MLE of a Multinomial.
3. Prove Theorem 15.5.
4. (Computer Experiment. Write a function to generate `nsim` observations from a Multinomial(n, p) distribution.

5. (Computer Experiment. Write a function to generate `nsim` observations from a Multivariate normal with given mean μ and covariance matrix Σ.

6. (Computer Experiment. Generate 1000 random vectors from a $N(\mu, \Sigma)$ distribution where

$$\mu = \begin{pmatrix} 3 \\ 8 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 2 & 6 \\ 6 & 2 \end{pmatrix}.$$

   Plot the simulation as a scatterplot. Find the distribution of $X_2|X_1 = x_1$ using theorem 15.5. In particular, what is the formula for $\mathbb{E}(X_2|X_1 = x_1)$? Plot $\mathbb{E}(X_2|X_1 = x_1)$ on your scatterplot. Find the correlation ρ between X_1 and X_2. Compare this with the sample correlations from your simulation. Find a 95 per cent confidence interval for ρ. Estimate the covariance matrix Σ.

7. Generate 100 random vectors from a multivariate Normal with mean $(0, 2)^T$ and variance

$$\begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix}.$$

   Find a 95 per cent confidence interval for the correlation ρ. What is the true value of ρ?

16
Inference about Independence

In this chapter we address the following questions:
(1) How do we test if two random variables are independent?
(2) How do we estimate the strength of dependence between two random variables?

When Y and Z are not independent, we say that they are **dependent** or **associated** or **related.** If Y and Z are associated, it does not imply that Y causes Z or that Z causes Y. If Y does cause Z then changing Y will change the distribution of Z, otherwise it will not change the distribution of Z.

Recall that we write $Y \amalg Z$ to mean that Y and Z are independent.

For example, quitting smoking Y will reduce your probability of heart disease Z. In this case Y does cause Z. As another example, owning a TV Y is associated with having a lower incidence of starvation Z. This is because if you own a TV you are less likely to live in an impoverished nation. But giving a starving person will not cause them to stop being hungry. In this case, Y and Z are associated but the relationship is not causal.

We defer detailed discussion of the important question of causation until a later chapter.

16.1 Two Binary Variables

Suppose that Y and Z are both binary. Consider a data set $(Y_1, Z_1), \ldots, (Y_n, Z_n)$. Represent the data as a two-by-two table:

	$Y=0$	$Y=1$	
$Z=0$	X_{00}	X_{01}	$X_{0\cdot}$
$Z=1$	X_{10}	X_{11}	$X_{1\cdot}$
	$X_{\cdot 0}$	$X_{\cdot 1}$	$n = X_{\cdot\cdot}$

where the X_{ij} represent counts:

$$X_{ij} = \text{number of observations for which } Y = i \text{ and } Z = j.$$

The dotted subscripts denote sums. For example, $X_{i\cdot} = \sum_j X_{ij}$. This is a convention we use throughout the remainder of the book. Denote the corresponding probabilities by:

	$Y=0$	$Y=1$	
$Z=0$	p_{00}	p_{01}	$p_{0\cdot}$
$Z=1$	p_{10}	p_{11}	$p_{1\cdot}$
	$p_{\cdot 0}$	$p_{\cdot 1}$	1

where $p_{ij} = \mathbb{P}(Z = i, Y = j)$. Let $X = (X_{00}, X_{01}, X_{10}, X_{11})$ denote the vector of counts. Then $X \sim \text{Multinomial}(n, p)$ where $p = (p_{00}, p_{01}, p_{10}, p_{11})$. It is now convenient to introduce two new parameters.

Definition 16.1 *The* **odds ratio** *is defined to be*

$$\psi = \frac{p_{00}p_{11}}{p_{01}p_{10}}. \tag{16.1}$$

The **log odds ratio** *is defined to be*

$$\gamma = \log(\psi). \tag{16.2}$$

Theorem **16.2** *The following statements are equivalent:*

1. $Y \amalg Z$.
2. $\psi = 1$.
3. $\gamma = 0$.
4. *For* $i, j \in \{0, 1\}$,
$$p_{ij} = p_{i\cdot}p_{\cdot j}. \tag{16.3}$$

Now consider testing

$$H_0 : \ Y \amalg Z \ \text{ versus } H_1 : \ Y \not\amalg Z.$$

First we consider the likelihood ratio test. Under H_1, $X \sim$ Multinomial(n, p) and the MLE is $\widehat{p} = X/n$. Under H_0, we again have that $X \sim$ Multinomial(n, p) but p is subject to the constraint

$$p_{ij} = p_{i\cdot}p_{\cdot j}, \quad j = 0, 1.$$

This leads to the following test.

Theorem **16.3 (Likelihood Ratio Test for Independence in a 2-by-2 table)** *Let*

$$T = 2\sum_{i=0}^{1}\sum_{j=0}^{2} X_{ij} \log\left(\frac{X_{ij}X_{\cdot\cdot}}{X_{i\cdot}X_{\cdot j}}\right). \tag{16.4}$$

Under H_0, $T \rightsquigarrow \chi^2_1$. *Thus, an approximate level* α *test is obtained by rejecting* H_0 *then* $T > \chi^2_{1,\alpha}$.

Another popular test for independence is Pearson's χ^2 test.

Theorem **16.4 (Pearson's χ^2 test for Independence in a 2-by-2 table)** *Let*

$$U = \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{(X_{ij} - E_{ij})^2}{E_{ij}} \tag{16.5}$$

where

$$E_{ij} = \frac{X_{i\cdot}X_{\cdot j}}{n}.$$

Under H_0, $U \rightsquigarrow \chi^2_1$. Thus, an approximate level α test is obtained by rejecting H_0 then $U > \chi^2_{1,\alpha}$.

Here is the intuition for the Pearson test. Under H_0, $p_{ij} = p_{i\cdot}p_{\cdot j}$, so the maximum likelihood estimator of p_{ij} under H_0 is

$$\widehat{p}_{ij} = \widehat{p}_{i\cdot}\widehat{p}_{\cdot j} = \frac{X_{i\cdot}}{n}\frac{X_{\cdot j}}{n}.$$

Thus, the expected number of observations in the (i,j) cell is

$$E_{ij} = n\widehat{p}_{ij} = \frac{X_{i\cdot}X_{\cdot j}}{n}.$$

The statistic U compares the observed and expected counts.

Example 16.5 *The following data from Johnson and Johnson (1972) relate tonsillectomy and Hodgkins disease. (The data are actually from a case-control study; we postpone discussion of this point until the next section.)*

	Hodgkins Disease	*No Disease*	
Tonsillectomy	*90*	*165*	*255*
No Tonsillectomy	*84*	*307*	*391*
Total	*174*	*472*	*646*

We would like to know if tonsillectomy is related to Hodgkins disease. The likelihood ratio statistic is $T = 14.75$ and the p-value is $\mathbb{P}(\chi^2_1 > 14.75) = .0001$. The χ^2 statistic is $U = 14.96$

and the p-value is $\mathbb{P}(\chi^2_1 > 14.96) = .0001$. *We reject the null hypothesis of independence and conclude that tonsillectomy is associated with Hodgkins disease. This does not mean that tonsillectomies cause Hodgkins disease. Suppose, for example, that doctors gave tonsillectomies to the most seriosuly ill patients. Then the assocation between tonsillectomies and Hodgkins disease may be due to the fact that those with tonsillectomies were the most ill patients and hence more likely to have a serious disease.*

We can also estimate the strength of dependence by estimating the odds ratio ψ and the log-odds ratio γ.

Theorem **16.6** *The* MLE *'s of* ψ *and* γ *are*

$$\widehat{\psi} = \frac{X_{00}X_{11}}{X_{01}X_{10}}, \quad \widehat{\gamma} = \log \widehat{\psi}. \tag{16.6}$$

The asymptotic stanrad errors (computed from the delta method) are

$$\widehat{\mathsf{se}}\,(\widehat{\gamma}) = \sqrt{\frac{1}{X_{00}} + \frac{1}{X_{01}} + \frac{1}{X_{10}} + \frac{1}{X_{11}}} \tag{16.7}$$

$$\widehat{\mathsf{se}}\,(\widehat{\psi}) = \widehat{\psi}\widehat{\mathsf{se}}\,(\widehat{\gamma}). \tag{16.8}$$

Remark 16.7 *For small sample sizes,* $\widehat{\psi}$ *and* $\widehat{\gamma}$ *can have a very large variance. In this case, we often use the modified estimator*

$$\widehat{\psi} = \frac{\left(X_{00} + \frac{1}{2}\right)\left(X_{11} + \frac{1}{2}\right)}{\left(X_{01} + \frac{1}{2}\right)\left(X_{10} + \frac{1}{2}\right)}. \tag{16.9}$$

Yet another test for indepdnence is the Wald test for $\gamma = 0$ given by $W = (\widehat{\gamma} - 0)/\widehat{\mathsf{se}}\,(\widehat{\gamma})$. A $1-\alpha$ confidence interval for γ is $\widehat{\gamma} \pm z_{\alpha/2}\widehat{\mathsf{se}}\,(\widehat{\gamma})$. A $1-\alpha$ confidence interval for ψ can be obtained in two ways. First, we could use $\widehat{\psi} \pm z_{\alpha/2}\widehat{\mathsf{se}}\,(\widehat{\psi})$. Second, since $\psi = e^{\gamma}$ we could use

$$\exp\left\{\widehat{\gamma} \pm z_{\alpha/2}\widehat{\mathsf{se}}\,(\widehat{\gamma})\right\}. \tag{16.10}$$

This second method is usually more accurate.

Example 16.8 *In the previous example,*

$$\widehat{\psi} = \frac{90 \times 307}{165 \times 84} = 1.99$$

and

$$\widehat{\gamma} = \log(1.99) = .69.$$

So tonsillectomy patients were twice as likely to have Hodgkins disease. The standard error of $\widehat{\gamma}$ *is*

$$\sqrt{\frac{1}{90} + \frac{1}{84} + \frac{1}{165} + \frac{1}{307}} = .18.$$

The Wald statistic is $W = .69/.18 = 3.84$ *whose p-value is* $\mathbb{P}(|Z| > 3.84) = .0001$*, the same as the other tests. A 95 per cent confidence interval for* γ *is* $\widehat{\gamma} \pm 2(.18) = (.33, 1.05)$*. A 95 per cent confidence interval for* ψ *is* $(e^{.33}, e^{1.05}) = (1.39, 2.86)$*.*

16.2 Interpreting The Odds Ratios

Suppose event A as probability $P(A)$. The odds of A are defined as

$$\text{odds}(A) = \frac{P(A)}{1 - P(A)}.$$

It follows that

$$\mathbb{P}(A) = \frac{\text{odds}(A)}{1 + \text{odds}(A)}.$$

Let E be the event that someone is exposed to something (smoking, radiation, etc) and let D be the event that they get a disease. The odds of getting the disease given that you are exposed are

$$\text{odds}(D|E) = \frac{\mathbb{P}(D|E)}{1 - \mathbb{P}(D|E)}$$

and the odds of getting the disease given that you are not exposed are

$$\text{odds}(D|E^c) = \frac{\mathbb{P}(D|E^c)}{1 - \mathbb{P}(D|E^c)}.$$

The *odds ratio* is defined to be

$$\psi = \frac{\text{odds}(D|E)}{\text{odds}(D|E^c)}.$$

If $\psi = 1$ then disease probability is the same for exposed and unexposed. This implies that these events are independent. Recall that the log-odds ratio is defined as $\gamma = \log(\psi)$. Independence corresponds to $\gamma = 0$.

Consider this table of probabilities:

	D^c	D	
E^c	p_{00}	p_{01}	$p_{0\cdot}$
E	p_{10}	p_{11}	$p_{1\cdot}$
	$p_{\cdot 0}$	$p_{\cdot 1}$	1

Denote the data by

	D^c	D	
E^c	X_{00}	X_{01}	$X_{0\cdot}$
E	X_{10}	X_{11}	$X_{1\cdot}$
	$X_{\cdot 0}$	$X_{\cdot 1}$	$X_{\cdot\cdot}$

Now

$$\mathbb{P}(D|E) = \frac{p_{11}}{p_{10}+p_{11}} \quad \text{and} \quad \mathbb{P}(D|E^c) = \frac{p_{01}}{p_{00}+p_{01}}$$

and so

$$\text{odds}(D|E) = \frac{p_{11}}{p_{10}} \quad \text{and} \quad \text{odds}(D|E^c) = \frac{p_{01}}{p_{00}}$$

and therefore,

$$\psi = \frac{p_{11}p_{00}}{p_{01}p_{10}}.$$

To estimate the parameters, we have to first consider how the data were collected. There are three methods.

Multinomial Sampling. We draw a sample from the population and, for each person, record their exposure and disease status. In this case, $X = (X_{00}, X_{01}, X_{10}, X_{11}) \sim \text{Multinomial}(n, p)$.

We then estimate the probabilities in the table by $\widehat{p}_{ij} = X_{ij}/n$ and

$$\widehat{\psi} = \frac{\widehat{p}_{11}\widehat{p}_{00}}{\widehat{p}_{01}\widehat{p}_{10}} = \frac{X_{11}X_{00}}{X_{01}X_{10}}.$$

PROSPECTIVE SAMPLING. (COHORT SAMPLING). We get some exposed and unexposed people and count the number with disease in each group. Thus,

$$\begin{aligned} X_{01} &\sim \text{Binomial}(X_{0\cdot}, P(D|E^c)) \\ X_{11} &\sim \text{Binomial}(X_{1\cdot}, P(D|E)). \end{aligned}$$

We should really write $x_{0\cdot}$ and $x_{1\cdot}$ instead of $X_{0\cdot}$ and $X_{1\cdot}$ since in this case, these are fixed not random but for notational simplicity I'll keep using capital letters. We can estimate $P(D|E)$ and $P(D|E^c)$ but we cannot estimate all the probabilities in the table. Still, we can estimate ψ since ψ is a function of $P(D|E)$ and $P(D|E^c)$. Now

$$\widehat{P}(D|E) = \frac{X_{11}}{X_{1\cdot}}$$

and

$$\widehat{P}(D|E^c) = \frac{X_{01}}{X_{0\cdot}}.$$

Thus,

$$\widehat{\psi} = \frac{X_{11}X_{00}}{X_{01}X_{10}}$$

just as before.

CASE-CONTROL (RETROSPECTIVE) SAMPLING. Here we get some diseased and non-diseased people and we observe how many are exposed. This is much more efficient if the disease is rare. Hence,

$$\begin{aligned} X_{10} &\sim \text{Binomial}(X_{\cdot 0}, P(E|D^c)) \\ X_{11} &\sim \text{Binomial}(X_{\cdot 1}, P(E|D)). \end{aligned}$$

From these data we can estimate $\mathbb{P}(E|D)$ and $\mathbb{P}(E|D^c)$. Surprisingly, we can also still estimate ψ. To understand why, note

that

$$\mathbb{P}(E|D) = \frac{p_{11}}{p_{01}+p_{11}}, \quad 1-\mathbb{P}(E|D) = \frac{p_{01}}{p_{01}+p_{11}}, \quad \text{odds}(E|D) = \frac{p_{11}}{p_{01}}.$$

By a similar argument,

$$\text{odds}(E|D^c) = \frac{p_{10}}{p_{00}}.$$

Hence,

$$\frac{\text{odds}(E|D)}{\text{odds}(E|D^c)} = \frac{p_{11}p_{00}}{p_{01}p_{10}} = \psi.$$

From the data, we form the following estimates:

$$\widehat{P}(E|D) = \frac{X_{11}}{X_{\cdot 1}}, \quad 1-\widehat{P}(E|D) = \frac{X_{01}}{X_{\cdot 1}}, \quad \widehat{\text{odds}}(E|D) = \frac{X_{11}}{X_{01}}, \quad \widehat{\text{odds}}(E|D^c) = \frac{X_{10}}{X_{00}}.$$

Therefore,

$$\widehat{\psi} = \frac{X_{00}X_{11}}{X_{01}X_{10}}.$$

So in all three data collection methods, the estimate of ψ turns out to be the same.

It is tempting to try to estimate $\mathbb{P}(D|E) - \mathbb{P}(D|E^c)$. In a case-control design, this quantity is not estimable. To see this, we apply Bayes' theorem to get

$$\mathbb{P}(D|E) - \mathbb{P}(D|E^c) = \frac{\mathbb{P}(E|D)\mathbb{P}(D)}{\mathbb{P}(E)} - \frac{\mathbb{P}(E^c|D)\mathbb{P}(D)}{\mathbb{P}(E^c)}.$$

Because of the way we obtained the data, $\mathbb{P}(D)$ is not estimable from the data. However, we can estimate $\xi = \mathbb{P}(D|E)/\mathbb{P}(D|E^c)$, which is called the **relative risk**, under the **rare disease assumption.**

Theorem **16.9** *Let* $\xi = \mathbb{P}(D|E)/\mathbb{P}(D|E^c)$. *Then*

$$\frac{\psi}{\xi} \to 1$$

as $\mathbb{P}(D) \to 0$.

Thus, under the rare disease assumption, the relative risk is approximately the same as the odds ratio and, as we have seen, we can estimate the ods ratio.

In a randomized experiment, we can interpret a strong association, that is $\psi \neq 1$, as a causal relationship. In an observational (non-randomized) study, the association can be due to other unobserved **confounding** variables. We'll discuss causation in more detail later.

16.3 Two Discrete Variables

Now suppose that $Y \in \{1, \ldots, I\}$ and $Z \in \{1, \ldots, J\}$ are two discrete variables. The data can be represented as an $I - by - J$ table of counts:

	$Y=1$	$Y=2$	$\cdots$	$Y=j$	$\cdots$	$Y=J$	
$Z=1$	X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1J}	$X_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Z=i$	X_{i1}	X_{i2}	$\cdots$	X_{ij}	$\cdots$	X_{iJ}	$X_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$Z=I$	X_{I1}	X_{I2}	$\cdots$	X_{Ij}	$\cdots$	X_{IJ}	$X_{I\cdot}$
	$X_{\cdot 1}$	$X_{\cdot 2}$	$\cdots$	$X_{\cdot j}$	$\cdots$	$X_{\cdot J}$	n

Consider testing $H_0 : Y \amalg Z$ versus $H_1 : Y \not\amalg Z$.

Theorem **16.10** *Let*

$$T = 2\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}\log\left(\frac{X_{ij}X_{\cdot\cdot}}{X_{i\cdot}X_{\cdot j}}\right). \tag{16.11}$$

The limiting distribution of T under the null hypothesis of independence is χ^2_ν where $\nu = (I-1)(J-1)$. Pearson's χ^2 test statistic is

$$U = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}. \tag{16.12}$$

Asymptotically, under H_0, U has a χ^2_ν distribution where $\nu = (I-1)(J-1)$.

Example 16.11 *These data are from a study by Hancock et al (1979). Patients with Hodkins disease are classified by their response to treatment and by histological type.*

Type	*Positive Response*	*Partial Response*	*No Response*	
LP	*74*	*18*	*12*	*104*
NS	*68*	*16*	*12*	*96*
MC	*154*	*54*	*58*	*266*
LD	*18*	*10*	*44*	*72*

The χ^2 test statistic is 75.89 with $2\times 3 = 6$ degrees of freedom. The p-value is $\mathbb{P}(\chi^2_6 > 75.89) \approx 0$. The likelihood ratio test statistic is 68.30 with $2\times 3 = 6$ degrees of freedom. The p-value is $\mathbb{P}(\chi^2_6 > 68.30) \approx 0$. Thus there is strong evidence that reponse to treatment and histological type are associated.

There are a variety of ways to quantify the strength of dependence between two discrete variables Y and Z. Most of them are not very intuitive. The one we shall use is not standard but is more interpretable.

We define

$$\delta(Y,Z) = \max_{A,B} |\mathbb{P}_{Y,Z}(Y \in A, Z \in B) - \mathbb{P}_Y(Y \in A) - \mathbb{P}_Z(Z \in B)| \tag{16.13}$$

where the maximum is over all pairs of events A and B.

Theorem **16.12** *Properties of δ:*

1. $0 \leq \delta(Y,Z) \leq 1$.
2. $\delta(Y,Z) = 0$ *if and only if* $Y \amalg Z$.
3. *The following identity holds:*

$$\delta(X,Y) = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left|p_{ij} - p_{i\cdot}\, p_{\cdot j}\right|. \tag{16.14}$$

4. *The* MLE *of δ is*

$$\delta(X,Y) = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left|\widehat{p}_{ij} - \widehat{p}_{i\cdot}\,\widehat{p}_{\cdot j}\right| \tag{16.15}$$

where

$$\widehat{p}_{ij} = \frac{X_{ij}}{n}, \quad \widehat{p}_{i\cdot} = \frac{X_{i\cdot}}{n}, \quad \widehat{p}_{\cdot j} = \frac{X_{\cdot j}}{n}.$$

The interpretation of δ is this: if one person makes probability statements assuming independence and another person makes probability statements without assuming independence, their probability statements may differ by as much as δ. Here is a suggested scale for interpreting δ:

$0 < \delta \leq .01$	negligible association
$.01 < \delta \leq .05$	non-negligible association
$.05 < \delta \leq .10$	substantial association
$\delta > .10$	very strong association

A confidence interval for δ can be obtained by bootstrapping. The steps for bootstrapping are:

1. Draw $X^* \sim \text{Multinomial}(n, \widehat{p})$;
2. Compute $\widehat{p}^*_{ij},\ \widehat{p}^*_{i\cdot},\ \widehat{p}^*_{\cdot j}$.
3. Compute δ^*.
4. Repeat.

Now we use any of the methods we learned earlier for constructing bootstrap confidence intervals. However, we should not use a Wald interval in this case. The reason is that if $Y \amalg Z$ then $\delta = 0$ and we are on the boundary of the parameter space. In this case, the Wald method is not valid.

Example 16.13 *Returning to Example 16.11 we find that* $\widehat{\delta} = .11$. *Using a pivotal bootstrap with* $10,000$ *bootstrap samples, a 95 per cent confidence interval for* δ *is (.09,.22). Our conclusion is that the association between histological type and response is substantial.*

16.4 Two Continuous Variables

Now suppose that Y and Z are both continuous. If we assume that the joint distribution of Y and Z is bivariate Normal, then we mesure the dependence between Y and Z by means of the correaltion coefficient ρ. Tests, estimates and confidence intervals for ρ in the Normal case are given in the previous chapter. If we do not assume Normality then we need a nonparametric method for assessing dependence.

Recall that the correaltion is

$$\rho = \frac{\mathbb{E}((X_1 - \mu_1)(X_2 - \mu_2))}{\sigma_1 \sigma_2}.$$

A nonparametric estimator of ρ is the plug-in estimator which is

$$\widehat{\rho} = \frac{\sum_{i=1}^n (X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)}{\sqrt{\sum_{i=1}^n (X_{1i} - \overline{X}_1)^2 \sum_{i=1}^n (X_{2i} - \overline{X}_2)^2}}$$

which is just the sample correlation. A confidence interval can be constructed using the bootstrap. A test for $\rho = 0$ can be based on the Wald test using the bootstrap to estimate the standard error.

The plug-in approach is useful for large samples. For small samples, we measure the correlation using the **Spearman rank correaltion coefficient** $\widehat{\rho}_S$. We simply replace the data by their ranks – ranking each variable separately – then we compute the correaltion coefficient of the ranks. To test the null hypothesis that $\rho_S = 0$ we need the distribution of $\widehat{\rho}_S$ under the null hypothesis. This can be obtained easily by simulation. We fix the ranks of the first variable as $1, 2, \ldots, n$. The ranks of the second variable are chosen at random from the set of $n!$ possible orderings. Then we compute the correlation. This is repeated many times and the resulting distribution $\mathbb{P}_0$ is the null distribution of $\widehat{\rho}_S$. The p-value for the test is $\mathbb{P}_0(|R| > |\widehat{\rho}_S|)$ where R is drawn from the null distribution $\mathbb{P}_0$.

Example 16.14 *The following data (Snedecor and Cochran, 1980, p. 191) are systolic blood pressure X_1 and are diastolic blood pressure X_2 in millimeters:*

X_1	*100*	*105*	*110*	*110*	*120*	*120*	*125*	*130*	*130*	*150*	*170*	*195*
X_2	*65*	*65*	*75*	*70*	*78*	*80*	*75*	*82*	*80*	*90*	*95*	*90*

The sample correlation is $\widehat{\rho} = .88$. The bootstrap standard error is .046 and the Wald test statistic is $.88/.046 = 19.23$. The p-value is near 0 giving strong evidence that the correlation is not 0. The 95 per cent pivotal bootstrap confidence interval is (.78,.94). Because the sample size is small, consider Spearman's rank correlation. In ranking the data, we will use average ranks if there are ties. So if the third and fourth lowest numbers are the same, they each get rank 3.5. The ranks of the data are:

X_1	*1*	*2*	*3.5*	*3.5*	*5.5*	*5.5*	*7*	*8.5*	*8.5*	*10*	*11*	*12*
X_2	*1.5*	*1.5*	*4.5*	*3*	*6*	*7.5*	*4.5*	*9*	*7.5*	*10.5*	*12*	*10.5*

The rank correlation is $\widehat{\rho}_S = .94$ *and the p-value for testing the null hypothesis that there is no correaltion is* $\mathbb{P}_0(|R| > .94) \approx 0$ *which is was obtained by simulation.*

16.5 One Continuous Variable and One Discrete

Suppose that $Y \in \{1, \ldots, I\}$ is discrete and Z is continuous. Let $F_i(z) = \mathbb{P}(Z \leq z | Y = i)$ denote the CDF of Z conditional on $Y = i$.

Theorem **16.15** *When* $Y \in \{1, \ldots, I\}$ *is discrete and* Z *is continuous, then* $Y \amalg Z$ *if and only if* $F_1 = \cdots = F_I$.

It follows from the previous theorem that to test for independence, we need to test

$$H_0 : F_1 = \cdots = F_I \quad \text{versus} \quad H_1 : \text{not } H_0.$$

For simplicity, we consider the case where $I = 2$. To test the null hypothesis that $F_1 = F_2$ we will use the **two sample Kolmogorov-Smirnov test**. Let n_1 denote the nimber of observations for which $Y_i = 1$ and let n_2 denote the nimber of observations for which $Y_i = 2$. Let

$$\widehat{F}_1(z) = \frac{1}{n_1} \sum_{i=1}^{n} I(Z_i \leq z) I(Y_i = 1)$$

and

$$\widehat{F}_2(z) = \frac{1}{n_2} \sum_{i=1}^{n} I(Z_i \leq z) I(Y_i = 2)$$

denote the empirical distribution function of Z given $Y = 1$ and $Y = 2$ respectively. Define the test statistic

$$D = \sup_x |\widehat{F}_1(x) - \widehat{F}_2(x)|.$$

Theorem **16.16** *Let*

$$H(t) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2t^2}.$$

Under the null hypthesis that $F_1 = F_2$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{\frac{n_1 n_2}{n_1+n_2}}D \le t\right) = H(t).$$

It follows from the theorem that an approximate level α test is obtained by rejecting H_0 when

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}D > H^{-1}(1-\alpha).$$

16.6 Bibliographic Remarks

Johnson, S.K. and Johnson, R.E. (1972). New England Journal of Medicine. 287. 1122-1125.

Hancock, B.W. (1979). Clinical Oncology, 5, 283-297.

16.7 Exercises

1. Prove Theorem 16.2.
2. Prove Theorem 16.3.
3. Prove Theorem 16.9.
4. Prove equation (16.14).
5. The New York Times (January 8, 2003, page A12) reported the following data on death sentencing and race, from a study in Maryland:

The data here are an approximate re-creation using the information in the article.

	Death Sentence	No Death Sentence
Black Victim	14	641
White Victim	62	594

Analyze the data using the tools from this Chapter. Interpret the results. Explain why, based only on this information, you can't make causal conclusions. (The authors of the study did use much more information in their full report.)

6. Analyze the data on the variables Age and Financial Status from:

   http://lib.stat.cmu.edu/DASL/Datafiles/montanadat.html

7. Estimate the correlation between temperature and latitude using the data from

   http://lib.stat.cmu.edu/DASL/Datafiles/USTemperatures.html

   Use the correlation coefficient and the Spearman rank correaltion. Provide estimates, tests and confidence intervals.

8. Test whether calcium intake and drop in blood pressure are associated. Use the data in

   http://lib.stat.cmu.edu/DASL/Datafiles/Calcium.html

17

Undirected Graphs and Conditional Independence

Graphical models are a class of multivariate statistical models that are useful for representing independence relations. They are also useful develop parsimonious models for multivariate data.

To see why parsimonious models are useful in the multivariate setting, consider the problem of estimating the joint distribution of several discrete random variables. Two binary variables Y_1 and Y_2 can be represented as a two-by-two table which corresponds to a multinomial with four categories. Similarly, k binary variables $Y_1, \ldots, Y_k$ correspond to a multinomial with $N = 2^k$ categories. When k is even moderately large, $N = 2^k$ will be huge. It can be shown in that case that the MLE is a poor estimator. The reason is that the data are **sparse**: there are not enough data to estimate so many parameters. Graphical models often require fewer parameters and may lead to estimators with smaller risk. There are two main types of graphical models: undirected and directed. Here, we introduce undirected graphs. We'll discuss directed graphs later.

17.1 Conditional Independence

Underlying graphical models is the concept of conditional independence.

Definition 17.1 *Let X, Y and Z be discrete random variables. We say that X and Y are* **conditionally independent given** *Z, written $X \amalg Y \mid Z$, if*

$$\mathbb{P}(X = x, Y = y | Z = z) = \mathbb{P}(X = x | Z = z)\mathbb{P}(Y = y | Z = z) \tag{17.1}$$

for all x, y, z. If X, Y and Z are continuous random variables, we say that X and Y are conditionally independent given Z if

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

for all x, y and z.

Intuitively, this means that, once you know Z, Y provides no extra information about X.

The conditional independence relation satisfies some basic properties.

Theorem **17.2** *The following implications hold:*

$$\begin{aligned}
X \amalg Y \mid Z &\implies Y \amalg X \mid Z \\
X \amalg Y \mid Z \text{ and } U = h(X) &\implies U \amalg Y \mid Z \\
X \amalg Y \mid Z \text{ and } U = h(X) &\implies X \amalg Y \mid (Z, U) \\
X \amalg Y \mid Z \text{ and } X \amalg W | (Y, Z) &\implies X \amalg (W, Y) \mid Z \\
X \amalg Y \mid Z \text{ and } X \amalg Z \mid Y &\implies X \amalg (Y, Z).
\end{aligned}$$

The last property requires the assumption that all events have positive probability; the first four do not.

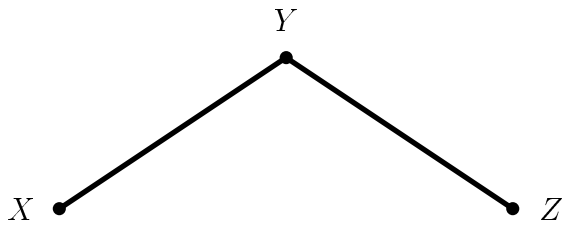


FIGURE 17.1. A graph with vertices $V = \{X, Y, Z\}$. The edge set is $E = \{(X, Y), (Y, Z)\}$.

17.2 Undirected Graphs

An **undirected graph** $\mathcal{G} = (V, E)$ has a finite set V of **vertices (or nodes)** and a set E of **edges (or arcs)** consisting of pairs of vertices. The vertices correspond to random variables $X, Y, Z, \ldots$ and edges are written as unordered pairs. For example, $(X, Y) \in E$ means that X and Y are joined by an edge.

An example of a graph is in Figure 17.1.

Two vertices are **adjacent**, written $X \sim Y$, if there is an edge between them. In Figure 17.1, X and Y are adjacent but X and Z are not adjacent. A sequence $X_0, \ldots, X_n$ is called a **path** if $X_{i-1} \sim X_i$ for each i. In Figure 17.1, X, Y, Z is a path. A graph is **complete** if there is an edge between every pair of vertices. A subset $U \subset V$ of vertices together with their edges is called a **subgraph.**

If A, B and C are three distinct subsets if V, we say that C **separates** A **and** B if every path from a variable in A to a variable in B intersects a variable in C. In Figure 17.2 $\{Y, W\}$ and $\{Z\}$ are separated by $\{X\}$. Also, W and Z are separated by $\{X, Y\}$.

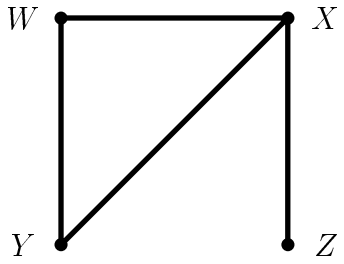


FIGURE 17.2. $\{Y, W\}$ and $\{Z\}$ are separated by $\{X\}$. Also, W and Z are separated by $\{X, Y\}$.

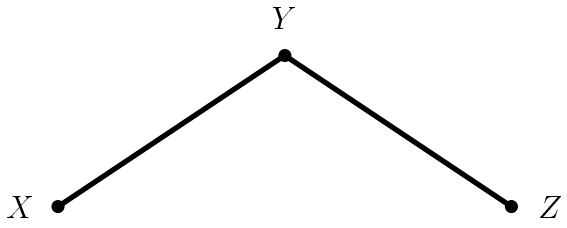


FIGURE 17.3. $X \amalg Z|Y$.

17.3 Probability and Graphs

Let V be a set of random variables with distribution $\mathbb{P}$. Construct a graph with one vertex for each random variable in V. Suppose we omit the edge between a pair of variables if they are independent given the rest of the variables:

$$\text{no edge between } X \text{ and } Y \iff X \amalg Y|\text{rest}$$

where "rest" refers to all the other variables besides X and Y. This type of graph is called a **pairwise Markov graph**. Some examples are shown in Figures 17.3, 17.4 17.6 and 17.5.

The graph encodes a set of pairwise conditional independence relations. These relations imply other conditional independence

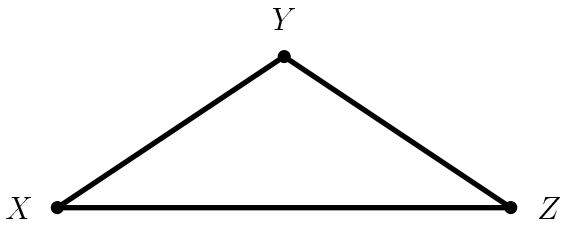


FIGURE 17.4. No implied independence relations.

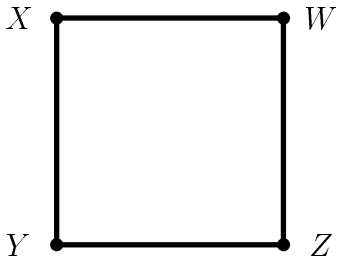


FIGURE 17.5. $X \amalg Z|\{Y, W\}$ and $Y \amalg W|\{X, Z\}$.

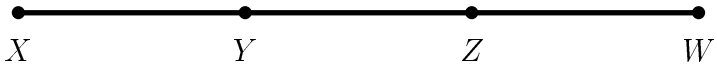


FIGURE 17.6. Pairwise independence implies that $X \amalg Z|\{Y, W\}$. But is $X \amalg Z|Y$?

relations. How can we figure out what they are? Fortunately, we can read these other conditional independence relations directly from the graph as well, as is explained in the next theorem.

Theorem **17.3** *Let $\mathcal{G} = (V, E)$ be a pairwise Markov graph for a distribution $\mathbb{P}$. Let A, B and C be distinct subsets of V such that C separates A and B. Then $A \amalg B | C$.*

Remark 17.4 *If A and B are not connected (i.e. there is no path from A to B) then we may regard A and B as being separated by the empty set. Then Theorem 17.3 implies that $A \amalg B$.*

The independence condition in Theorem 17.3 is called the **global Markov property.** We thus see that the pairwise and global Markov properties are equivalent. Let us state this more precisely. Given a graph $\mathcal{G}$, let $M_{\text{pair}}(\mathcal{G})$ be the set of distributions which satisfy the pairwise Markov property: thus $\mathbb{P} \in M_{\text{pair}}(\mathcal{G})$ if, under $\mathbb{P}$, $X \amalg Y | \text{rest}$ if and only if there is no edge between X and Y. Let $M_{\text{global}}(\mathcal{G})$ be the set of distributions which satisfy the global Markov property: thus $\mathbb{P} \in M_{\text{pair}}(\mathcal{G})$ if, under $\mathbb{P}$, $A \amalg B | C$ if and only if C separates A and B.

Theorem **17.5** *Let $\mathcal{G}$ be a graph. Then, $M_{\text{pair}}(\mathcal{G}) = M_{\text{global}}(\mathcal{G})$.*

This theorem allows us to construct graphs using the simpler pairwise property and then we can deduce other independence relations using the global Markov property. Think how hard this would be to do algebraically. Returning to 17.6, we now see that $X \amalg Z | Y$ and $Y \amalg W | Z$.

Example 17.6 *Figure 17.7 implies that $X \amalg Y$, $X \amalg Z$ and $X \amalg (Y, Z)$.*

Example 17.7 *Figure 17.8 implies that $X \amalg W | (Y, Z)$ and $X \amalg Z | Y$.*

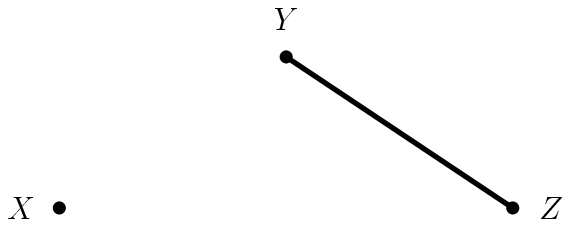


FIGURE 17.7. $X \amalg Y$, $X \amalg Z$ and $X \amalg (Y, Z)$.

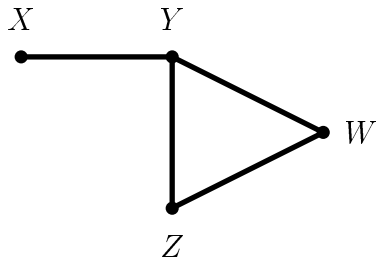


FIGURE 17.8. $X \amalg W|(Y, Z)$ and $X \amalg Z|Y$.

17.4 Fitting Graphs to Data

Given a data set, how do we find a graphical model that fits the data. Some authors have devoted whole books to this subject. We will only treat the discrete case and we will consider a method based on **log-linear models** which are the subject of the next chapter.

17.5 Bibliographic Remarks

Thorough treatments of undirected graphs can be found in Whittaker (1990) and Lauritzen (1996). Some of the exercises below are adapted from Whittaker (1990).

17.6 Exercises

1. Consider random variables (X_1, X_2, X_3). In each of the following cases, draw a graph that has the given independence relations.

   (a) $X_1 \amalg X_3 \mid X_2$.

   (b) $X_1 \amalg X_2 \mid X_3$ and $X_1 \amalg X_3 \mid X_2$.

   (c) $X_1 \amalg X_2 \mid X_3$ and $X_1 \amalg X_3 \mid X_2$ and $X_2 \amalg X_3 \mid X_1$.

2. Consider random variables (X_1, X_2, X_3, X_4). In each of the following cases, draw a graph that has the given independence relations.

   (a) $X_1 \amalg X_3 \mid X_2, X_4$ and $X_1 \amalg X_4 \mid X_2, X_3$ and $X_2 \amalg X_4 \mid X_1, X_3$.

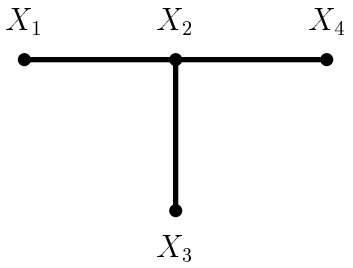


FIGURE 17.9.

FIGURE 17.10.

(b) $X_1 \amalg X_2 \mid X_3, X_4$ and $X_1 \amalg X_3 \mid X_2, X_4$ and $X_2 \amalg X_3 \mid X_1, X_4$.

(c) $X_1 \amalg X_3 \mid X_2, X_4$ and $X_2 \amalg X_4 \mid X_1, X_3$.

3. A conditional independence between a pair of variables is **minimal** if it is not possible to use the Separation Theorem to eliminate any variable from the conditioning set, i.e. from the right hand side of the bar (Whittaker 1990). Write down the minimal conditional independencies from: (a) Figure 17.9; (b) Figure 17.10; (c) Figure 17.11; (d) Figure 17.12.

4. Here are breast cancer data on diagnostic center (X_1), nuclear grade (X_2), and survival (X_3):

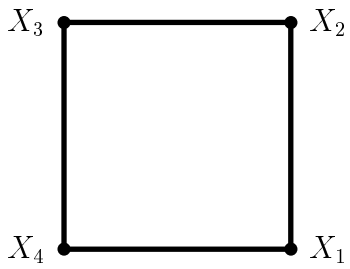


FIGURE 17.11.

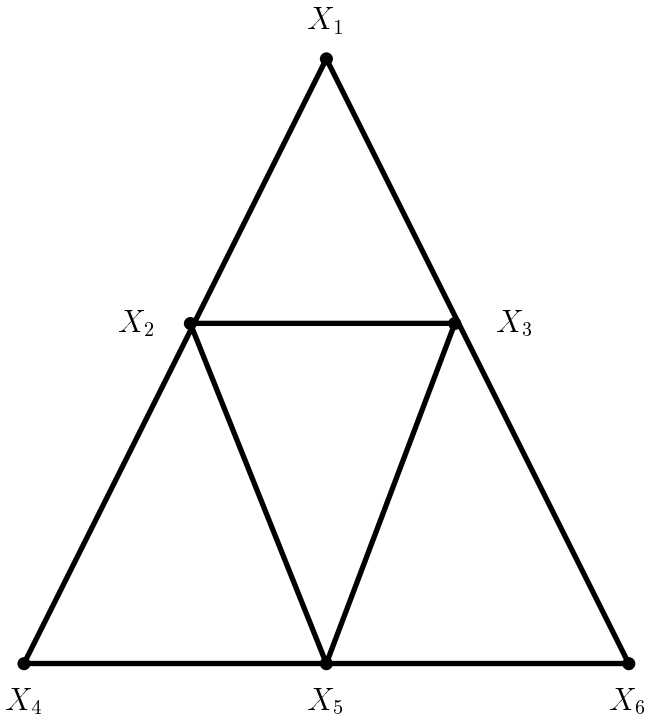


FIGURE 17.12.

	X_2 X_3	malignant died	malignant survived	benign died	benign survived
X_1	Boston	35	59	47	112
	Glamorgan	42	77	26	76

(a) Treat this as a multinomial and find the maximum likelihood estimator.

(b) If someone has a tumour classified as benign at the Glamorgan clinic, what is the estimated probability that they will die? Find the standard error for this estimate.

(c) Test the following hypotheses:

$$X_1 \amalg X_2 | X_3 \quad \text{versus} \quad X_1 \not\amalg X_2 | X_3$$
$$X_1 \amalg X_3 | X_2 \quad \text{versus} \quad X_1 \not\amalg X_3 | X_2$$
$$X_2 \amalg X_3 | X_1 \quad \text{versus} \quad X_2 \not\amalg X_3 | X_1$$

Based on the results of your tests, draw and interpret the resulting graph.

18
Loglinear Models

In this chapter we study **loglinear models** which are useful for modelling multivariate discrete data. There is a strong connection between loglinear models and undirected graphs. Parts of this Chapter draw on the material in Whittaker (1990).

18.1 The Loglinear Model

Let $X = (X_1, \ldots, X_m)$ be a random vector with probability function

$$f(x) = \mathbb{P}(X = x) = \mathbb{P}(X_1 = x_1, \ldots, X_m = x_m)$$

where $x = (x_1, \ldots, x_m)$. Let r_j be the number of values that X_j takes. Without loss of generality, we can assume that $X_j \in \{0, 1, \ldots, r_j - 1\}$. Suppose now that we have n such random vectors. We can think of the data as a sample from a Multinomial with $N = r_1 \times r_2 \times \cdots \times r_m$ categories. The data can be represented as counts in a $r_1 \times r_2 \times \cdots \times r_m$ table. Let $p = (p_1, \ldots, p_N)$ denote the multinomial parameter.

Let $S = \{1, \ldots, m\}$. Given a vector $x = (x_1, \ldots, x_m)$ and a subset $A \subset S$, let $x_A = (x_j : \ j \in A)$. For example, if $A = \{1, 3\}$ then $x_A = (x_1, x_3)$.

Theorem **18.1** *The joint probability function $f(x)$ of a single random vector $X = (X_1, \ldots, X_m)$ can be written as*

$$\log f(x) = \sum_{A \subset S} \psi_A(x) \tag{18.1}$$

where the sum is over all subsets A of $S = \{1, \ldots, m\}$ and the ψ's satisfy the following conditions:

1. *$\psi_\emptyset(x)$ is a constant;*
2. *For every $A \subset S$, $\psi_A(x)$ is only a function of x_A and not the rest of the $x_j's$.*
3. *If $i \in A$ and $x_i = 0$, then $\psi_A(x) = 0$.*

The formula in equation (18.1) is called the **log-linear expansion** of f. Note that this is the probability function for a single draw. Each $\psi_A(x)$ will depend on some unknown parameters β_A. Let $\beta = (\beta_A : \ A \subset S)$ be the set of all these parameters. We will write $f(x) = f(x; \beta)$ when we want to estimate the dependence on the unknown parameters β.

In terms of the multinomial, the parameter space is

$$\mathcal{P} = \Big\{ p = (p_1, \ldots, p_N) : \ p_j \geq 0, \ \sum_{j=1}^{N} p_j = 1 \Big\}.$$

This is an $N - 1$ dimensional space. In the log-linear represenation, the parameter space is

$$\Theta = \Big\{ \beta = (\beta_1, \ldots, \beta_N) : \ \beta = \beta(p), p \in \mathcal{P} \Big\}$$

where $\beta(p)$ is the set of β values associated with p. The set Θ is a $N - 1$ dimensinal surface in $\mathbb{R}^N$. We can always go back and forth betwee the two parameterizations we can write $\beta = \beta(p)$ and $p = p(\beta)$.

Example 18.2 *Let* $X \sim \text{Bernoulli}(p)$ *where* $0 < p < 1$. *We can write the probability mass function for* X *as*

$$f(x) = p^x(1-p)^{1-x} = p_1^x\, p_2^{1-x}$$

for $x = 0, 1$, *where* $p_1 = p$ *and* $p_2 = 1 - p$. *Hence,*

$$\log f(x) = \psi_\emptyset(x) + \psi_1(x)$$

where

$$\begin{aligned} \psi_\emptyset(x) &= \log(p_2) \\ \psi_1(x) &= x \log\left(\frac{p_1}{p_2}\right). \end{aligned}$$

Notice that $\psi_\emptyset(x)$ *is a constant (as a function of* x*) and* $\psi_1(x) = 0$ *when* $x = 0$. *Thus the three conditions of the Theorem hold. The loglinear parameters are*

$$\beta_0 = \log(p_2), \quad \beta_1 = \log\left(\frac{p_1}{p_2}\right).$$

The original, multinomial parameter space is $\mathcal{P} = \{(p_1, p_2) : p_j \geq 0, p_1 + p_2 = 1\}$. *The log-linear parameter space is*

$$\Theta = \{(\beta_0, \beta_1) \in \mathbb{R}^2 : \; e^{\beta_0+\beta_1} + e^{\beta_0} = 1.\}$$

Given (p_1, p_2) *we can solve for* (β_0, β_1). *Conversely, given* (β_0, β_1) *we can solve for* (p_1, p_2). ■

Example 18.3 *Let* $X = (X_1, X_2)$ *where* $X_1 \in \{0, 1\}$ *and* $X_2 \in \{0, 1, 2\}$. *The joint distribution of* n *such random vectors is a multinomial with 6 categories. The multinomial parameters can be written as a 2-by-3 table as follows:*

multinomial		x_2	*0*	*1*	*2*
x_1	*0*		p_{00}	p_{01}	p_{02}
	1		p_{10}	p_{11}	p_{12}

The n data vectors can be summarized as counts:

data		x_2	*0*	*1*	*2*
x_1	*0*		C_{00}	C_{01}	C_{02}
	1		C_{10}	C_{11}	C_{12}

For $x = (x_1, x_2)$, the log-linear expansion takes the form

$$\log f(x) = \psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_{12}(x)$$

where

$$\begin{aligned}
\psi_\emptyset(x) &= \log p_{00} \\
\psi_1(x) &= x_1 \log\left(\frac{p_{10}}{p_{00}}\right) \\
\psi_2(x) &= I(x_2 = 1)\log\left(\frac{p_{01}}{p_{00}}\right) + I(x_2 = 2)\log\left(\frac{p_{02}}{p_{00}}\right) \\
\psi_{12}(x) &= I(x_1 = 1, x_2 = 1)\log\left(\frac{p_{11}p_{00}}{p_{01}p_{10}}\right) + I(x_1 = 1, x_2 = 2)\log\left(\frac{p_{12}p_{00}}{p_{02}p_{10}}\right).
\end{aligned}$$

Convince yourself that the three conditions on the ψ's of the theorem are satisfied. The six parameters of this model are:

$$\begin{array}{lll}
\beta_1 = \log p_{00} & \beta_2 = \log\left(\frac{p_{10}}{p_{00}}\right) & \beta_3 = \log\left(\frac{p_{01}}{p_{00}}\right) \\
\beta_4 = \log\left(\frac{p_{02}}{p_{00}}\right) & \beta_5 = \log\left(\frac{p_{11}p_{00}}{p_{01}p_{10}}\right) & \beta_6 = \log\left(\frac{p_{12}p_{00}}{p_{02}p_{10}}\right).
\end{array}$$

■

The next Theorem gives an easy way to check for conditional independence in a loglinear model.

Theorem 18.4 *Let (X_a, X_b, X_c) be a partition of a vectors $(X_1, \ldots, X_m)$. Then $X_b \amalg X_c | X_a$ if and only if all the ψ-terms in the log-linear expansion that have at least one coordinate in b and one coordinate in c are 0.*

To prove this Theorem, we will use the following Lemma whose proof follows easily from the definition of conditional independence.

Lemma 18.5 *A partition (X_a, X_b, X_c) satisfies $X_b \amalg X_c | X_a$ if and only if $f(x_a, x_b, x_c) = g(x_a, x_b)h(x_a, x_c)$ for some functions g and h*

PROOF. (Theorem 18.4.) Suppose that ψ_t is 0 whenever t has coordinates in b and c. Hence, ψ_t is 0 if $t \not\subset a \bigcup b$ or $t \not\subset a \bigcup c$. Therefore

$$\log f(x) = \sum_{t \subset a \bigcup b} \psi_t(x) + \sum_{t \subset a \bigcup c} \psi_t(x) - \sum_{t \subset a} \psi_t(x).$$

Exponentiating, we see that the joint density is of the form $g(x_a, x_b)h(x_a, x_c)$. By Lemma 18.5, $X_b \amalg X_c | X_a$. The converse follows by reversing the argument. ■

18.2 Graphical Log-Linear Models

A log-linear model is **graphical** if missing terms correspond only to conditional independence constraints.

Definition 18.6 *Let $\log f(x) = \sum_{A \subset S} \psi_A(x)$ be a log-linear model. Then f is* **graphical** *if all ψ-terms are non-zero except for any pair of coordinates not in the edge set for some graph $\mathcal{G}$. In other words, $\psi_A(x) = 0$ if and only if $\{i, j\} \subset A$ and (i, j) is not an edge.*

Here is way to think about the definition above:

> If you can add a term to the model and the graph does not change, then the model is not graphical.

Example 18.7 *Consider the graph in Figure 18.1.*

The graphical log-linear model that corresponds to this graph is

$$\begin{aligned} \log f(x) &= \psi_\emptyset + \psi_1(x) + \psi_2(x) + \psi_3(x) + \psi_4(x) + \psi_5(x) \\ &+ \psi_{12}(x) + \psi_{23}(x) + \psi_{25}(x) + \psi_{34}(x) + \psi_{35}(x) + \psi_{45}(x) + \psi_{235}(x) + \psi_{345}(x). \end{aligned}$$

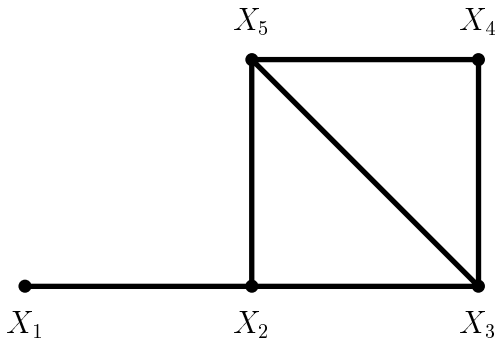


FIGURE 18.1. Graph for Example 18.7.

Let's see why this model is graphical. The edge $(1,5)$ *is missing in the graph. Hence any term containing that pair of indices is omitted from the model. For example,*

$$\psi_{15},\ \psi_{125},\ \psi_{135},\ \psi_{145},\ \psi_{1235},\ \psi_{1245},\ \psi_{1345},\ \psi_{12345}$$

are all omitted. Similarly, the edge $(2,4)$ *is missing and hence*

$$\psi_{24},\ \psi_{124},\ \psi_{234},\ \psi_{245},\ \psi_{1234},\ \psi_{1245},\ \psi_{2345},\ \psi_{12345}$$

are all omitted. There are other missing edges as well. You can check that the model omits all the corresponding ψ *terms. Now consider the model*

$$\begin{aligned}\log f(x) &= \psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_3(x) + \psi_4(x) + \psi_5(x) \\ &+ \psi_{12}(x) + \psi_{23}(x) + \psi_{25}(x) + \psi_{34}(x) + \psi_{35}(x) + \psi_{45}(x).\end{aligned}$$

This is the same model except that the three way interactions were removed. If we draw a graph for this model, we will get the same graph. For example, no ψ *terms contain* $(1,5)$ *so we omit the edge between* X_1 *and* X_5*. But this is not graphical since it has extra terms omitted. The independencies and graphs for the two models are the same but the latter model has other constraints besides conditional independence constraints. This is not a bad*

thing. It just means that if we are only concerned about presence or absence of conditional independences, then we need not consider such a model. The presence of the three-way interaction ψ_{235} *means that the strength of association between* X_2 *and* X_3 *varies as a function of* X_5*. Its absence indicates that this is not so.* ■

18.3 Hierarchical Log-Linear Models

There is a set of log-linear models that is larger than the set of graphical models and that are used quite a bit. These are the hierarchical log-linear models.

Definition 18.8 *A log-linear model is* **hierarchical** *if* $\psi_a = 0$ *and* $a \subset t$ *implies that* $\psi_t = 0$.

Lemma 18.9 *A graphical model is hierarchical but the reverse need not be true.*

Example 18.10 *Let*

$$\log f(x) = \psi_\emptyset(x) + \psi_1(x) + \psi_2(x) + \psi_3(x) + \psi_{12}(x) + \psi_{13}(x).$$

The model is hierarchical; its graph is given in Figure 18.2. The model is graphical because all terms involving (1,3) are omitted. It is also hierarchical. ■

Example 18.11 *Let*

$$\log f(x) = \psi_\emptyset(x)+\psi_1(x)+\psi_2(x)+\psi_3(x)+\psi_{12}(x)+\psi_{13}(x)+\psi_{23}(x).$$

The model is hierarchical. It is not graphical. The graph corresponding to this model is complete; see Figure 18.3. It is not graphical because $\psi_{123}(x) = 0$ *which does not correspond to any pairwise conditional independence.* ■

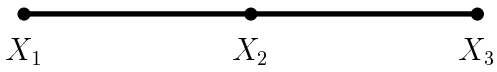


FIGURE 18.2. Graph for Example 18.10.

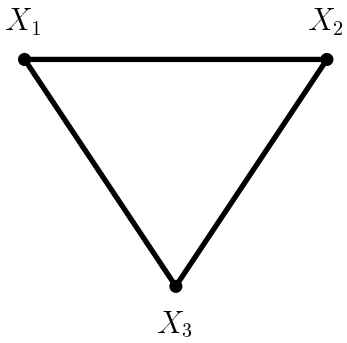


FIGURE 18.3. The graph is complete. The model is hierarchical but not graphical.

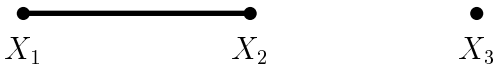


FIGURE 18.4. The model for this graph is not hierarchical.

Example 18.12 *Let*

$$\log f(x) = \psi_\emptyset(x) + \psi_3(x) + \psi_{12}(x).$$

The graph corresponding is in Figure 18.4. This model is not hierarchical since $\psi_2 = 0$ *but* ψ_{12} *is not. Since it is not hierarchical, it is not graphical either.* ■

18.4 Model Generators

Hierarchical models can be written succinctly using **generators**. This is most easily explained by example. Suppose that $X = (X_1, X_2, X_3)$. Then, $M = 1.2 + 1.3$ stands for

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_3 + \psi_{12} + \psi_{13}.$$

The formula $M = 1.2+1.3$ says: "include ψ_{12} and ψ_{13}." We have to also include the lower order terms or it won't be hierarchical. The generator $M = 1.2.3$ is the **saturated** model

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_3 + \psi_{12} + \psi_{13} + \psi_{23} + \psi_{123}.$$

The saturated models corresponds to fitting an unconstrained multinomial. Consider $M = 1 + 2 + 3$ which means

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_3.$$

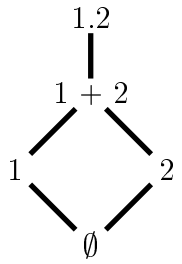


FIGURE 18.5. The lattice with two variables.

This is the mutual independence model. Finally, consider $M = 1.2$ which has log-linear expansion

$$\log f = \psi_\emptyset + \psi_1 + \psi_2 + \psi_{12}.$$

This model makes $X_3|X_2 = x_2, X_1 = x_1$ a uniform distribution.

18.5 Lattices

Hierarchical models can be organized into something called a **lattice**. This is the set of all hierarchical models partially ordered by inclusion. The set of all hierarchical models for two variables can be illustrated as in Figure 18.5.

$M = 1.2$ is the saturated model, $M = 1 + 2$ is the independence model, $M = 1$ is independence plus $X_2|X_1$ is uniform, $M = 2$ is independence plus $X_1|X_2$ is uniform, $M = 0$ is the uniform distribution.

The lattice of trivariate models is shown in figure 18.6.

18.6 Fitting Log-Linear Models to Data

Let β denote all the parameters in a log-linear model M. The loglikelihood for β is

$$\ell(\beta) = \sum_j x_j \log p_j(\beta)$$

saturated (graphical)		1.2.3	
two-way		1.2 + 1.3 + 2.3	
graphical	1.2 + 1.3	1.2 + 2.3	1.3 + 2.3
graphical	1.2 + 3	1.3 + 2	2.3 + 1
mutual independence (graphical)		1 + 2 + 3	
conditional uniform	1.2	1.3	2.3
independence	1 + 2	1 + 3	2 + 3
	1	2	3
uniform		$\emptyset$	

FIGURE 18.6. The lattice of models for three variables.

where the sum is over the cells and $p(\beta)$ denotes the cell probabilities corresponding to β. The MLE $\widehat{\beta}$ generally has to be found numerically. The model with all possible ψ-terms is called the **saturated model.** We can also fit any **sub-model** which corresponds to setting some subset of ψ terms to 0.

Definition 18.13 *For any submodel M, define the* **deviance** $\text{dev}(M)$ *by*

$$\text{dev}(M) = 2(\widehat{\ell}_{sat} - \widehat{\ell}_M)$$

where $\widehat{\ell}_{sat}$ is the log-likelihood of the saturated model evaluated at the MLE *and $\widehat{\ell}_M$ is the log-likelihood of the model M evaluated at its* MLE .

Theorem **18.14** *The deviance is the likelihood ratio test statistic for*

$$H_0 : \text{the model is M} \quad \text{versus} \quad H_1 : \text{the model is not M.}$$

Under H_0, $dev(M) \xrightarrow{d} \chi^2_\nu$ with ν degrees of freedom equal to the difference in the number of parameters between the saturated model and M.

One way to find a good model is to use the deviance to test every sub-model. Every model that is not rejected by this test is then considered a plausible model. However, this is not a good strategy for two reasons. First, we will end up doing many tests which means that there is ample opportunity for making Type I and Type II errors. Second, we will end up using models where we failed to reject H_0. But we might fail to reject H_0 due to low power. The result is that we end up with a bad model just due to low power.

There are many model searching strategies. A common approach is to use some form of *penalized likelihood.* One version of penalized is the AIC that we used in regression. For any model

M define

$$\text{AIC}(M) = -2\Big(\widehat{\ell}(M) - |M|\Big) \tag{18.2}$$

where $|M|$ is the number of parameters. Here is the explanation of where AIC comes from.

Consider a set of models $\{M_1, M_2, \ldots\}$. Let $\widehat{f}_j(x)$ denote the estimated probability function obtained by using the maximum likelihood estimator ofr mode M_j. Thus, $\widehat{f}_j(x) = \widehat{f}(x; \widehat{\beta}_j)$ where $\widehat{\beta}_j$ is the MLE of the set of parameters β_j for model M_j. We will use the loss function $D(f, \widehat{f})$ where

$$D(f, g) = \sum_x f(x) \log\left(\frac{f(x)}{g(x)}\right)$$

is the Kullback-Leibler distance between two probability functions. The corresponding risk function is $R(f, \widehat{f}) = \mathbb{E}(D(f, \widehat{f})$. Notice that $D(f, \widehat{f}) = c - A(f, \widehat{f})$ where $c = \sum_x f(x) \log f(x)$ deos not depend on $\widehat{f}$ and

$$A(f, \widehat{f}) = \sum_x f(x) \log \widehat{f}(x).$$

Thus, minimizing the risk is equivalent to maximizing $a(f, \widehat{f}) \equiv \mathbb{E}(A(f, \widehat{f}))$.

It is tempting to estimate $a(f, \widehat{f})$ by $\sum_x \widehat{f}(x) \log \widehat{f}(x)$ but, just as the training error in regression is a highly biased estimate of prediction risk, it is also the case that $\sum_x \widehat{f}(x) \log \widehat{f}(x)$ is a highly biased estimate of $a(f, \widehat{f})$. In fact, the bias is approximately equal to $|M_j|$. Thus:

Theorem **18.15** *$AIC(M_j)$ is an approximately unbiased estimate of $a(f, \widehat{f})$.*

After finding a "best model" this way we can draw the corresponding graph. We can also check the overall fit of the selected model using the deviance as described above.

Example 18.16 *Data on breast cancer from Morrison et al (1973) were presented in homework question 4 of Chapter 17. The data are on diagnostic center (X_1), nuclear grade (X_2), and survival (X_3): (Morrison et al 1973):*

	X_2	*malignant*	*malignant*	*benign*	*benign*
	X_3	*died*	*survived*	*died*	*survived*
X_1	*Boston*	*35*	*59*	*47*	*112*
	Glamorgan	*42*	*77*	*26*	*76*

The saturated log-linear model is:

Coefficient	*Standard Error*	*Wald Statistic*	*p-value*	
(Intercept)	*3.56*	*0.17*	*21.03*	*0.00* ***
center	*0.18*	*0.22*	*0.79*	*0.42*
grade	*0.29*	*0.22*	*1.32*	*0.18*
survival	*0.52*	*0.21*	*2.44*	*0.01* *
center×grade	*-0.77*	*0.33*	*-2.31*	*0.02* *
center×survival	*0.08*	*0.28*	*0.29*	*0.76*
grade×survival	*0.34*	*0.27*	*1.25*	*0.20*
center×grade×survival	*0.12*	*0.40*	*0.29*	*0.76*

The best sub-model, selected using AIC and backward searching is

Coefficient	*Standard Error*	*Wald Statistic*	*p-value*	
(Intercept)	*3.52*	*0.13*	*25.62*	*¡ 0.00* ***
center	*0.23*	*0.13*	*1.70*	*0.08*
grade	*0.26*	*0.18*	*1.43*	*0.15*
survival	*0.56*	*0.14*	*3.98*	*6.65e-05* ***
center×grade	*-0.67*	*0.18*	*-3.62*	*0.00* ***
grade×survival	*0.37*	*0.19*	*1.90*	*0.05*

The graph for this model M is shown in Figure 18.7. To test the fit of this model, we compute the deviance of M which is 0.6. The appropriate χ^2 has $8 - 6 = 2$ degrees of freedom. The p-value is $\mathbb{P}(\chi^2_2 > .6) = .74$. So we have no evidence to suggest that the model is a poor fit. ■

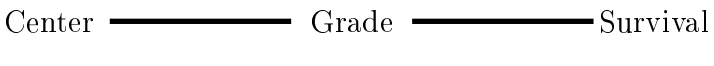


FIGURE 18.7. The graph for Example 18.16.

The example above is a toy example. With so few variables, there is little to be gained by searching for a best model. The overall multinomial MLE would be sufficient. More importantly, we are probably interested in predicting survivial from grade and center in which case this should really be treated as a regresion problem with outcome $Y = \textit{survival}$ and covariates center and grade.

Example 18.17 *Here is a synthetic example. We generate* $n = 100$ *random vectors* $X = (X_1, \ldots, X_5)$ *of length 5. We generated the data as follows:*

$$X_1 \sim \text{Bernoulli}(1/2)$$

and

$$X_j | X_1, \ldots, X_{j-1} \sim \begin{cases} 1/4 & \text{if } X_{j-1} = 0 \\ 3/4 & \text{if } X_{j-1} = 1. \end{cases}$$

It follows that

$$\begin{aligned} f(x_1, \ldots, x_5) &= \left(\frac{1}{2}\right)\left(\frac{3}{4}\right)^{x1}\left(\frac{3}{4}\right)^{1-x1}\left(\frac{3}{4}\right)^{x2}\left(\frac{3}{4}\right)^{1-x2}\left(\frac{3}{4}\right)^{x3}\left(\frac{3}{4}\right)^{1-x3}\left(\frac{3}{4}\right)^{x4}\left(\frac{3}{4}\right)^{1-x4} \\ &= \left(\frac{1}{2}\right)\left(\frac{3}{4}\right)^{x1+x2+x3+x4}\left(\frac{3}{4}\right)^{4-x1-x2-x3-x4}. \end{aligned}$$

We estimated f *using three methods: (i) maximum likelihood treating this as a multinomial with 32 categories, (ii) maximum likelihood from the best loglinear model using AIC and forward selection and (iii) maximum likelihood from the best loglinear*

model using BIC and forward selection. We estimated the risk by simulating the example 100 times. The average risks were:

Method	*Risk*
MLE	*0.63*
AIC	*0.54*
BIC	*0.53*

In this example, there is little difference between AIC and BIC. Both are better than maximum likelihood. ■

18.7 Bibliographic Remarks

A classic references on loglinear models is Bishop, Fienberg and Holland (1975). See also Whittaker (1990) from which some of the exercises are borrowed.

18.8 Exercises

1. Solve for the $p'_{ij}s$ in terms of the β's in Example 18.3.
2. Repeat example 18.17 using 7 covariates and $n = 1000$. To avoid numerical problems, replace any zero count with a one.
3. Prove Lemma 18.5.
4. Prove Lemma 18.9.
5. Consider random variables (X_1, X_2, X_3, X_4). Suppose the log-density is

$$\log f(x) = \psi_\emptyset(x) + \psi_{12}(x) + \psi_{13}(x) + \psi_{24}(x) + \psi_{34}(x).$$

   (a) Draw the graph G for these variables.

   (b) Write down all independence and conditional independence relations implied by the graph.

(c) Is this model graphical? Is it hierarchical?

6. Suppose that parameters $p(x_1, x_2, x_3)$ are proportional to the following values:

	x_2	0	0	1	1
	x_3	0	1	0	1
x_1	0	2	8	4	16
	1	16	128	32	256

Find the ψ-terms for the log-linear expansion. Comment on the model.

7. Let $X_1, \ldots, X_4$ be binary. Draw the independence graphs corresponding to the following log-linear models. Also, identify whether each is graphical and/or hierarchical (or neither).

(a) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4$

(b) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 78x_2x_4 + 3x_3x_4 + 32x_2x_3x_4$

(c) $\log f = 7 + 11x_1 + 2x_2 + 1.5x_3 + 17x_4 + 12x_2x_3 + 3x_3x_4 + x_1x_4 + 2x_1x_2$

(d) $\log f = 7 + 5055x_1x_2x_3x_4$

19
Causal Inference

In this Chapter we discuss causation. Roughly speaking, the statement "X causes Y" means that changing the value of X will change the distribution of Y. When X causes Y, X and Y will be associated but the reverse is not, in general, true. Association does not necessarily imply causation. We will consider two frameworks for discussing causation. The first uses the notation of **counterfactual** random variables. The second, presented in the next Chapter, uses **directed acyclic graphs.**

19.1 The Counterfactual Model

Suppose that X is a binary treatment variable where $X = 1$ means "treated" and where $X = 0$ means "not treated." We are using the word "treatment" in a very broad sense: treatment might refer to a medication or something like smoking. An alternative to "treated/not treated" is "exposed/not exposed" but we shall use the former.

Let Y be some outcome variable such as presence or absence of disease. To distinguish the statement "X is associated Y" from the statement "X causes Y" we need to enrich our probabilistic vocabulary. We will decompose the response Y into a more fine-grained object.

We introduce two new random variables (C_0, C_1), called **potential outcomes** with the following interpretation: C_0 is the outcome if the subject is not treated $(X = 0)$ and C_1 is the outcome if the subject is treated $(X = 1)$. Hence,

$$Y = \begin{cases} C_0 & \text{if } X = 0 \\ C_1 & \text{if } X = 1. \end{cases}$$

We can express the relationship between Y and (C_0, C_1) more succintly by

$$Y = C_X. \tag{19.1}$$

This equation is called the **consistency relationship**.

Here is a toy data set to make the idea clear:

X	Y	C_0	C_1
0	4	4	*
0	7	7	*
0	2	2	*
0	8	8	*
1	3	*	3
1	5	*	5
1	8	*	8
1	9	*	9

The asterisks denote unobserved values. When $X = 0$ we don't observe C_1 in which case we say that C_1 is a **counterfactual** since it is the outcome you would have had if, counter to the fact, you had been treated $(X = 1)$. Similarly, when $X = 1$ we don't observe C_0 and we say that C_0 is **counterfactual**.

Notice that there are four types of subjects:

Type	C_0	C_1
Survivors	1	1
Responders	0	1
Anti-repsonders	1	0
Doomed	0	0

Think of the potential outcomes (C_0, C_1) as hidden variables that contain all the relavent information about the subject.

Define the **average causal effect** or **average treatment effect** to be

$$\theta = \mathbb{E}(C_1) - \mathbb{E}(C_0). \tag{19.2}$$

The parameter θ has the following interpretation: θ is the mean if everyone were treated $(X = 1)$ minus the mean if everyone were not treated $(X = 0)$. There are other ways of measuring the causal effect. For example, if C_0 and C_1 are binary, we define the **causal odds ratio**

$$\frac{\mathbb{P}(C_1 = 1)}{\mathbb{P}(C_1 = 0)} \div \frac{\mathbb{P}(C_0 = 1)}{\mathbb{P}(C_0 = 0)}$$

and the **causal relative risk**

$$\frac{\mathbb{P}(C_1 = 1)}{\mathbb{P}(C_0 = 1)}.$$

The main ideas will be the same whatever causal effect we use. For simplicity, we shall work with the average causal effect θ.

Define the **association** to be

$$\alpha = \mathbb{E}(Y\,|X = 1) - \mathbb{E}(Y\,|X = 0). \tag{19.3}$$

Again, we could use odds ratios or other summaries if we wish.

Theorem **19.1 (Association is not equal to Causation)** *In general, $\theta \neq \alpha$.*

Example 19.2 *Suppose the whole population is as follows:*

X	Y	C_0	C_1
0	0	0	0^*
0	0	0	0^*
0	0	0	0^*
0	0	0	0^*
1	1	1^*	1
1	1	1^*	1
1	1	1^*	1
1	1	1^*	1

Again, the asterisks denote unobserved values. Notice that $C_0 = C_1$ *for every subject, thus, this treatment has no effect. Indeed,*

$$\begin{aligned}\theta &= \mathbb{E}(C_1) - \mathbb{E}(C_0) = \frac{1}{8}\sum_{i=1}^{8} C_{1i} - \frac{1}{8}\sum_{i=1}^{8} C_{0i} \\ &= \frac{1+1+1+1+0+0+0+0}{8} - \frac{1+1+1+1+0+0+0+0}{8} \\ &= 0.\end{aligned}$$

Thus the average causal effect is 0. The observed data are X*'s and* Y*'s only from which we can estimate the association:*

$$\begin{aligned}\alpha &= \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0) \\ &= \frac{1+1+1+1}{4} - \frac{0+0+0+0}{4} = 1.\end{aligned}$$

Hence, $\theta \neq \alpha$.

To add some intutition to this example, imagine that the outcome variable is 1 if "healthy" and 0 if "sick". Suppose that $X = 0$ *means that the subject does not take vitamin C and that* $X = 1$ *means that the subject does take vitamin C. Vitamin C has no causal effect since* $C_0 = C_1$*. In this example there are two types of people: healthy people* $(C_0, C_1) = (1, 1)$ *and unhealthy people* $(C_0, C_1) = (0, 0)$*. Healthy people tend to take vitamin C while unhealty people don't. It is this association between* (C_0, C_1) *and* X *that creates an association between* X *and* Y*. If we only had data on* X *and* Y *we would conclude that* X

and Y are associated. Suppose we wrongly interpret this causally and conclude that vitamin C prevents illness. Next we might encourage everyone to take vitamin C. If most people comply the population will look something like this:

X	Y	C_0	C_1
0	0	0	0^*
1	0	0	0^*
1	0	0	0^*
1	0	0	0^*
1	1	1^*	1
1	1	1^*	1
1	1	1^*	1
1	1	1^*	1

Now $\alpha = (4/7) - (0/1) = 4/7$. We see that α went down from 1 to 4/7. Of course, the causal effect never changed but the naive observer who does not distinguish association and causation will be confused because his advice seems to have made things worse instead of better. ■

In the last example, $\theta = 0$ and $\alpha = 1$. It is not hard to create examples in which $\alpha > 0$ and yet $\theta < 0$. The fact that the association and causal effects can have different signs is very confusing to many people including well educated statisticians.

The example makes it clear that, in general, we cannot use the association to estimate the causal effect θ. The reason that $\theta \neq \alpha$ is that (C_0, C_1) was not independent of X. That is, treatment assigment was not independent of person type.

Can we ever estimate the causal effect? The answer is: sometimes. In particular, random asigment to treatment makes it possible to estimate θ.

Theorem **19.3** *Suppose we randomly assign subjects to treatment and that $\mathbb{P}(X = 0) > 0$ and $\mathbb{P}(X = 1) > 0$. Then $\alpha = \theta$. Hence, any consistent estimator of α is a consistent estimator*

of θ. In particular, a consistent estimator is

$$\begin{aligned}\widehat{\theta} &= \widehat{\mathbb{E}}(Y|X=1) - \widehat{\mathbb{E}}(Y|X=0) \\ &= \overline{Y}_1 - \overline{Y}_0\end{aligned}$$

is a consistent estimator of θ, where

$$\overline{Y}_1 = \frac{1}{n_1}\sum_{i:X_i=1} Y_i, \qquad \overline{Y}_0 = \frac{1}{n_0}\sum_{i:X_i=0} Y_i,$$

$n_1 = \sum_{i=1}^n X_i$ *and* $n_0 = \sum_{i=1}^n (1-X_i)$.

PROOF. Since X is randomly assigned, X is independent of (C_0, C_1). Hence,

$$\begin{aligned}\theta &= \mathbb{E}(C_1) - \mathbb{E}(C_0) \\ &= \mathbb{E}(C_1|X=1) - \mathbb{E}(C_0|X=0) \qquad \text{since } X \amalg (C_0, C_1) \\ &= \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0) \qquad \text{since } Y = C_X \\ &= \alpha.\end{aligned}$$

The consistency follows from the law of large numbers. ■

If Z is a covariate, we define the **conditional causal effect** by

$$\theta_z = \mathbb{E}(C_1|Z=z) - \mathbb{E}(C_0|Z=z).$$

For example, if Z denotes gender with values $Z=0$ (women) and $Z=1$ (men), then θ_0 is the causal effect among women and θ_1 is the causal effect among men. In a randomized experiment, $\theta_z = \mathbb{E}(Y|X=1, Z=z) - \mathbb{E}(Y|X=0, Z=z)$ and we can estimate the conditional causal effect using appropriate sample averages.

Summary

Random variables: (C_0, C_1, X, Y).
Consistency relationship: $Y = C_X$.
Causal Effect: $\theta = \mathbb{E}(C_1) - \mathbb{E}(C_0)$.
Association: $\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)$.
Random assigment $\implies (C_0, C_1) \amalg X \implies \theta = \alpha$.

19.2 Beyond Binary Treatments

Let us now generalize beyond the binary case. Suppose that $X \in \mathcal{X}$. For example, X could be the dose of a drug in which case $X \in \mathbb{R}$. The counterfactual vector (C_0, C_1) now becomes the **counterfactual process**

$$\{C(x) : \ x \in \mathcal{X}\} \tag{19.4}$$

where $C(x)$ is the outcome a subject would have if he received dose x. The observed reponse is given by the consistency relation

$$Y \equiv C(X). \tag{19.5}$$

See Figure 19.1. The **causal regression function** is

$$\theta(x) = \mathbb{E}(C(x)). \tag{19.6}$$

The regression function, which measures association, is $r(x) = \mathbb{E}(Y|X = x)$.

Theorem **19.4** *In general, $\theta(x) \neq r(x)$. However, when X is randomly assigned, $\theta(x) = r(x)$.*

Example 19.5 *An example in which $\theta(x)$ is constant but $r(x)$ is not constant is shown in Figure 19.2. The figure shows the counterfactuals processes for four subjects. The dots represent*

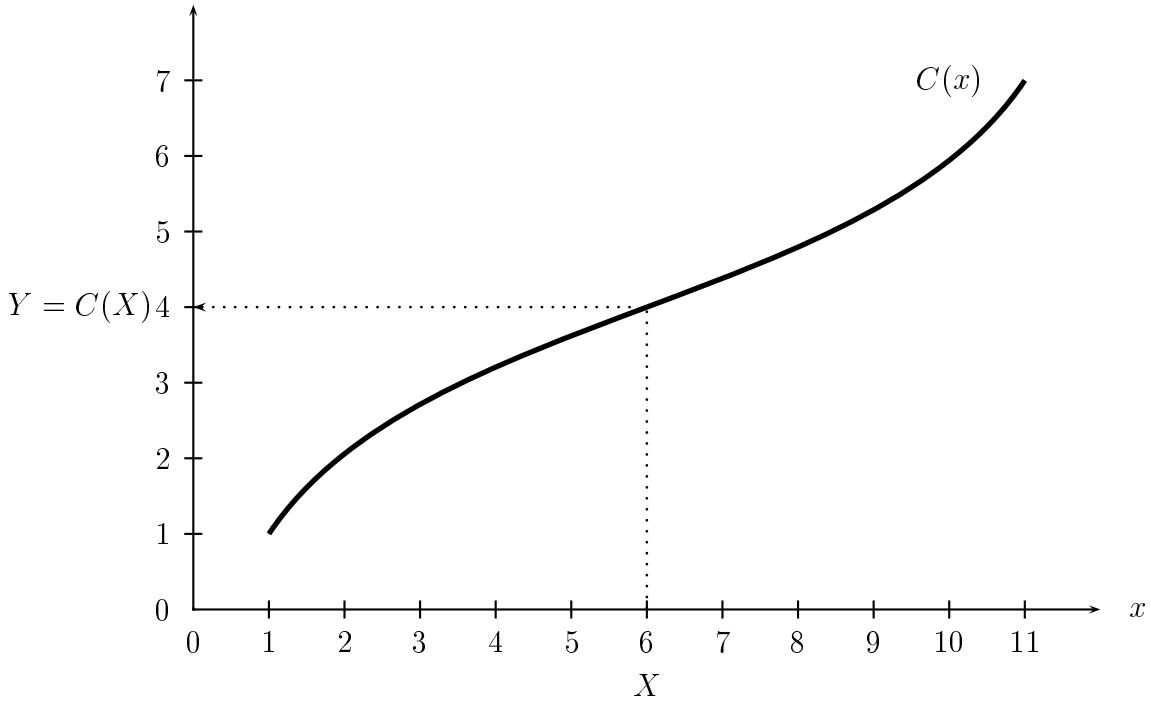


FIGURE 19.1. A counterfactual process $C(x)$. The outcome Y is the value of the curve $C(x)$ evaluated at the observed dose X.

their X values X_1, X_2, X_3, X_4. Since $C_i(x)$ is constant over x for all i, there is no causal effect and hence

$$\theta(x) = \frac{C_1(x) + C_2(x) + C_3(x) + C_4(x)}{4}$$

is constant. Changing the dose x will not change anyone's outcome. The four dots in the lower plot represent the observed data points $Y_1 = C_1(X_1), Y_2 = C_2(X_2), Y_3 = C_3(X_3), Y_4 = C_4(X_4)$. The dotted line represents the regression $r(x) = \mathbb{E}(Y|X = x)$. Although there is no causal effect, there is an association since the regression curve $r(x)$ is not constant. ■

19.3 Observational Studies and Confounding

A study in which treatment (or exposure) is not randomly assigned, is called an **observational study.** In these studies, subjects select their own value of the exposure X. Many of the health studies your read about in the newspaper are like this. As we saw, association and causation could in general be quite different. This descrepancy occurs in non-randomized studies because the potential outcome C is not independent of treatment X. However, suppose we could find groupings of subjects such that, within groups, X and $\{C(x) : x \in \mathcal{X}\}$ are independent. This would happen if the subjects are very similar within groups. For example, suppose we find people who are very similar in age, gender, educational background and ethnic background. Among these people we might feel it is reasonable to assume that the choice of X is essentially random. These other variables are called **confounding variables**. If we denote these other variables collectively as Z, then we can express this idea by saying that

A more precise definition of confounding is given in the next Chapter.

$$\{C(x) : x \in \mathcal{X}\} \amalg X|Z. \tag{19.7}$$

Equation (19.7) means that, within groups of Z, the choice of treatment X does not depend on type, as represented by $\{C(x) :$

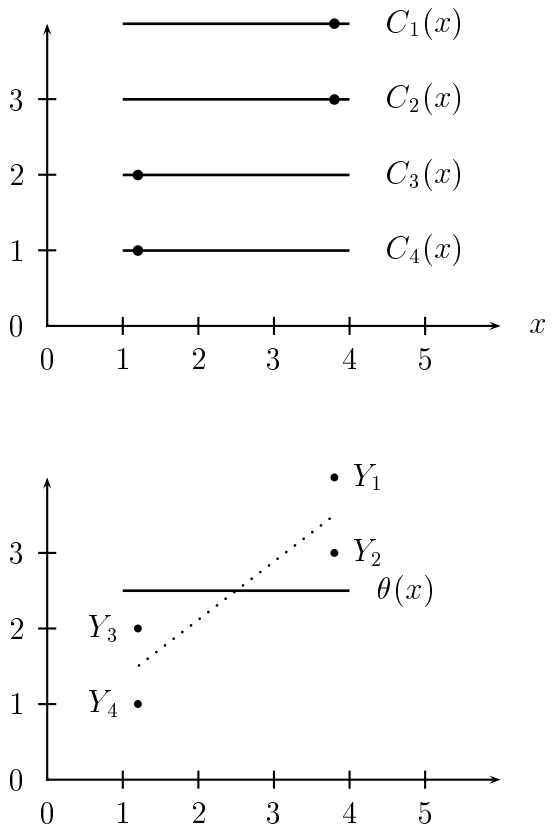


FIGURE 19.2. The top plot shows the counterfactual process $C(x)$ for four subjects. The dots represent their X values. Since $C_i(x)$ is constant over x for all i, there is no causal effect. Changing the dose will not change anyone's outcome. The lower plot shows the causal regression function $\theta(x) = (C_1(x) + C_2(x) + C_3(x) + C_4(x))/4$. The four dots represent the observed data points $Y_1 = C_1(X_1), Y_2 = C_2(X_2), Y_3 = C_3(X_3), Y_4 = C_4(X_4)$. The dotted line represents the regression $r(x) = \mathbb{E}(Y|X = x)$. There is no causal effect since $C_i(x)$ is constant for all i. But there is an association since the regression curve $r(x)$ is not constant.

$x \in \mathcal{X}\}$. If (19.7) holds and we observe Z then we say that there is **no unmeasured confounding**.

Theorem **19.6** *Suppose that (19.7) holds. Then,*

$$\theta(x) = \int \mathbb{E}(Y|X=x, Z=z)dF_Z(z)dz. \tag{19.8}$$

If $\widehat{r}(x,z)$ is a consistent estimate of the regression function $\mathbb{E}(Y|X = x, Z = z)$, then a consistent estimate of $\theta(x)$ is

$$\widehat{\theta}(x) = \frac{1}{n}\sum_{i=1}^{n} \widehat{r}(x, Z_i).$$

In particular, if $r(x,z) = \beta_0 + \beta_1 x + \beta_2 z$ is linear, then a consistent estimate of $\theta(x)$ is

$$\widehat{\theta}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 \overline{Z}_n \tag{19.9}$$

where $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ are the least squares estimators.

Remark 19.7 *It is useful to compare equation (19.8) to $\mathbb{E}(Y|X = x)$ which can be written as $\mathbb{E}(Y|X = x) = \int \mathbb{E}(Y|X = x, Z = z)dF_{Z|X}(z|x)$.*

Epidemiologists call (19.8) the **adjusted treatment effect**. The process of computing adjusted treatment effects is called **adjusting (or controlling) for confounding.** The selection of what confounders Z to measure and control for requires scientific insight. Even after adjusting for confounders, we cannot be sure that there are not other confounding variables that we missed. This is why observational studies must be treated with healthy skepticism. Results from observational studies start to become believable when: (i) the results are replicated in many studies, (ii) each of the studies controlled for plausible confounding variables, (iii) there is a plausible scientific explanation for the existence of a causal relationship.

A good example is smoking and cancer. Numerous studies have shown a relationship between smoking and cancer even after adjusting for many confouding variables. Moreover, in laboratory studies, smoking has been shown to damage lung cells. Finally, a causal link between smoking and cancer has been found in randomized animal studies. It is this collection of evidence over many years that makes this a convincing case. One single observational study is not, by itself, strong evidence. Remember that when you read the newspaper.

19.4 Simpson's Paradox

Simpson's paradox is a puzzling phenomenon that is dicussed in most statistics texts. Unfortunately, it is explained incorrectly in most statistics texts. The reason is that it is nearly impossible to explain the paradox without using counterfactuals (or directed acyclic graphs).

Let X be a binary treatment variable, Y a binary outcome and Z a third binary variable such as gender. Suppose the joint distribution of X, Y, Z is

	$Y=1$	$Y=0$	$Y=1$	$Y=0$
$X=1$	.1500	.2250	.1000	.0250
$X=0$	.0375	.0875	.2625	.1125
	$Z=1$ (men)		$Z=0$ (women)	

The marginal distribution for (X, Y) is

	$Y=1$	$Y=0$	
$X=1$	.25	.25	.50
$X=0$	.30	.20	.50
	.55	.45	1

Now,

$$\begin{aligned}\mathbb{P}(Y=1|X=1)-\mathbb{P}(Y=1|X=0) &= \frac{\mathbb{P}(Y=1,X=1)}{\mathbb{P}(X=1)}-\frac{\mathbb{P}(Y=1,X=0)}{\mathbb{P}(X=0)}\\ &= \frac{.25}{.50}-\frac{.30}{.50}\\ &= -0.1.\end{aligned}$$

We might naively interpret this to mean that the treatment is bad for you since $\mathbb{P}(Y=1|X=1) < \mathbb{P}(Y=1|X=0)$. Furthermore, among men,

$$\begin{aligned}\mathbb{P}(Y=1|X=1,Z=1) &- \mathbb{P}(Y=1|X=0,Z=1)\\ &= \frac{\mathbb{P}(Y=1,X=1,Z=1)}{\mathbb{P}(X=1,Z=1)}-\frac{\mathbb{P}(Y=1,X=0,Z=1)}{\mathbb{P}(X=0,Z=1)}\\ &= \frac{.15}{.3750}-\frac{.0375}{.1250}\\ &= 0.1.\end{aligned}$$

Among women,

$$\begin{aligned}\mathbb{P}(Y=1|X=1,Z=0) &- \mathbb{P}(Y=1|X=0,Z=0)\\ &= \frac{\mathbb{P}(Y=1,X=1,Z=0)}{\mathbb{P}(X=1,Z=0)}-\frac{\mathbb{P}(Y=1,X=0,Z=0)}{\mathbb{P}(X=0,Z=0)}\\ &= \frac{.1}{.1250}-\frac{.2625}{.3750}\\ &= 0.1.\end{aligned}$$

To summarize, we seem to have the following information:

Mathematical Statement	English Statement?
$\mathbb{P}(Y=1\|X=1) < \mathbb{P}(Y=1\|X=0)$	treatment is harmful
$\mathbb{P}(Y=1\|X=1,Z=1) > \mathbb{P}(Y=1\|X=0,Z=1)$	treatment is beneficial to men
$\mathbb{P}(Y=1\|X=1,Z=0) > \mathbb{P}(Y=1\|X=0,Z=0)$	treatment is beneficial to women

Clearly, something is amiss. There can't be a treatment which is good for men, good for women but bad overall. This is nonsense. The problem is with the set of English statements in the table. Our translation from math into english is specious.

> **The inequality** $\mathbb{P}(Y=1|X=1) < \mathbb{P}(Y=1|X=0)$ **does not mean that treatment is harmful.**

The phrase "treatment is harmful" should be written mathematically as $\mathbb{P}(C_1=1) < \mathbb{P}(C_0=1)$. The phrase "treatment is harmful for men" should be written $\mathbb{P}(C_1=1|Z=1) < \mathbb{P}(C_0=1|Z=1)$. The three mathematical statements in the table are not at all contradictory. It is only the translation into English that is wrong.

Let us now show that a real Simpson's paradox cannot happen, that is, there cannot be a treatment that is beneficial for men and women but harmful overall. Suppose that treatment is beneficial for both sexes. Then

$$\mathbb{P}(C_1=1|Z=z) > \mathbb{P}(C_0=1|Z=z)$$

for all z. It then follows that

$$\begin{aligned}\mathbb{P}(C_1=1) &= \sum_z \mathbb{P}(C_1=1|Z=z)\mathbb{P}(Z=z) \\ &> \sum_z \mathbb{P}(C_0=1|Z=z)\mathbb{P}(Z=z) \\ &= \mathbb{P}(C_0=1).\end{aligned}$$

Hence, $\mathbb{P}(C_1=1) > \mathbb{P}(C_0=1)$ so treatment is beneficial overall. No paradox.

19.5 Bibliographic Remarks

The use of potential outcomes to clarify causation is due mainly to Jerzy Neyman and Don Rubin. Later developments are due to Jamie Robins, Paul Rosenbaum and others. A parallel development took place in econometrics by various people including Jim Heckman and Charles Manski. Currently, there are no textbook treatments of causal inference from the potential outcomes viewpoint.

19.6 Exercises

1. Create an example like Example 19.2 in which $\alpha > 0$ and $\theta < 0$.

2. Prove Theorem 19.4.

3. Suppose you are given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ from an observational study, where $X_i \in \{0, 1\}$ and $Y_i \in \{0, 1\}$. Although it is not possible to estimate the causal effect θ, it is possible to put bounds on θ. Find upper and lower bounds on θ that can be consistently estimated from the data. Show that the bounds have width 1. Hint: Note that $\mathbb{E}(C_1) = \mathbb{E}(C_1|X = 1)\mathbb{P}(X = 1) + \mathbb{E}(C_1|X = 0)\mathbb{P}(X = 0)$.

4. Suppose that $X \in \mathbb{R}$ and that, for each subject i, $C_i(x) = \beta_{1i}x$. Each subject has their own slope β_{1i}. Construct a joint distribution on (β_1, X) such that $\mathbb{P}(\beta_1 > 0) = 1$ but $\mathbb{E}(Y|X = x)$ is a decreasing function of x, where $Y = C(X)$. Interpret. Hint: Write $f(\beta_1, x) = f(\beta_1)f(x|\beta_1)$. Choose $f(x|\beta_1)$ so that when β_1 is large, x is small and when β_1 is small x is large.

20
Directed Graphs

20.1 Introduction

Directed graphs are similar to undirected graphs except that there are arrows between vertices instead of edges. Like undirected graphs, directed graphs can be used to represent independence relations. They can also be used as an alternative to counterfactuals to represent causal relationships. Some people use the phrase **Bayesian network** to refer to a directed graph endowed with a probability distribution. This is a poor choice of terminology. Statistical inference for directed graphs can be performed using frequentist or Bayesian methods so it is misleading to call them Bayesian networks.

20.2 DAG's

A **directed graph** $\mathcal{G}$ consists of a set of vertices V and an edge set E of ordered pairs of variables. If $(X, Y) \in E$ then there is an arrow pointing from X to Y. See Figure 20.1.

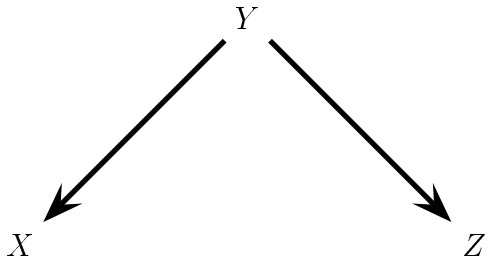


FIGURE 20.1. A directed graph with $V = \{X, Y, Z\}$ and $E = \{(Y, X), (Y, Z)\}$.

If an arrow connects two variables X and Y (in either direction) we say that X and Y are **adjacent.** If there is an arrow from X to Y then X is a **parent** of Y and Y is a **child** of X. The set of all parents of X is denoted by π_X or $\pi(X)$. A **directed path** from X to Y is a set of vertices beginning with X, ending with Y such that each pair is connected by an arrow and all the arrows point in the same direction as follows:

$$X \longrightarrow \cdots \longrightarrow Y \quad \text{or} \quad X \longleftarrow \cdots \longleftarrow Y$$

A sequence of adjacent vertices staring with X and ending with Y but ignoring the direction of the arrows is called an **undirected path.** The sequence $\{X, Y, Z\}$ in Figure 20.1 is an undirected path. X is an **ancestor** of Y is there is a directed path from X to Y. We also say that Y is a **descendant** of X.

A configuration of the form:

$$X \longrightarrow Y \longleftarrow Z$$

is called a **collider**. A configuration not of that form is called a **non-collider**, for example,

$$X \longrightarrow Y \longrightarrow Z$$

or

$$X \longleftarrow Y \longrightarrow Z$$

A directed path that starts and ends at the same variable is called a **cycle.** A directed graph is **acyclic** if it has no cycles. In this case we say that the graph is a **directed acyclic graph** or **DAG**. From now on, we only deal with graphs that are DAG's.

20.3 Probability and DAG's

Let $\mathcal{G}$ be a DAG with vertices $V = (X_1, \ldots, X_k)$.

Definition 20.1 *If $\mathbb{P}$ is a distribution for V with probability function p, we say that If $\mathbb{P}$ is* **Markov to** $\mathcal{G}$ *or that $\mathcal{G}$* **represents** $\mathbb{P}$ *if*

$$p(v) = \prod_{i=1}^{k} p(x_i \mid \pi_i) \tag{20.1}$$

where π_i are the parents of X_i. The set of distributions represented by $\mathcal{G}$ is denoted by $M(\mathcal{G})$.

Example 20.2 *For the DAG in Figure 20.2, $\mathbb{P} \in M(\mathcal{G})$ if and only if its probability function p has the form*

$$p(x, y, z, w) = p(x)p(y)p(z \mid x, y)p(w \mid z). \quad \blacksquare$$

The following theorem says that $\mathbb{P} \in M(\mathcal{G})$ if and only if the **Markov Condition** holds. Roughly speaking, the Markov Condition says that every variable W is independent of the "past" given its parents.

Theorem **20.3** *A distribution $\mathbb{P} \in M(\mathcal{G})$ if and only if the following* **Markov Condition** *holds: for every variable W,*

$$W \amalg \widetilde{W} \mid \pi_W \tag{20.2}$$

where $\widetilde{W}$ denotes all the other variables except the parents and descendants of W.

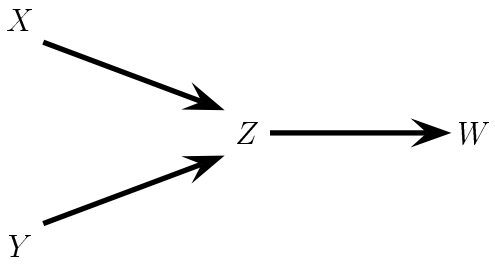


FIGURE 20.2. Another DAG.

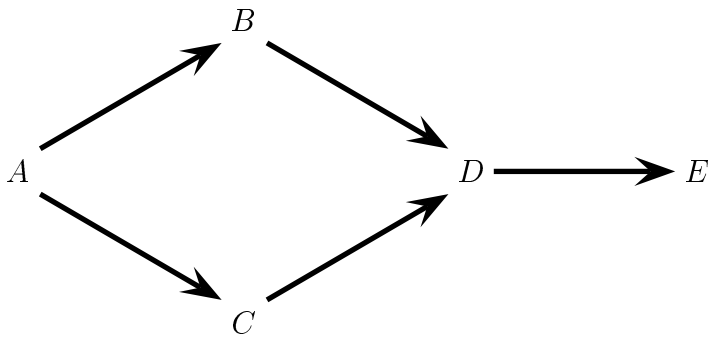


FIGURE 20.3. Yet another DAG.

Example 20.4 *In Figure 20.2, the Markov Condition implies that*

$$X \amalg Y \quad \text{and} \quad W \amalg \{X, Y\} \mid Z. \quad \blacksquare$$

Example 20.5 *Consider the DAG in Figure 20.3. In this case probability function must factor like*

$$p(a, b, c, d, e) = p(a)p(b|a)p(c|a)p(d|b, c)p(e|d).$$

The Markov Condition implies the following independence relations:

$$D \amalg A \mid \{B, C\}, \quad E \amalg \{A, B, C\} \mid D \quad \text{and} \quad B \amalg C \mid A \quad \blacksquare$$

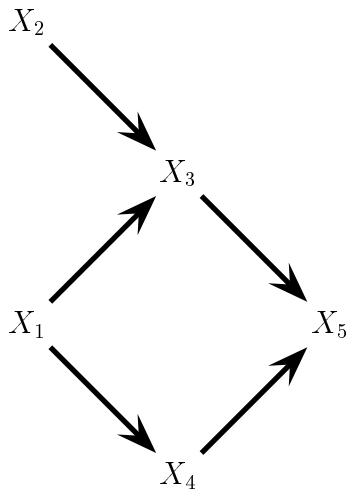


FIGURE 20.4. And yet another DAG.

20.4 More Independence Relations

With undirected graphs, we saw that pairwise independence relations implied other independence relations. Luckily, we were able to deduce these other relations from the graph. The situation is similar for DAG's. The Markov Condition allows us to list some independence relations. These relations may logically imply other other independence relations. Consider the DAG in Figure 20.4. The Markov Condition implies:

$$X_1 \amalg X_2, \qquad X_2 \amalg \{X_1, X_4\}, \qquad X_3 \amalg X_4 \mid \{X_1, X_2\},$$

$$X_4 \amalg \{X_2, X_3\} \mid X_1, \qquad X_5 \amalg \{X_1, X_2\} \mid \{X_3, X_4\}$$

It turns out (but it is not obvious) that these conditions imply that

$$\{X_4, X_5\} \amalg X_2 \mid \{X_1, X_3\}.$$

How do we find these extra independence relations? The answer is "d-separation." d-separation can be summarized by three rules. Consider the four DAG's in Figure 20.5 and the DAG in Figure 20.6. The first 3 DAG's in Figure 20.5 are non-colliders. The DAG in the lower right of Figure 20.5 is a collider. The DAG in Figure 20.6 is a collider with a descendent.

In what follows, we implicitly assume that $\mathbb{P}$ is **faithful** to $\mathcal{G}$ which means that $\mathbb{P}$ has no extra independence relations other than those logically implied by the Markov Condition.

FIGURE 20.5. The first three DAG's have no colliders. The fourth DAG in the lower right corner has a collider at Y.

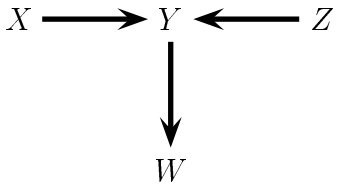


FIGURE 20.6. A collider with a descendent.

The rules of d-separation.

Consider the DAG's in Figures 20.5 and 20.6.
Rule (1). In a non-collider, X and Z are **d-connected**, but they are **d-separated** given Y.
Rule (2). If X and Z collide at Y then X and Z are **d-separated** but they are **d-connected** given Y.
Rule (3). Conditioning on the descendant of a collider has the same effect as conditioning on the collider. Thus in Figure 20.6, X and Z are **d-separated** but they are **d-connected** given W.

Here is a more formal definition of d-separation. Let X and Y be distinct vertices and let W be a set of vertices not containing X or Y. Then X and Y are **d-separated given** W if there

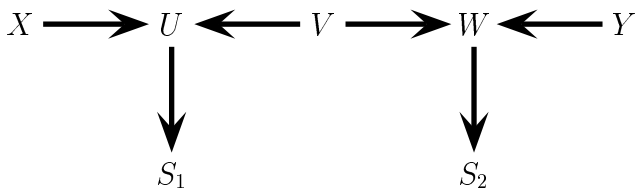


FIGURE 20.7. d-separation explained.

exists no undirected path U between X and Y such that (i) every collider on U has a descendant in W and (ii) no other vertex on U is in W. If U, V and W are distinct sets of vertices and U and V are not empty, then U and V are d-separated given W if for every $X \in U$ and $Y \in V$, X and Y are d-separated given W. Vertices that are not d-separated are said to be d-connected.

Example 20.6 *Consider the DAG in Figure 20.7. From the d-seperation rules we conclude that:*

X and Y are d-separated (given the empty set)
X and Y are d-connected given $\{S_1, S_2\}$
X and Y are d-separated given $\{S_1, S_2, V\}$

Theorem **20.7 (Spirtes, Glymour and Scheines)** *Let A, B and C be disjoint sets of vertices. Then $A \amalg B \mid C$ if and only if A and B are d-separated by C.*

Example 20.8 *The fact that conditioning on a collider creates dependence might not seem intuitive. Here is a whimsical example from Jordan (2003) that makes this idea more palatable. Your friend appears to be late for a meeting with you. There are two explanations: she was abducted by aliens or you forgot to set your watch ahead one hour for daylight savings time. See Figure 20.8* Aliens *and* Watch *are blocked by a collider which implies they are marginally independent. This seems reasonable since, before we know anything about your friend being late, we would expect these variables to be independent. We would also expect that $\mathbb{P}(Aliens = yes|Late = Yes) > \mathbb{P}(Aliens = yes)$;*

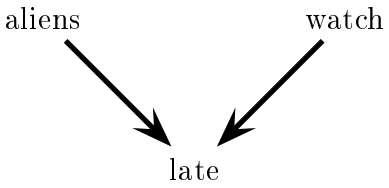


FIGURE 20.8. Jordan's alien example. Was your friend kidnapped by aliens or did you forget to set your watch?

learning that your friend is late certainly increases the probability that she was abducted. But when we learn that you forgot to set your watch properly, we would lower the chance that your friend was abducted. Hence, $\mathbb{P}(Aliens = yes|Late = Yes) \neq \mathbb{P}(Aliens = yes|Late = Yes, Watch = no)$. *Thus,* Aliens *and* Watch *are dependent given* Late.

Graphs that look different may actually imply the same independence relations. If $\mathcal{G}$ is a DAG, we let $\mathcal{I}(\mathcal{G})$ denote all the independence statements implied by $\mathcal{G}$. Two DAG's $\mathcal{G}_1$ and $\mathcal{G}_2$ for the same variables V are **Markov equivalent** if $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$. Given a DAG $\mathcal{G}$, let skeleton($\mathcal{G}$) denote the undirected graph obtained by replacing the arrows with undirected edges.

Theorem **20.9** *Two DAG's* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *are Markov equivalent if and only if (i)* skeleton($\mathcal{G}_1$) = skeleton($\mathcal{G}_2$) *and (ii)* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *have the same colliders.*

Example 20.10 *The first three DAG's in Figure 20.5 are Markov equivalent. The DAG in the lower right of the Figure is not Markov equivalent to the others.* ■

20.5 Estimation for DAG's

Let $\mathcal{G}$ be a DAG. Assume that all the variables $V = (X_1, \ldots, X_m)$ are discrete. The probability function can be written

$$p(v) = \prod_{i=1}^{k} p(x_i \mid \pi_i). \tag{20.3}$$

To estimate $p(v)$ we need to estimate $p(x_i \mid \pi_i)$ for each i. Think of the parents of X_i as one discrete variable $\widetilde{X}_i$ with many levels. For example, suppose that X_3 has parents X_1 and X_2 and that $X_1 \in \{0,1,2\}$ while $X_2 \in \{0,1\}$. We can regard the parents X_1 and X_2 as a single variable $\widetilde{X}_3$ defined by

$$\widetilde{X}_3 = \begin{cases} 1 & \text{if } X_1 = 0, X_2 = 0 \\ 2 & \text{if } X_1 = 0, X_2 = 1 \\ 3 & \text{if } X_1 = 1, X_2 = 0 \\ 4 & \text{if } X_1 = 1, X_2 = 1 \\ 5 & \text{if } X_1 = 2, X_2 = 0 \\ 6 & \text{if } X_1 = 2, X_2 = 1. \end{cases}$$

Hence we can write

$$p(v) = \prod_{i=1}^{k} p(x_i \mid \widetilde{x}_i). \tag{20.4}$$

Theorem **20.11** *Let $V_1, \ldots, V_n$ be* IID *random vectors from distribution p given in (20.4). The maximum likelihood estimator of p is*

$$\widehat{p}(v) = \prod_{i=1}^{k} \widehat{p}(x_i \mid \widetilde{x}_i) \tag{20.5}$$

where

$$\widehat{p}(x_i \mid \widetilde{x}_i) = \frac{\#\{i:\ X_i = x_i \text{ and } \widetilde{X}_i = \widetilde{x}_i\}}{\#\{i:\ \widetilde{X}_i = \widetilde{x}_i\}}.$$

Example 20.12 *Let $V = (X, Y, Z)$ have the DAG in the top left of Figure 20.5. Given n observations $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$, the* MLE *of $p(x, y, z) = p(x)p(y|x)p(z|y)$ is*

$$\widehat{p}(x) = \frac{\#\{i:\ X_i = x\}}{n},$$

$$\widehat{p}(y|x) = \frac{\#\{i:\ X_i = x \text{ and } Y_i = y\}}{\#\{i:\ Y_i = y\}},$$

and

$$\widehat{p}(z|y) = \frac{\#\{i:\ Y_i = y \text{ and } Z_i = z\}}{\#\{i:\ Z_i = z\}}.$$

It is possible to extend these ideas to continuous random variables as well. For example, we might use some parametric model $p(x|\pi_x;\theta_x)$ for each condiitonal density. The likelihood function is then

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(V_i;\theta) = \prod_{i=1}^{n}\prod_{j=1}^{m} p(X_{ij}|\pi_j;\theta_j)$$

where X_{ij} is the value of X_j for the i^{th} data point and θ_j are the parameters for the j^{th} conditional density. We can then proceed using maximum likelihood.

So far, we have assumed that the structure of the DAG is given. One can also try to estimate the structure of the DAG itself from the data. For example, we could fit every possible DAG using maximum likelihood and use AIC (or some other method) to choose a DAG. However, there are many possible DAG's so you would need much data for such a method to be reliable. Producing a valid, accurate confidence set for the DAG structure would require astronomical sample sizes. DAG's are thus most useful for encoding conditional independence information rather than discovering it.

20.6 Causation Revisited

We discussed causation in Chapter 19 using the idea of counterfactual random variables. A different approach to causation uses DAG's. The two approaches are mathematically equivalent though they appear to be quite different. In Chapter 19, the extra element we added to clarify causation was the idea of a counterfactual. In the DAG approach, the extra element is the idea of **intervention**. Consider the DAG in Figure 20.9.

The probability function for a distribution consistent with this DAG has the form $p(x,y,z) = p(x)p(y|x)p(z|x,y)$. Here is pseudo-code for generating from this distribution:

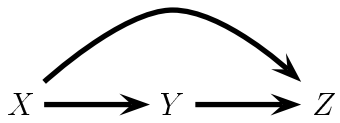


FIGURE 20.9. Conditioning versus intervening.

$$
\begin{array}{rcl}
\text{for } i & = & 1, \ldots, n: \\
x_i & \leftarrow & p_X(x_i) \\
y_i & \leftarrow & p_{Y|X}(y_i|x_i) \\
z_i & \leftarrow & p_{Z|X,Y}(z_i|x_i, y_i)
\end{array}
$$

Suppose we repeat this code many times yielding data $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$. Among all the times that we observe $Y = y$, how often is $Z = z$? The answer to this question is given by the conditional distribution of $Z|Y$. Specifically,

$$
\begin{aligned}
\mathbb{P}(Z = z|Y = y) &= \frac{\mathbb{P}(Y = y, Z = z)}{\mathbb{P}(Y = y)} = \frac{p(y, z)}{p(y)} \\
&= \frac{\sum_x p(x, y, z)}{p(y)} = \frac{\sum_x p(x)\, p(y|x)\, p(z|x, y)}{p(y)} \\
&= \sum_x p(z|x, y) \frac{p(y|x)\, p(x)}{p(y)} = \sum_x p(z|x, y) \frac{p(x, y)}{p(y)} \\
&= \sum_x p(z|x, y)\, p(x|y).
\end{aligned}
$$

Now suppose we **intervene** by changing the computer code. Specifically, suppose we fix Y at the value y. The code now looks like this:

$$\begin{aligned} \text{set } Y &= y \\ \text{for } i &= 1, \ldots, n \\ x_i &\leftarrow p_X(x_i) \\ z_i &\leftarrow p_{Z|X,Y}(z_i|x_i, y) \end{aligned}$$

Having **set** $Y = y$, how often was $Z = z$? To answer, note that the intervention has changed the joint probability to be

$$p^*(x, z) = p(x)p(z|x, y).$$

The answer to our question is given by the marginal distribution

$$p^*(z) = \sum_x p^*(x, z) = \sum_x p(x)p(z|x, y).$$

We shall denote this as $\mathbb{P}(Z = z|Y := y)$ or $p(z|Y := y)$. We call $\mathbb{P}(Z = z|Y = y)$ **conditioning by observation** or **passive conditioning.** We call $\mathbb{P}(Z = z|Y := y)$ **conditioning by intervention** or **active conditioning.**

Passive conditioning is used to answer a predictive question like:

"Given that Joe smokes, what is the probability he will get lung cancer?"

Active conditioning is used to answer a causal question like:

"If Joe quits smoking, what is the probability he will get lung cancer?"

Consider a pair $(\mathcal{G}, \mathbb{P})$ where $\mathcal{G}$ is a DAG and $\mathbb{P}$ is a distribution for the variables V of the DAG. Let p denote the probability function for $\mathbb{P}$. Consider intervening and fixing a variable X to be equal to x. We represent the intervention by doing two things:

(1) Create a new DAG $\mathcal{G}^*$ by removing all arrows pointing into X;

(2) Create a new distribution $p^*(v) = \mathbb{P}(V = v|X := x)$ by removing the term $p(x|\pi_X)$ from $p(v)$.

The new pair $(\mathcal{G}^*, P^*)$ represents the intervention "set $X = x$."

Example 20.13 *You may have noticed a correlation between rain and having a wet lawn, that is, the variable "Rain" is not independent of the variable "Wet Lawn" and hence* $p_{R,W}(r,w) \neq p_R(r)p_W(w)$ *where* R *denotes Rain and* W *denotes Wet Lawn. Consider the following two DAGS:*

$$\text{Rain} \longrightarrow \text{Wet Lawn} \qquad \text{Rain} \longleftarrow \text{Wet Lawn}.$$

The first DAG implies that $p(w,r) = p(r)p(w|r)$ *while the second implies that* $p(w,r) = p(w)p(r|w)$ *No matter what the joint distribution* $p(w,r)$ *is, both graphs are correct. Both imply that* R *and* W *are not independent. But, intuitively, if we want a graph to indicate causation, the first graph is right and the second is wrong. Throwing water on your lawn doesn't cause rain. The reason we feel the first is correct while the second is wrong is because the interventions implied by the first graph are correct.*

Look at the first graph and form the intervention $W = 1$ *where 1 denotes "wet lawn." Following the rules of intervention, we break the arrows into* W *to get the modified graph:*

$$\text{Rain} \qquad \boxed{\textbf{set} \ \text{Wet Lawn} = 1}$$

with distribution $p^*(r) = p(r)$. *Thus* $\mathbb{P}(R = r \mid W := w) = \mathbb{P}(R = r)$ *tells us that "wet lawn" does not cause rain.*

Suppose we (wrongly) assume that the second graph is the correct causal graph and form the intervention $W = 1$ *on the second graph. There are no arrows into* W *that need to be broken so the intervention graph is the same as the original graph. Thus* $p^*(r) = p(r|w)$ *which would imply that changing "wet" changes "rain." Clearly, this is nonsense.*

Both are correct probability graphs but only the first is correct causally. We know the correct causal graph by using background knowledge.

Remark 20.14 *We could try to learn the correct causal graph from data but this is dangerous. In fact it is impossible with two variables. With more than two variables there are methods that can find the causal graph under certain assumptions but they are large sample methods and, furthermore, there is no way to ever know if the sample size you have is large enough to make the methods reliable.*

We can use DAG's to represent confouding variables. If X is a treatment and Y is an outcome, a confounding variable Z is a variable with arrows into both X and Y; see Figure 20.10. It is easy to check, using the formalism of interventions, that the following facts are true.

In a randomized study, the arrow betwen Z and X is broken. In this case, even with Z unobserved (represented by enclosing Z in a circle), the causal relationship between X and Y is estimable because it can be shown that $\mathbb{E}(Y|X := x) = \mathbb{E}(Y|X = x)$ which does not involve the unobserved Z. In an observational study, with all confounders observed, we get $\mathbb{E}(Y|X := x) = \int \mathbb{E}(Y|X = x, Z = z)dF_Z(z)$ as in formula (19.8). If Z is unobserved then we cannot estimate the causal effect because $\mathbb{E}(Y|X := x) = \int \mathbb{E}(Y|X = x, Z = z)dF_Z(z)$ involves the unobserved Z. We can't just use X and Y since in this case. $\mathbb{P}(Y = y|X = x) \neq \mathbb{P}(Y = y|X := x)$ which is just another way of saying that causation is not association.

In fact, we can make a precise connection between DAG's and counterfactuals as follows. Suppose that X and Y are binary. Define the confounding variable Z by

$$Z = \begin{cases} 1 & \text{if } (C_0, C_1) = (0,0) \\ 2 & \text{if } (C_0, C_1) = (0,1) \\ 3 & \text{if } (C_0, C_1) = (1,0) \\ 4 & \text{if } (C_0, C_1) = (1,1). \end{cases}$$

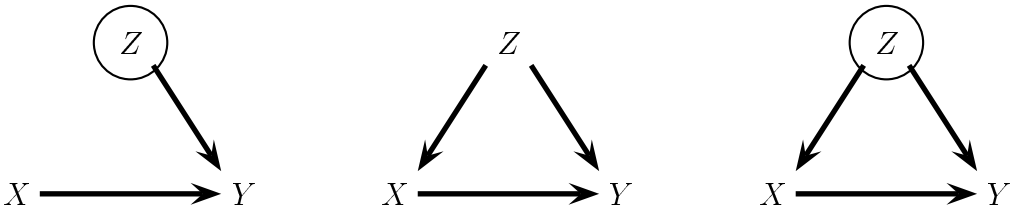


FIGURE 20.10. Randomized study; Observational study with measured confounders; Observational study with unmeasured confounders. The circled variables are unobserved.

From this, you can make the correspondence between the DAG approach and the counterfactual approach explicit. I leave this for the interested reader.

20.7 Bibliographic Remarks

There are a number of texts on DAG's including Edwards (1996) and Jordan (2003). The first use of DAG's for representing causal relationships was by Wright (1934). Modern tratments are contained in Spirtes, Glymour, and Scheines (1990) and Pearl (2000). Robins, Scheines, Spirtes and Wasserman (2003) discuss the problems with estimating causal structure from data.

20.8 Exercises

1. Consider the three DAG's in Figure 20.5 without a collider. Prove that $X \amalg Z|Y$.

2. Consider the DAG in Figure 20.5 with a collider. Prove that $X \amalg Z$ and that X and Z are dependent given Y.

3. Let $X \in \{0,1\}$, $Y \in \{0,1\}$, $Z \in \{0,1,2\}$. Suppose the distribution of (X,Y,Z) is Markov to:

$$X \longrightarrow Y \longrightarrow Z$$

Create a joint distribution $p(x, y, z)$ that is Markov to this DAG. Generate 1000 random vectors from this distribution. Estimate the distribution from the data using maximum likelihood. Compare the estimated distribution to the true distribution. Let $\theta = (\theta_{000}, \theta_{001}, \ldots, \theta_{112})$ where $\theta_{rst} = \mathbb{P}(X = r, Y = s, Z = t)$. Use the bootstrap to get standard errors and 95 per cent confidence intervals for these 12 parameters.

4. Let $V = (X, Y, Z)$ have the following joint distribution

$$\begin{aligned} X &\sim \text{Bernoulli}\left(\frac{1}{2}\right) \\ Y \mid X = x &\sim \text{Bernoulli}\left(\frac{e^{4x-2}}{1+e^{4x-2}}\right) \\ Z \mid X = x, Y = y &\sim \text{Bernoulli}\left(\frac{e^{2(x+y)-2}}{1+e^{2(x+y)-2}}\right). \end{aligned}$$

(a) Find an expression for $\mathbb{P}(Z = z \mid Y = y)$. In particular, find $\mathbb{P}(Z = 1 \mid Y = 1)$.

(b) Write a program to simulate the model. Conduct a simulation and compute $\mathbb{P}(Z = 1 \mid Y = 1)$ empirically. Plot this as a function of the simulation size N. It should converge to the theoretical value you computed in (a).

(c) Write down an expression for $\mathbb{P}(Z = 1 \mid Y := y)$. In particular, find $\mathbb{P}(Z = 1 \mid Y := 1)$.

(d) Modify your program to simulate the intervention "set $Y = 1$." Conduct a simulation and compute $\mathbb{P}(Z = 1 \mid Y := 1)$ empirically. Plot this as a function of the simulation size N. It should converge to the theoretical value you computed in (c).

5. This is a continuous, Gaussian version of the last question. Let $V = (X, Y, Z)$ have the following joint distribution

$$\begin{aligned} X &\sim \text{Normal } (0, 1) \\ Y \mid X = x &\sim \text{Normal } (\alpha x, 1) \\ Z \mid X = x, Y = y &\sim \text{Normal } (\beta y + \gamma x, 1). \end{aligned}$$

Here, α, β and γ are fixed parameters. Economists refer to models like this as *structural equation models*.

(a) Find an explicit expression for $f(z \mid y)$ and $\mathbb{E}(Z \mid Y = y) = \int z f(z \mid y) dz$.

(b) Find an explicit expression for $f(z \mid Y := y)$ and then find $\mathbb{E}(Z \mid Y := y) \equiv \int z f(z \mid Y := y) dy$. Compare to (b).

(c) Find the joint distribution of (Y, Z). Find the correlation ρ between Y and Z.

(d) Suppose that X is not observed and we try to make causal conclusions from the marginal distribution of (Y, Z). (Think of X as unobserved confounding variables.) In particular, suppose we declare that Y causes Z if $\rho \neq 0$ and

we declare that Y does not cause Z if $\rho = 0$. Show that this will lead to erroneous conclusions.

(e) Suppose we conduct a randomized experiment in which Y is randomly assigned. To be concrete, suppose that

$$\begin{aligned} X &\sim \text{Normal}(0,1) \\ Y &\sim \text{Normal}(\alpha,1) \\ Z \mid X = x, Y = y &\sim \text{Normal}(\beta y + \gamma x, 1). \end{aligned}$$

Show that the method in (d) now yields correct conclusions i.e. $\rho = 0$ if and only if $f(z \mid Y := y)$ does not depend on y.

21
Nonparametric Curve Estimation

In this Chapter we discuss nonparametric estimation of probability density functions and regression functions which we refer to as **curve estimation**.

In Chapter 8 we saw that it is possible to consistently estimate a cumulative distribution function F without making any assumptions about F. If we want to estimate a probability density function $f(x)$ or a regression function $r(x) = \mathbb{E}(Y|X = x)$ the situation is different. We cannot estimate these functions consistently without making some smoothness assumptions. Correspondingly, we will perform some sort of smoothing operation on the data.

A simple example of a density estimator is a **histogram**, which we discuss in detail in Section 21.2. To form a histogram estimator of a density f, we divide the real line to disjoint sets called **bins**. The histogram estimator is piecewise constant function where the height of the function is proportional to number of observations in each bin; see Figure 21.3. The number of bins is an example of a **smoothing parameter**. If we smooth too

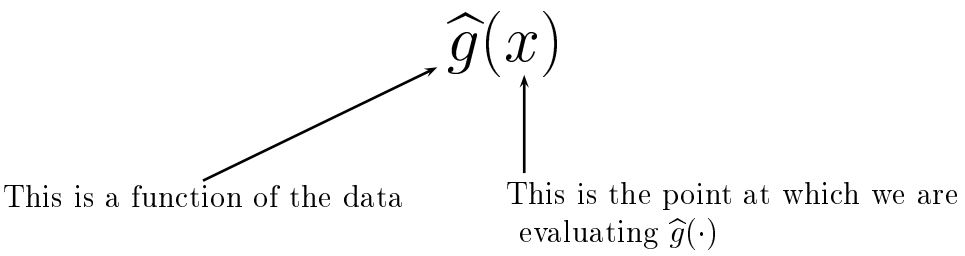


FIGURE 21.1. A curve estimate $\widehat{g}$ is random because it is a function of the data. The point x at which we evaluate $\widehat{g}$ is not a random variable.

much (large bins) we get a highly biased estimator while if we smooth too little (small bins) we get a highly variable estimator. Much of curve estimation is concerned with trying to optimally balance variance and bias.

21.1 The Bias-Variance Tradeoff

Let g denote an unknown function and let $\widehat{g}_n$ denote an estimator of g. Bear in mind that $\widehat{g}_n(x)$ is a random function evaluated at a point x. $\widehat{g}_n$ is random because it depends on the data. Indeed, we could be more explicit and write $\widehat{g}_n(x) = h_x(X_1, \ldots, X_n)$ to show that $\widehat{g}_n(x)$ is a function of the data $X_1, \ldots, X_n$ and that the function could be different for each x. See Figure 21.1.

As a loss function, we will use the **integrated squared error (ISE)**:

$$L(g, \widehat{g}_n) = \int (g(u) - \widehat{g}_n(u))^2 \, du. \tag{21.1}$$

The **risk** or **mean integrated squared error (MISE)** is

$$R(f, \widehat{f}) = \mathbb{E}\Big(L(g, \widehat{g})\Big). \tag{21.2}$$

Lemma 21.1 *The risk can be written as*

$$R(g, \widehat{g}_n) = \int b^2(x)\, dx + \int v(x)\, dx \tag{21.3}$$

where

$$b(x) = \mathbb{E}(\widehat{g}_n(x)) - g(x) \tag{21.4}$$

is the bias of $\widehat{g}_n(x)$ *at a fixed* x *and*

$$v(x) = \mathbb{V}(\widehat{g}_n(x)) = \mathbb{E}\Big((\widehat{g}_n(x) - \mathbb{E}(\widehat{g}_n(x))^2)\Big) \tag{21.5}$$

is the variance of $\widehat{g}_n(x)$ *at a fixed* x.

In summary,

$$\text{RISK} = \text{BIAS}^2 + \text{VARIANCE}. \tag{21.6}$$

When the data are over-smoothed, the bias term is large and the variance is small. When the data are under-smoothed the opposite is true; see Figure 21.2. This is called the **bias-variance trade-off**. Minimizing risk corresponds to balancing bias and variance.

21.2 Histograms

Let $X_1, \ldots, X_n$ be IID on $[0, 1]$ with density f. The restriction to $[0, 1]$ is not crucial; we can always rescale the data to be on this interval. Let m be an integer and define **bins**

$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \ \ldots, \ B_m = \left[\frac{m-1}{m}, 1\right]. \tag{21.7}$$

Define the **binwidth** $h = 1/m$, let ν_j be the number of observations in B_j, let $\widehat{p}_j = \nu_j/n$ and let $p_j = \int_{B_j} f(u)du$.

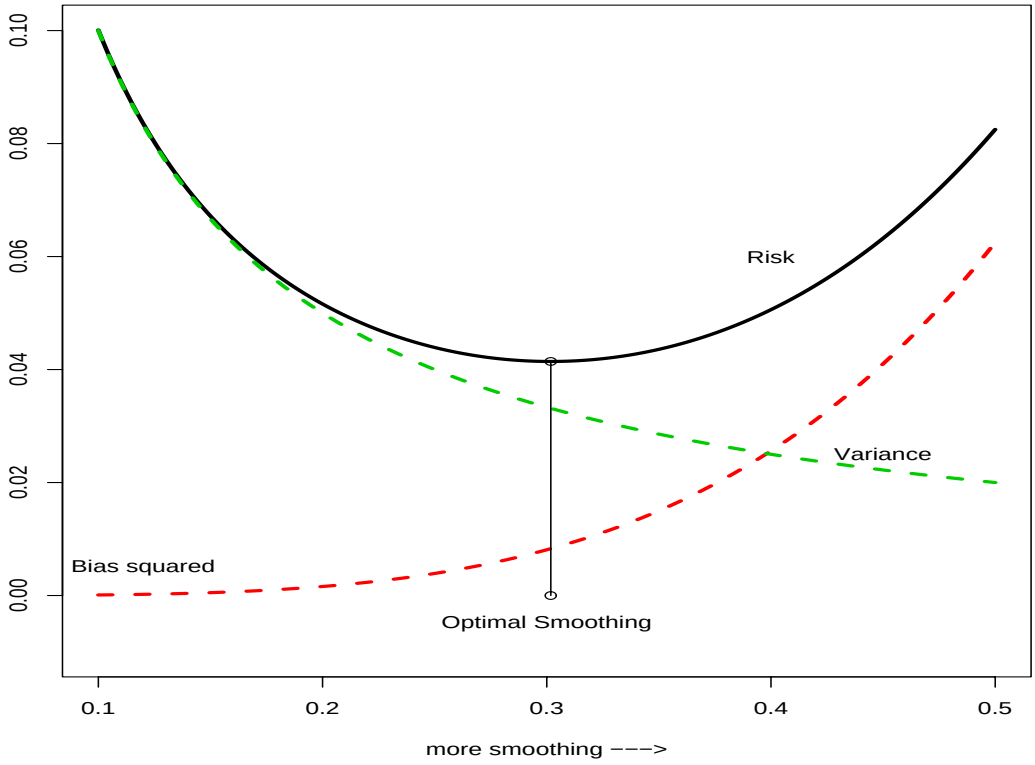


FIGURE 21.2. The Bias-Variance trade-off. The bias increases and the variance decreases with the amount of smoothing. The optimal amount of smoothing, indicated by the vertical line, minimizes the risk = bias^2 + variance.

The **histogram estimator** is defined by

$$\widehat{f}_n(x) = \begin{cases} \widehat{p}_1/h & x \in B_1 \\ \widehat{p}_2/h & x \in B_2 \\ \vdots & \vdots \\ \widehat{p}_m/h & x \in B_m \end{cases}$$

which we can write more succinctly as

$$\widehat{f}_n(x) = \sum_{j=1}^{n} \frac{\widehat{p}_j}{h} I(x \in B_j). \tag{21.8}$$

To understand the motivation for this estimator, let $p_j = \int_{B_j} f(u)du$ and note that, for $x \in B_j$ and h small,

$$\widehat{f}_n(x) = \frac{\widehat{p}_j}{h} \approx \frac{p_j}{h} = \frac{\int_{B_j} f(u)du}{h} \approx \frac{f(x)h}{f(x)} = f(x).$$

Example 21.2 *Figure 21.3 shows three different histograms based on $n = 1,266$ data points from an astronomical sky survey. Each data point represents the distance from us to a galaxy. The galaxies lie on a "pencilbeam" pointing directly from the Earth out into space. Because of the finite speed of light, looking at galaxies farther and farther away corresponds to looking back in time. Choosing the right number of bins involves finding a good tradeoff between bias and variance. We shall see later that the top left histogram has too few bins resulting in oversmoothing and too much bias. The bottom left histogram has too many bins resulting in undersmoothing and too few bins. The top right histogram is just right. The histogram reveals the presence of clusters of galaxies. Seeing how the size and number of galaxy clusters varies with time, helps cosmologists understand the evolution of the universe.* ■

The mean and variance of $\widehat{f}_n(x)$ are given in the following Theorem.

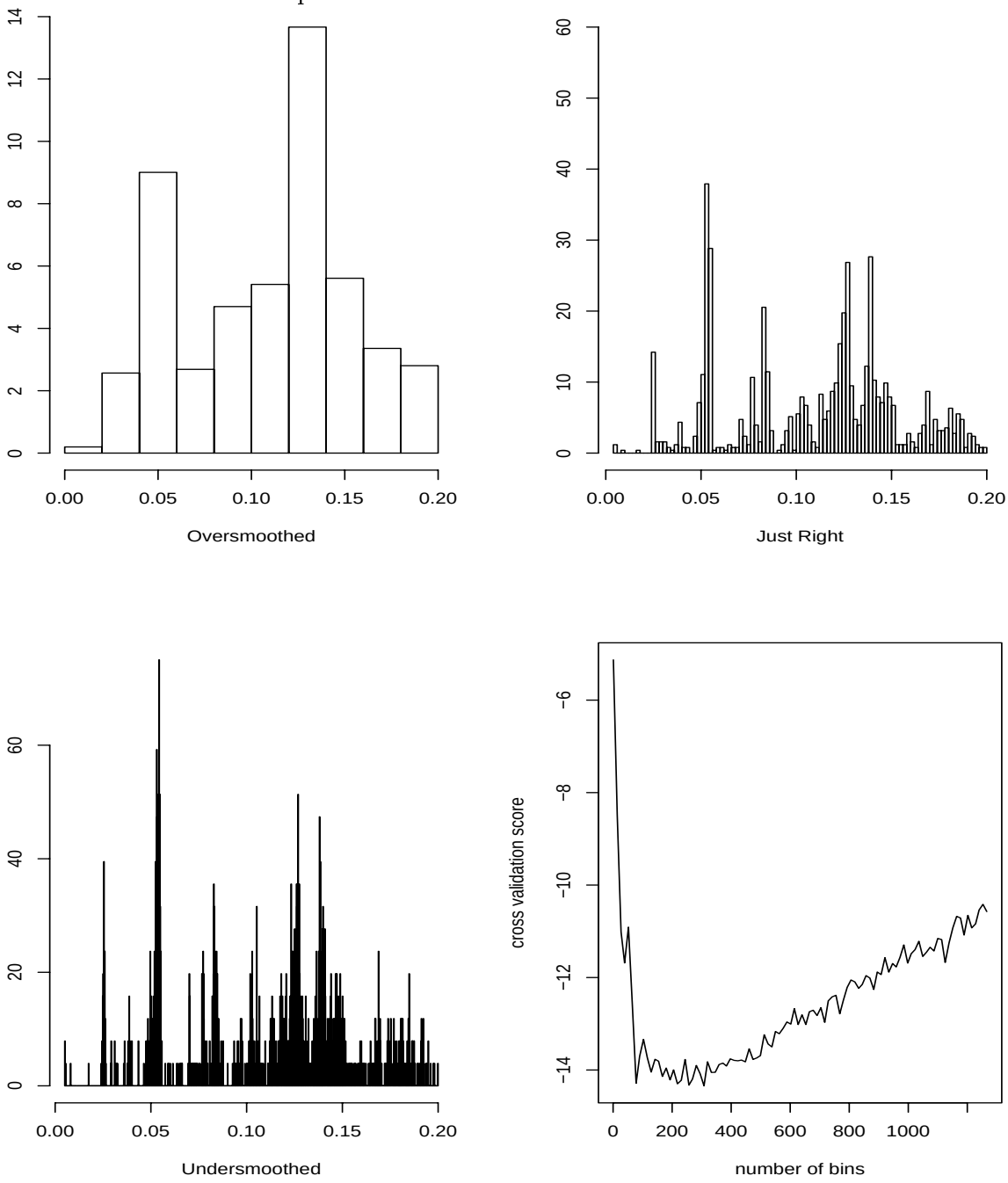


FIGURE 21.3. Three versions of a histogram for the astronomy data. The top left histogram has too few bins. The bottom left histogram has too many bins. The top right histogram is just right. The lower, right plot shows the estimated risk versus the number of bins.

Theorem **21.3** *Consider fixed x and fixed m, and let B_j be the bin containing x. Then,*

$$\mathbb{E}(\widehat{f}_n(x)) = \frac{p_j}{h} \quad \text{and} \quad \mathbb{V}(\widehat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}. \tag{21.9}$$

Let's take a closer look at the bias-variance tradeoff using equation (21.9) . Consider some $x \in B_j$. For any other $u \in B_j$,

$$f(u) \approx f(x) + (u-x)f'(x)$$

and so

$$\begin{aligned} p_j = \int_{B_j} f(u)du &\approx \int_{B_j} \Big(f(x) + (u-x)f'(x)\Big)du \\ &= f(x)h + hf'(x)\left(h\left(j - \frac{1}{2}\right) - x\right). \end{aligned}$$

Therefore, the bias $b(x)$ is

$$\begin{aligned} b(x) &= \mathbb{E}(\widehat{f}_n(x)) - f(x) = \frac{p_j}{h} - f(x) \\ &\approx \frac{f(x)h + hf'(x)\left(h\left(j-\frac{1}{2}\right) - x\right)}{h} - f(x) \\ &= f'(x)\left(h\left(j - \frac{1}{2}\right) - x\right). \end{aligned}$$

If $\widetilde{x}_j$ is the center of the bin, then

$$\begin{aligned} \int_{B_j} b^2(x)\,dx &\approx \int_{B_j} (f'(x))^2 \left(h\left(j-\frac{1}{2}\right) - x\right)^2 dx \\ &\approx (f'(\widetilde{x}_j))^2 \int_{B_j} \left(h\left(j-\frac{1}{2}\right) - x\right)^2 dx \\ &= (f'(\widetilde{x}_j))^2 \frac{h^3}{12}. \end{aligned}$$

Therefore,

$$\begin{aligned} \int_0^1 b^2(x)dx &= \sum_{j=1}^m \int_{B_j} b^2(x)dx \approx \sum_{j=1}^m (f'(\widetilde{x}_j))^2 \frac{h^3}{12} \\ &= \frac{h^2}{12} \sum_{j=1}^m h\,(f'(\widetilde{x}_j))^2 \approx \frac{h^2}{12} \int_0^1 (f'(x))^2 dx. \end{aligned}$$

Note that this increases as a function of h. Now consider the variance. For h small, $1 - p_j \approx 1$, so

$$\begin{aligned} v(x) &\approx \frac{p_j}{nh^2} \\ &= \frac{f(x)h + hf'(x)\left(h\left(j - \frac{1}{2}\right) - x\right)}{nh^2} \\ &\approx \frac{f(x)}{nh} \end{aligned}$$

where we have kept only the dominant term. So,

$$\int_0^1 v(x)dx \approx \frac{1}{nh}.$$

Note that this decreases with h. Putting all this together, we get:

Theorem 21.4 *Suppose that* $\int (f'(u))^2 du < \infty$. *Then*

$$R(\widehat{f}_n, f) \approx \frac{h^2}{12}\int (f'(u))^2 du + \frac{1}{nh}. \tag{21.10}$$

The value h^* *that minimizes (21.10) is*

$$h^* = \frac{1}{n^{1/3}}\left(\frac{6}{\int (f'(u))^2 du}\right)^{1/3}. \tag{21.11}$$

With this choice of binwidth,

$$R(\widehat{f}_n, f) \approx \frac{C}{n^{2/3}} \tag{21.12}$$

where $C = (3/4)^{2/3}\left(\int (f'(u))^2 du\right)^{1/3}$.

Theorem 21.4 is quite revealing. We see that with an optimally chosen binwidth, the MISE decreases to 0 at rate $n^{-2/3}$. By comparison, most parametric estimators converge at rate

n^{-1}. The slower rate of convergence is the price we pay for being nonparametric. The formula for the optimal binwidth h^* is of theoretical interest but it is not useful in practice since it depends on the unknown function f.

A practical way to choose the binwidth is to estimate the risk function and minimize over h. Recall that the loss function, which we now write as a function of h, is

$$\begin{aligned} L(h) &= \int (\widehat{f}_n(x) - f(x))^2 \, dx \\ &= \int \widehat{f}_n^2(x) \, dx - 2 \int \widehat{f}_n(x) f(x) dx + \int f^2(x) \, dx. \end{aligned}$$

The last term does not depend on the binwidth h so minimizing the risk is equivalent to minimizing the expected value of

$$J(h) = \int \widehat{f}_n^2(x) \, dx - 2 \int \widehat{f}_n(x) f(x) dx.$$

We shall refer to $\mathbb{E}(J(h))$ as the risk, although it differs from the true risk by the constant term $\int f^2(x) \, dx$.

Definition 21.5 *The* **cross-validation estimator of risk** *is*

$$\widehat{J}(h) = \int \left(\widehat{f}_n(x) \right)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_{(-i)}(X_i) \qquad (21.13)$$

where $\widehat{f}_{(-i)}$ is the histogram estimator obtained after removing the i^{th} observation. We refer to $\widehat{J}(h)$ as the cross-validation score or estimated risk.

Theorem **21.6** *The cross-validation estimator is nearly unbiased:*

$$\mathbb{E}(\widehat{J}(x)) \approx \mathbb{E}(J(x)).$$

In principle, we need to recompute the histogram n times to compute $\widehat{J}(h)$. Moreover, this has to be done for all values of h. Fortunately, there is a shortcut formula.

Theorem **21.7** *The following identity holds:*

$$\widehat{J}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)} \sum_{j=1}^{m} \widehat{p}_j^2. \tag{21.14}$$

Example 21.8 *We used cross-validation in the astronomy example. The cross-validation function is quite flat near its minimum. Any m in the range of 73 to 310 is an approximate minimizer but the resulting histogram does not change much over this range. The histogram in the top right plot in Figure 21.3 was constructed using $m = 73$ bins. The bottom right plot shows the estimated risk, or more precisely, $\widehat{A}$, plotted versus the number of bins.*

Next we want some sort of confidence set for f. Suppose $\widehat{f}_n$ is a histogram with m bins and binwidth $h = 1/m$. We cannot realistically make confidence statements about the fine details of the true density f. Instead, we shall make confidence statements about f at the resolution of the histogram. To this end, define

$$\widetilde{f}(x) = \frac{p_j}{h} \quad \text{for } x \in B_j \tag{21.15}$$

where $p_j = \int_{B_j} f(u)du$ which is a "histogramized" version of f.

Theorem **21.9** *Let $m = m(n)$ be the number of bins in the histogram $\widehat{f}_n$. Assume that $m(n) \to \infty$ and $m(n) \log n/n \to 0$ as $n \to \infty$. Define*

$$\begin{aligned} \ell(x) &= \left(\max\left\{ \sqrt{\widehat{f}_n(x)} - c, 0 \right\} \right)^2 \\ u(x) &= \left(\sqrt{\widehat{f}_n(x)} + c \right)^2 \end{aligned} \tag{21.16}$$

where

$$c = \frac{z_{\alpha/(2m)}}{2}\sqrt{\frac{m}{n}}. \tag{21.17}$$

Then,

$$\lim_{n\to\infty} \mathbb{P}\Big(\ell(x) \le \widetilde{f}(x) \le u(x) \ \text{ for all } x\Big) \ge 1-\alpha. \tag{21.18}$$

PROOF. Here is an outline of the proof. A rigorous proof requires fairly sophisticated tools. From the central limit theorem, $\widehat{p}_j \approx N(p_j, p_j(1-p_j)/n)$. By the delta method, $\sqrt{\widehat{p}_j} \approx N(\sqrt{p_j}, 1/(4n))$. Moreover, it can be shown that the $\sqrt{\widehat{p}_j}$'s are approximately independent. Therefore,

$$2\sqrt{n}\Big(\sqrt{\widehat{p}_j} - \sqrt{p}_j\Big) \approx Z_j \tag{21.19}$$

where $Z_1, \ldots, Z_m \sim N(0,1)$. Let A be the event that $\ell(x) \le \widetilde{f}(x) \le u(x)$ for all x. So,

$$\begin{aligned}
A &= \Big\{\ell(x) \le \widetilde{f}(x) \le u(x) \ \text{ for all } x\Big\} \\
&= \Big\{\sqrt{\ell(x)} \le \sqrt{\widetilde{f}(x)} \le \sqrt{u(x)} \ \text{ for all } x\Big\} \\
&= \Big\{\sqrt{\widehat{f}_n(x)} - c \le \sqrt{\widetilde{f}(x)} \le \sqrt{\widehat{f}_n(x)} + c \ \text{ for all } x\Big\} \\
&= \Big\{\max_x \Big|\sqrt{\widehat{f}_n(x)} - \sqrt{\widetilde{f}(x)}\Big| \le c\Big\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}(A^c) &= \mathbb{P}\left(\max_x \left|\sqrt{\widehat{f}_n(x)} - \sqrt{\widetilde{f}(x)}\right| > c\right) = \mathbb{P}\left(\max_j \left|\sqrt{\frac{\widehat{p}_j}{h}} - \sqrt{\frac{p_j}{h}}\right| > c\right) \\
&= \mathbb{P}\left(\max_j \left|\sqrt{\widehat{p}_j} - \sqrt{p_j}\right| > c\sqrt{h}\right) = \mathbb{P}\left(\max_j 2\sqrt{n}\left|\sqrt{\widehat{p}_j} - \sqrt{p_j}\right| > 2c\sqrt{hn}\right) \\
&= \mathbb{P}\left(\max_j 2\sqrt{n}\left|\sqrt{\widehat{p}_j} - \sqrt{p_j}\right| > z_{\alpha/(2m)}\right) \\
&\approx \mathbb{P}\left(\max_j |Z_j| > z_{\alpha/(2m)}\right) \le \sum_{j=1}^{m} \mathbb{P}\left(|Z_j| > z_{\alpha/(2m)}\right)
\end{aligned}$$

$$= \sum_{j=1}^{m} \frac{\alpha}{m} = \alpha. \quad \blacksquare$$

Example 21.10 *Figure 21.4 shows a 95 per cent confidence envelope for the astronomy data. We see that even with over 1,000 data points, there is still substantial uncertainty.* ■

21.3 Kernel Density Estimation

Histograms are discontinuous. **Kernel density estimators** are smoother and they converge faster to the true density than histograms.

Let $X_1, \ldots, X_n$ denote the observed data, a sample from f. In this chapter, a **kernel** is defined to be any smooth function K such that $K(x) \geq 0$, $\int K(x)\,dx = 1$, $\int xK(x)dx = 0$ and $\sigma_K^2 \equiv \int x^2 K(x)dx > 0$. Two examples of kernels are the **Epanechnikov kernel**

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2/5)/\sqrt{5} & |x| < \sqrt{5} \\ 0 & otherwise \end{cases} \tag{21.20}$$

and the Gaussian (Normal) kernel $K(x) = (2\pi)^{-1/2} e^{-x^2/2}$.

Definition 21.11 *Given a kernel K and a positive number h, called the* **bandwidth**, *the* **kernel density estimator** *is defined to be*

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right). \tag{21.21}$$

An example of a kernel density estimator is show in Figure 21.5. The kernel estimator effectively puts a smoothed out lump of mass of size $1/n$ over each data point X_i. The bandwidth h

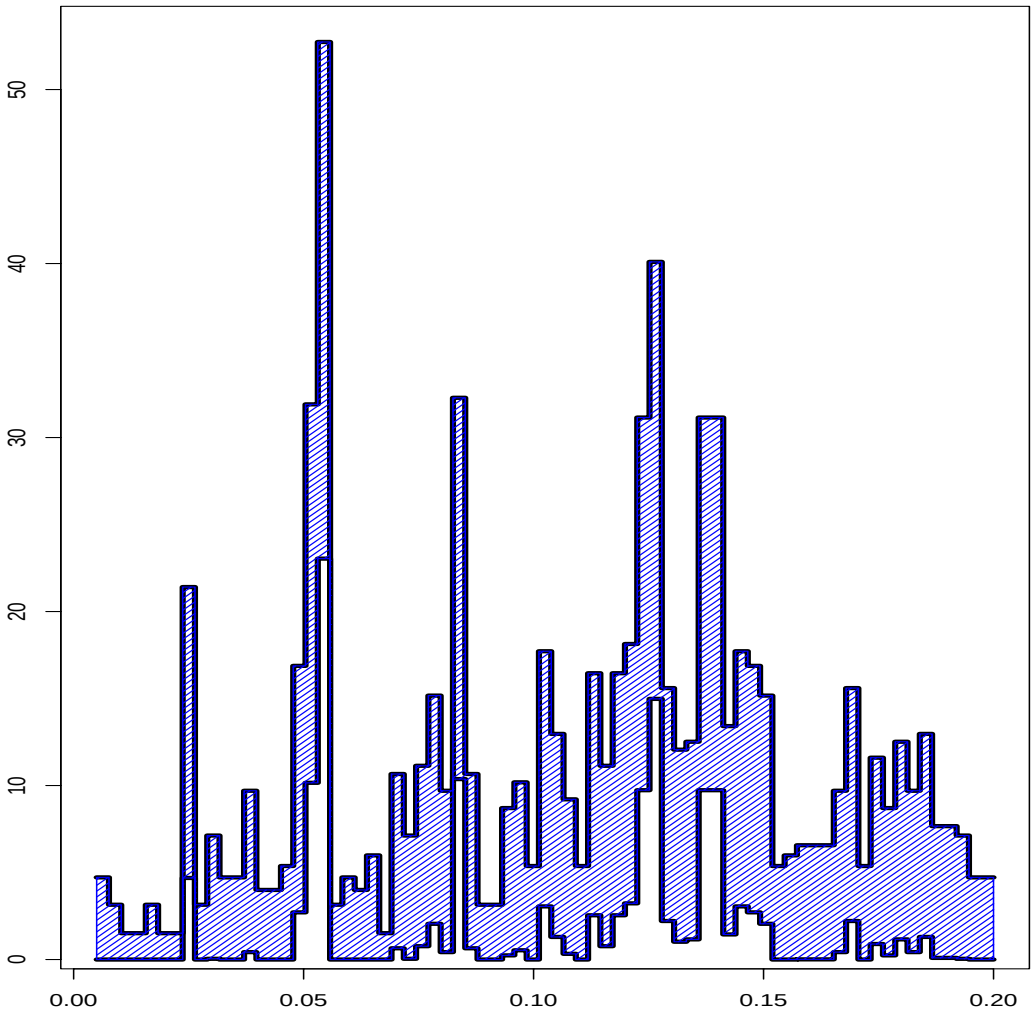


FIGURE 21.4. 95 per cent confidence envelope for astronomy data using $m = 73$ bins.

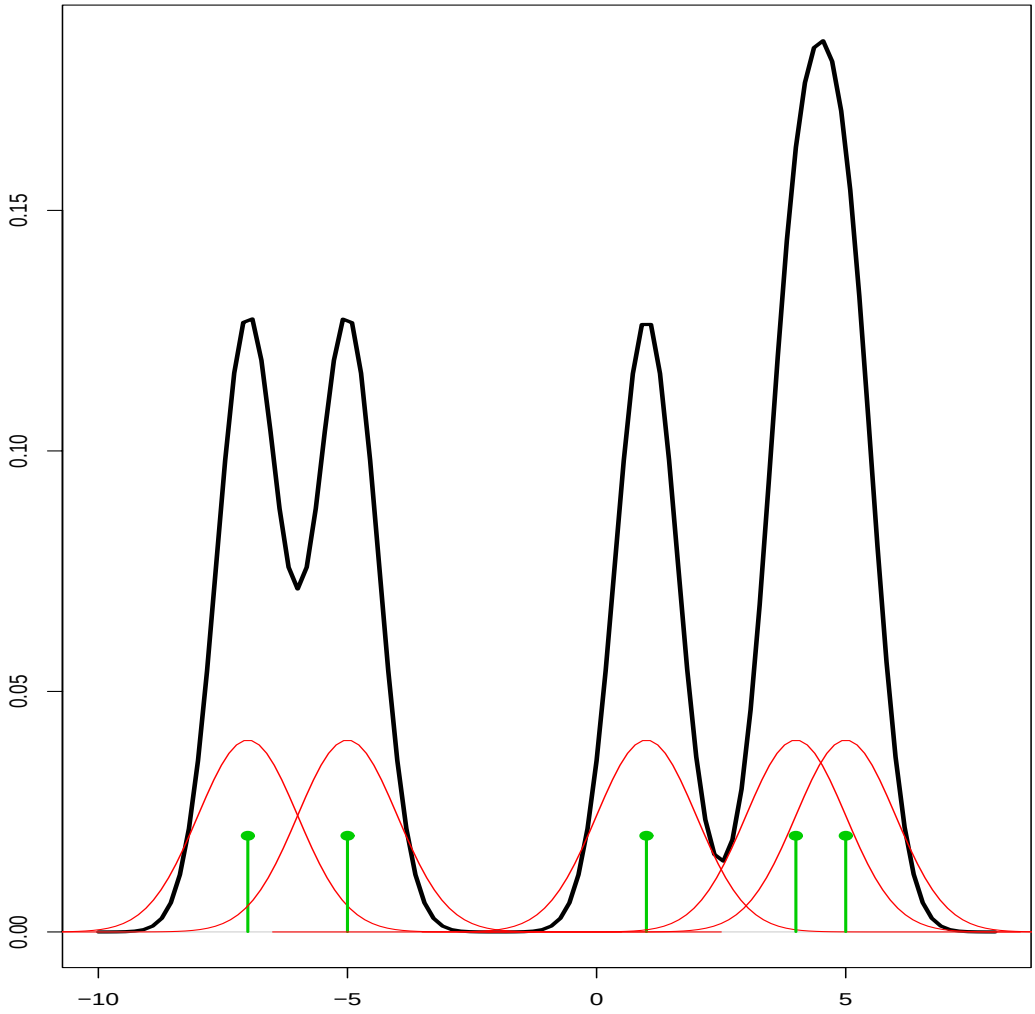


FIGURE 21.5. A kernel density estimator $\widehat{f}$. At each point x, $\widehat{f}(x)$ is the average of the kernels centered over the data points X_i. The data points are indicated by short vertical bars.

controls the amount of smoothing. When h is close to 0, $\widehat{f}_n$ consists of a set of spikes, one at each data point. The height of the spikes tends to infinity as $h \to 0$. When $h \to \infty$, $\widehat{f}_n$ tends to a uniform density.

Example 21.12 *Figure 21.6 shows kernel density estimators for the astronomy data using for three different bandwidths. In each case we used a Gaussian kernel. The properly smoothed kernel density estimator in the top right panel shows similar structure as the histogram. However, it is easier to see the clusters in the kernel estimator.* ■

To construct a kernel density estimator, we need to choose a kernel K and a bandwidth h. It can be shown theoretically and empirically that the choice of K is not crucial. However, the choice of bandwidth h is very important. As with the histogram, we can make a theoretical statement about how the risk of the estimator depends on the bandwidth.

Theorem **21.13** *Under weak assumptions on f and K,*

$$R(f, \widehat{f}_n) \approx \frac{1}{4}\sigma_K^4 h^4 \int (f''(x))^2 + \frac{\int K^2(x)dx}{nh} \tag{21.22}$$

where $\sigma_K^2 = \int x^2 K(x)dx$. The optimal bandwidth is

$$h^* = \frac{c_1^{-2/5} c_2^{1/5} c_3^{-1/5}}{n^{1/5}} \tag{21.23}$$

where $c_1 = \int x^2 K(x)dx$, $c_2 = \int K(x)^2 dx$ and $c_3 = \int (f''(x))^2 dx$. With this choice of bandwidth,

$$R(f, \widehat{f}_n) \approx \frac{c_4}{n^{4/5}}$$

for some constant $c_4 > 0$.

PROOF. Write $K_h(x, X) = h^{-1} K\left((x - X)/h\right)$ and $\widehat{f}_n(x) = n^{-1} \sum_i K_h(x, X_i)$. Thus, $\mathbb{E}[\widehat{f}_n(x)] = \mathbb{E}[K_h(x, X)]$ and $\mathbb{V}[\widehat{f}_n(x)] =$

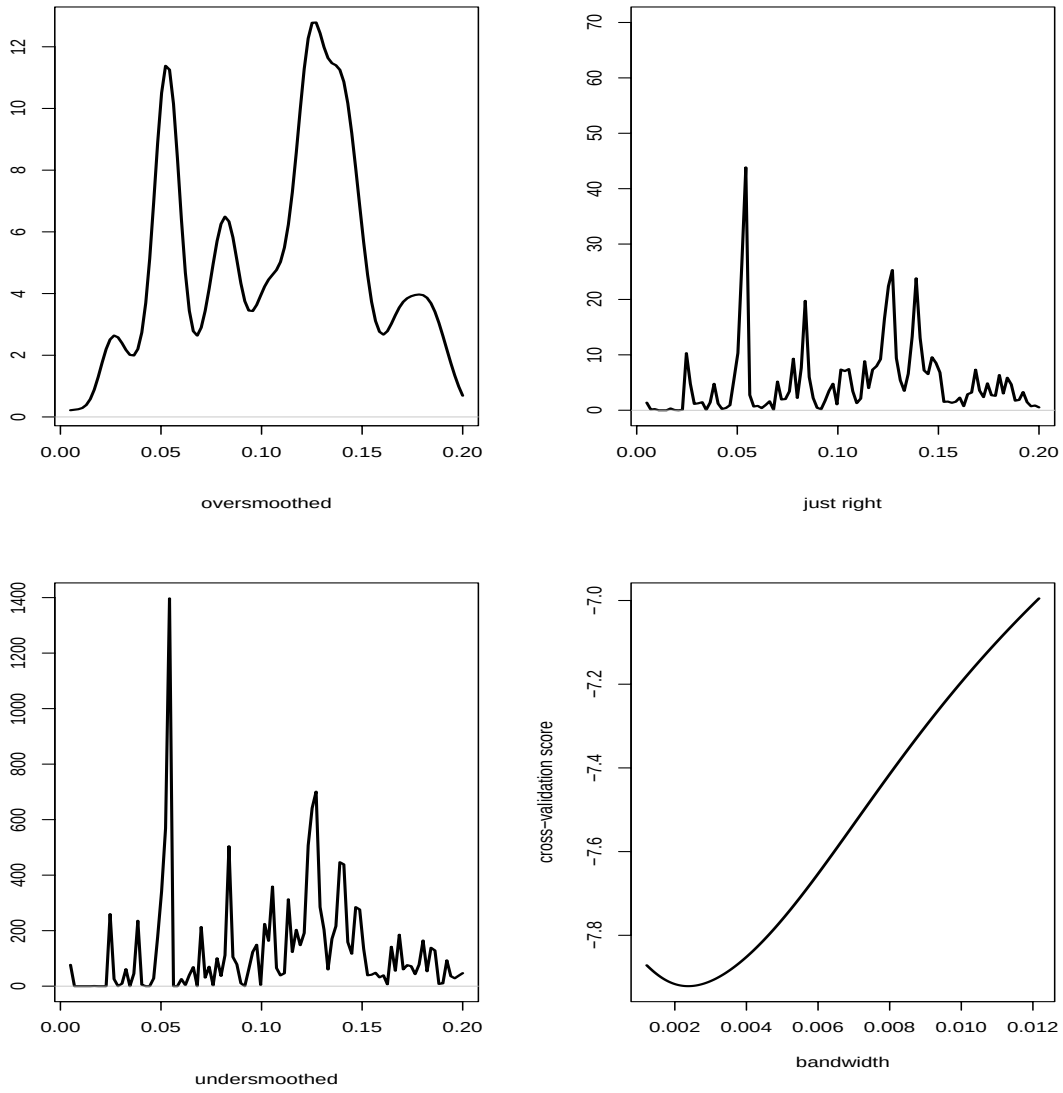


FIGURE 21.6. Kernel density estimators and estimated risk for the astronomy data.

$n^{-1}\mathbb{V}[K_h(x,X)]$. Now,

$$\begin{aligned}\mathbb{E}[K_h(x,X)] &= \int \frac{1}{h}K\left(\frac{x-t}{h}\right) f(t)\,dt \\ &= \int K(u) f(u-hu)\,du \\ &= \int K(u)\left[f(u) - hf'(u) + \frac{1}{2}f''(u) + \cdots\right] du \\ &= f(u) + \frac{1}{2}h^2 f''(u)\int u^2 K(u)\,du\cdots\end{aligned}$$

since $\int K(x)\,dx = 1$ and $\int x\,K(x)\,dx = 0$. The bias is

$$\mathbb{E}[K_h(x,X)] - f(x) \approx \frac{1}{2}\sigma_k^2 h^2 f''(x).$$

By a similar calculation,

$$\mathbb{V}[\widehat{f}_n(x)] \approx \frac{f(x)\int K^2(x)\,dx}{n\,h_n}.$$

The second result follows from integrating the squared bias plus the variance. ■

We see that kernel estimators converge at rate $n^{-4/5}$ while histograms converge at the slower rate $n^{-2/3}$. It can be shown that, under weak assumptions, there does not exist a nonparametric estimator that converges faster than $n^{-4/5}$.

The expression for h^* depends on the unknown density f which makes the result of little practical use. As with the histograms, we shall use cross-validation to find a bandwidth. Thus, we estimate the risk (up to a constant) by

$$\widehat{J}(h) = \int \widehat{f}^2(x)dz - 2\frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{-i}(X_i) \tag{21.24}$$

where $\widehat{f}_{-i}$ is the kernel density estimator after omitting the i^{th} observation.

Theorem **21.14** *For any $h > 0$,*

$$\mathbb{E}\left[\widehat{J}(h)\right] = \mathbb{E}\left[J(h)\right].$$

Also,

$$\widehat{J}(h) \approx \frac{1}{hn^2}\sum_i\sum_j K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0) \qquad (21.25)$$

where $K^*(x) = K^{(2)}(x) - 2K(x)$ and $K^{(2)}(z) = \int K(z-y)K(y)dy$. In particular, if K is a N(0,1) Gaussian kernel then $K^{(2)}(z)$ is the $N(0,2)$ density.

For large data sets, $\widehat{f}$ and (21.25) can be computed quickly using the fast Fourier transform.

We then choose the bandwidth h_n that minimizes $\widehat{J}(h)$. A justification for this method is given by the following remarkable theorem due to Stone.

Theorem **21.15 (Stone's Theorem.)** *Suppose that f is bounded. Let $\widehat{f}_h$ denote the kernel estimator with bandwidth h and let h_n denote the bandwidth chosen by cross-validation. Then,*

$$\frac{\int\left(f(x) - \widehat{f}_{h_n}(x)\right)^2 dx}{\inf_h \int\left(f(x) - \widehat{f}_h(x)\right)^2 dx} \xrightarrow{\text{P}} 1. \qquad (21.26)$$

Example 21.16 *The top right panel of Figure 21.6 is based on cross-validation. These data are rounded which problems for cross-validation. Specifically, it causes the minimizer to be $h = 0$. To overcome this problem, we added a small amount of random Normal noise to the data. The result is that $\widehat{J}(h)$ is very smooth with a well defined minimum.*

Remark 21.17 *Do not assume that, if the estimator $\widehat{f}$ is wiggly, then cross-validation has let you down. The eye is not a good judge of risk.*

To construct confidence bands, we use something similar to histograms although the details are more complicated. The version described here is from Chaudhuri and Marron (1999).

Confidence Band for Kernel Density Estimator

1. Choose an evenly spaced grid of points $\mathcal{V} = \{v_1, \ldots, v_N\}$. For every $v \in \mathcal{V}$, define

$$Y_i(v) = \frac{1}{h} K\left(\frac{v - X_i}{h}\right). \tag{21.27}$$

   Note that $\widehat{f}_n(v) = \overline{Y}_n(v)$, the average of the $Y_i(v)'s$.

2. Define

$$\mathsf{se}\,(v) = \frac{s(v)}{\sqrt{n}} \tag{21.28}$$

   where $s^2(v) = (n-1)^{-1} \sum_{i=1}^n (Y_i(v) - \overline{Y}_n(v))^2$.

3. Compute the effective sample size

$$ESS(v) = \frac{\sum_{i=1}^n K\left(\frac{(v-X_i)}{h}\right)}{K(0)}. \tag{21.29}$$

   Roughly, this is the number of data points being averaged together to form $\widehat{f}_n(v)$.

4. Let $\mathcal{V}^* = \{v \in \mathcal{V} : \ ESS(v) \geq 5\}$. Now define the number of independent blocks m by

$$\frac{1}{m} = \frac{\overline{ESS}}{n} \tag{21.30}$$

   where $\overline{ESS}$ is the average of ESS over $\mathcal{V}^*$.

5. Let

$$q = \Phi^{-1}\left(\frac{1 + (1-\alpha)^{1/m}}{2}\right) \tag{21.31}$$

   and define

$$\ell(v) = \widehat{f}_n(v) - q\,\mathsf{se}\,(v) \quad \text{and} \quad u(v) = \widehat{f}_n(v) + q\,\mathsf{se}\,(v) \tag{21.32}$$

   where we replace $\ell(v)$ with 0 if it becomes negative.

Then,

$$\mathbb{P}\Big\{\ell(v) \leq f(v) \leq u(v) \ \text{ for all } v\Big\} \approx 1 - \alpha.$$

Example 21.18 *Figure 21.7 shows approximate 95 per cent confidence bands for the astronomy data.* ■

Suppose now that the data $X_i = (X_{i1}, \ldots, X_{id})$ are d-dimensional. The kernel estimator can easily be generalized to d dimensions. Let $h = (h_1, \ldots, h_d)$ be a vector of bandwidths and define

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) \tag{21.33}$$

where

$$K_h(x - X_i) = \frac{1}{nh_1 \cdots h_d}\left\{\prod_{j=1}^{d} K\left(\frac{x_i - X_{ij}}{h_j}\right)\right\} \tag{21.34}$$

where $h_1, \ldots, h_d$ are bandwidths. For simplicity, we might take $h_j = s_j h$ where s_j is the standard deviation of the j^{th} variable. There is now only a single bandwidth h to choose. Using calculations like those in the one-dimensional case, the risk is given by

$$\begin{aligned} R(f, \widehat{f}_n) &\approx \frac{1}{4}\sigma_K^4 \left[\sum_{j=1}^{d} h_j^4 \int f_{jj}^2(x)dx + \sum_{j \neq k} h_j^2 h_k^2 \int f_{jj} f_{kk} dx\right] \\ &\quad + \frac{\left(\int K^2(x)dx\right)^d}{nh_1 \cdots h_d} \end{aligned}$$

where f_{jj} is the second partial derivative of f. The optimal bandwidth satisfies $h_i \approx c_1 n^{-1/(4+d)}$) leading to a risk of order $n^{-4/(4+d)}$. From this fact, we see that the risk increases quickly with dimension, a problem usually called the **curse of dimensionality**. To get a sense of how serious this problem is, consider the following table from Silverman (1986) which shows the sample size required to ensure a relative mean squared error less than 0.1 at 0 when the density is multivariate normal and the optimal bandwidth is selected.

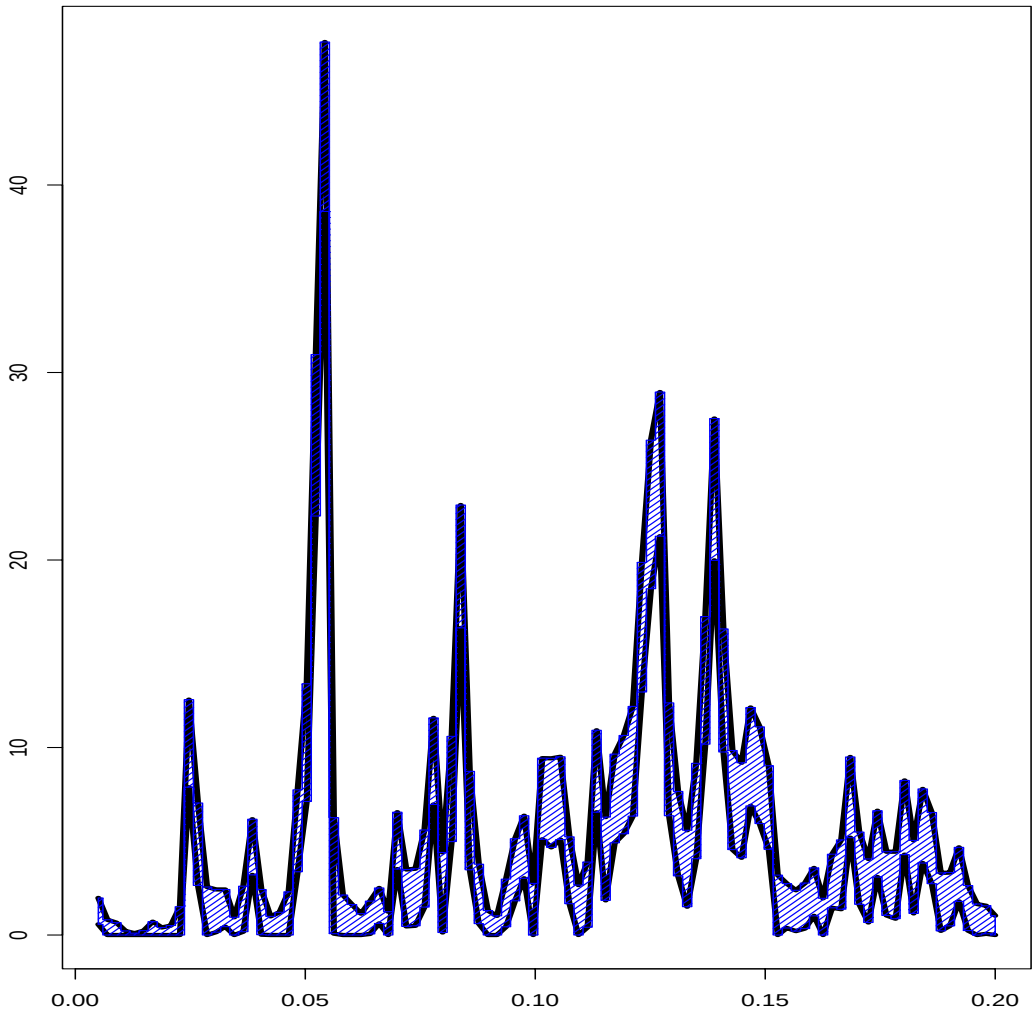


FIGURE 21.7. 95 per cent Confidence bands for kernel density estimate for the astronomy data.

Dimension	Sample Size
1	4
2	19
3	67
4	223
5	768
6	2790
7	10,700
8	43,700
9	187,000
10	842,000

This is bad news indeed. It says that having 842,000 observations in a ten dimensional problem is really like having 4 observations.

21.4 Nonparametric Regression

Consider pairs of points $(x_1, Y_1), \ldots, (x_n, Y_n)$ related by

$$Y_i = r(x_i) + \epsilon_i \tag{21.35}$$

where $\mathbb{E}(\epsilon_i) = 0$ and we are treating the x_i's as fixed. In nonparametric regression, we want to estimate the regression function $r(x) = \mathbb{E}(Y|X = x)$.

There are many nonparametric regression estimators. Most involve estimating $r(x)$ by taking some sort of weighted average of the Y_i's, giving higher weight to those points near x. In particular, the **Nadaraya-Watson kernel estimator** is defined by

$$\widehat{r}(x) = \sum_{i=1}^{n} w_i(x) Y_i \tag{21.36}$$

where the weights $w_i(x)$ are given by

$$w_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}. \tag{21.37}$$

The form of this estimator comes from first estimating the joint density $f(x,y)$ using kernel density estimation and then inserting this into:

$$r(x) = \mathbb{E}(Y|X=x) = \int y f(y|x) dy = \frac{\int y f(x,y) dy}{\int f(x,y) dy}.$$

Theorem 21.19 *Suppose that $\mathbb{V}(\epsilon_i) = \sigma^2$. The risk of the Nadaraya-Watson kernel estimator is*

$$\begin{aligned} R(\widehat{r}_n, r) &\approx \frac{h^4}{4}\left(\int x^2 K^2(x) dx\right)^4 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2 dx \\ &\quad + \int \frac{\sigma^2 \int K^2(x) dx}{nhf(x)} dx. \end{aligned} \tag{21.38}$$

The optimal bandwidth decreases at rate $n^{-1/5}$ and with this choice the risk decreases at rate $n^{-4/5}$.

In practice, to choose the bandwidth h we minimize the cross validation score

$$\widehat{J}(h) = \sum_{i=1}^{n} (Y_i - \widehat{r}_{-i}(x_i))^2 \tag{21.39}$$

where $\widehat{r}_{-i}$ is the estimator we get by omitting the i^{th} variable. An approximation to $\widehat{J}$ is given by

$$\widehat{J}(h) = \sum_{i=1}^{n} (Y_i - \widehat{r}(x_i))^2 \frac{1}{\left(1 - \frac{K(0)}{\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right)}\right)^2}. \tag{21.40}$$

Example 21.20 *Figures 21.8 shows cosmic microwave background (CMB) data from BOOMERaNG (Netterfield et al. 2001), Maxima (Lee et al. 2001) and DASI (Halverson 2001). The data consist of n pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$ where x_i is called the multipole moment and Y_i is the estimated power spectrum of the temperature fluctuations. What your seeing are sound waves in the cosmic microwave background radiation which is the heat, left over from the big bang. If $r(x)$ denotes the true power spectrum then*

$$Y_i = r(x_i) + \epsilon_i$$

where ϵ_i is a random error with mean 0. The location and size of peaks in $r(x)$ provides valuable clues about the behavior of the early universe. Figure 21.8 shows the fit based on cross-validation as well as an undersmoothed and oversmoothed fit. The cross-validation fit shows the presence of three well defined peaks, as predicted by the physics of the big bang. ■

The procedure for finding confidence bands is similar to that for density estimation. However, we first need to estimate σ^2. Suppose that the x_i's are ordered. Assuming $r(x)$ is smooth, we have $r(x_{i+1}) - r(x_i) \approx 0$ and hence

$$Y_{i+1} - Y_i = \Big[r(x_{i+1}) + \epsilon_{i+1}\Big] - \Big[r(x_i) + \epsilon_i\Big] \approx \epsilon_{i+1} - \epsilon_i$$

and hence

$$\mathbb{V}(Y_{i+1} - Y_i) \approx \mathbb{V}(\epsilon_{i+1} - \epsilon_i) = \mathbb{V}(\epsilon_{i+1}) + \mathbb{V}(\epsilon_i) = 2\sigma^2.$$

We can thus use the average of the $n-1$ differences $Y_{i+1} - Y_i$ to estimate σ^2. Hence, define

$$\widehat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2. \tag{21.41}$$

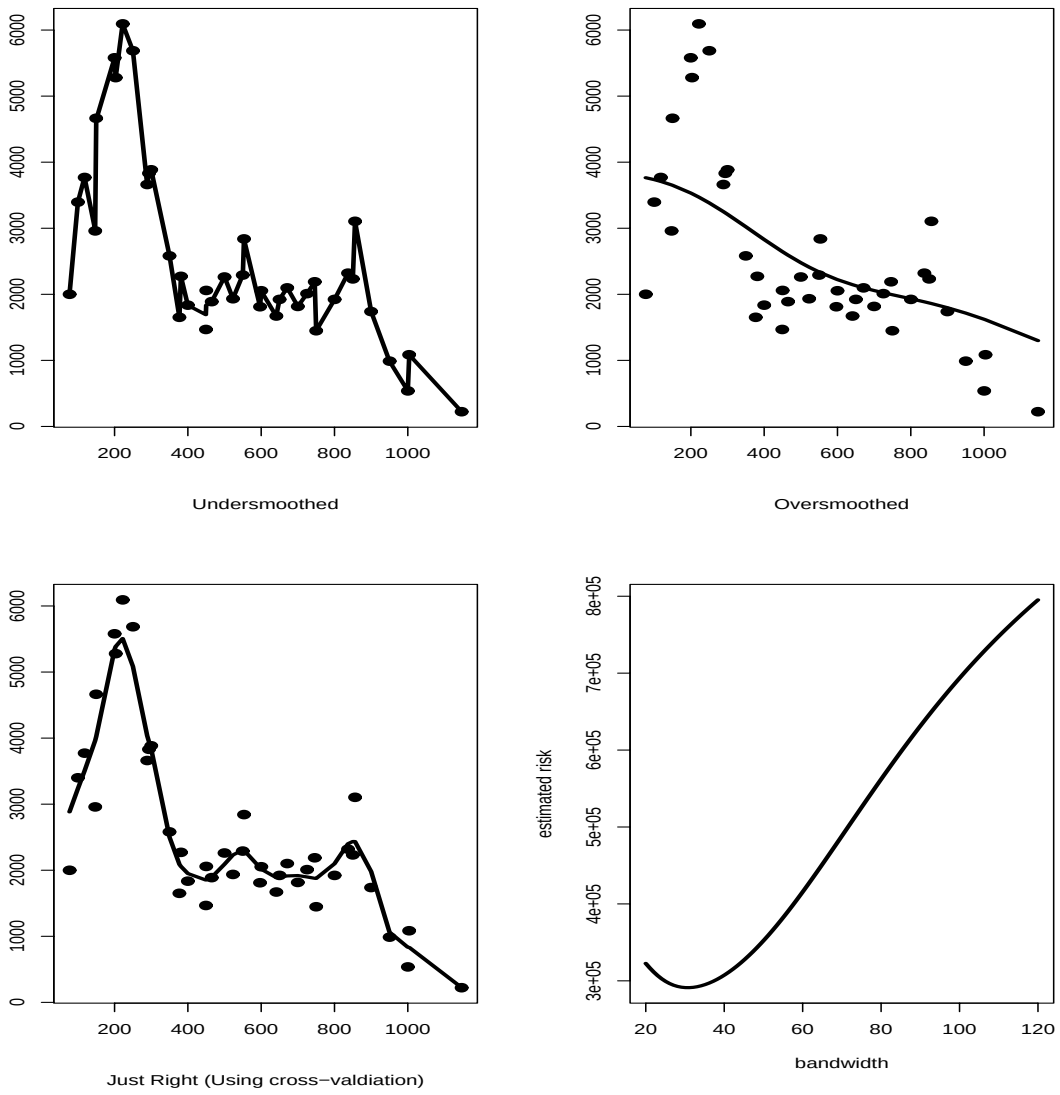


FIGURE 21.8. Regression analysis of the CMB data. The first fit is undersmoothed, the second is oversmoothed and the third is based on cross-validation. The last panel shows the estimated risk versus the bandwidth of the smoother. The data are from BOOMERaNG, Maxima and DASI.

Confidence Bands for Kernel Regression

Follow the same procedure as for kernel density estimators except, change the definition of $\mathsf{se}(v)$ to

$$\mathsf{se}(v) = \widehat{\sigma}\sqrt{\sum_{i=1}^{n} w^2(x_i)} \tag{21.42}$$

where $\widehat{\sigma}$ is defined in (21.41).

Example 21.21 *Figure 21.9 shows a 95 per cent confidence envelope for the CMB data. We see that we are highly confidence of the existence and position of the first peak. We are more uncertain about the second and third peak. At the time of this writing, more accurate data are becoming available that apparently provide sharper estimates of the second and third peak.* ■

The extension to multiple regressors $X = (X_1, \ldots, X_p)$ is straightforward. As with kernel density estimation we just replace the kernel with a multivariate kernel. However, the same caveats about the curse of dimensionality apply. In some cases, we might consider putting some restrictions on the regression function which will then reduce the curse of dimensionality. For example, **additive regression** is based on the model

$$Y = \sum_{j=1}^{p} r_j(X_j) + \epsilon. \tag{21.43}$$

Now we only need to fit p one-dimensional functions. The model can be enriched by adding various interactions, for example,

$$Y = \sum_{j=1}^{p} r_j(X_j) + \sum_{j<k} r_{jk}(X_j X_k) + \epsilon. \tag{21.44}$$

Additive models are usually fit by an algorithm called **backfitting**.

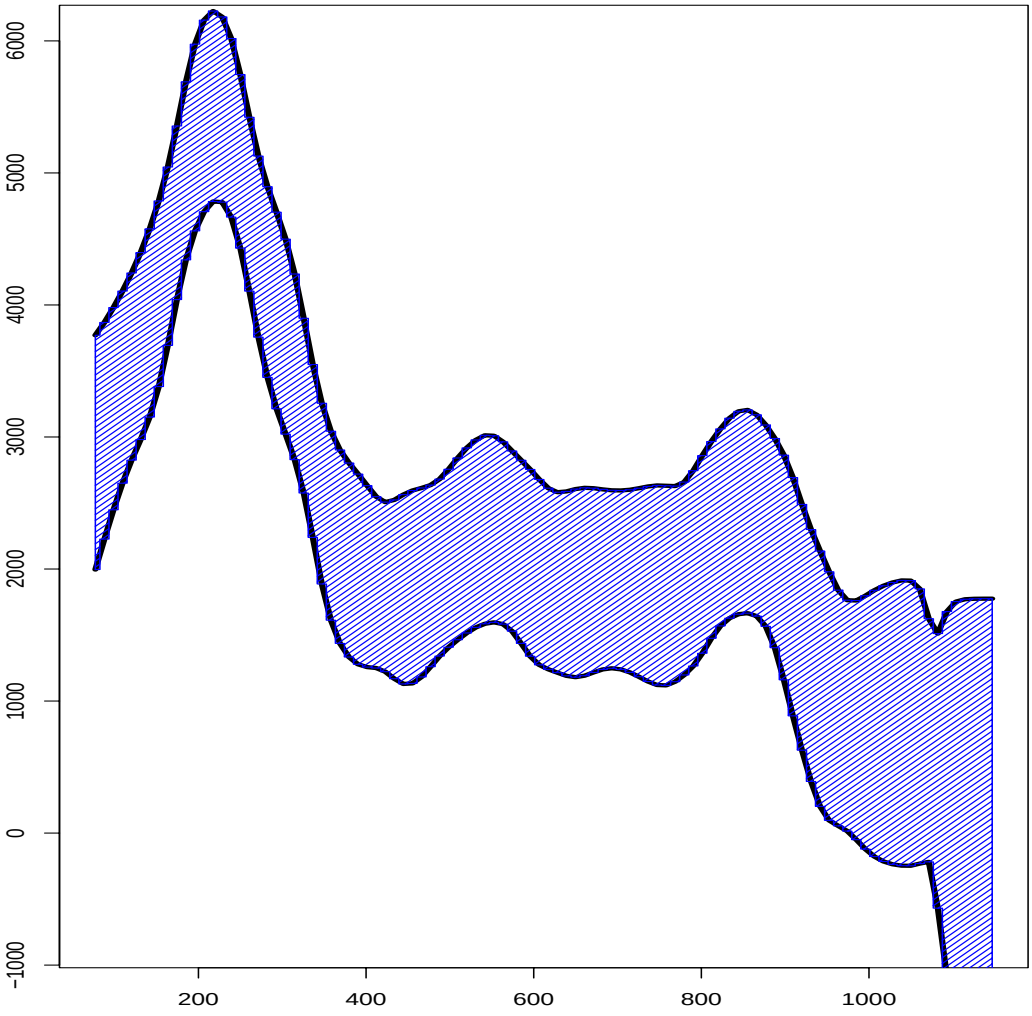


FIGURE 21.9. 95 per cent confidence envelope for the CMB data.

Backfitting

1. Initialize $r_1(x_1), \ldots, r_p(x_p)$.
2. For $j = 1, \ldots, p$:
   (a) Let $\epsilon_i = Y_i - \sum_{s \neq j} r_s(x_i)$.
   (b) Let r_j be the function estimate obtained by regressing the ϵ_i's on the j^{th} covariate.
3. If converged STOP. Else, go back to step 2.

Additive models have the advantage that they avoid the curse of dimensionality and they can be fit quickly but they have one disadvantage: the model is not fully nonparametric. In other words, the true regression function $r(x)$ may not be of the form (21.43).

21.5 Appendix: Confidence Sets and Bias

The confidence bands we computed are not for the density function or regression function but rather, for the smoothed function. For example, the confidence band for a kernel density estimate with bandwidth h is a band for the function one gets by smoothing the true function with a kernel with the same bandwidth. Getting a confidence set for the true function is complicated for reasons we now explain.

First, let's review the parametric case. When estimating a scalar quantity θ with an estimator $\widehat{\theta}$, the usual confidence interval is of the form $\widehat{\theta} \pm z_{\alpha/2} s_n$ where $\widehat{\theta}$ is the maximum likelihood estimator and $s_n = \sqrt{\mathbb{V}(\widehat{\theta})}$ is the estimated standard error of the estimator. Under standard regularity conditions, $\widehat{\theta} \approx N(\theta, s_n)$

and

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{\theta} - 2\,s_n \;\leq\; \theta \;\leq\; \widehat{\theta} + 2\,s_n\right) = 1-\alpha.$$

But let us take a closer look.

Let $b_n = E(\widehat{\theta}_n) - \theta$ We know that $\widehat{\theta} \approx N(\theta, s_n)$ but a more accurate statement is $\widehat{\theta} \approx N(\theta + b_n, s_n)$. We usually ignore the bias term b_n in large sample calculations but let's keep track of it. The coverage of the usual confidence interval is

$$\begin{aligned}
\mathbb{P}\left(\widehat{\theta} - z_{\alpha/2}s_n \leq \theta \leq \widehat{\theta} + z_{\alpha/2}s_n\right) &= \mathbb{P}\left(-z_{\alpha/2} \leq \frac{\widehat{\theta}-\theta}{s_n} \leq z_{\alpha/2}\right) \\
&= \mathbb{P}\left(-z_{\alpha/2} \leq \frac{N\left(b_n, s_n\right)}{s_n} \leq z_{\alpha/2}\right) \\
&= \mathbb{P}\left(-z_{\alpha/2} \leq N\left(\frac{b_n}{s_n}, 1\right) \leq z_{\alpha/2}\right).
\end{aligned}$$

In a well behaved parametric model, s_n is of size $n^{-1/2}$ b_n is of size n^{-1}. Hence, $b_n/s_n \to 0$ so the last displayed probability statement becomes $\mathbb{P}(-z_{\alpha/2} \leq N(0,1) \leq z_{\alpha/2}) = 1-\alpha$. What makes parametric confidence intervals have the right coverage is the fact that $b_n/s_n \to 0$.

The situation is more complicated for kernel methods. Consider estimating a density $f(x)$ at a single point x with a kernel density estimator. Since $\widehat{f}(x)$ is a sum if iid random variables, the central limit theorem implies that

$$\widehat{f}(x) \approx N\left(f(x) + b_n(x), \frac{c_2 f(x)}{nh}\right) \tag{21.45}$$

where

$$b_n(x) = \frac{1}{2}h^2 f''(x)c_1 \tag{21.46}$$

is the bias, $c_1 = \int x^2 K(x)dx$ and $c_2 = \int K^2(x)dx$. The estimated standard error is

$$s_n(x) = \left\{\frac{c_2\widehat{f}(x)}{nh}\right\}^{1/2}. \tag{21.47}$$

Suppose we use the usual interval $\widehat{f}(x) \pm z_{\alpha/2} s_n$. Arguing as above, the coverage is approximately,

$$\mathbb{P}\left(-z_{\alpha/2} \leq N\left(\frac{b_n(x)}{s_n(x)}, 1\right) \leq z_{\alpha/2}\right).$$

The optimal bandwidth is of the form $h = cn^{-1/5}$ for some constant c. If we plug $h = cn^{-1/5}$ into (21.46) and(21.47) we see that $\mathrm{b_n(x)}/s_n(x)$ does not tend to 0. Thus, the confidence interval will have coverage less than $1 - \alpha$.

21.6 Bibliographic Remarks

Two very good books on curve estimation are Scott (1992) and Silverman (1986). I drew heavily on those books for this Chapter.

21.7 Exercises

1. Let $X_1, \ldots, X_n \sim f$ and let $\widehat{f}_n$ be the kernel density estimator using the boxcar kernel:

$$K(x) = \begin{cases} 1 & -\frac{1}{2} < x < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

(a) Show that

$$\mathbb{E}(\widehat{f}(x)) = \frac{1}{h}\int_{x-(h/2)}^{x+(h/2)} f(y)dy$$

and

$$\mathbb{V}(\widehat{f}(x)) = \frac{1}{nh^2}\left[\int_{x-(h/2)}^{x+(h/2)} f(y)dy - \left(\int_{x-(h/2)}^{x+(h/2)} f(y)dy\right)^2\right].$$

(b) Show that if $h \to 0$ and $nh \to \infty$ as $n \to \infty$ then $\widehat{f}_n(x) \xrightarrow{\text{P}} f(x)$.

2. Get the data on fragments of glass collected in forensic work. You need to download the MASS library for R then:

```
library(MASS)
data(fgl)
x <- fgl[[1]]
help(fgl)
```

The data are also on my website. Estimate the density of the first variable (refractive index) using a histogram and use a kernel density estimator. Use cross-validation to choose the amount of smoothing. Experiment with different binwidths and bandwidths. Comment on the similarities and differences. Construct 95 per cent confidence bands for your estimators.

3. Consider the data from question 2. Let Y be refractive index and let x be aluminium content (the fourth variable). Do a nonparametric regression to fit the model $Y = f(x) + \epsilon$. Use cross-validation to estimate the bandwidth. Construct 95 per cent confidence bands for your estimate.

4. Prove Lemma 21.1.

5. Prove Theorem 21.3.

6. Prove Theorem 21.7.

7. Prove Theorem 21.14.

8. Consider regression data $(x_1, Y_1), \ldots, (x_n, Y_n)$. Suppose that $0 \leq x_i \leq 1$ for all i. Define bins B_j as in equation (21.7). For $x \in B_j$ define

$$\widehat{r}_n(x) = \overline{Y}_j$$

where $\overline{Y}_j$ is the mean of all the Y_i's corresponding to those x_i's in B_j. Find the approximate risk of this estimator. From this expression for the risk, find the optimal bandwidth. At what rate does the risk go to zero?

9. Show that with suitable smoothness assumptions on $r(x)$, $\widehat{\sigma}^2$ in equation (21.41) is a consistent estimator of σ^2.

22

Smoothing Using Orthogonal Functions

In this Chapter we study a different approach to nonparametric curve estimation based on **orthogonal functions**. We begin with a brief introduction to the theory of orthogonal functions. Then we turn to density estimation and regression.

22.1 Orthogonal Functions and L_2 Spaces

Let $v = (v_1, v_2, v_3)$ denote a three dimensional vector, that is, a list of three real numbers. Let $\mathcal{V}$ denote the set of all such vectors. If a is a scalar (a number) and v is a vector, we define $av = (av_1, av_2, av_3)$. The sum of vectors v and w is defined by $v + w = (v_1 + w_1, v_2 + w_2, v_3 + w_3)$. The **inner product** between two vectors v and w is defined by $\langle v, w \rangle = \sum_{i=1}^{3} v_i w_i$. The **norm (or length)** of a vector v is defined by

$$||v|| = \sqrt{\langle v, v \rangle} = \sqrt{\sum_{i=1}^{3} v_i^2}. \tag{22.1}$$

Two vectors are **orthogonal (or perpendicular)** if $\langle v, w\rangle = 0$. A set of vectors are orthogonal if each pair in the set is orthogonal. A vector is **normal** if $||v|| = 1$.

Let $\phi_1 = (1,0,0)$, $\phi_2 = (0,1,0)$, $\phi_3 = (0,0,1)$. These vectors are said to be an **orthonormal basis** for $\mathcal{V}$ since they have the following properties:
(i) they are orthogonal;
(ii) they are normal;
(iii) they form a basis for $\mathcal{V}$ which means that if $v \in \mathcal{V}$ then v can be written as a linear combination of ϕ_1, ϕ_2, ϕ_3:

$$v = \sum_{j=1}^{3} \beta_j \phi_j \quad \text{where} \quad \beta_j = \langle \phi_j, v\rangle. \tag{22.2}$$

For example, if $v = (12,3,4)$ then $v = 12\phi_1 + 3\phi_2 + 4\phi_3$. There are other orthonormal bases for $\mathcal{V}$, for example,

$$\psi_1 = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right),\ \psi_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right),\ \psi_3 = \left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}\right).$$

You can check that these three vectors also form an orthonormal basis for $\mathcal{V}$. Again, if v is any vector then we can write

$$v = \sum_{j=1}^{3} \beta_j \psi_j \quad \text{where} \quad \beta_j = \langle \psi_j, v\rangle.$$

For example, if $v = (12,3,4)$ then

$$v = 10.97\psi_1 + 6.36\psi_2 + 2.86\psi_3.$$

Now we make the leap from vectors to functions. Basically, we just replace vectors with functions and sums with integrals. Let $L_2(a,b)$ denote all functions defined on the interval $[a,b]$ such that $\int_a^b f(x)^2 dx < \infty$:

$$L_2[a,b] = \left\{ f : [a,b] \to \mathbb{R}, \quad \int_a^b f(x)^2 dx < \infty \right\}. \tag{22.3}$$

We sometimes write L_2 instead of $L_2(a, b)$. The inner product between two functions $f, g \in L_2$ is defined by $\int f(x)g(x)dx$. The norm of f is

$$||f|| = \sqrt{\int f(x)^2 dx}. \tag{22.4}$$

Two functions are orthogonal if $\int f(x)g(x)dx = 0$. A function is normal if $||f|| = 1$.

A sequence of functions $\phi_1, \phi_2, \phi_3, \ldots$ is **orthonormal** if $\int \phi_j^2(x)dx = 1$ for each j and $\int \phi_i(x)\phi_j(x)dx = 0$ for $i \neq j$. An orthonormal sequence is **complete** if the only function that is orthogonal to each ϕ_j is the zero function. In this case, the functions $\phi_1, \phi_2, \phi_3, \ldots$ form in basis, meaning that if $f \in L_2$ then f can be written as

$$f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x), \quad \text{where} \quad \beta_j = \int_a^b f(x)\phi_j(x)dx. \tag{22.5}$$

The equality in the displayed equation means that $\int (f(x) - f_n(x))^2 dx \to 0$ where $f_n(x) = \sum_{j=1}^{n} \beta_j \phi_j(x)$.

Parseval's relation says that

$$||f||^2 \equiv \int f^2(x)dx = \sum_{j=1}^{\infty} \beta_j^2 \equiv ||\beta||^2 \tag{22.6}$$

where $\beta = (\beta_1, \beta_2, \ldots)$.

Example 22.1 *An example of an orthonormal basis for $L_2[0, 1]$ is the* **cosine basis** *defined as follows. Let $\phi_1(x) = 1$ and for $j \geq 2$ define*

$$\phi_j(x) = \sqrt{2}\cos((j-1)\pi x). \tag{22.7}$$

The first six functions are plotted in Figure 22.1. ■

Example 22.2 *Let*

$$f(x) = \sqrt{x(1-x)}\,\sin\left(\frac{2.1\pi}{(x+.05)}\right)$$

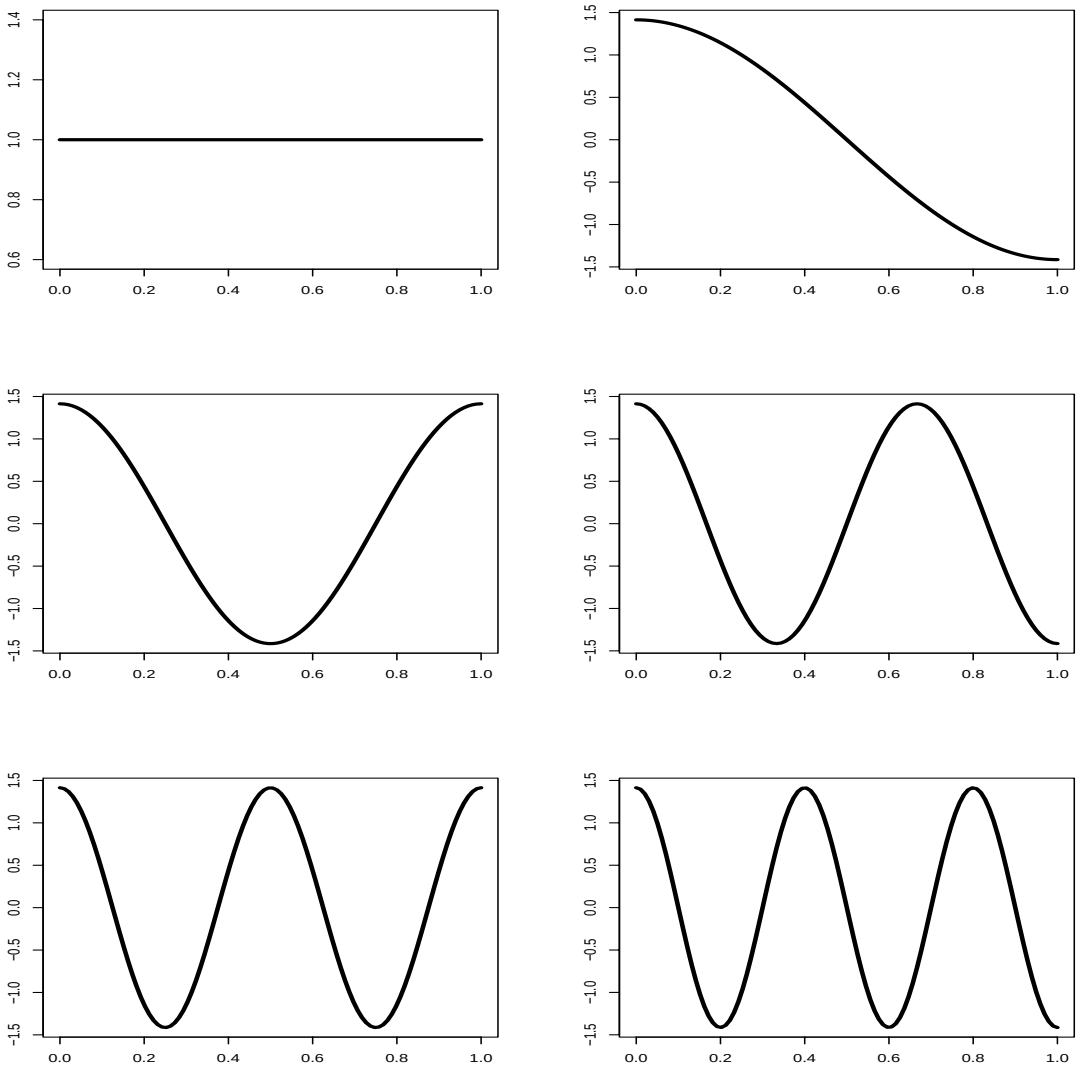


FIGURE 22.1. The first six functions in the cosine basis.

which is called the "doppler function." Figure 22.2 shows f (top left) and its approximation

$$f_J(x) = \sum_{j=1}^{J} \beta_j \phi_j(x)$$

with J equal to 5 (top right), 20 (bottom left) and 200 (bottom right). As J increases we see that $f_J(x)$ gets closer to $f(x)$. The coefficients $\beta_j = \int_0^1 f(x)\phi_j(x)dx$ were computed numerically. ■

Example 22.3 *The* **Legendre polynomials** *on $[-1, 1]$ are define by*

$$P_j(x) = \frac{1}{2^j j!} \frac{d^j}{dx^j}(x^2 - 1)^j, \quad j = 0, 1, 2, \ldots \tag{22.8}$$

It can be shown that these functions are complete and orthogonal and that

$$\int_{-1}^{1} P_j^2(x)dx = \frac{2}{2j+1}. \tag{22.9}$$

It follows that the functions $\phi_j(x) = \sqrt{(2j+1)/2}P_j(x)$, $j = 0, 1, \ldots$ form an orthonormal basis for $L_2(-1, 1)$. The first few Legendre polynomials are

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= \frac{1}{2}\Big(3x^2 - 1\Big) \\ P_3(x) &= \frac{1}{2}\Big(5x^3 - 3x\Big). \end{aligned}$$

These polynomials may be constructed explicitly using the following recursive relation:

$$P_{j+1}(x) = \frac{(2j+1)xP_j(x) - jP_{j-1}(x)}{j+1}. \quad ■ \tag{22.10}$$

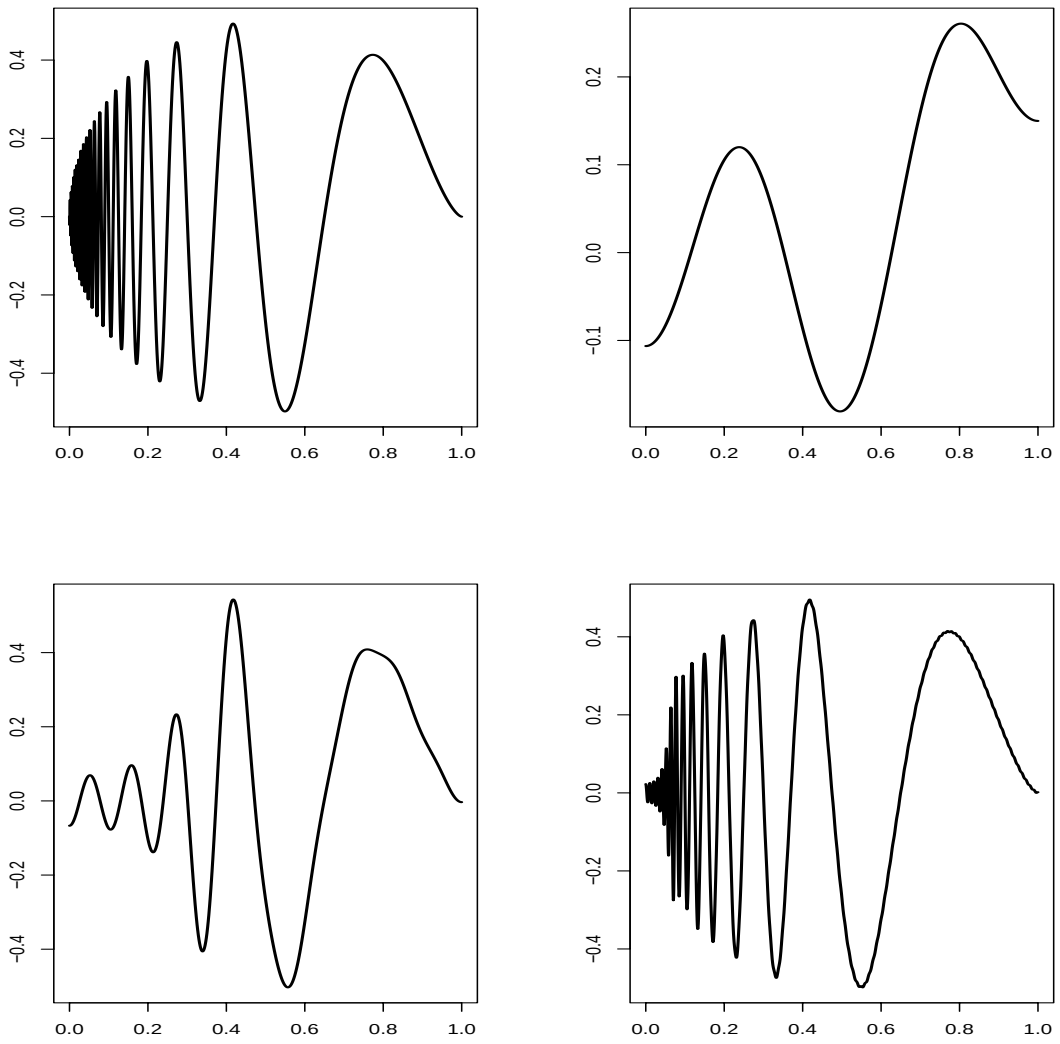


FIGURE 22.2. Approximating the doppler function $f(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{(x+.05)}\right)$ with its expansion in the cosine basis. The function f (top left) and its approximation $f_J(x) = \sum_{j=1}^{J} \beta_j \phi_j(x)$ with J equal to 5 (top right), 20 (bottom left) and 200 (bottom right). The coefficients $\beta_j = \int_0^1 f(x)\phi_j(x)dx$ were computed numerically.

The coefficients $\beta_1, \beta_2, \ldots$ are related to the smoothness of the function f. To see why, note that if f is smooth, then its derivatives will be finite. Thus we expect that, for some k, $\int_0^1 (f^{(k)}(x))^2 dx < \infty$ where $f^{(k)}$ is ther k^{th} derivative of f. Now consider the cosine basis (22.7) and let $f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x)$. Then,

$$\int_0^1 (f^{(k)}(x))^2 dx = 2 \sum_{j=1}^{\infty} \beta_j^2 (\pi(j-1))^{2k}.$$

The only way that $\sum_{j=1}^{\infty} \beta_j^2 (\pi(j-1))^{2k}$ can be finite is if the β_j's get small when j gets large. To summarize:

> If the function f is smooth then the coeffcients β_j will be small when j is large.

For the rest of this chapter, assume we are using the cosine basis unless otherwise specified.

22.2 Density Estimation

Let $X_1, \ldots, X_n$ be IID observations from a distribution on $[0,1]$ with density f. Assuming $f \in L_2$ we can write

$$f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x)$$

where $\phi_1, \phi_2, \ldots$ is an orthonormal basis. Define

$$\widehat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i). \tag{22.11}$$

Theorem **22.4** *The mean and variance of $\widehat{\beta}_j$ are*

$$\mathbb{E}\left(\widehat{\beta}_j\right) = \beta_j \tag{22.12}$$

$$\mathbb{V}\left(\widehat{\beta}_j\right) = \frac{\sigma_j^2}{n} \tag{22.13}$$

where

$$\sigma_j^2 = \mathbb{V}(\phi_j(X_i)) = \int (\phi_j(x) - \beta_j)^2 f(x) dx. \tag{22.14}$$

PROOF. We have

$$\begin{aligned} \mathbb{E}\left(\widehat{\beta}_j\right) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\phi_j(X_i)\right) \\ &= \mathbb{E}\left(\phi_j(X_1)\right) \\ &= \int \phi_j(x) f(x) dx = \beta_j. \end{aligned}$$

The calculation for the variance is similar. ■

Hence, $\widehat{\beta}_j$ is an unbiased estimate of β_j. It is tempting to estimate f by $\sum_{j=1}^{\infty} \widehat{\beta}_j \phi_j(x)$ but this turns out to have a very high variance. Instead, consider the estimator

$$\widehat{f}(x) = \sum_{j=1}^{J} \widehat{\beta}_j \phi_j(x). \tag{22.15}$$

The number of terms J is a smoothing parameter. Increasing J will decrease bias while increasing variance. For technical reasons, we restrict J to lie in the range

$$1 \leq J \leq p$$

where $p = p(n) = \sqrt{n}$. To emphasize the dependence of the risk function on J, we write the risk function as $R(J)$.

Theorem 22.5 *The risk of $\widehat{f}$ is given by*

$$R(J) = \sum_{j=1}^{J} \frac{\sigma_j^2}{n} + \sum_{j=J+1}^{\infty} \beta_j^2. \tag{22.16}$$

In kernel estimation, we used cross-validation to estimate the risk. In the orthogonal function approach, we instead use the risk estimator

$$\widehat{R}(J) = \sum_{j=1}^{J} \frac{\widehat{\sigma}_j^2}{n} + \sum_{j=J+1}^{p} \left(\widehat{\beta}_j^2 - \frac{\widehat{\sigma}_j^2}{n}\right)_+ \tag{22.17}$$

where $a_+ = \max\{a, 0\}$ and

$$\widehat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\phi_j(X_i) - \widehat{\beta}_j\right)^2. \tag{22.18}$$

To motivate this estimator, note that $\widehat{\sigma}_j^2$ is an unbiased estmate of σ_j^2 and $\widehat{\beta}_j^2 - \widehat{\sigma}_j^2$ is an unbiased estimator of β_j^2. We take the positive part of the latter term since we know that β_j^2 cannot be negative. We now choose $1 \leq \widehat{J} \leq p$ to minimize $\widehat{R}(\widehat{f}, f)$. Here is a summary.

Summary of Orthogonal Function Density Estimation

1. Let
$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i).$$

2. Choose $\widehat{J}$ to minimize $\widehat{R}(J)$ over $1 \leq J \leq p = \sqrt{n}$ where
$$\widehat{R}(J) = \sum_{j=1}^{J}\frac{\widehat{\sigma}_j^2}{n} + \sum_{j=J+1}^{p}\left(\widehat{\beta}_j^2 - \frac{\widehat{\sigma}_j^2}{n}\right)_+$$
and
$$\widehat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\phi_j(X_i) - \widehat{\beta}_j\right)^2.$$

3. Let
$$\widehat{f}(x) = \sum_{j=1}^{\widehat{J}}\widehat{\beta}_j\phi_j(x).$$

The estimator $\widehat{f}_n$ can be negative. If we are interested in exploring the shape of f, this is not a problem. However, if we need our estimate to be a probability density function, we

can truncate the estimate and then normalize it. That, we take $\widehat{f}^* = \max\{\widehat{f}_n(x), 0\} / \int_0^1 \max\{\widehat{f}_n(u), 0\} du$.

Now let us construct a confidence band for f. Suppose we estimate f using J orthogonal functions. We are essentially estimating $\widetilde{f}(x) = \sum_{j=1}^{J} \beta_j \phi_j(x)$ not the true density $f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x)$. Thus, the confidence band should be regarded as a band for $\widetilde{f}(x)$.

Theorem **22.6** *An approximate $1-\alpha$ confidence band for $\widetilde{f}$ is $(\ell(x), u(x))$ where*

$$\ell(x) = \widehat{f}_n(x) - c, \quad u(x) = \widehat{f}_n(x) + c \tag{22.19}$$

where

$$c = \frac{JK^2}{\sqrt{n}} \sqrt{1 + \frac{\sqrt{2} z_\alpha}{\sqrt{J}}} \tag{22.20}$$

and

$$K = \max_{1 \le j \le J} \max_x |\phi_j(x)|.$$

For the cosine basis, $K = \sqrt{2}$.

Example 22.7 *Let*

$$f(x) = \frac{5}{6}\phi(x; 0, 1) + \frac{1}{6} \sum_{j=1}^{5} \phi(x; \mu_j, .1)$$

where $\phi(x; \mu, \sigma)$ denotes a Normal density with mean μ and standard deviation σ, and $(\mu_1, \ldots, \mu_5) = (-1, -1/2, 0, 1/2, 1)$. Marron and Wand (1990) call this "the claw" although I prefer the "Bart Simpson." Figure 22.3 shows the true density as well as the estimated density based on $n = 5,000$ observations and a 95 per cent confidence band. The density has been rescaled to have most of its mass between 0 and 1 using the transformation $y = (x + 3)/6$.

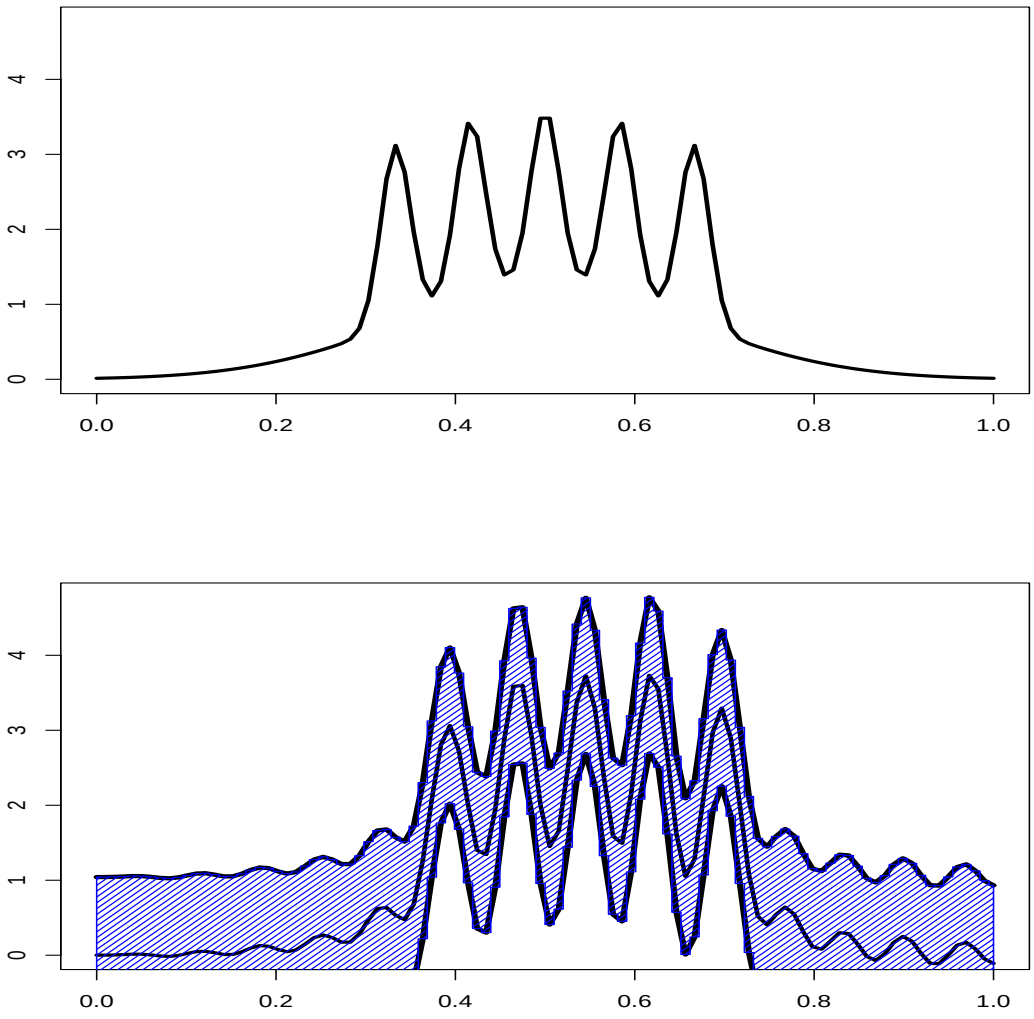


FIGURE 22.3. The top plot is the true density for the Bart Simpson distribution (rescaled to have most of its mass between 0 and 1). The bottom plot is the orthogonal function density estimate and 95 per cent confidence band.

22.3 Regression

Consider the regression model

$$Y_i = r(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{22.21}$$

where the ϵ_i are independent with mean 0 and variance σ^2. We will initially focus on the special case where $x_i = i/n$. We assume that $r \in L_2[0,1]$ and hence we can write

$$r(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x) \quad \text{where} \quad \beta_j = \int_0^1 r(x)\phi_j(x)dx \tag{22.22}$$

where $\phi_1, \phi_2, \ldots$ where is an orthonormal basis for $[0,1]$.

Define

$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i\, \phi_j(x_i), \quad j = 1, 2, \ldots \tag{22.23}$$

Since $\widehat{\beta}_j$ is an average, the central limit theorem tells us that $\widehat{\beta}_j$ will be approximately normally distributed.

Theorem **22.8**

$$\widehat{\beta}_j \approx N\left(\beta_j, \frac{\sigma^2}{n}\right). \tag{22.24}$$

PROOF. The mean of $\widehat{\beta}_j$ is

$$\begin{aligned} \mathbb{E}(\widehat{\beta}_j) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i)\phi_j(x_i) = \frac{1}{n}\sum_{i=1}^{n} r(x_i)\phi_j(x_i) \\ &\approx \int r(x)\phi_j(x)dx = \beta_j \end{aligned}$$

where the approximate equality follows from the definition of a Riemann integral: $\sum_i \Delta_n h(x_i) \to \int_0^1 h(x)dx$ where $\Delta_n = 1/n$. The variance is

$$\mathbb{V}(\widehat{\beta}_j) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(Y_i)\phi_j^2(x_i)$$

$$\begin{aligned} &= \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\phi_j^2(x_i) = \frac{\sigma^2}{n}\frac{1}{n}\sum_{i=1}^{n}\phi_j^2(x_i) \\ &\approx \frac{\sigma^2}{n}\int \phi_j^2(x)dx = \frac{\sigma^2}{n} \end{aligned}$$

since $\int \phi_j^2(x)dx = 1$. ■

As we did for density estimation, we will estimate r by

$$\widehat{r}(x) = \sum_{j=1}^{J}\widehat{\beta}_j\phi_j(x).$$

Let

$$R(J) = \mathbb{E}\int (r(x) - \widehat{r}(x))^2\, dx$$

be the risk of the estimator.

Theorem 22.9 *The risk $R(J)$ of the estimator $\widehat{r}_n(x) = \sum_{j=1}^{J}\widehat{\beta}_j\phi_j(x)$ is*

$$R(J) = \frac{J\sigma^2}{n} + \sum_{j=J+1}^{\infty}\beta_j^2. \tag{22.25}$$

To estimate for $\sigma^2 = Var(\epsilon_i)$ we use

$$\widehat{\sigma}^2 = \frac{n}{k}\sum_{i=n-k+1}^{n}\widehat{\beta}_j^2 \tag{22.26}$$

where $k = n/4$. To motivate this estimator, recall that if f is smooth then $\beta_j \approx 0$ for large j. So, for $j \geq k$, $\widehat{\beta}_j \approx N(0, \sigma^2/n)$. Thus, $\widehat{\beta}_j \approx \sigma\widehat{\beta}_j/\sqrt{n}$ for for $j \geq k$, where $\widehat{\beta}_j \sim N(0,1)$. Therefore,

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{n}{k}\sum_{i=n-k+1}^{n}\widehat{\beta}_j^2 \\ &\approx \frac{n}{k}\sum_{i=n-k+1}^{n}\left(\frac{\sigma}{\sqrt{n}}\widehat{\beta}_j\right)^2 \end{aligned}$$

$$= \frac{\sigma^2}{k} \sum_{i=n-k+1}^{n} \widehat{\beta}_j^2$$
$$= \frac{\sigma^2}{k} \chi_k^2$$

since a sum of k Normals has a χ_k^2 distribution. Now $\mathbb{E}(\chi_k^2) = k$ and hence $\mathbb{E}(\widehat{\sigma}^2) \approx \sigma^2$. Also, $\mathbb{V}(\chi_k^2) = 2k$ and hence $\mathbb{V}(\widehat{\sigma}^2) \approx (\sigma^4/k^2)(2k) = (2\sigma^4/k) \to 0$ as $n \to \infty$. Thus we expect $\widehat{\sigma}^2$ to be a consistent estimator of σ^2. There is nothing special about the choice $k = n/4$. Any k that increases with n at an appropriate rate will suffice.

We estimate the risk with

$$\widehat{R}(J) = J\frac{\widehat{\sigma}^2}{n} + \sum_{j=J+1}^{n} \left(\widehat{\beta}_j^2 - \frac{\widehat{\sigma}^2}{n}\right)_+ . \tag{22.27}$$

We are now ready to give a complete description of the method, which Beran (2000) calls REACT (Risk Estimation and Adaptation by Coordinate Transformation).

Orthogonal Series Regression Estimator

1. Let
$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_j(x_i), \quad j = 1, \ldots, n.$$

2. Let
$$\widehat{\sigma}^2 = \frac{n}{k}\sum_{i=n-k+1}^{n} \widehat{\beta}_j^2 \tag{22.28}$$
where $k \approx n/4$.

3. For $1 \leq J \leq n$, compute the risk estimate
$$\widehat{R}(J) = J\frac{\widehat{\sigma}^2}{n} + \sum_{j=J+1}^{n}\left(\widehat{\beta}_j^2 - \frac{\widehat{\sigma}^2}{n}\right)_+.$$

4. Choose $\widehat{J} \in \{1, \ldots n\}$ to minimize $\widehat{R}(J)$.

5. Let
$$\widehat{f}(x) = \sum_{j=1}^{\widehat{J}} \widehat{\beta}_j \phi_j(x).$$

Example 22.10 *Figure 22.4 shows the doppler function f and $n = 2,048$ observations generated from the model*
$$Y_i = r(x_i) + \epsilon_i$$
where $x_i = i/n$, $\epsilon_i \sim N(0, (.1)^2)$. Then figure shows the true function, the data, the estimated function and the estimated risk. The estimated figure was based on $\widehat{J} = 234$ terms.

Finally, we turn to confidence bands. As before, these bands are not really for the true function $r(x)$ but rather for the smoothed version of the function $\widetilde{r}(x) = \sum_{j=1}^{\widehat{J}} \beta_j \phi_j(x)$.

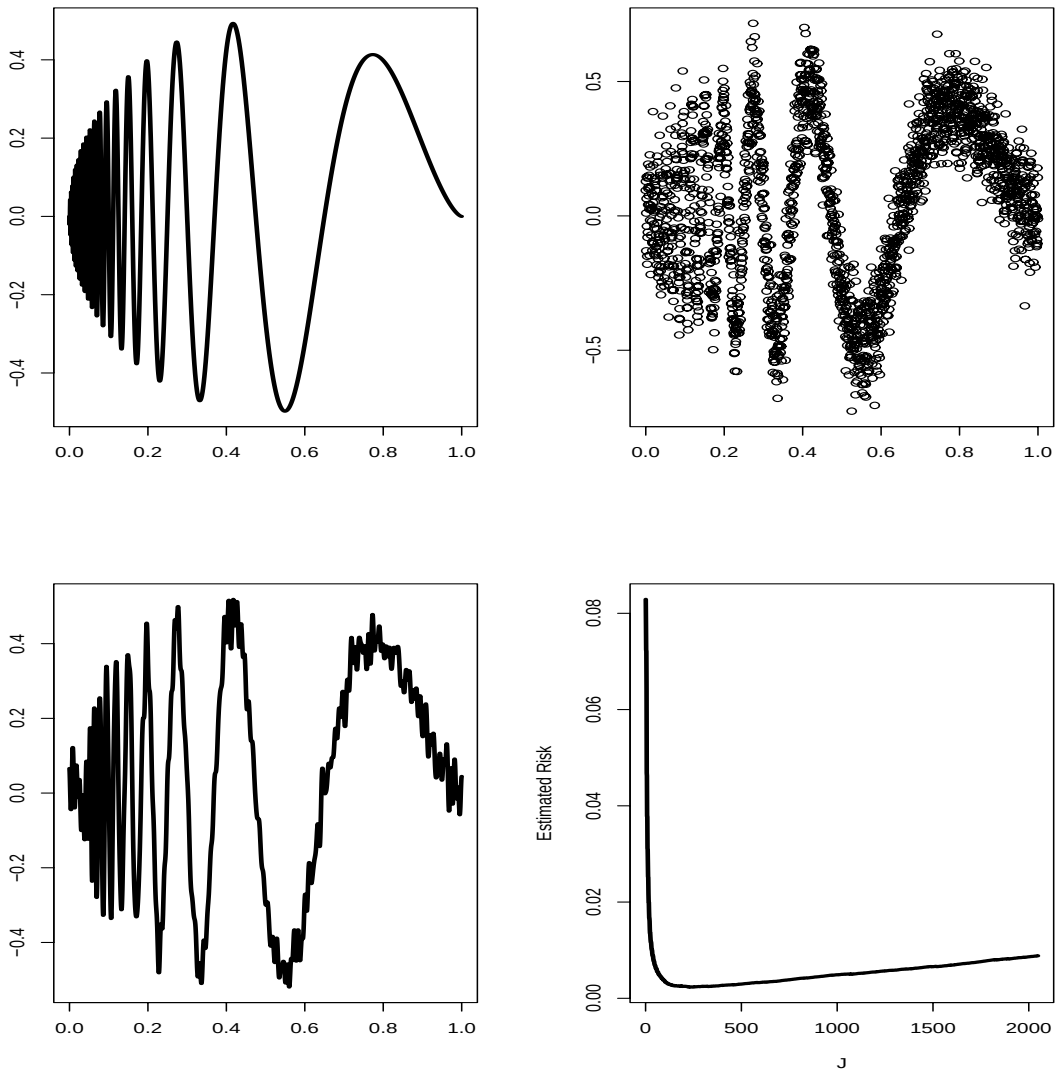


FIGURE 22.4. The doppler test function.

Theorem **22.11** *Suppose the estimate $\widehat{r}$ is based on J terms and $\widehat{\sigma}$ is defined as in equation (22.28). Assume that $J < n - k + 1$. An approximate $1-\alpha$ confidence band for $\widetilde{r}$ is (ℓ, u) where*

$$\ell(x) = \widehat{r}_n(x) - K\sqrt{J}\sqrt{\frac{z_\alpha \widehat{\tau}}{\sqrt{n}} + \frac{J\widehat{\sigma}^2}{n}} \tag{22.29}$$

$$u(x) = \widehat{r}_n(x) + K\sqrt{J}\sqrt{\frac{z_\alpha \widehat{\tau}}{\sqrt{n}} + \frac{J\widehat{\sigma}^2}{n}} \tag{22.30}$$

where

$$K = \max_{1\le j\le J} \max_x |\phi_j(x)|$$

and

$$\widehat{\tau}^2 = \frac{2J\widehat{\sigma}^4}{n}\left(1 + \frac{J}{k}\right) \tag{22.31}$$

and $k = n/4$ is from equation (22.28). In the cosine basis, we may use $K = \sqrt{2}$.

Example 22.12 *Figure 22.5 shows the confidence envelope for the doppler signal. Th first plot is based on $J = 234$ (the value of J that minimizes the estimated risk). The second is based on $J = 45 \approx \sqrt{n}$. Larger J yields a higher resolution estimator at the cost of large confidence bands. Smaller J yields a lower resolution estimator but has tighter confidence bands.* ■

So far, we have assumed that the x_i's are of the form $\{1/n, 2/n, \ldots, 1\}$. If the x_i's are on interval $[a, b]$ then we can rescale them so that are in the interval $[0, 1]$. If the x_i's are not equally spaced the methods we have discussed still apply so long as the x_i's "fill out" the interval [0,1] in such a way so as to not be too clumped together. If we want to treat the x_i's as random instead of fixed, then the method needs significant modifications which we shall not deal with here.

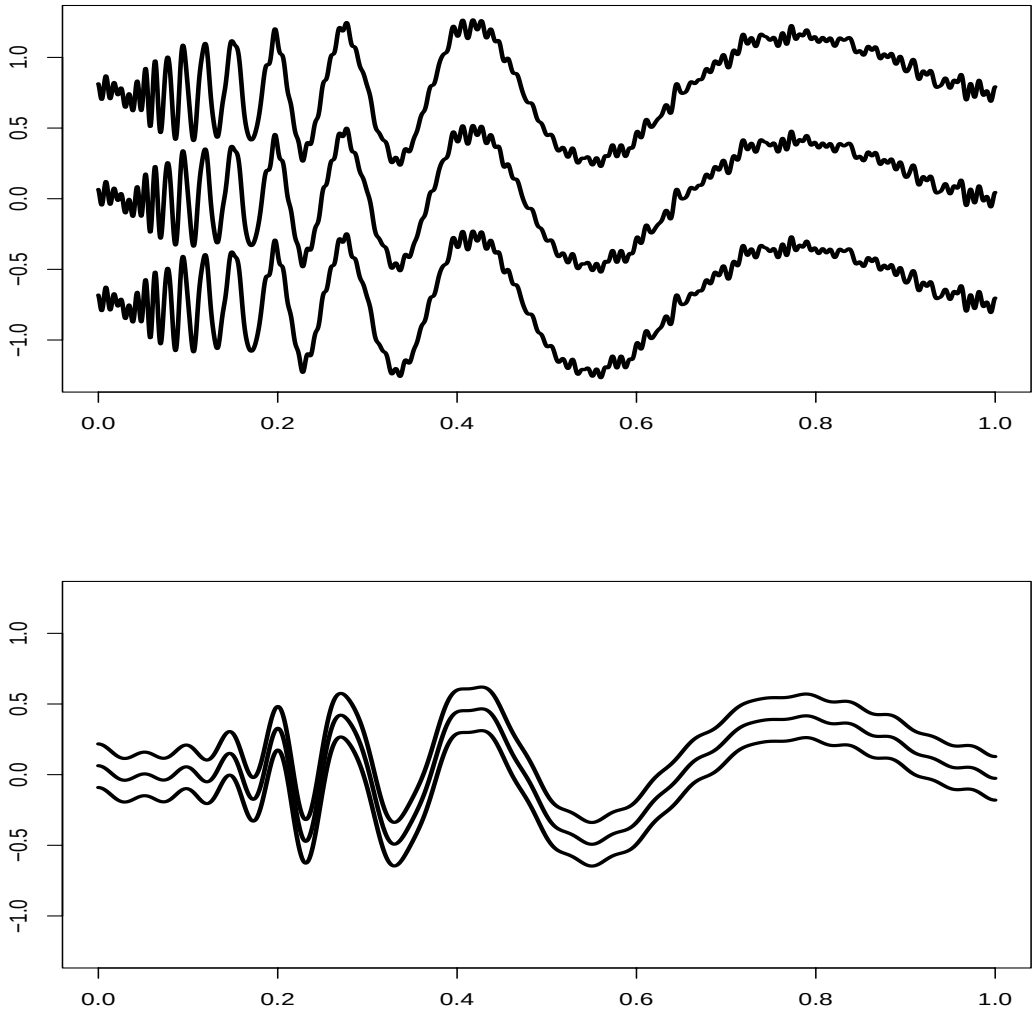


FIGURE 22.5. The confidence envelope for the doppler test function using $n = 2,048$ observations. The top plot shows the estimate and envelope using $J = 234$ terms. The bottom plot shows the estimate and envelope using $J = 45$ terms.

22.4 Wavelets

Suppose there is a sharp jump in a regression function f at some point x but that f is otherwise very smooth. Such a function f is said to be **spatially inhomogeneous**. See Figure 22.6 for an example.

It is hard to estimate f using the methods we have discussed so far. If we use a cosine basis and only keep low order terms, we will miss the peak; if we allow higher order terms we will find the peak but we will make the rest of the curve very wiggly. Similar comments apply to kernel regression. If we use a large bandwidth, then we will smooth out the peak; if we use a small bandwidth, then we will find the peak but we will make the rest of the curve very wiggly.

One way to estimate inhomogeneous functions is to use a more carefully chosen basis that allows us to place a "blip" in some small region without adding wiggles elsewhere. In this section, we describe a special class of bases called **wavelets**, that are aimed at fixing this problem. Statistical inference suing wavelets is a large and active area. We will just discuss a few of the main ideas to get a flavor of this approach.

We start with a particular wavelet called the **Haar wavelet.** The **Haar father wavelet** or **Haar scaling function** is defined by

$$\phi(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{22.32}$$

The **mother Haar wavelet** is defined by

$$\psi(x) = \begin{cases} -1 & \text{if } 0 \leq x \leq \frac{1}{2}, \\ 1 & \text{if } \frac{1}{2} < x \leq 1. \end{cases} \tag{22.33}$$

For any integers j and k define

$$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k) \quad \text{and} \quad \psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k). \tag{22.34}$$

The function $\psi_{j,k}$ has the same shape as ψ but it has been rescaled by a factor of $2^{j/2}$ and shifted by a factor of k.

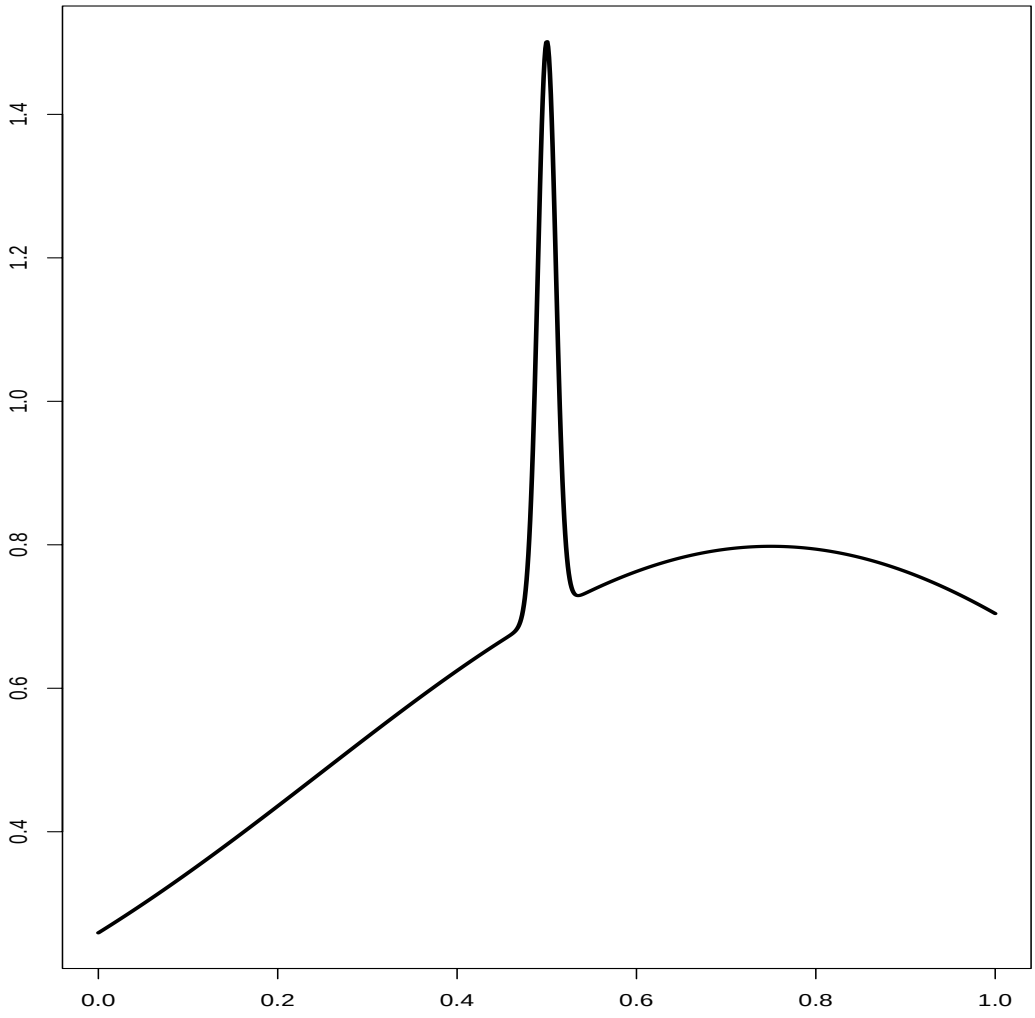


FIGURE 22.6. An inhomogeneous function. The function is smooth except for a sharp peak in one place.

See Figure 22.7 for some examples of Haar wavelets. Notice that for large j, $\psi_{j,k}$ is a very localized function. This makes it possible to add a blip to a function in one place without adding wiggles elsewhere. Increasing j is like looking in a microscope at increasing degrees of resolution. In technical terms, we say that wavelets provide a **multiresolution analysis** of $L_2(0,1)$.

Let

$$W_j = \{\psi_{jk},\ k = 0, 1, \ldots, 2^j - 1\}$$

be the set of rescaled and shifted mother wavelets at resolution j.

Theorem **22.13** *The set of functions*

$$\Big\{\phi, W_0, W_1, W_2, \ldots, \Big\}$$

is an orthonormal basis for $L_2(0,1)$.

It follows from this theorem that we can expand any function $f \in L_2(0,1)$ in this basis. Because each W_j is itself a set of functions, we write the expansion as a double sum:

$$f(x) = \alpha\, \phi(x) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \beta_{j,k} \psi_{j,k}(x) \tag{22.35}$$

where

$$\alpha = \int_0^1 f(x)\phi(x)\, dx, \quad \beta_{j,k} = \int_0^1 f(x)\psi_{j,k}(x)\, dx.$$

We call α the **scaling coefficient** and the $\beta_{j,k}$'s are called the **detail coefficients**. We call the finite sum

$$\widetilde{f}(x) = \alpha\phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \beta_{j,k} \psi_{j,k}(x) \tag{22.36}$$

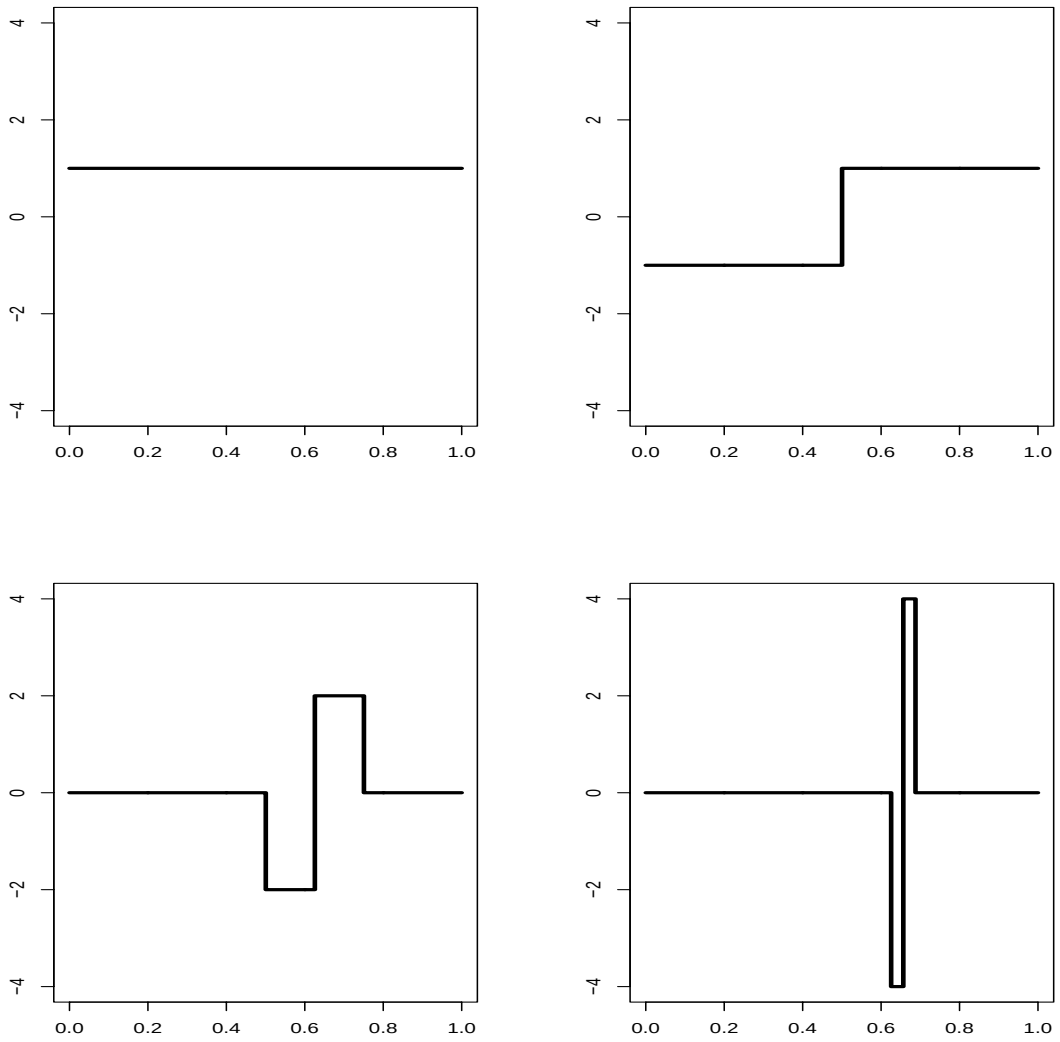


FIGURE 22.7. Some Haar wavelets. Top left: the father wavelet $\phi(x)$; top right: the mother wavelet $\psi(x)$; bottom left: $\psi_{2,2}(x)$; bottom right: $\psi_{4,10}(x)$.

the **resolution** J approximation to f. The total number of terms in this sum is

$$1 + \sum_{j=0}^{J-1} 2^j = 1 + 2^J - 1 = 2^J.$$

Example 22.14 *Figure 22.8 shows the doppler signal, and its resolution J approximation for $J = 3, 5$ and $J = 8$.* ■

Haar wavelets are localized, meaning that they are zero outside an interval. But they are not smooth. This raises the question of whether there exist smooth, localized wavelets that from an orthonormal basis. In 1988, Ingrid Daubechie showed that such wavelets do exist. These smooth wavelets are difficult to describe. They can be constructed numerically but there is no closed form formual for the smoother wavelets. To keep things simple, we will continue to use Haar wavelets. For your interest, figure 22.9 shows an example of a smooth wavelet called the "symmlet 8."

We can now use wavelets to do density estimation and regression. We shall only discuss the regression problem $Y_i = r(x_i) + \sigma\epsilon_i$ where $\epsilon_i \sim N(0,1)$ and $x_i = i/n$. To simplify the discussion we assume that $n = 2^J$ for some J.

There is one major difference between estimation using wavelets instead of a cosine (or polynomial) basis. With the cosine basis, we used all the terms $1 \leq j \leq J$ for some J. With wavelets, we use a method called **thresholding** where we keep a term in the function approximation if its coefficient is large, otherwise, we throw we don't use that term. There are many versions of thresholding. The simplest is called hard, universal thresholding.

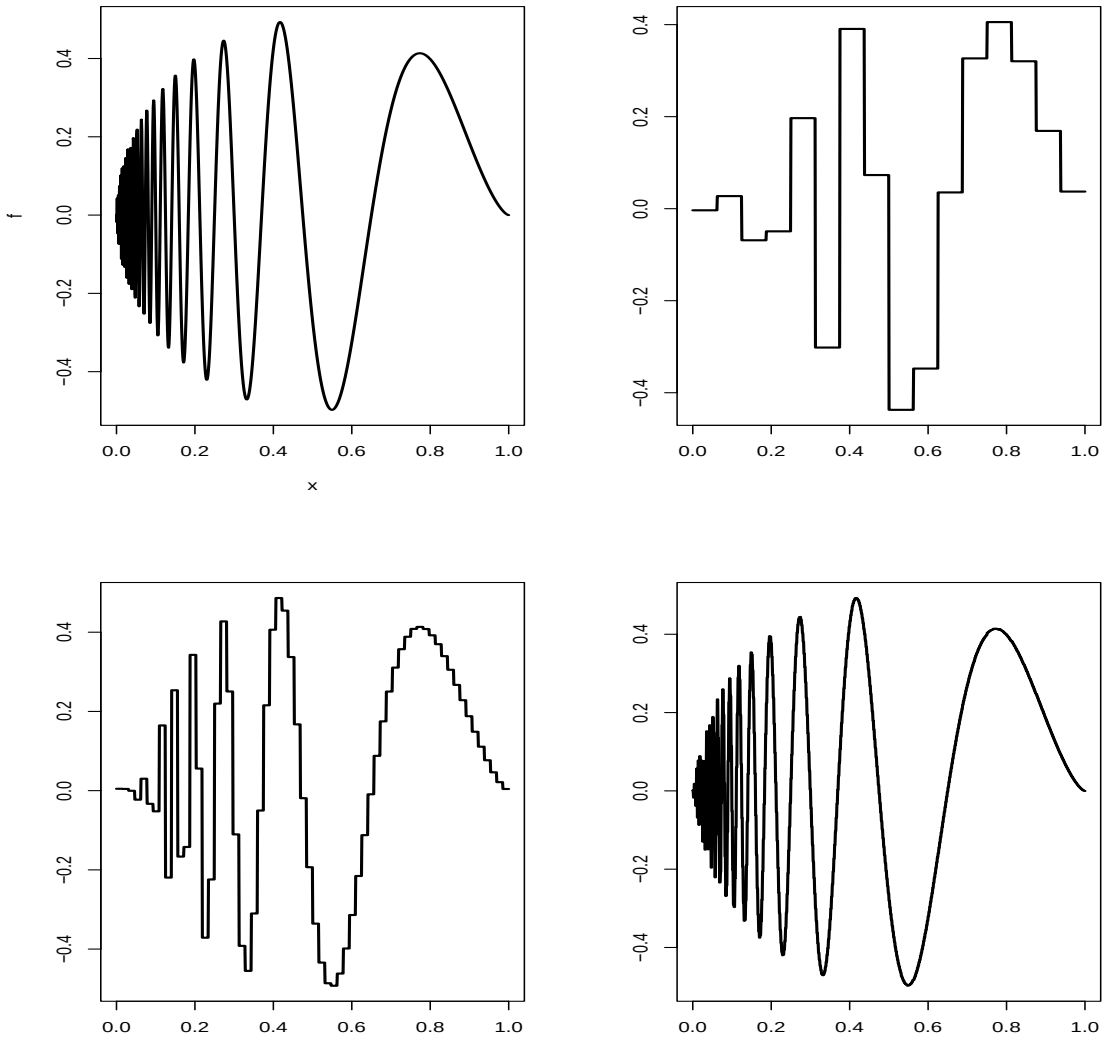


FIGURE 22.8. The doppler signal and its reconstruction $\widetilde{f}(x) = \alpha\phi(x) + \sum_{j=0}^{J-1}\sum_k \beta_{j,k}\psi_{j,k}(x)$ based on $J = 3$, $J = 5$ and $J = 8$.

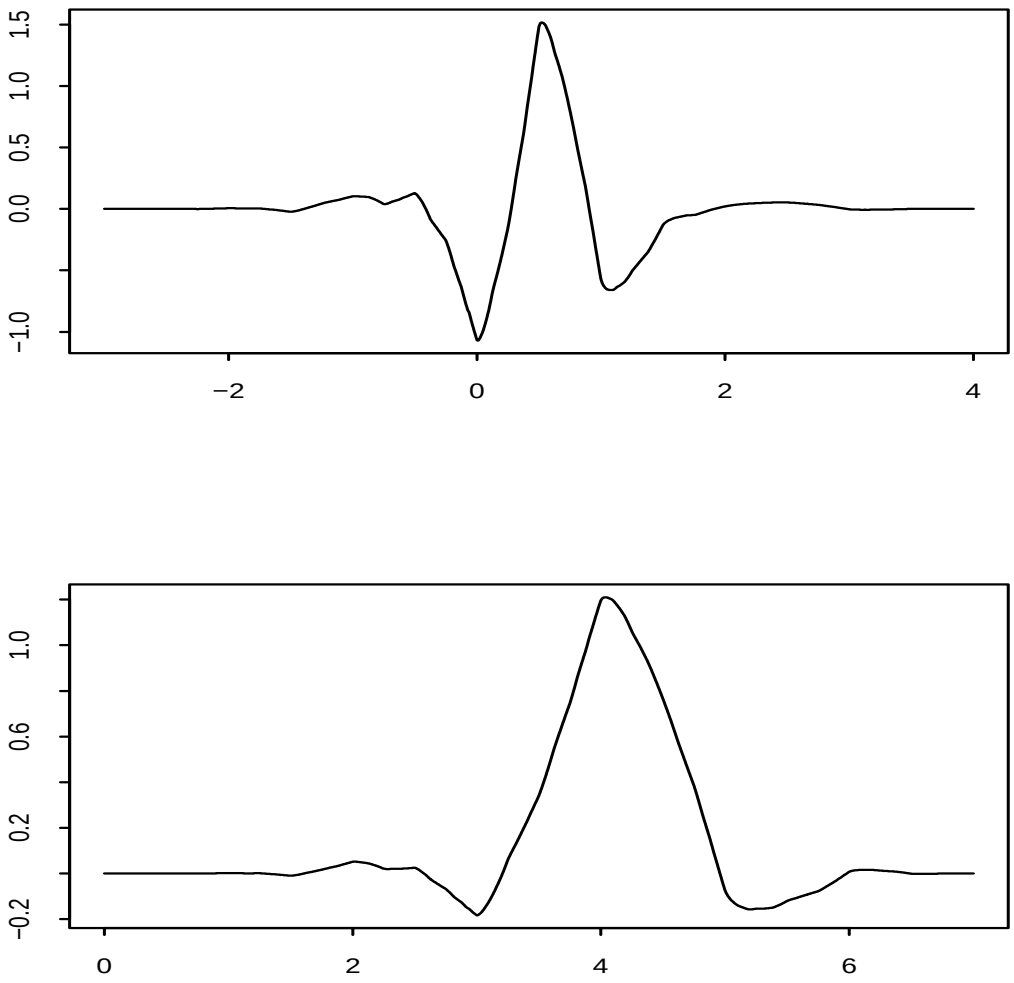


FIGURE 22.9. The symmlet 8 wavelet. The top plot is the mother and the bottom plot is the father.

Haar Wavelet Regression

1. Let $J = \log_2(n)$ and define

$$\widehat{\alpha} = \frac{1}{n}\sum_i \phi_k(x_i)Y_i \quad \text{and} \quad D_{j,k} = \frac{1}{n}\sum_i \psi_{j,k}(x_i)Y_i \tag{22.37}$$

   for $0 \le j \le J-1$.

2. Estimate σ by

$$\widehat{\sigma} = \sqrt{n} \times \frac{\text{median}\left(|D_{J-1,k}| : \ k = 0, \ldots, 2^{J-1}-1\right)}{0.6745}. \tag{22.38}$$

3. Apply universal thresholding:

$$\widehat{\beta}_{j,k} = \left\{ \begin{array}{ll} D_{j,k} & \text{if } |D_{j,k}| > \widehat{\sigma}\sqrt{\frac{2\log n}{n}} \\ 0 & \text{otherwise.} \end{array} \right\} \tag{22.39}$$

4. Set

$$\widehat{f}(x) = \widehat{\alpha}\phi(x) + \sum_{j=j_0}^{J-1}\sum_{k=0}^{2^j-1} \widehat{\beta}_{j,k}\psi_{j,k}(x).$$

In practice, we do not compute S_k and $D_{j,k}$ using (22.37). Instead, we use the **discrete wavelet transform (DWT)** which is very fast. For Haar wavelets, the DWT works as follows.

DWT for Haar Wavelets

Let y be the vector of Y_i's (length n) and let $J = \log_2(n)$. Create a list D with elements

$$D[[0]],\ \ldots,\ D[[J-1]].$$

Set:

$$temp \ <\!\!- \ y/\sqrt{n}.$$

Then do:

$$\begin{aligned}
for(j \ \ in \ \ & (J-1):0)\{ \\
m \ <\!\!- \ & 2^j \\
I \ <\!\!- \ & (1:m) \\
D[[j]] \ <\!\!- \ & \Big(temp[2*I] - temp[(2*I)-1]\Big)/\sqrt{2} \\
temp \ <\!\!- \ & \Big(temp[2*I] + temp[(2*I)-1]\Big)/\sqrt{2} \\
& \}
\end{aligned}$$

Remark 22.15 *If you use a programming language that does not allow a 0 index for a list, then index the list as*

$$D[[1]],\ \ldots,\ D[[J]]$$

and replace the second last line with

$$D[[j+1]] <\!\!-\Big(temp[2*I] - temp[(2*I)-1]\Big)/\sqrt{2}$$

The estimate for σ probably looks strange. It is similar to the estimate we used for the cosine basis but it is designed to be insensitive to sharp peaks in the function.

To understand the intuition behind universal thresholding, consider what happens when there is no signal, that is, when $\beta_{j,k} = 0$ for all j and k.

Theorem **22.16** *Suppose that $\beta_{j,k} = 0$ for all j and k and let $\widehat{\beta}_{j,k}$ be the universal threshold estimator. Then*

$$\mathbb{P}(\widehat{\beta}_{j,k} = 0 \text{ for all } j,k) \to 1$$

as $n \to \infty$.

PROOF. To simplify the proof, assume that σ is known. Now $D_{j,k} \approx N(0, \sigma^2/n)$. We will need **Mill's inequality**: if $Z \sim N(0,1)$ then $\mathbb{P}(|Z| > t) \leq (c/t)e^{-t^2/2}$ where $c = \sqrt{2/\pi}$ is a constant. Thus,

$$\begin{aligned}
\mathbb{P}(\max |D_{j,k}| > \lambda) &\leq \sum_{j,k} \mathbb{P}(|D_{j,k}| > \lambda) \\
&\leq \sum_{j,k} \mathbb{P}\left(\frac{\sqrt{n}|D_{j,k}|}{\sigma} > \frac{\sqrt{n}\lambda}{\sigma} \right) \\
&\leq \sum_{j,k} \frac{c\sigma}{\lambda\sqrt{n}} \exp\left\{ -\frac{1}{2}\frac{n\lambda^2}{\sigma^2} \right\} \\
&= \frac{c}{\sqrt{2\log n}} \to 0. \quad \blacksquare
\end{aligned}$$

Example 22.17 *Consider $Y_i = r(x_i) + \sigma\epsilon_i$ where f is the doppler signal, $\sigma = .1$ and $n = 2048$. Figure 22.10 shows the data and the estimated function using universal thresholding. Of course, the estimate is not smooth since Haar wavelets are not smooth. Nonetheless, the estimate is quite accurate.* $\blacksquare$

22.5 Bibliographic Remarks

A reference for orthogonal function methods is Efromovich (1999). See also Beran (2000) and Beran and Dümbgen (1998). An introduction to wavelets is given in Ogden (1997). A more advanced treatment can be found in Härdle, Kerkyacharian, Picard and Tsybakov (1998). The theory of statistical estimation

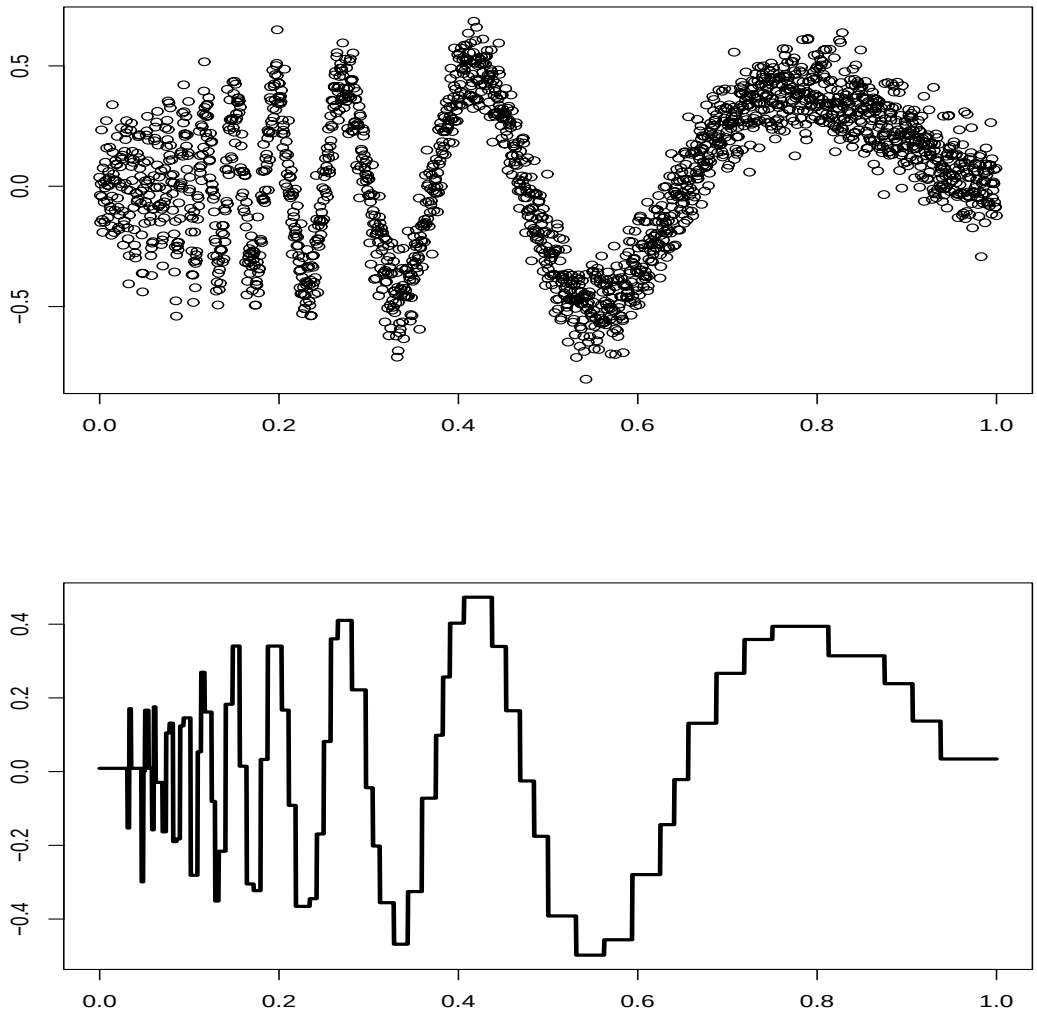


FIGURE 22.10. Estimate of the Doppler function using Haar wavelets and universal thresholding.

using wavelets has been developed by many authors, especially David Donoho and Ian Johnstone. There is a series of papers especially worth reading: Donoho and Johnstone (1994), Donoho and Johnstone (1995a), Donoho and Johnstone (1995b), and Donoho and Johnstone (1998).

22.6 Exercises

1. Prove Theorem 22.5.

2. Prove Theorem 22.9.

3. Let

$$\psi_1 = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right), \ \psi_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right), \ \psi_3 = \left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{2}{\sqrt{6}}\right).$$

   Show that these vectors have norm 1 and are orthogonal.

4. Prove Parseval's relation equation (22.6).

5. Plot the first five Legendre polynomials. Very, numerically, that they are orthonormal.

6. Expand the following functions in the cosine basis on $[0, 1]$. For (a) and (b), find the coefficients β_j analytically. For (c) and (d), find the coefficients β_j numerically, i.e.

$$\beta_j = \int_0^1 f(x)\phi_j(x) \approx \frac{1}{N}\sum_{r=1}^{N} f\left(\frac{r}{N}\right)\phi_j\left(\frac{r}{N}\right)$$

   for some large integer N. Then plot the partial sum $\sum_{j=1}^{n} \beta_j \phi_j(x)$ for increasing values of n.

   (a) $f(x) = \sqrt{2}\cos(3\pi x)$.

   (b) $f(x) = \sin(\pi x)$.

   (c) $f(x) = \sum_{j=1}^{11} h_j K(x - t_j)$ where $K(t) = (1 + \text{sign}(t))/2$,

$(t_j) = (.1, .13, .15, .23, .25, .40, .44, .65, .76, .78, .81)$,

$(h_j) = (4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2)$.

(d) $f = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{(x+.05)}\right)$.

7. Consider the glass fragments data. Let Y be refractive index and let X be aluminium content (the fourth variable).

   (a) Do a nonparametric regression to fit the model $Y = f(x)+\epsilon$ using the cosine basis method. The data are not on a regular grid. Ignore this when estimating the function. (But do sort the data first.) Provide a function estimate, an estimate of the risk and a confidence band.

   (b) Use the wavelet method to estimate f.

8. Show that the Haar wavelets are orthonormal.

9. Consider again the doppler signal:

$$f(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x+0.05}\right).$$

   Let $n = 1024$ and let $(x_1, \ldots, x_n) = (1/n, \ldots, 1)$. Generate data

$$Y_i = f(x_i) + \sigma\epsilon_i$$

   where $\epsilon_i \sim N(0,1)$.

   (a) Fit the curve using the cosine basis method. Plot the function estimate and confidence band for $J = 10, 20, \ldots, 100$.

   (b) Use Haar wavelets to fit the curve.

10. (Haar density Estimation.) Let $X_1, \ldots, X_n \sim f$ for some density f on $[0,1]$. Let's consider constructing a wavelet histogram. Let ϕ and ψ be the Haar father and mother wavelet. Write

$$f(x) \approx \phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}(x)$$

where $J \approx \log_2(n)$. Let

$$\widehat{\beta}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \psi_{j,k}(X_i).$$

(a) Show that $\widehat{\beta}_{j,k}$ is an unbiased estimate of $\beta_{j,k}$.

(b) Define the Haar histogram

$$\widehat{f}(x) = \phi(x) + \sum_{j=0}^{B}\sum_{k=0}^{2^j-1} \widehat{\beta}_{j,k}\psi_{j,k}(x)$$

for $0 \leq B \leq J-1$.

(c) Find an approximate expression for the MSE as a function of B.

(d) Generate $n = 1000$ observations from a Beta (15,4) density. Estimate the density using the Haar histogram. Use leave-one-out cross vaidation to choose B.

11. In this question, we will explore the motivation for equation (22.38). Let $X_1, \ldots, X_n \sim N(0, \sigma^2)$. Let

$$\widehat{\sigma} = \sqrt{n} \times \frac{\text{median}\left(|X_1|, \ldots, |X_n|\right)}{0.6745}.$$

(a) Show that $\mathbb{E}(\widehat{\sigma}) = \sigma$.

(b) Simulate $n = 100$ observations from a N(0,1) distribution. Compute $\widehat{\sigma}$ as well as the usual estimate of σ. Repeat 1000 times and compare the MSE.

(c) Repeat (b) but add some outliers to the data. To do this, simulate each observation from a N(0,1) with probability .95 and simulate each observation from a N(0,10) with probability .95.

12. Repeat question 6 using the Haar basis.

23
Classification

23.1 Introduction

The problem of predicting a discrete random variable Y from another random variable X is called **classification**, **supervised learning**, **discrimination** or **pattern recognition.**

In more detail, consider IID data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where

$$X_i = (X_{i1}, \ldots, X_{id}) \in \mathcal{X} \subset \mathbb{R}^d$$

is a d-dimensional vector and Y_i takes values in some finite set $\mathcal{Y}$. A **classification rule** is a function $h : \mathcal{X} \to \mathcal{Y}$. When we observe a new X, we predict Y to be $h(X)$.

Example 23.1 *Here is a an example with fake data. Figure 23.1 shows 100 data points. The covariate $X = (X_1, X_2)$ is 2-dimensional and the outcome $Y \in \mathcal{Y} = \{0, 1\}$. The Y values are indicated on the plot with the triangles representing $Y = 1$ and the squares representing $Y = 0$. Also shown is a linear classification rule represented by the solid line. This is a rule of the form*

$$h(x) = \begin{cases} 1 & \text{if } a + b_1 x_1 + b_2 x_2 > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Everything above the line is classified as a 0 and everything below the line is classified as a 1. ■

Example 23.2 *The Coronary Risk-Factor Study (CORIS) data involve 462 males between the ages of 15 and 64 from three rural areas in South Africa (Rousseauw et al 1983, Hastie and Tibshirani, 1987). The outcome Y is the presence ($Y = 1$) or absence ($Y = 0$) of coronary heart disease. There are 9 covariates: systolic blood pressure, cumulative tobacco (kg), ldl (low density lipoprotein cholesterol), adiposity, famhist (family history of heart disease), typea (type-A behavior), obesity, alcohol (current alcohol consumption), and age. Figure 23.2 shows the outcomes and two of the covariates, systolic blood pressure and tobaco consumption. People with/without heart disease are coded with triangles and squares. Also shown is a linear decision boundary which we will explain shortly. In this example, the groups are much harder to tell apart. In fact, 141 people are misclassified using this classification rule.*

At this point, it is worth revisiting the Statistics/Data Mining dictionary:

Statistics	Computer Science	Meaning
classification	supervised learning	predicting a discrete Y from X
data	training sample	$(X_1, Y_1), \ldots, (X_n, Y_n)$
covariates	features	the X_i's
classifier	hypothesis	map $h : \mathcal{X} \to \mathcal{Y}$
estimation	learning	finding a good classifier

In most cases in this Chapter, we deal with the case $\mathcal{Y} = \{0, 1\}$.

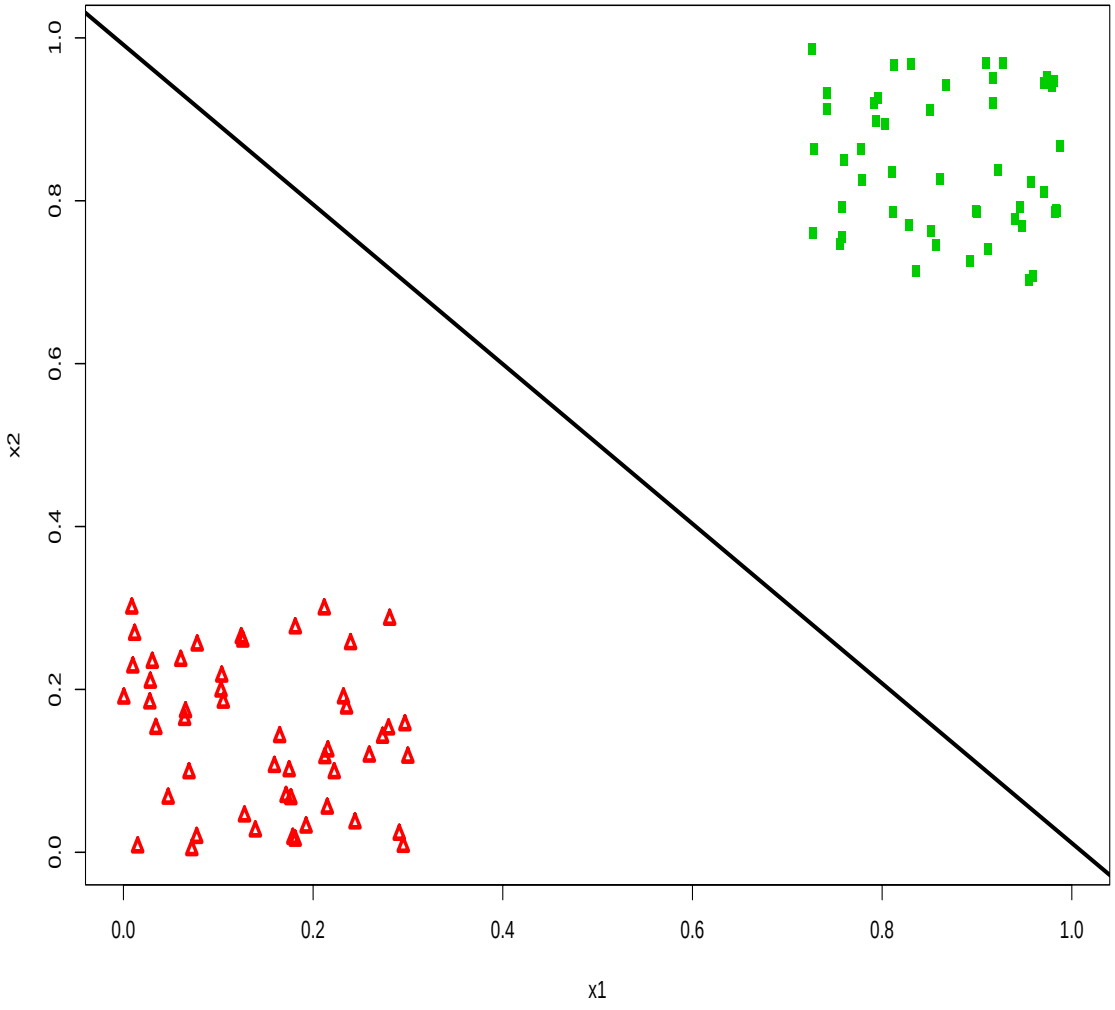


FIGURE 23.1. Two covariates and a linear decision boundary. $\triangle$ means $Y = 1$. $\square$ means $Y = 0$. You won't see real data like this.

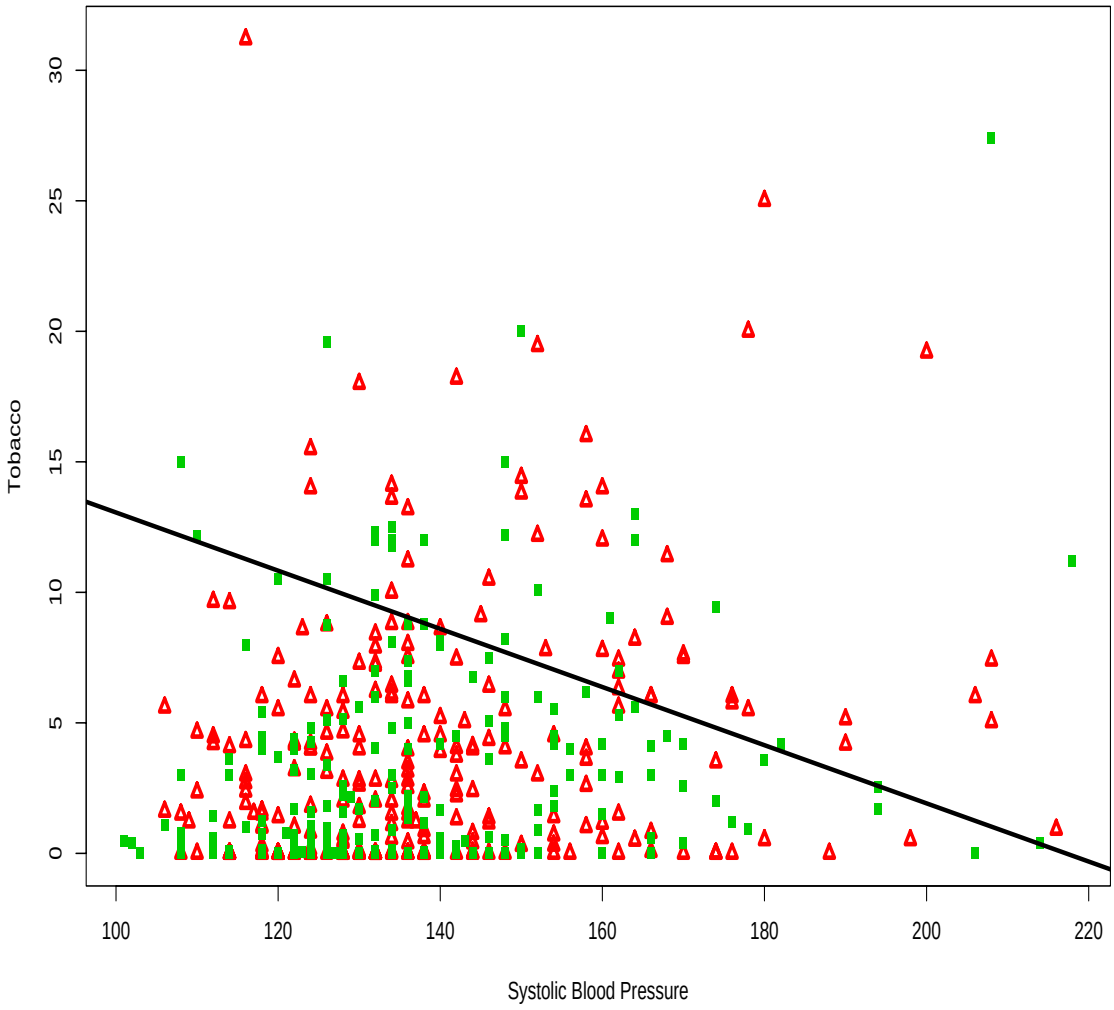


FIGURE 23.2. Two covariates and and a linear decision boundary with data from the Coronary Risk-Factor Study (CORIS). $\triangle$ means $Y = 1$. $\square$ means $Y = 0$. This is real life.

23.2 Error Rates and The Bayes Classifier

Our goal is to find a classification rule h that makes accurate predictions. We start with the following definitions.

One can use other measures of error as well.

Definition 23.3 *The* **true error rate** *of a classifier h is*

$$L(h) = \mathbb{P}(\{h(X) \neq Y\}) \tag{23.1}$$

and the **empirical error rate** *or* **training error rate** *is*

$$\widehat{L}_n(h) = \frac{1}{n}\sum_{i=1}^{n} I(h(X_i) \neq Y_i). \tag{23.2}$$

First we consider the special case where $\mathcal{Y} = \{0, 1\}$. Let

$$r(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$$

denote the **regression function**. From Bayes' theorem we have that

$$r(x) = \frac{\pi f_1(x)}{\pi f_1(x) + (1-\pi) f_0(x)} \tag{23.3}$$

where

$$\begin{aligned} f_0(x) &= f(x|Y = 0) \\ f_1(x) &= f(x|Y = 1) \end{aligned}$$

and $\pi = \mathbb{P}(Y = 1)$.

Definition 23.4 *The* **Bayes classification rule** h^* *is defined to be*

$$h^*(x) = \begin{cases} 1 & \text{if } r(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{23.4}$$

The set $\mathcal{D}(h) = \{x : \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 0|X = x)\}$ is called the **decision boundary.**

Warning! The Bayes rule has nothing to do with Bayesian inference. We could estimate the Bayes rule using either frequentist or Bayesian methods.

The Bayes' rule may be written in several equivalent forms:

$$h^*(x) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X=x) > \mathbb{P}(Y=0|X=x) \\ 0 & \text{otherwise} \end{cases} \tag{23.5}$$

and

$$h^*(x) = \begin{cases} 1 & \text{if } \pi f_1(x) > (1-\pi) f_0(x) \\ 0 & \text{otherwise.} \end{cases} \tag{23.6}$$

Theorem **23.5** *The Bayes rule is optimal, that is, if h is any other classification rule then $L(h^*) \leq L(h)$.*

The Bayes rule depends on unknown quantities so we need to use the data to find some approximation to the Bayes rule. At the risk of oversimplifying, there are three main approaches:

1. Empirical Risk Minimization. Choose a set of classifiers $\mathcal{H}$ and find $\widehat{h} \in \mathcal{H}$ that minimizes some estimate of $L(h)$.

2. Regression. Find an estimate $\widehat{r}$ of the regression function r and define

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

3. Density Estimation. Estimate f_0 from the X_i's for which $Y_i = 0$, estimate f_1 from the X_i's for which $Y_i = 1$ and let $\widehat{\pi} = n^{-1} \sum_{i=1}^n Y_i$. Define

$$\widehat{r}(x) = \widehat{\mathbb{P}}(Y=1|X=x) = \frac{\widehat{\pi}\widehat{f}_1(x)}{\widehat{\pi}\widehat{f}_1(x) + (1-\widehat{\pi})\widehat{f}_0(x)}$$

and

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Now let us generalize to the case where Y takes on more than two values as follows.

Theorem **23.6** *Suppose that* $Y \in \mathcal{Y} = \{1, \ldots, K\}$. *The optimal rule is*

$$h(x) = \text{argmax}_k \, \mathbb{P}(Y = k|X = x) \quad (23.7)$$

$$= \text{argmax}_k \pi_k \, f_k(x) \quad (23.8)$$

where

$$\mathbb{P}(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_r f_r(x)\pi_r}, \quad (23.9)$$

$\pi_r = P(Y = r)$, $f_r(x) = f(x|Y = r)$ *and* argmax_k *means "the value of* k *that maximizes that expression."*

23.3 Gaussian and Linear Classifiers

Perhaps the simplest approach to classification is to use the density estimation strategy and assume a parametric model for the densities. Suppose that $\mathcal{Y} = \{0, 1\}$ and that $f_0(x) = f(x|Y = 0)$ and $f_1(x) = f(x|Y = 1)$ are both multivariate Gaussians:

$$f_k(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right\}, \; k = 0, 1.$$

Thus, $X|Y = 0 \sim N(\mu_0, \Sigma_0)$ and $X|Y = 1 \sim N(\mu_1, \Sigma_1)$.

Theorem **23.7** *If* $X|Y=0 \sim N(\mu_0, \Sigma_0)$ *and* $X|Y=1 \sim N(\mu_1, \Sigma_1)$*, then the Bayes rule is*

$$h^*(x) = \begin{cases} 1 & \text{if } r_1^2 < r_0^2 + 2\log\left(\frac{\pi_1}{\pi_0}\right) + \log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) \\ 0 & \text{otherwise} \end{cases} \tag{23.10}$$

where

$$r_i^2 = (x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i), \quad i=1,2 \tag{23.11}$$

is the **Manalahobis distance.** *An equivalent way of expressing the Bayes' rule is*

$$h(x) = \text{argmax}_k \delta_k(x)$$

where

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k \tag{23.12}$$

and $|A|$ *denotes the determinant of a matrix* A*.*

The decision boundary of the above classifier is quadratic so this procedure is often called **quadratic discriminant analysis (QDA)**. In practice, we use sample estimates of $\pi, \mu_1, \mu_2, \Sigma_0, \Sigma_1$ in place of the true value, namely:

$$\begin{aligned}
\widehat{\pi}_0 &= \frac{1}{n}\sum_{i=1}^{n}(1-Y_i), \quad \widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i \\
\widehat{\mu}_0 &= \frac{1}{n_0}\sum_{i:\ Y_i=0} X_i, \quad \widehat{\mu}_1 = \frac{1}{n_1}\sum_{i:\ Y_i=1} X_i \\
S_0 &= \frac{1}{n_0}\sum_{i:\ Y_i=0}(X_i-\widehat{\mu}_0)(X_i-\widehat{\mu}_0)^T, \quad S_1 = \frac{1}{n_1}\sum_{i:\ Y_i=1}(X_i-\widehat{\mu}_1)(X_i-\widehat{\mu}_1)^T
\end{aligned}$$

where $n_0 = \sum_i (1-Y_i)$ and $n_1 = \sum_i Y_i$.

A simplification occurs if we assume that $\Sigma_0 = \Sigma_0 = \Sigma$. In that case, the Bayes rule is

$$h(x) = \text{argmax}_k \delta_k(x) \tag{23.13}$$

where now

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} + \log \pi_k. \tag{23.14}$$

The parameters are estimated as before except that the MLE of Σ is

$$S = \frac{n_0 S_0 + n_1 S_1}{n_0 + n_1}.$$

The classification rule is

$$h^*(x) = \begin{cases} 1 & \text{if } \delta_1(x) > \delta_0(x) \\ 0 & \text{otherwise} \end{cases} \tag{23.15}$$

where

$$\delta_j(x) = x^T S^{-1} \widehat{\mu}_j - \frac{1}{2} \widehat{\mu}_j^T S^{-1} \widehat{\mu}_j + \log \widehat{\pi}_j$$

is called the **discriminant function**. The decision boundary $\{x : \delta_0(x) = \delta_1(x)\}$ is linear so this method is called **linear discrimination analysis (LDA).**

Example 23.8 *Let us return to the South African heart disease data. The decision rule in Figure 23.2 was obtained by linear discrimination. The outcome was*

	classified as 0	*classified as 1*
$y = 0$	*277*	*25*
$y = 1$	*116*	*44*

The observed misclassification rate is $141/462 = .31$. *Including all the covariates reduces the error rate to .27. The results from quadratic discrimination are*

	classified as 0	*classified as 1*
$y = 0$	*272*	*30*
$y = 1$	*113*	*47*

which has about the same error rate $143/462 = .31$. *Including all the covariates reduces the error rate to .26. In this example, there is little advantage to QDA over LDA.*

Now we generalize to the case where Y takes on more than two values.

Theorem **23.9** *Suppose that* $Y \in \{1, \ldots, K\}$. *If* $f_k(x) = f(x|Y = k)$ *is Gaussian, the Bayes rule is*

$$h(x) = \text{argmax}_k \delta_k(x)$$

where

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + \log \pi_k. \tag{23.16}$$

If the variances of the Gaussians are equal then

$$\delta_k(x) = x^T \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} + \log \pi_k. \tag{23.17}$$

We estimate $\delta_k(x)$ by by inserting estimates of μ_k, Σ_k and π_k.

There is another version of linear discriminant analysis due to Fisher. The idea is to first reduce the dimension of covariates to one dimension by projecting the data onto a line. Algebraically, this meansd replacing the covariate $X = (X_1, \ldots, X_d)$ with a linear combination $U = w^T X = \sum_{j=1}^d w_j X_j$. The goal is to choose the vector $w = (w_1, \ldots, w_d)$ that "best separates the data." Then we perform classification with the new covariate Z instead of X.

We need define what we mean by separation of the groups. We would like the two groups to have means that are far apart relative to their spread. Let μ_j denote the mean of X for Y_j and let Σ be the variance matrix of X. Then $\mathbb{E}(U|Y = j) = \mathbb{E}(w^T X|Y = j) = w^T \mu_j$ and $\mathbb{V}(U) = w^T \Sigma w$. Define the separa-

The quantity J arises in physics, where it is called the Rayleigh coefficient.

tion by

$$\begin{aligned} J(w) &= \frac{(\mathbb{E}(U|Y=0) - \mathbb{E}(U|Y=1))^2}{w^T \Sigma w} \\ &= \frac{(w^T\mu_0 - w^T\mu_1)^2}{w^T\Sigma w} \\ &= \frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T w}{w^T\Sigma w}. \end{aligned}$$

We estimate J as follows. Let $n_j \sum_{i=1}^n I(Y_i = j)$ be the number of observations in group j, let $\overline{X}_j$ be the sample mean vector of the X's for group j, and let S_j be the sample covariance matrix in group j. Define

$$\widehat{J}(w) = \frac{w^T S_B w}{w^T S_W w} \tag{23.18}$$

where

$$\begin{aligned} S_B &= (\overline{X}_0 - \overline{X}_1)(\overline{X}_0 - \overline{X}_1) \\ S_W &= \frac{(n_0-1)S_0 + (n_1-1)S_1}{(n_0-1)+(n_1-1)}. \end{aligned}$$

Theorem **23.10** *The vector*

$$w = S_W^{-1}(\overline{X}_0 - \overline{X}_1) \tag{23.19}$$

is a minimizer of $\widehat{J}(w)$. *We call*

$$U = w^T X = (\overline{X}_0 - \overline{X}_1)^T S_W^{-1} X \tag{23.20}$$

the **Fisher linear discriminant function**. *The midpoint* m *between* $\overline{X}_0$ *and* $\overline{X}_1$ *is*

$$m = \frac{1}{2}(\overline{X}_0 + \overline{X}_1) = \frac{1}{2}(\overline{X}_0 - \overline{X}_1)^T S_B^{-1}(\overline{X}_0 + \overline{X}_1) \tag{23.21}$$

Fisher's classification rule is

$$\begin{aligned} h(x) &= \begin{cases} 0 & \text{if } w^T X \geq m \\ 1 & \text{if } w^T X < m \end{cases} \\ &= \begin{cases} 0 & \text{if } (\overline{X}_0 - \overline{X}_1)^T S_W^{-1} x \geq m \\ 1 & \text{if } (\overline{X}_0 - \overline{X}_1)^T S_W^{-1} x < m. \end{cases} \end{aligned}$$

Fisher's rule is the same as the Bayes linear classifier in equation (23.14) when $\widehat{\pi} = 1/2$.

23.4 Linear Regression and Logistic Regression

A more direct approach to classification is to estimate the regression function $r(x) = \mathbb{E}(Y|X = x)$ without bothering to estimste the densities f_k. For the rest of this section, we will only consider the case where $\mathcal{Y} = \{0, 1\}$. Thus, $r(x) = \mathbb{P}(Y = 1|X = x)$ and once we have an estimate $\widehat{r}$, we will use the classification rule

$$h(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{23.22}$$

The simplest regression model is the linear regression model

$$Y = r(x) + \epsilon = \beta_0 + \sum_{j=1}^{d} \beta_j X_j + \epsilon \tag{23.23}$$

where $\mathbb{E}(\epsilon) = 0$. This model can't be correct since it does not force $Y = 0$ or 1. Nonetheless, it can sometimes lead to a decent classifier.

Recall that the least squares estimate of $\beta = (\beta_0, \beta_1, \ldots, \beta_d)^T$ minimizes the residual sums of squares

$$RSS(\beta) = \sum_{i=1}^{n} \Big(Y_i - \beta_0 - \sum_{j=1}^{d} X_{ij}\beta_j \Big)^2 .$$

Let us briefly review what that estimator is. Let $\mathbf{X}$ denote the $N \times (d + 1)$ matrix of the form

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1d} \\ 1 & X_{21} & \ldots & X_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \ldots & X_{nd} \end{bmatrix} .$$

Also let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. Then,

$$RSS(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta)$$

and the model can be written as

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. The least squares solution $\widehat{\beta}$ that minimizes RSS can be found by elementary calculus and is given by

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y.$$

The predicted values are

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta}.$$

Now we use (23.22) to classify, where $\widehat{r}(x) = \widehat{\beta}_0 + \sum_j \widehat{\beta}_j x_j$.

It seems sensible to use a regression method that takes into account that $Y \in \{0, 1\}$. The most common method for doing so is **logistic regression.** Recall that the model is

$$r(x) = P(Y = 1|X = x) = \frac{e^{\beta_0 + \sum_j \beta_j x_j}}{1 + e^{\beta_0 + \sum_j \beta_j x_j}}. \tag{23.24}$$

We may write this is

$$\text{logit } P(Y = 1|X = x) = \beta_0 + \sum_j \beta_j x_j$$

where logit $(a) = \log(a/(1-a))$. Under this model, each Y_i is a Bernoulli with success probability

$$p_i(\beta) = \frac{e^{\beta_0 + \sum_j \beta_j X_{ij}}}{1 + e^{\beta_0 + \sum_j \beta_j X_{ij}}}.$$

The likelihood function for the data set is

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} p_i(\beta)^{Y_i}(1 - p_i(\beta))^{1-Y_i}.$$

We obtain the MLE numerically.

Example 23.11 *Let us return to the heart disease data. A logistic regression yields the following estimates and Wald tests for the coefficients:*

Covariate	*Estimate*	*Std. Error*	*Wald statistic*	*p-value*
Intercept	*-6.145*	*1.300*	*-4.738*	*0.000*
sbp	*0.007*	*0.006*	*1.138*	*0.255*
tobacco	*0.079*	*0.027*	*2.991*	*0.003*
ldl	*0.174*	*0.059*	*2.925*	*0.003*
adiposity	*0.019*	*0.029*	*0.637*	*0.524*
famhist	*0.925*	*0.227*	*4.078*	*0.000*
typea	*0.040*	*0.012*	*3.233*	*0.001*
obesity	*-0.063*	*0.044*	*-1.427*	*0.153*
alcohol	*0.000*	*0.004*	*0.027*	*0.979*
age	*0.045*	*0.012*	*3.754*	*0.000*

Are surprised by the fact that systolic blood pressure is not significant or by the minus sign for the obesity coefficient? If yes, then you are confusing correlation and causation. There are lots of unobserved confounding variables so the coefficients above cannot be interpreted causally. The fact that blood pressure is not significant does not mean that blood pressure is not an important cause *of heart disease. It means that it is not an important* predictor *of heart disease relative to the other variables in the model. The error rate, using this model for classification, is .27. The error rate from a linear regression is .26.*

We can get a better classifier by fitting a richer model. For example we could fit

$$\text{logit } P(Y=1|X=x) = \beta_0 + \sum_j \beta_j x_j + \sum_{j,k} \beta_{jk} x_j x_k. \quad (23.25)$$

More generally, we could add terms of up to order r for some integer r. Large values of r give a more complicated model which should fit the data better. But there is a bias-variance tradeoff which we'll discuss later.

Example 23.12 *If we use model (23.25) for the heart disease data with $r = 2$, the error rate is reduced to .22.*

The logistic regression model can easily be extended to k groups but we shall not give the details here.

23.5 Relationship Between Logistic Regression and LDA

LDA and logistic regression are almost the same thing. If we assume that each group is Gaussian with the same covariance matrix then we saw that

$$\begin{aligned}\log\left(\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}\right) &= \log\left(\frac{\pi_0}{\pi_1}\right) - \frac{1}{2}(\mu_0+\mu_1)^T\Sigma^{-1}(\mu_1-\mu_0) + x^T\Sigma^{-1}(\mu_1-\mu_0) \\ &\equiv \alpha_0 + \alpha^T x.\end{aligned}$$

On the other hand, the logistic model is, by assumption,

$$\log\left(\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}\right) = \beta_0 + \beta^T x.$$

These are the same model since they both lead to classification rules that are linear in x. The difference is in how we estimate the parameters.

The joint density of a single observation is $f(x,y) = f(x|y)f(y) = f(y|x)f(x)$. In LDA we estimated the whole joint distribution by estimating $f(x|y)$ and $f(y)$; specifically, we estimated $f_k(x) = f(x|Y=k)$, $\pi_1 = f_Y(1)$ and $\pi_0 = f_Y(0)$. We maximized the likelihood

$$\prod_i f(x_i,y_i) = \underbrace{\prod_i f(x_i|y_i)}_{\text{Gaussian}} \underbrace{\prod_i f(y_i)}_{\text{Bernoulli}}. \tag{23.26}$$

In logistic regression we maximized the conditional likelihood $\prod_i f(y_i|x_i)$ but we ignored the second term $f(x_i)$:

$$\prod_i f(x_i,y_i) = \underbrace{\prod_i f(y_i|x_i)}_{\text{logistic}} \underbrace{\prod_i f(x_i)}_{\text{ignored}}. \tag{23.27}$$

Since classification only requires knowing $f(y|x)$, we don't really need to estimate the whole joint distribution. Here is an analogy: we can estimate the mean μ of a distribution with the sample mean $\overline{X}$ without bothering to estimate the whole density function. Logistic regression leaves the marginal distribution $f(x)$ un-specified so it is more nonparametric than LDA.

To summarize: LDA and logistic regression both lead to a linear classification rule. In LDA we estimate the entire joint distribution $f(x, y) = f(x|y)f(y)$. In logistic regression we only estimate $f(y|x)$ and we don't bother estimating $f(x)$.

23.6 Density Estimation and Naive Bayes

Recall that the Bayes rule is $h(x) = \text{argmax}_k \ \pi_k \, f_k(x)$. If we can estimate π_k and f_k then we can estimate the Bayes classification rule. Estimating π_k is easy but what about f_k? We did this previously by assuming f_k was Gaussian. Another strategy is to estimate f_k with some nonparametric density estimator $\widehat{f_k}$ such as a kernel estimator. But if $x = (x_1, \ldots, x_d)$ is high dimensional, nonparametric density estimation is not very reliable. This problem is ameliorated if we assume that $X_1, \ldots, X_d$ are independent, for then, $f_k(x_1, \ldots, x_d) = \prod_{j=1}^{d} f_{kj}(x_j)$. This reduces the problem to d one-dimensional density estimation problems, within each of the k groups. The resulting classifier is called **the naive Bayes classifier.** The assumption that the components of X are independent is usually wrong yet the resulting classifier might still be accurate. Here is a summary of the steps in the naive Bayes classifier:

The Naive Bayes Classifier

1. For each group k, compute an estimate $\widehat{f}_{kj}$ of the density f_{kj} for X_j, using the data for which $Y_i = k$.
2. Let
$$\widehat{f}_k(x) = \widehat{f}_k(x_1, \ldots, x_d) = \prod_{j=1}^{d} \widehat{f}_{kj}(x_j).$$
3. Let
$$\widehat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} I(Y_i = k)$$
where $I(Y_i = k) = 1$ if $Y_i = k$ and $I(Y_i = k) = 0$ if $Y_i \neq k$.
4. Let
$$h(x) = \text{argmax}_k \, \widehat{\pi}_k \, \widehat{f}_k(x).$$

The naive Bayes classifier is especially popular when x is high dimensional and discrete. In that case, $\widehat{f}_{kj}(x_j)$ is especially simple.

23.7 Trees

Trees are classification methods that partition the covariate space $\mathcal{X}$ into disjoint pieces and then classify the observations according to which partition element they fall in. As the name implies, the classifier can be represented as a tree.

For illustration, suppose there are two covariates, $X_1 =$ age and $X_2 =$ blood pressure. Figure 23.3 shows a classification tree using these variables.

The tree is used in the following way. If a subject has Age ≥ 50 then we classify him as $Y = 1$. If a subject has Age < 50 then we check his blood pressure. If blood pressure is < 100

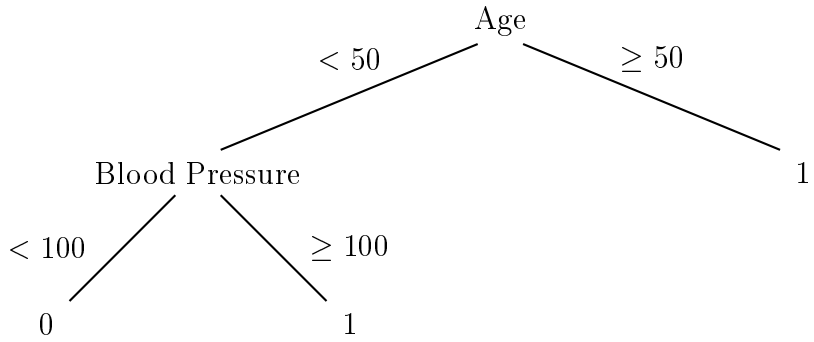


FIGURE 23.3. A simple classification tree.

then we classify him as $Y = 1$, otherwise we classify him as $Y = 0$. Figure 23.4 shows the same classifier as a partition of the covariate space.

Here is how a tree is constructed. For simplicity, we focus on the case in which $\in \mathcal{Y} = \{0, 1\}$. First, suppose there is only a single covariate X. We choose a split point t that divides the real line into two sets $A_1 = (-\infty, t]$ and $A_2 = (t, \infty)$. Let $\widehat{p}_s(j)$ be the proportion of observations in A_s such that $Y_i = j$:

$$\widehat{p}_s(j) = \frac{\sum_{i=1}^n I(Y_i = j, X_i \in A_s)}{\sum_{i=1}^n I(X_i \in A_s)} \tag{23.28}$$

for $s = 1, 2$ and $j = 0, 1$. The **impurity** of the split t is defined to be

$$I(t) = \sum_{s=1}^{2} \gamma_s \tag{23.29}$$

where

$$\gamma_s = 1 - \sum_{j=0}^{1} \widehat{p}_s(j)^2. \tag{23.30}$$

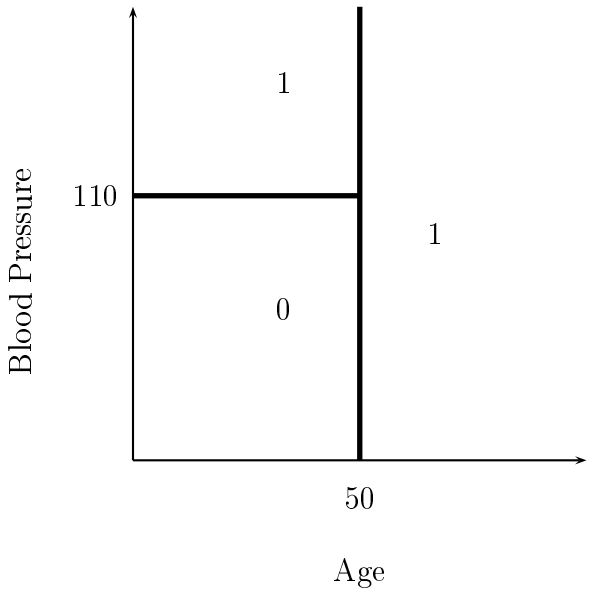


FIGURE 23.4. Partition representation of classification tree.

This measure of impurity is known as the **Gini index**. If a partition element A_s contains all 0's or all 1's, then $\gamma_s = 0$. Otherwise, $\gamma_s > 0$. We choose the split point t to minimize the impurity. (Other indices of impurity besides can be used besides the Gini index.)

When there are several covariates, we choose whichever covariate and split that leads to the lowest impurity. This process is continued until some stopping criterion is met. For example, we might stop when every partition element has fewer than n_0 data points, where n_0 is some fixed number. The bottom nodes of the tree are called the **leaves**. Each leaf is assigned a 0 or 1 depending on whether there are more data points with $Y = 0$ or $Y = 1$ in that partition element.

This procedure is easily generalized to the case where $Y \in \{1, \ldots, K\}$. We simply define the impurity by

$$\gamma_s = 1 - \sum_{j=1}^{k} \widehat{p}_s(j)^2 \tag{23.31}$$

where $\widehat{p}_i(j)$ is the proportion of observations in the partition element for which $Y = j$.

Example 23.13 *Figure 23.5 shows a classification tree for the heart disease data. The misclassification rate is .21. Suppose we do the same example but only using two covariates: tobacco and age. The misclassification rate is then .29. The tree is shown in Figure 23.6. Because there are only two covariates, we can also display the tree as a partition of $\mathcal{X} = \mathbb{R}^2$ as shown in Figure 23.7.* ■

Our description of how to build trees is incomplete. If we keep splitting until there are few cases in each leaf of the tree, we are likely to overfit the data. We should choose the complexity of the tree in such a way that the estimated true error rate is low. In the next section, we discuss estimation of the error rate.

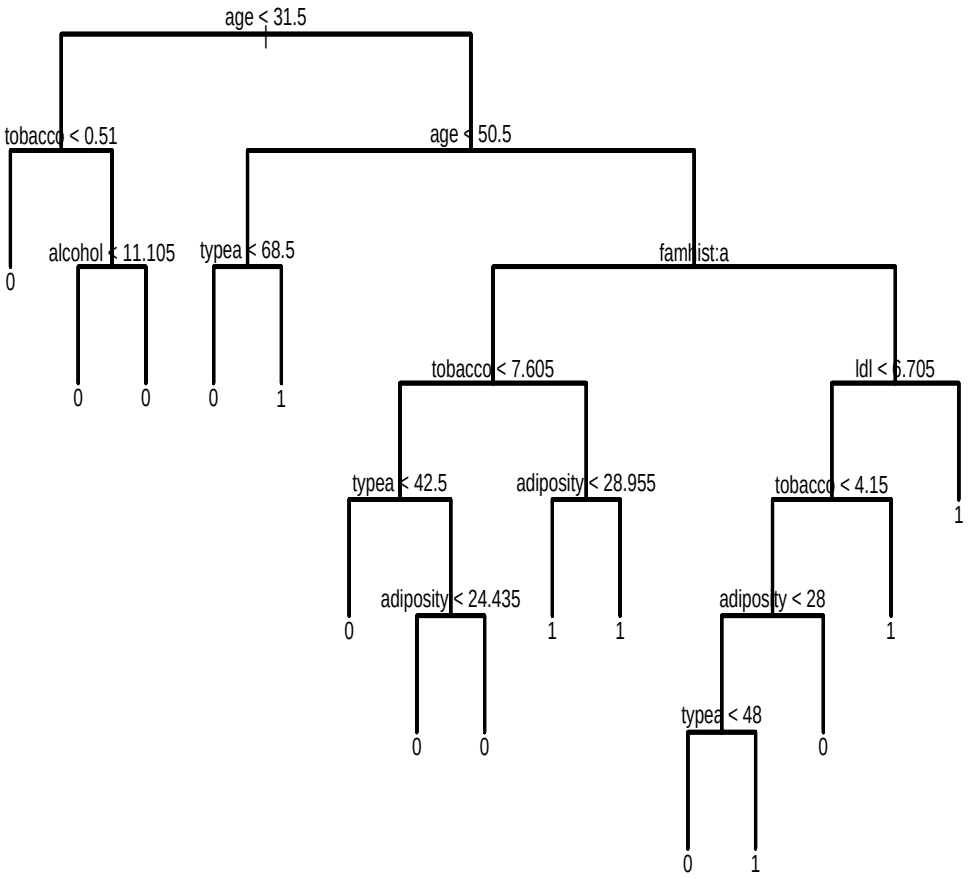


FIGURE 23.5. A classification tree for the heart disease data.

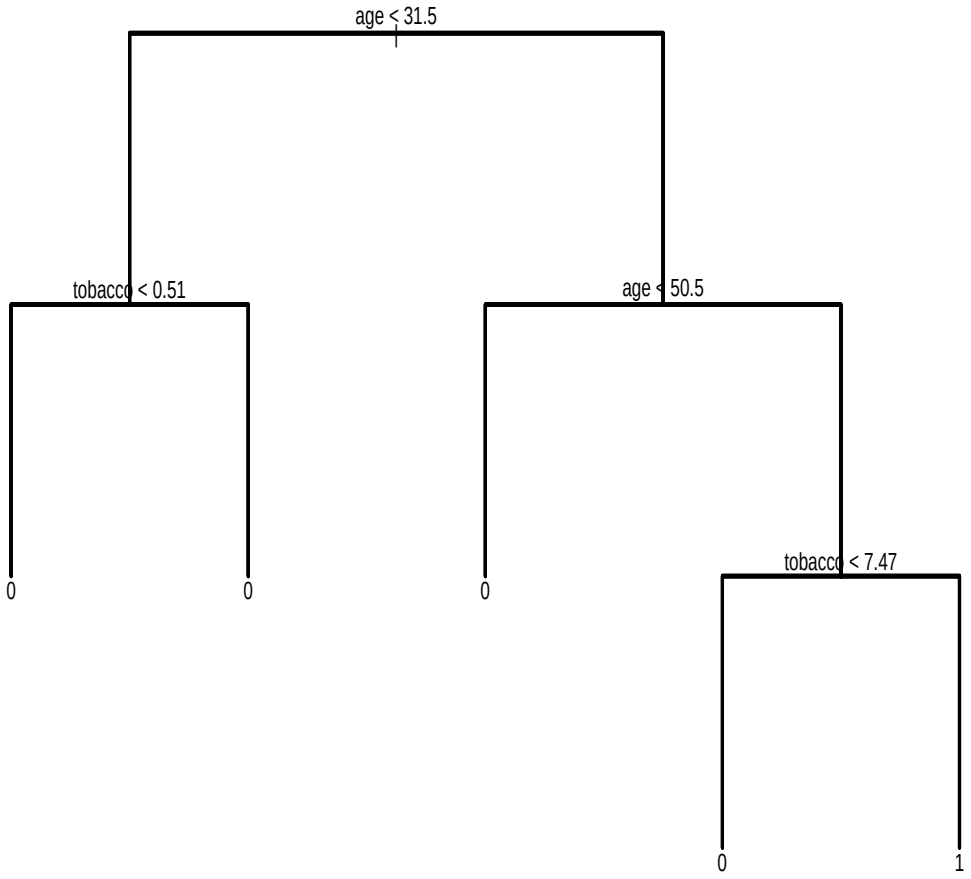


FIGURE 23.6. A classification tree for the heart disease data using two covariates.

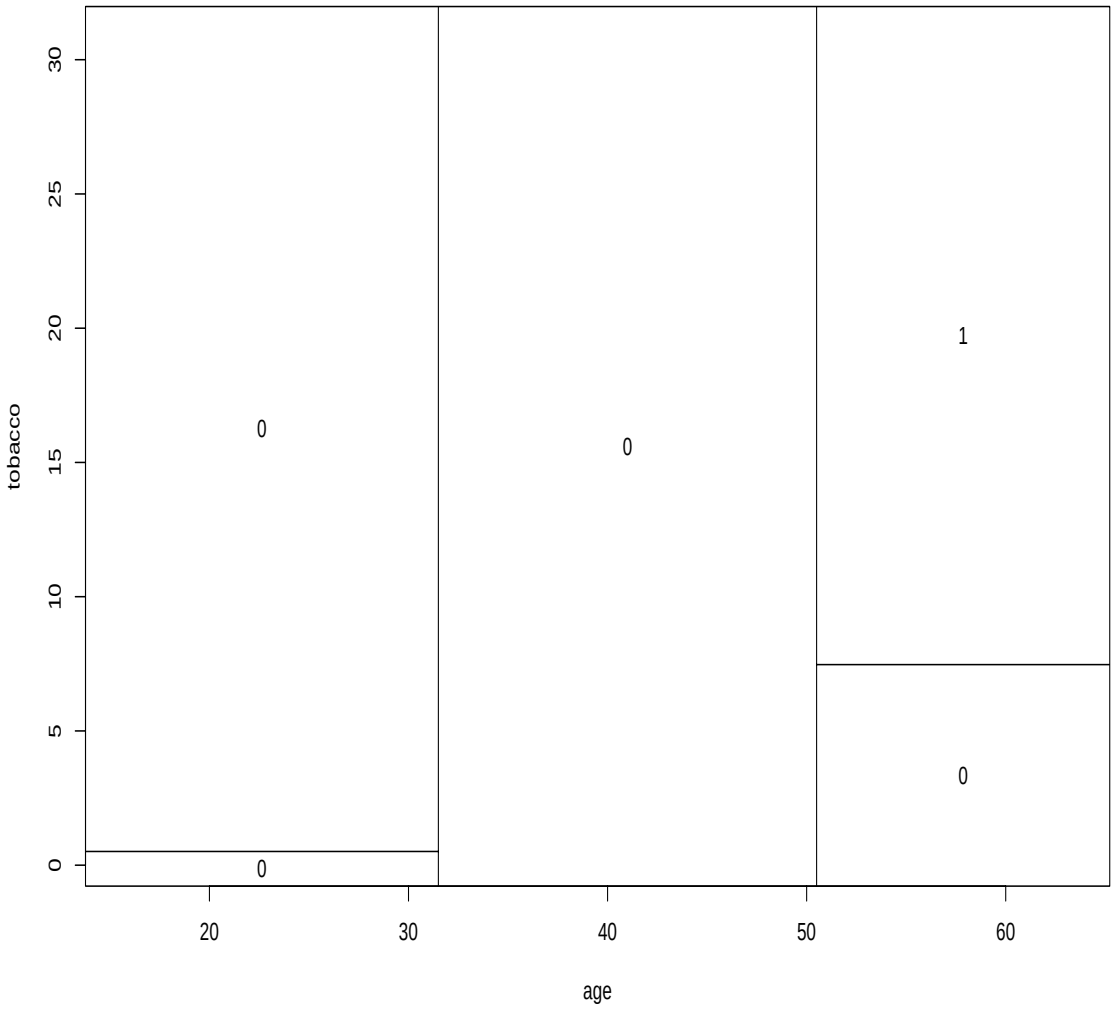


FIGURE 23.7. The tree pictured as a partition of the covariate space.

23.8 Assessing Error Rates and Choosing a Good Classifier

How do we choose a good classifier? We would like to have a classifier h with a low true error rate $L(h)$. Usually, we can't use the training error rate $\widehat{L}_n(h)$ as an estimate of the true error rate because it is biased downward.

Example 23.14 *Consider the heart disease data again. Suppose we fit a sequence logistic regression models. In the first model we include one covariate. In the second model we include two covariates and so on. The ninth model includes all the covariates. We can go even further. Let's also fit a tenth model that includes all nine covariates plus the first covariate squared. Then we fit an eleventh model that includes all nine covariates plus the first covariate squared and the second covariate squared. Continuing this way we will get a sequence of 18 classifiers of increasing complexity. The solid line in Figure 23.8 shows the observed classification error which steadily decreases as we make the model more complex. If we keep going, we can make a model with zero observed classification error. The dotted line shows the* **cross-validation estimate** *of the error rate (to be explained shortly) which is a better estimate of the true error rate than the observed classification error. The estimated error decreases for a while then increases. This is the bias-variance tradeoff phenomenon we have seen before.* ■

There are many ways to estimate the error rate. We'll consider two: **cross-validation** and **probability inequalities.**

CROSS-VALIDATION. The basic idea of cross-validation, which we have already encountered in curve estmation, is to leave out some of the data when fitting a model. The simplest version of cross-validation involves randomply splitting the data into two

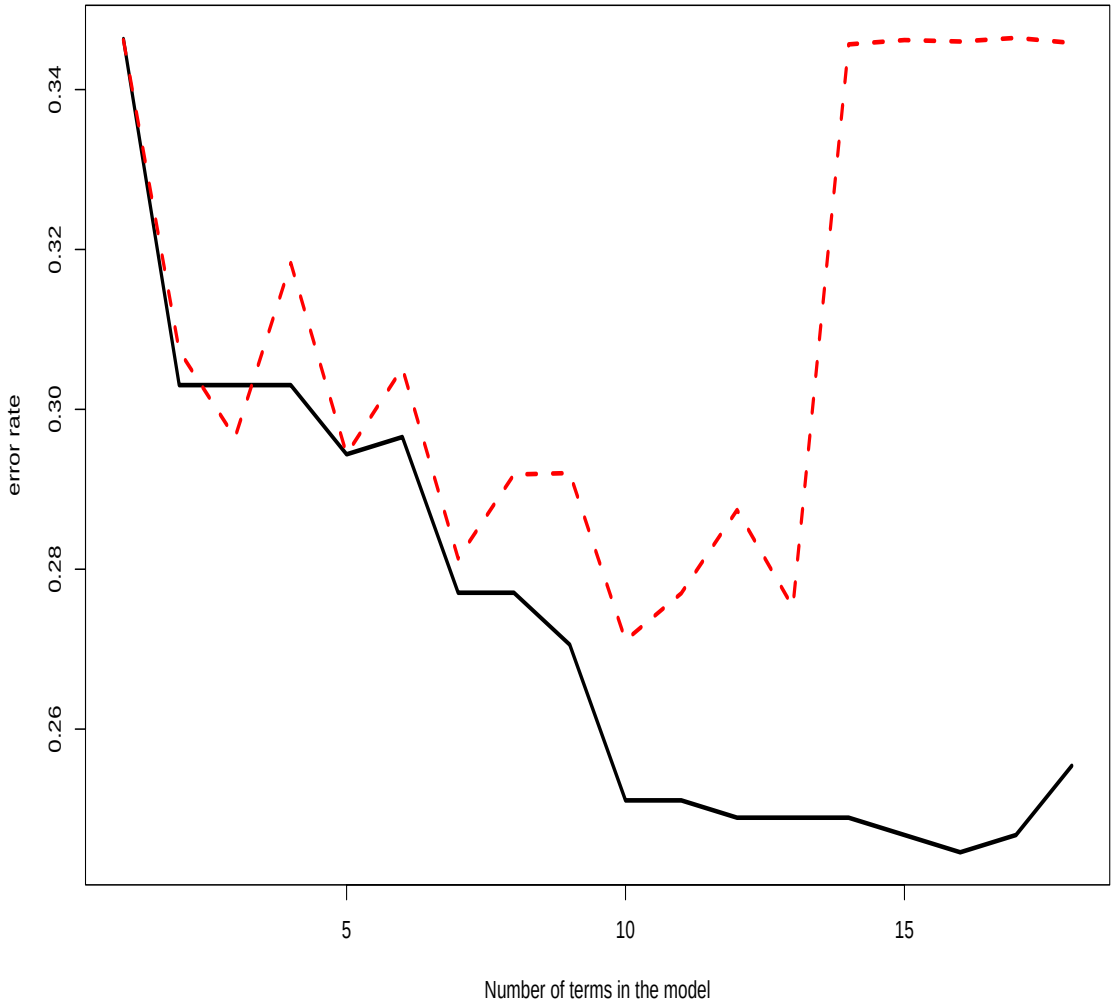


FIGURE 23.8. Solid line is the observed error rate and dashed line is the cross-validation estimate of true error rate.

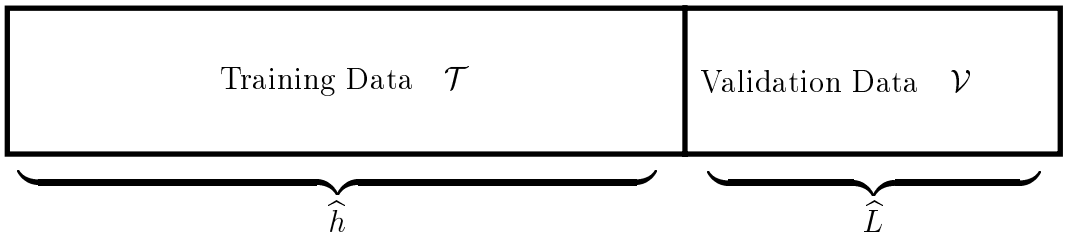


FIGURE 23.9. Cross-validation.

pieces: the **training set** $\mathcal{T}$ and the **validation set** $\mathcal{V}$. Often, about 10 per cent of the data might be set aside as the validation set. The classifier h is constructed from the training set. We then estimate the error by

$$\widehat{L}(h) = \frac{1}{m} \sum_{X_i \in \mathcal{V}} I(h(X_i) \neq Y_I). \tag{23.32}$$

where m is the size of the validation set. See Figure 23.9.

Another approach to cross-validation is **K-fold cross-validation** which is obtained from the following algorithm.

K-fold cross-validation.

1. Randomly divide the data into K chunks of approximately equal size. A common choice is $K = 10$.
2. For k = 1 to K do the following:
   (a) Delete chunk k from the data.
   (b) Compute the classifier $\widehat{h}_{(k)}$ from the rest of the data.
   (c) Use $\widehat{h}_{(k)}$ to the predict the data in chunk k. Let $\widehat{L}_{(k)}$ denote the observed error rate.
3. Let

$$\widehat{L}(h) = \frac{1}{K}\sum_{k=1}^{K}\widehat{L}_{(k)}. \tag{23.33}$$

This is how dotted line in Figure 23.8 was computed. An alternative approach to K-fold cross-validation is to split the data into two pieces called the **training set** and the **test set.** All the models are fit on the training set. The models are then used to do predictions on the test set and the errors are estimated from these predictions. When the clasification procedure is time consuming and there are lots of data, this approach is preferred to K-fold cross-validation.

Cross-validation can be applied to any classification method. To apply it to trees, one begins by fitting an initial tree. Smaller trees are obtained by pruning tree, that is removing splits from the tree. We can do this for trees of various sizes where size refers to the number of terminal nodes on the tree. Cross-validation is then used to estimate error rate as a function of tree size.

Example 23.15 *Figure 23.10 shows the cross-validation plot for a classification tree fitted to the heart disease data. The error*

is shown as a function of tree size. Figure 23.11 shows the tree with the size chosen by minimizing the cross-validation error. ■

PROBABILITY INEQUALITIES. Another approach to estimating the error rate is to find a confidence interval for $\widehat{L}_n(h)$ using probability inequalities. This method is useful in the context of **empirical risk minimization**.

Let $\mathcal{H}$ be a set of classifiers, for example, all linear classifiers. Empirical risk minimization means choosing the classifier $\widehat{h} \in \mathcal{H}$ to minimize the training error $\widehat{L}_n(h)$, also called the empirical risk. Thus,

$$\widehat{h} = \text{argmin}_{h \in \mathcal{H}} \widehat{L}_n(h) = \text{argmin}_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_i I(h(X_i) \neq Y_i) \right). \tag{23.34}$$

Typically, $\widehat{L}_n(\widehat{h})$ underestimates the true error rate $L(\widehat{h})$ because $\widehat{h}$ was chosen to make $\widehat{L}_n(\widehat{h})$ small. Our goal is to assess how much underestimation is taking place. Our main tool for this analysis is **Hoeffding's inequality.** Recall that if $X_1, \ldots, X_n \sim$ Bernoulli(p), then, for any $\epsilon > 0$,

$$\mathbb{P}\left(|\widehat{p} - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2} \tag{23.35}$$

where $\widehat{p} = n^{-1} \sum_{i=1}^n X_i$.

First, suppose that $\mathcal{H} = \{h_1, \ldots, h_m\}$ consists of finitely many classifiers. For any fixed h, $\widehat{L}_n(h)$ converges in almost surely to $L(h)$ by the law of large numbers. We will now establish a stronger result.

Theorem **23.16 (Uniform Convergence.)** *Assume $\mathcal{H}$ is finite and has m elements. Then,*

$$\mathbb{P}\left(\max_{h \in \mathcal{H}} |\widehat{L}_n(h) - L(h)| > \epsilon \right) \leq 2me^{-2n\epsilon^2}.$$

PROOF. We will use Hoeffding's inequality and we will also use the fact that if $A_1, \ldots, A_m$ is a set of events then $\mathbb{P}(\bigcup_{i=1}^m A_i) \leq$

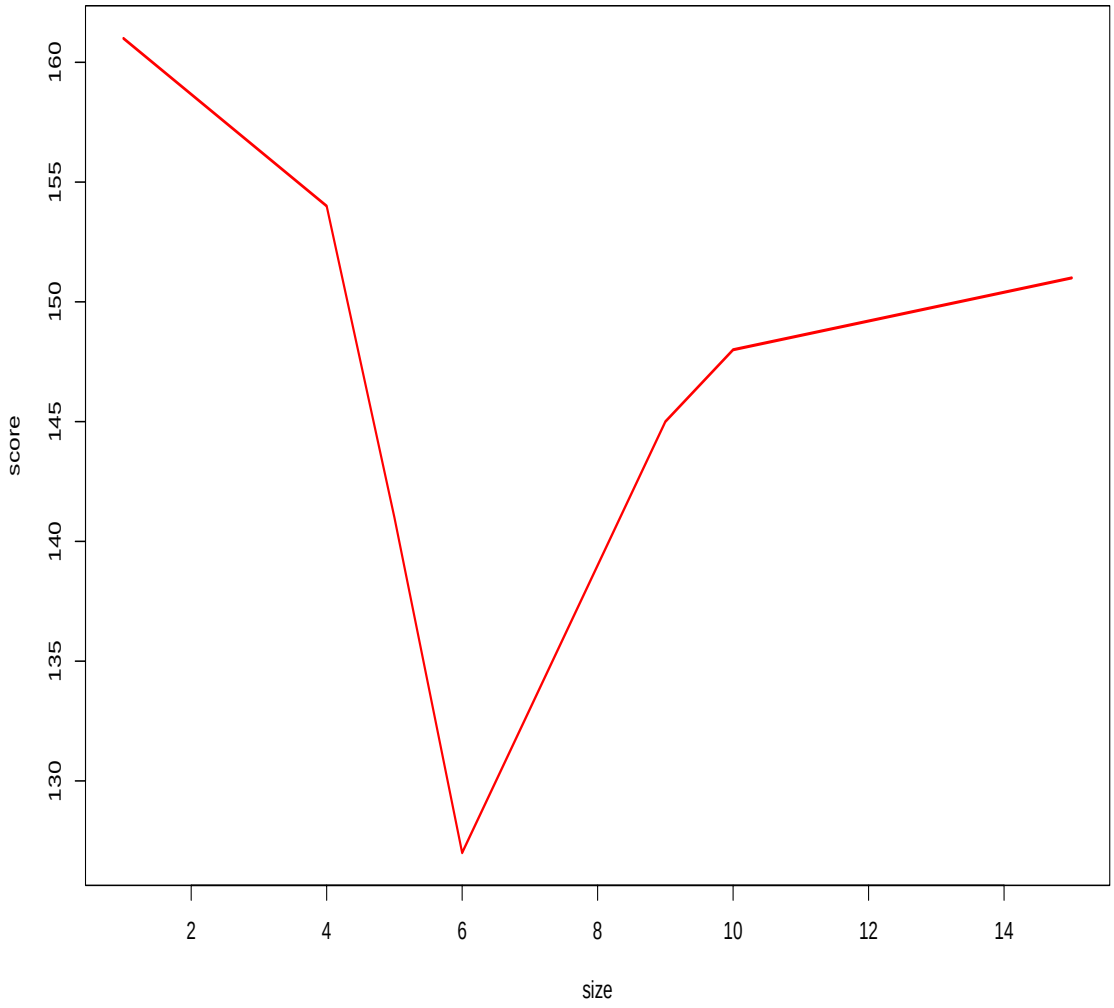


FIGURE 23.10. Cross validation plot

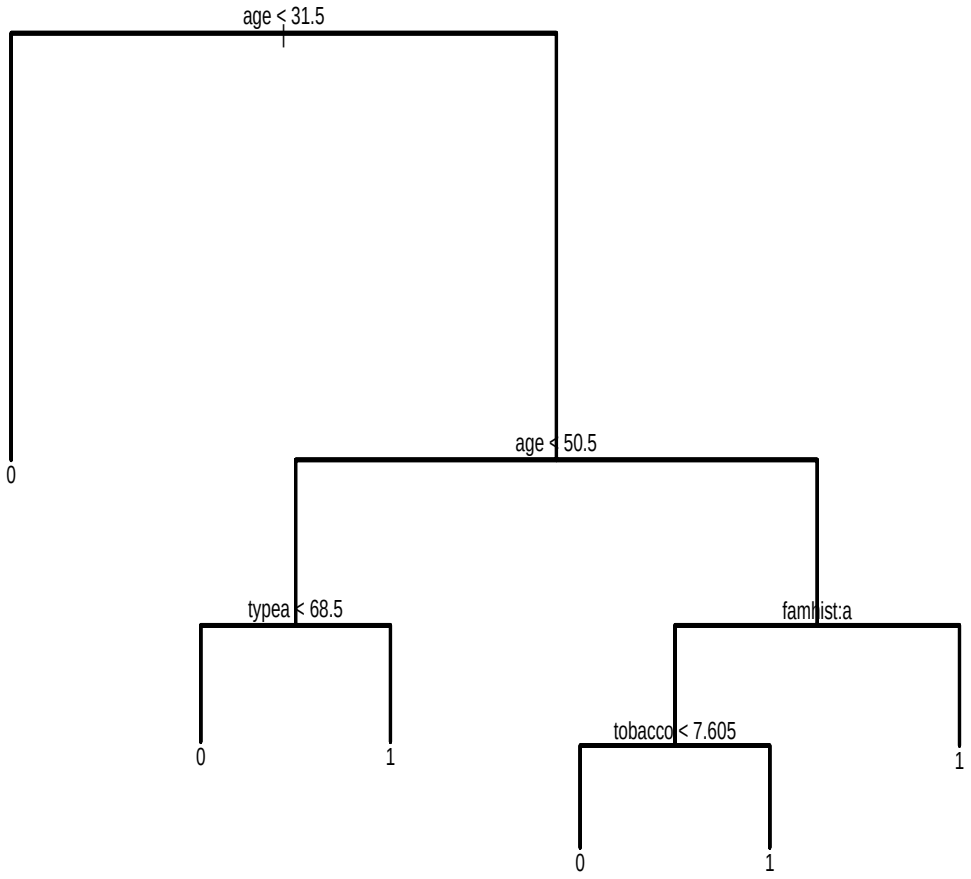


FIGURE 23.11. Smaller classification tree with size chosen by cross-validation.

$\sum_{i=1}^{m} \mathbb{P}(A_i)$. Now,

$$\begin{aligned}
\mathbb{P}\left(\max_{h\in\mathcal{H}} |\widehat{L}_n(h) - L(h)| > \epsilon\right) &= \mathbb{P}\left(\bigcup_{h\in\mathcal{H}} |\widehat{L}_n(h) - L(h)| > \epsilon\right) \\
&\leq \sum_{H\in\mathcal{H}} \mathbb{P}\left(|\widehat{L}_n(h) - L(h)| > \epsilon\right) \\
&\leq \sum_{H\in\mathcal{H}} 2e^{-2n\epsilon^2} = 2me^{-2n\epsilon^2}. \quad \blacksquare
\end{aligned}$$

Theorem **23.17** *Let*

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

Then $\widehat{L}_n(\widehat{h}) \pm \epsilon$ *is a* $1-\alpha$ *confidence interval for* $L(\widehat{h})$.

PROOF. This follows from the fact that

$$\begin{aligned}
P(|\widehat{L}_n(\widehat{h}) - L(\widehat{h})| > \epsilon) &\leq P\left(\max_{h\in\mathcal{H}} |\widehat{L}_n(\widehat{h}) - L(\widehat{h})| > \epsilon\right) \\
&\leq 2me^{-2n\epsilon^2} = \alpha. \quad \blacksquare
\end{aligned}$$

When $\mathcal{H}$ is large the confidence interval for $L(\widehat{h})$ is large. The more functions there are in $\mathcal{H}$ the more likely it is we have "overfit" which we compensate for by having a larger confidence interval.

In practice we usually use sets $\mathcal{H}$ that are infinite, such as the set of linear classifiers. To extend our analysis to these cases we want to be able to say something like

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} |\widehat{L}_n(h) - L(h)| > \epsilon\right) \leq \text{something not too big.}$$

All the other results followed from this inequality. One way to develop such a generalization is by way of the **Vapnik-Chervonenkis** or **VC dimension.** We now consider the main ideas in VC theory.

Let $\mathcal{A}$ be a class of sets. Give a finite set $F = \{x_1, \ldots, x_n\}$ let

$$N_{\mathcal{A}}(F) = \#\Big\{F \bigcap A : \ A \in \mathcal{A}\Big\} \tag{23.36}$$

be the number of subsets of F "picked out" by $\mathcal{A}$. Here $\#(B)$ denotes the number of elements of a set B. The **shatter coefficient** is defined by

$$s(\mathcal{A}, n) = \max_{F \in \mathcal{F}_n} N_{\mathcal{A}}(F) \tag{23.37}$$

where $\mathcal{F}_n$ consists of all finite sets of size n. Now let $X_1, \ldots, X_n \sim \mathbb{P}$ and let

$$\mathbb{P}_n(A) = \frac{1}{n}\sum_i I(X_i \in A)$$

denote the empirical probability measure. The following remarkable theorem bounds the distance between $\mathbb{P}$ and $\mathbb{P}_n$.

Theorem **23.18 (Vapnik and Chervonenkis (1971).)** *For any $\mathbb{P}$, n and $\epsilon > 0$,*

$$\mathbb{P}\left\{\sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)| > \epsilon\right\} \leq 8 s(\mathcal{A}, n) e^{-n\epsilon^2/32}. \tag{23.38}$$

The proof, though very elegant, is long and we omit it. If $\mathcal{H}$ is a set of classifiers, define $\mathcal{A}$ to be the class of sets of the form $\{x : \ h(x) = 1\}$. When then define $s(\mathcal{H}, n) = s(\mathcal{A}, n)$.

Theorem **23.19**

$$\mathbb{P}\left\{\sup_{h \in \mathcal{H}} |\widehat{L}_n(h) - L(h)| > \epsilon\right\} \leq 8 s(\mathcal{H}, n) e^{-n\epsilon^2/32}.$$

A $1 - \alpha$ confidence interval for $L(\widehat{h})$ is $\widehat{L}_n(\widehat{h}) \pm \epsilon_n$ where

$$\epsilon_n^2 = \frac{32}{n} \log\left(\frac{8 s(\mathcal{H}, n)}{\alpha}\right).$$

These theorems are only useful if the shatter coefficients do not grow too quickly with n. This is where VC dimension enters.

Definition 23.20 *The VC (Vapnik-Chervonenkis) dimension of a class of sets $\mathcal{A}$ is defined as follows. If $s(\mathcal{A}, n) = 2^n$ for all n set $VC(\mathcal{A}) = \infty$. Otherwise, define $VC(\mathcal{A})$ to be the largest k for which $s(\mathcal{A}, n) = 2^k$.*

Thus, the VC-dimension is the size of the largest finite set F than can be **shattered** by $\mathcal{A}$ meaning that $\mathcal{A}$ picks out each subset of F. If $\mathcal{H}$ is a set of classifiers we define $VC(\mathcal{H}) = VC(\mathcal{A})$ where $\mathcal{A}$ is the class of sets of the form $\{x : \ h(x) = 1\}$ as h varies in $\mathcal{H}$. The following theorem shows that if $\mathcal{A}$ has finite VC-dimension then the shatter coefficients grow as a polynomial in n.

Theorem **23.21** *If $\mathcal{A}$ has finite VC-dimension v, then*

$$s(\mathcal{A}, n) \leq n^v + 1.$$

Example 23.22 *Let $\mathcal{A} = \{(-\infty, a]; \ a \in \mathcal{R}\}$. The $\mathcal{A}$ shatters ever one point set $\{x\}$ but it shatters no set of the form $\{x, y\}$. Therefore, $VC(\mathcal{A}) = 1$.*

Example 23.23 *Let $\mathcal{A}$ be the set of closed intervals on the real line. Then $\mathcal{A}$ shatters $S = \{x, y\}$ but it cannot shatter sets with 3 points. Consider $S = \{x, y, z\}$ where $x < y < z$. One cannot find an interval A such that $A \bigcap S = \{x, z\}$. So, $VC(\mathcal{A}) = 2$.*

Example 23.24 *Let $\mathcal{A}$ be all linear half-spaces on the plane. Any three point set (not all on a line) can be shattered. No 4 point set can be shattered. Consider, for example, 4 points forming a diamond. Let T be the left and rightmost points. This can't be picked out. Other configurations can also be seen to be unshatterable. So $VC(\mathcal{A}) = 3$. In general, halfspaces in $\mathcal{R}^d$ have VC dimension $d + 1$.*

Example 23.25 *Let $\mathcal{A}$ be all rectangles on the plane with sides parallel to the axes. Any 4 point set can be shattered. Let S be a 5 point set. There is one point that is not leftmost, rightmost, uppermost, or lowermost. Let T be all points in S except this point. Then T can't be picked out. So $VC(\mathcal{A}) = 4$.*

Theorem **23.26** *Let x have dimension d and let $\mathcal{H}$ be th set of linear classifiers. The VC dimension of $\mathcal{H}$ is $d+1$. Hence, a $1-\alpha$ confidence interval for the true error rate is $\widehat{L}(\widehat{h}) \pm \epsilon$ where*

$$\epsilon_n^2 = \frac{32}{n} \log\left(\frac{8(n^{d+1}+1)}{\alpha}\right).$$

23.9 Support Vector Machines

In this section we consider a class of linear classifiers called **support vector machines**. Throughout this section, we assume that Y is binary. It will be convenient to label the outcomes as -1 and $+1$ instead of 0 and 1. A linear classifier can then be written as

$$h(x) = \text{sign}\Big(H(x)\Big)$$

where $x = (x_1, \ldots, x_d)$,

$$H(x) = a_0 + \sum_{i=1}^{d} a_i x_i$$

and

$$\text{sign}(z) = \begin{cases} -1 & \text{if } z < 0 \\ 0 & \text{if } z = 0 \\ 1 & \text{if } z > 0. \end{cases}$$

First, suppose that the data are **linearly separable**, that is, there exists a hyperplane that perfectly separates the two classes.

Lemma 23.27 *The data can be separated by some hyperplane if and only if there exists a hyperplane* $H(x) = a_0 + \sum_{i=1}^d a_i x_i$ *such that*

$$Y_i H(x_i) \geq 1, \quad i = 1, \ldots, n. \tag{23.39}$$

PROOF. Suppose the data can be separated by a hyperplane $W(x) = b_0 + \sum_{i=1}^d b_i x_i$. It follows that there exists some constant c such that $Y_i = 1$ implies $W(X_i) \geq c$ and $Y_i = -1$ implies $W(X_i) \leq -c$. Therefore, $Y_i W(X_i) \geq c$ for all i. Let $H(x) = a_0 + \sum_{i=1}^d a_i x_i$ where $a_j = b_j/c$. Then $Y_i H(X_i) \geq 1$ for all i. The reverse direction is straightforward. ■

In the separable case, there will be many separating hyperplanes. How should we choose one? Intuitively, it seems reasonable to choose the hyperplane "furthest" from the data in the sense that it separates the +1'a and -1's and maximizes the distance to the closest point. This hyperplane is called the **maximum margin hyperplane.** The margin is the distance to from the hyperplane to the nearest point. Points on the boundary of the margin are called **support vectors.** See Figure 23.12.

Theorem **23.28** *The hyperplane* $\widehat{H}(x) = \widehat{a}_0 + \sum_{i=1}^d \widehat{a}_i x_i$ *that separates the data and maximizes the margin is given by minimizing* $(1/2)\sum_{j=1}^d b_j^2$ *subject to (23.39).*

It turns out that this problem can be recast as a quadratic programming problem. Recall that $\langle X_i, X_k \rangle = X_i^T X_k$ is the inner product of X_i and X_k.

Theorem **23.29** *Let* $\widehat{H}(x) = \widehat{a}_0 + \sum_{i=1}^d \widehat{a}_i x_i$ *denote the optimal (largest margin) hyperplane. Then, for* $j = 1, \ldots, d$,

$$\widehat{a}_j = \sum_{i=1}^n \widehat{\alpha}_i Y_i X_j(i)$$

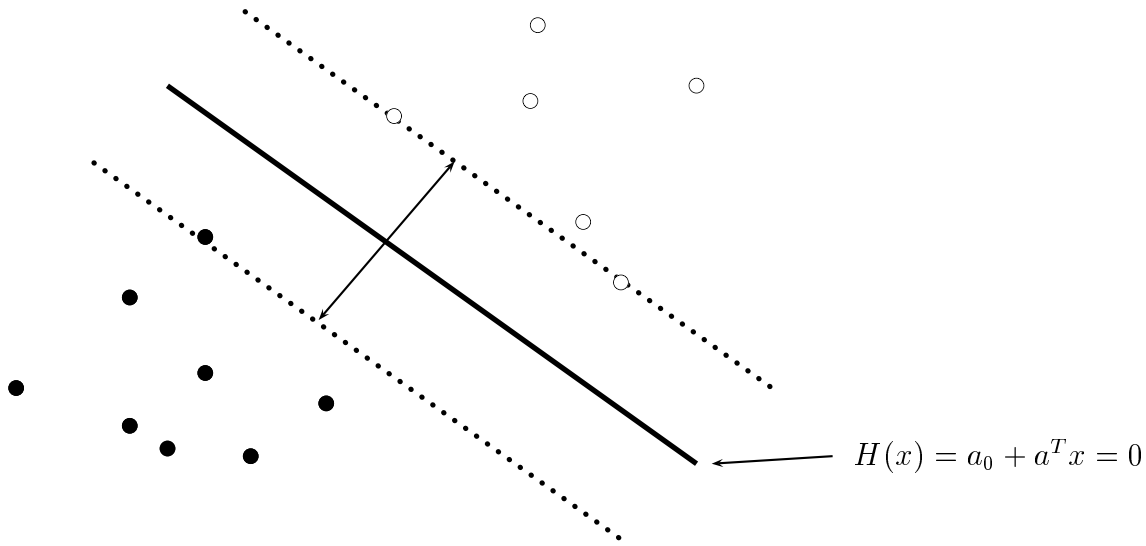


FIGURE 23.12. The hyperplane $H(x)$ has the largest margin of all hyperplanes that separate the two classes.

where $X_j(i)$ is the value of the covariate X_j for the i^{th} data point, and $\widehat{\alpha} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)$ is the vector that maximizes

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k Y_i Y_k \langle X_i, X_k \rangle \tag{23.40}$$

subject to

$$\alpha_i \geq 0$$

and

$$0 = \sum_i \alpha_i Y_i.$$

The points X_i for which $\widehat{\alpha} \neq 0$ are called **support vectors**. *$\widehat{a}_0$ can be found by solving*

$$\widehat{\alpha}_i \Big(Y_i (X_i^T \widehat{a} + \widehat{\beta}_0 \Big) = 0$$

for any support point X_i. $\widehat{H}$ may be written as

$$\widehat{H}(x) = \widehat{a}_0 + \sum_{i=1}^{n} \widehat{\alpha}_i Y_i \langle x, X_i \rangle.$$

There are many software packages that will solve this problem quickly. If there is no perfect linear classifier, then one allows overlap between the groups by replacing the condition (23.39) with

$$Y_i H(x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n. \tag{23.41}$$

The variables $\xi_1, \ldots, \xi_n$ are called **slack variables**.

We now maximize (23.40) subject to

$$0 \leq \xi_i \leq c, \quad i = 1, \ldots, n$$

and

$$\sum_{i=1}^{n} \alpha_i Y_i = 0.$$

The constant c is a tuning parameter that controls the amount of overlap.

23.10 Kernelization

There is a trick for improving a computationally simple classifier h called **kernelization**. The idea is to map the covariate X – which takes values in $\mathcal{X}$ – into a higher dimensional space $\mathcal{Z}$ and apply the classifier in the bigger space $\mathcal{Z}$. This can yield a more flexible classifier while retaining computationally simplicity.

The standard example of this idea is illustrated in Figure 23.13. The covariate $x = (x_1, x_2)$. The Y_i's can be separated into two groups using an ellipse. Define a mapping ϕ by

$$z = (z_1, z_2, z_3) = \phi(x) = (x_1^2, \sqrt{2}x_1 x_2, x_2^2).$$

This ϕ maps $\mathcal{X} = \mathbb{R}^2$ into $\mathcal{Z} = \mathbb{R}^3$. In the higher dimensional space $\mathcal{Z}$, the Y_i's are seperable by a linear decision boundary. In other words, a linear classifier in a higher dimenional space corresponds to a non-linear classifier in the original space.

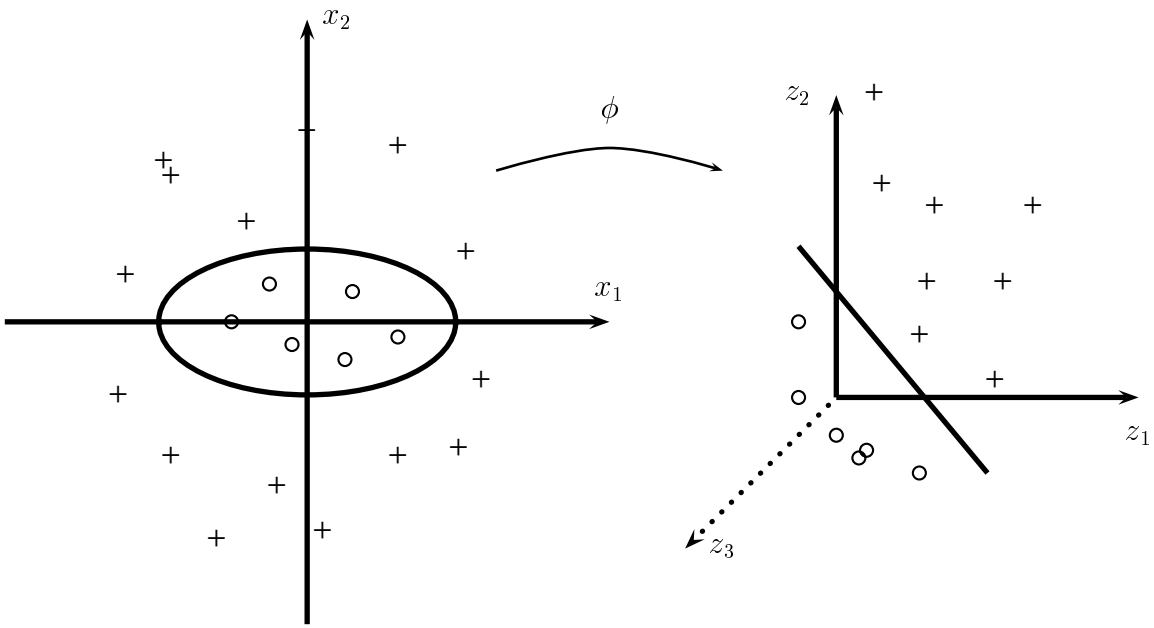


FIGURE 23.13. Kernelization. Mapping the covariates into a higher dimensional space can make a complicated decision boundary into a simpler decision boundary.

There is a potential drawback. If we signicantly expand the dimension of the problem, we might increase the computational burden. For example, x has dimension $d = 256$ and we wanted to use all fourth order terms, then $z = \phi(x)$ has dimension 183,181,376. What saves us is two observations. First, many classifiers do not require that we know the values of the individual points but, rather, just the inner product between pairs of points. Second, notice in our example that the inner product in $\mathcal{Z}$ can be written

$$\begin{aligned}\langle z, \widetilde{z}\rangle &= \langle \phi(x), \phi(\widetilde{x})\rangle \\ &= x_1^2\widetilde{x}_1^2 + 2x_1\widetilde{x}_1x_2\widetilde{x}_2 + x_2^2\widetilde{x}_2^2 \\ &= (\langle x, \widetilde{x}\rangle)^2 \\ &\equiv k(x, \widetilde{x}).\end{aligned}$$

Thus, we can compute $\langle z, \widetilde{z}\rangle$ without ever computing $Z_i = \phi(X_i)$.

To summarize, kernelization means involves finding a mapping $\phi : \mathcal{X} \to \mathcal{Z}$ and a classifier such that:

1. $\mathcal{Z}$ has higher dimension than $\mathcal{X}$ and so leads a richer set of classifiers.

2. The classifier only requires computing inner products.

3. There is a function K, called a kernel, such that $\langle \phi(x), \phi(\widetilde{x})\rangle = K(x, \widetilde{x})$.

4. Everywhere the term $\langle x, \widetilde{x}\rangle$ appears in the algorithm, replace it with $K(x, \widetilde{x})$.

In fact, we never need to construct the mapping ϕ at all. We only need to specify a kernel $k(x, \widetilde{x})$ that corresponds to $\langle \phi(x), \phi(\widetilde{x})\rangle$ for some ϕ. This raises an interesting question: given a function of two variables $K(x, y)$, does there exist a function $\phi(x)$ such that $K(x, y) = \langle \phi(x), \phi(y)\rangle$. The answer is provided

by **Mercer's theorem** which says, roughly, that if K is positive definite – meaning that

$$\int\int K(x,y)f(x)f(y)dxdy \geq 0$$

for square integrable functions f – then such a ϕ exists. Examples of commonly used kernels are:

$$\begin{array}{rrcl} \text{polynomial} & K(x,\widetilde{x}) & = & \Big(\langle x,\widetilde{x}\rangle + a\Big)^r \\ \text{sigmoid} & K(x,\widetilde{x}) & = & \tanh(a\langle x,\widetilde{x}\rangle + b) \\ \text{Gaussian} & K(x,\widetilde{x}) & = & \exp\Big(-||x-\widetilde{x}||^2/(2\sigma^2)\Big) \end{array}$$

Let us now see how we can use this trick in LDA and in support vector machines.

Recall that the Fisher linear discriminant method replaces X with $U = w^T X$ where w is chosen to maximize the Rayleigh coefficient

$$J(w) = \frac{w^T S_B w}{w^T S_W w},$$
$$S_B = (\overline{X}_0 - \overline{X}_1)(\overline{X}_0 - \overline{X}_1)^T$$

and

$$S_W = \left(\frac{(n_0-1)S_0}{(n_0-1)+(n_1-1)}\right) + \left(\frac{(n_1-1)S_1}{(n_0-1)+(n_1-1)}\right).$$

In the kernelized version, we replace X_i with $Z_i = \phi(X_i)$ and we find w to maximize

$$J(w) = \frac{w^T \widetilde{S}_B w}{w^T \widetilde{S}_W w} \tag{23.42}$$

where

$$\widetilde{S}_B = (\overline{Z}_0 - \overline{Z}_1)(\overline{Z}_0 - \overline{Z}_1)^T$$

and

$$S_W = \left(\frac{(n_0-1)\widetilde{S}_0}{(n_0-1)+(n_1-1)}\right) + \left(\frac{(n_1-1)\widetilde{S}_1}{(n_0-1)+(n_1-1)}\right).$$

Here, $\widetilde{S}_j$ is the sample of covariance of the Z_i's for which $Y = j$. However, to take advantage of kernelization, we need to re-express this in terms of inner products and then replace the inner products with kernels.

It can be shown that the maximizing vector w is a linear combination of the Z_i's. Hence we can write

$$w = \sum_{i=1}^{n} \alpha_i Z_i.$$

Also,

$$\overline{Z}_j = \frac{1}{n_j} \sum_{i=1}^{n} \phi(X_i) I(Y_i = j).$$

Therefore,

$$\begin{aligned} w^T \overline{Z}_j &= \Big(\sum_{i=1}^{n} \alpha_i Z_i\Big)^T \Big(\frac{1}{n_j} \sum_{i=1}^{n} \phi(X_i) I(Y_i = j)\Big) \\ &= \frac{1}{n_j} \sum_{i=1}^{n} \sum_{s=1}^{n} \alpha_i I(Y_s = j) Z_i^T \phi(X_s) \\ &= \frac{1}{n_j} \sum_{i=1}^{n} \alpha_i \sum_{s=1}^{n} I(Y_s = j) \phi(X_i) T \phi(X_s) \\ &= \frac{1}{n_j} \sum_{i=1}^{n} \alpha_i \sum_{s=1}^{n} I(Y_s = j) K(X_i, X_s) \\ &= \alpha^T M_j \end{aligned}$$

where M_j is a vector whose i^{th} component is

$$M_j(i) = \frac{1}{n_j} \sum_{s=1}^{n} K(X_i, X_s) I(Y_i = j).$$

It follows that

$$w^T \widetilde{S}_B w = \alpha^T M \alpha$$

where $M = (M_0 - M_1)(M_0 - M_1)^T$. By similar calcuations, we can write

$$w^T \widetilde{S}_W w = \alpha^T N \alpha$$

where

$$N = K_0 \left(I - \frac{1}{n_0}\mathbf{1} \right) K_0^T + K_1 \left(I - \frac{1}{n_1}\mathbf{1} \right) K_1^T,$$

I is the identity matrix, $\mathbf{1}$ is a matrix of all one's, and K_j is the $n \times n_j$ matrix with entries $(K_j)_{rs} = K(x_r, x_s)$ with x_s varying over the observations in group j. Hence, we now find α to maximize

$$J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T N \alpha}.$$

Notice that all the qauntities are expressed in terms of the kernel. Formally, the solution is $\alpha = N^{-1}(M_0 - M_1)$. However, N might be non-invertible. In this case on replaces N by $N + bI$ form some constant b. Finally, the projection onto the new subspace can be written as

$$U = w^T \phi(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x).$$

The support vector machine can similarly be kernelized. We simply replace $\langle X_i, X_j \rangle$ with $K(X_i, X_j)$. For example, instead of maximizing (23.40), we now maximize

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_i \alpha_k Y_i Y_k K(X_i, X_j). \tag{23.43}$$

The hyperplane can be written as $\widehat{H}(x) = \widehat{a}_0 + \sum_{i=1}^{n} \widehat{\alpha}_i Y_i K(X, X_i)$.

23.11 Other Classifiers

There are many other classifiers and space precludes a full dicussion of all of them. Let us briefly mention a few.

The **k-nearest-neighbors** classifier is very simple. Given a point x, find the k data points closest to x. Classify x using the majority vote of these k neighbors. Ties can be broke randomly. The parameter k can be chosen by cross-validation.

Bagging is a method for reducing the variability of a classifier. It is most helpful for highly non-linear classifiers such as a tree. We draw B boostrap samples from the data. The b^{th} bootstrap sample yields a classifier h_b. The final classifier is

$$\widehat{h}(x) = \begin{cases} 1 & \text{if } \frac{1}{B}\sum_{b=1}^{B} h_b(x) \geq \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Boosting is a method for starting with a simple classifier and gradually improving it by refitting the data giving higher weight to misclassified samples. Suppose that $\mathcal{H}$ is a collection of classifiers, for example, tress with onely one split. Assume that $Y_i \in \{-1, 1\}$ and that each h is such that $h(x) \in \{-1, 1\}$. We usually give equal weight to all data points in the methods we have discussed. But one can incorporate unequal weights quite easily in most algorithms. For example, in constructing a tree, we could replace the impurity measure with a weighted impurity measure. The original version of bossting, called AdaBoost, is as follows.

1. Set the weights $w_i = 1/n$, $i = 1, \ldots, n$.

2. For $j = 1, \ldots, J$, do the following steps:

   (a) Constructing a classifier h_j from the data using the weights $w_1, \ldots, w_n$.

   (b) Compute the weighted error estimate:

   $$\widehat{L}_j = \frac{\sum_{i=1}^{n} w_i I(Y_i \neq h_j(X_i))}{\sum_{i=1}^{n} w_i}.$$

   (c) Let $\alpha_j = \log((1 - \widehat{L}_j)/\widehat{L}_j)$.

   (d) Update the weights:

   $$w_i \longleftarrow w_i e^{\alpha_j I(Y_i \neq h_j(X_i))}$$

3. The final classifier is

$$\widehat{h}(x) = \text{sign}\Big(\sum_{j=1}^{J} \alpha_j h_j(x)\Big).$$

There is now an enormous literature trying to explain and improve on boosting. Whereas bagging is a variance reduction technique, boosting can be thought of as a bias reduction technique. We starting with a simple – and hence highly biased – classifier, and we gradually reduce the bias. The disadvantage of boosting is that the final classifier is quite complicated.

23.12 Bibliographic Remarks

An excellent reference on classification is Hastie, Tibshirani and Friedman (2001). For more on the theory, see Devroye, Györfi, Lugosi (1996) and Vapnik (2001).

23.13 Exercises

1. Prove Theorem 23.5.

2. Prove Theorem 23.7.

3. Download the spam data from:

   http://www-stat.stanford.edu/~tibs/ElemStatLearn/index.html

   The data file can also be found on the course web page. The data contain 57 covariates relating to email messages. Each email message was classified as spam (Y=1) or not spam (Y=0). The outcome Y is the last column in the file. The goal is to predict whether an email is spam or not.

(a) Construct classification rules using (i) LDA, (ii) QDA, (iii) logistic regression and (iv) a classification tree. For each, report the observed misclassification error rate and construct a 2-by-2 table of the form

	$\widehat{h}(x) = 0$	$\widehat{h}(x) = 1$
$Y = 0$		
$Y = 1$		

Hint: Here are some things to remember when using R.

(i) To fit a logistic regression and get the classified values:

```
d <- data.frame(x,y)
n <- nrow(x)
out <- glm(y ~ .,family=binomial,data=d)
p <- out$fitted.values
pred <- rep(0,n)
pred[p > .5] <- 1
table(y,pred)
```

(ii) To build a tree you must turn y into a factor and you must do it before you make the data frame:

```
y <- as.factor(y)
d <- data.frame(x,y)
library(tree)
out <- tree(y ~ .,data=d)
print(summary(out))
plot(out,type="u",lwd=3)
text(out)
pred <- predict(out,d,type="class")
table(y,pred)
```

(b) Use 5-fold cross-validation to estimate the prediction accuracy of LDA and logistic regression.

(c) Sometimes it helps to reduce the number of covariates. One strategy is to compare X_i for the spam and email group.. For each of the 57 covariates, test whether the mean of the covriate is the same or different between the two groups. Keep the 10 covariates with the smallest p-values. Try LDA and logstic regression using only these 10 variables.

4. Let $\mathcal{A}$ be the set of two-dimensional spheres. That is, $A \in \mathcal{A}$ if $A = \{(x,y) : (x-a)^2 + (y-b)^2 \leq c^2\}$ for some a, b, c. Find the VC dimension of $\mathcal{A}$.

5. Classify the spam data using support vector machines. Free software for the support vector machine is at

   http://svmlight.joachims.org/

   It is very easy to use. After installing, do the following. First, you must recode the Y_i's as +1 and -1. Then type

   ```
   svm_learn data output
   ```

   where "data" is the file with the data. It expects the data in the following form:

   ```
   1   1:3.4  2:.17  3:129  etc
   ```

   where the first number is Y (+1 or -1) and the rest of the line means: X_1 is 3.4, X_2 is .17, etc.

   If the file "test" contains test data, then type

```
svm_classify test output predictions
```

to get the predictions on the test data. There are many options available, as explained on the web site.

6. Use VC theory to get a confidence interval on the true error rate of the LDA classifier in question 3. Suppose that the covariate $X = (X_1, \ldots, X_d)$ has d dimensions. Assume that each variable has a continuous distribution.

7. Suppose that $X_i \in \mathbb{R}$ and that $Y_i = 1$ whenever $|X_i| \leq 1$ and $Y_i = 0$ whenever $|X_i| > 1$. Show that no linear classifier can perfectly classify these data. Show that the kernelized data $Z_i = (X_i, X_i^2)$ can be linearly separated.

8. Repeat question 5 using the kernel $K(x, \widetilde{x}) = (1 + x^T \widetilde{x})^p$. Choose p by cross-validation.

9. Apply the k nearest neighbors classifier to the "iris data." Get the data in R from the command data(iris). Choose k by cross-validation.

10. (Curse of Dimensionality.) Suppose that X has a uniform distribution on the d-dimensional cube $[-1/2, 1/2]^d$. Let R be the distance from the origin to the closest neighbor. Show that the median of R is

$$\left(\frac{\left(1 - \left(\frac{1}{2}\right)^{1/n}\right)}{v_d(1)} \right)^{1/d}$$

where

$$v_d(r) = r^d \frac{\pi^{d/2}}{\Gamma((d/2)+1)}$$

is the volume of a sphere of radius r. When does R exceed the edge of the cube when $n = 100, 1000, 10000$?

11. Fit a tree to the data in question 3. Now apply bagging and report your results.

12. Fit a tree that uses only one split on one variable to the data in question 3. Now apply boosting.

13. Let $r(x) = \mathbb{P}(Y = 1|X = x)$ and let $\widehat{r}(x)$ be an estimate of $r(x)$. Consider the classifier

$$h(x) = \begin{cases} 1 & \text{if } \widehat{r}(x) \geq 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Assume that $\widehat{r}(x) \approx N(\overline{r}(x), \sigma^2(x))$ for some functions $\overline{r}(x)$ and $\sigma^2(x)$). Show that

$$\mathbb{P}(Y \neq h(X)) \approx 1 - \Phi\left(\frac{\text{sign}\Big((\widehat{r}(x) - (1/2))(\overline{r}(x) - (1/2))\Big)}{\sigma(x)}\right)$$

where Φ is the standard Normal cdf. Regard $\text{sign}\Big((\widehat{r}(x) - (1/2))(\overline{r}(x) - (1/2))\Big)$ as a type of bias term. Explain the implications for the bias-variance trade-off in classification (Friedman, 1997).

24
Stochastic Processes

24.1 Introduction

Most of this book has focused on IID sequences of random variables. Now we consider sequences of dependent random variables. For example, daily temperatures will form a sequence of time-ordered random variables and clearly the temperature on one day is not independent of the temperature on the previous day.

A **stochastic process** $\{X_t : t \in T\}$ is a collection of random variables. We shall sometimes write $X(t)$ instead of X_t. The variables X_t take values in some set $\mathcal{X}$ called the **state space**. The set T is called the **index set** and for our purposes can be thought of as time. The index set can be discrete $T = \{0, 1, 2, \ldots\}$ or continuous $T = [0, \infty)$ depending on the application.

Example 24.1 (IID observations.) *A sequence of* IID *random variables can be written as* $\{X_t : t \in T\}$ *where* $T = \{1, 2, 3, \ldots, \}$.

Thus, a sequence of iid random variables is an example of a stochastic process. ■

Example 24.2 (The Weather.) *Let* $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$. *A typical sequence (depending on where you live) might be*

$$\text{sunny}, \text{sunny}, \text{cloudy}, \text{sunny}, \text{cloudy}, \text{cloudy}, \cdots$$

This process has a discrete state space and a discrete index set. ■

Example 24.3 (Stock Prices.) *Figure 24.1 shows the price of a stock over time. The price is monitored continuously so the index set* T *is not discrete. Price is discrete but, for practical purposes we can imagine treating it as a continuous variable.* ■

Example 24.4 (Empirical Distribution Function.) *Let* $X_1, \ldots, X_n \sim F$ *where* F *is some* CDF *on [0,1]. Let*

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

be the empirical CDF . *For any fixed value* t, $\widehat{F}_n(t)$ *is a random variable. But the whole empirical* CDF

$$\left\{ \widehat{F}_n(t) : \ t \in [0,1] \right\}$$

is a stochastic process with a continuous state space and a continuous index set. ■

We end this section by recalling a basic fact. If $X_1, \ldots, X_n$ are random variables then we can write the joint density as

$$\begin{aligned} f(x_1, \ldots, x_n) &= f(x_1) f(x_2|x_1) \cdots f(x_n|x_1, \ldots, x_{n-1}) \\ &= \prod_{i=1}^{n} f(x_i|\text{past}_i) \end{aligned} \tag{24.1}$$

where past_i refers to all the variables before X_i.

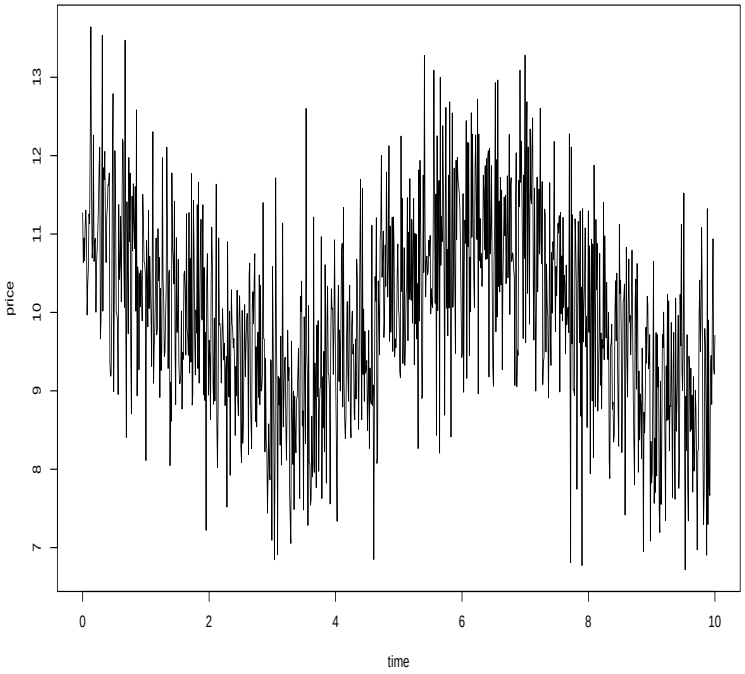


FIGURE 24.1. Stock price over ten week period.

24.2 Markov Chains

The simplest stochastic process is a Markov chain in which the distribution of X_t depends only on X_{t-1}. In this section we assume that the state space is discrete, either $\mathcal{X} = \{1, \ldots, N\}$ or $\mathcal{X} = \{1, 2, \ldots, \}$ and that the index set is $T = \{0, 1, 2, \ldots\}$. Typically, most authors write X_n instead of X_t when discussing Markov chains and I will do so as well.

Definition 24.5 *The process* $\{X_n : n \in T\}$ *is a* **Markov chain** *if*

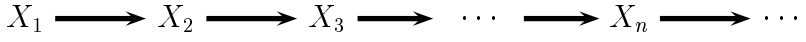
$$\mathbb{P}(X_n = x \mid X_0, \ldots, X_{n-1}) = \mathbb{P}(X_n = x \mid X_{n-1}) \quad (24.2)$$

for all n *and for all* $x \in \mathcal{X}$.

For a Markov chain, equation (24.1) simplifes to

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2|x_1) f(x_3|x_2) \cdots f(x_n|x_{n-1}).$$

A Markov chain can be represented by the following DAG:



Each variable has a single parent, namely, the previous observation.

The theory of Markov chains is a very rich and complex. We have to get through many definitions before we can do anything interesting. Our goal is to answer the following questions:

1. When does a Markov chain "settle down" into some sort of equilibrium?

2. How do we estimate the parameters of a Markov chain?

3. How can we construct Markov chains that converge to a given equilibrium distribution and why would we want to do that?

We will answer questions 1 and 2 in this chapter. We will answer question 3 in the next chapter. To understand question 1, look at the two chains in Figure 24.2. The first chain oscillates all over the place and will continue to do so forever. The second chain eventually settles into an equilibrium. If we constructed a histogram of the first process, it would keep changing as we got more and more observations. But a histogram from the second chain would eventually converge to some fixed distribution.

TRANSITION PROBABILITIES. The key quantities of a Markov chain are the probabilities of jumping from one state into another state.

Definition 24.6 *We call*

$$\mathbb{P}(X_{n+1} = j | X_n = i) \tag{24.3}$$

the **transition probabilities.** *If the transition probabilities do not change with time, we say the chain is* **homogeneous**. *In this case we define* $p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i)$. *The matrix* $\mathbf{P}$ *whose* (i, j) *element is* p_{ij} *is called the* **transition matrix.**

We will only consider homogeneous chains. Notice that $\mathbf{P}$ has two properties: (i) $p_{ij} \geq 0$ and (ii) $\sum_i p_{ij} = 1$. Each row is a probability mass function. A matrix with these properties is called a **stochastic matrix.**

Example 24.7 (Random Walk With Absorbing Barriers.) *Let* $\mathcal{X} = \{1, \ldots, N\}$. *Suppose you are standing at one of these points.*

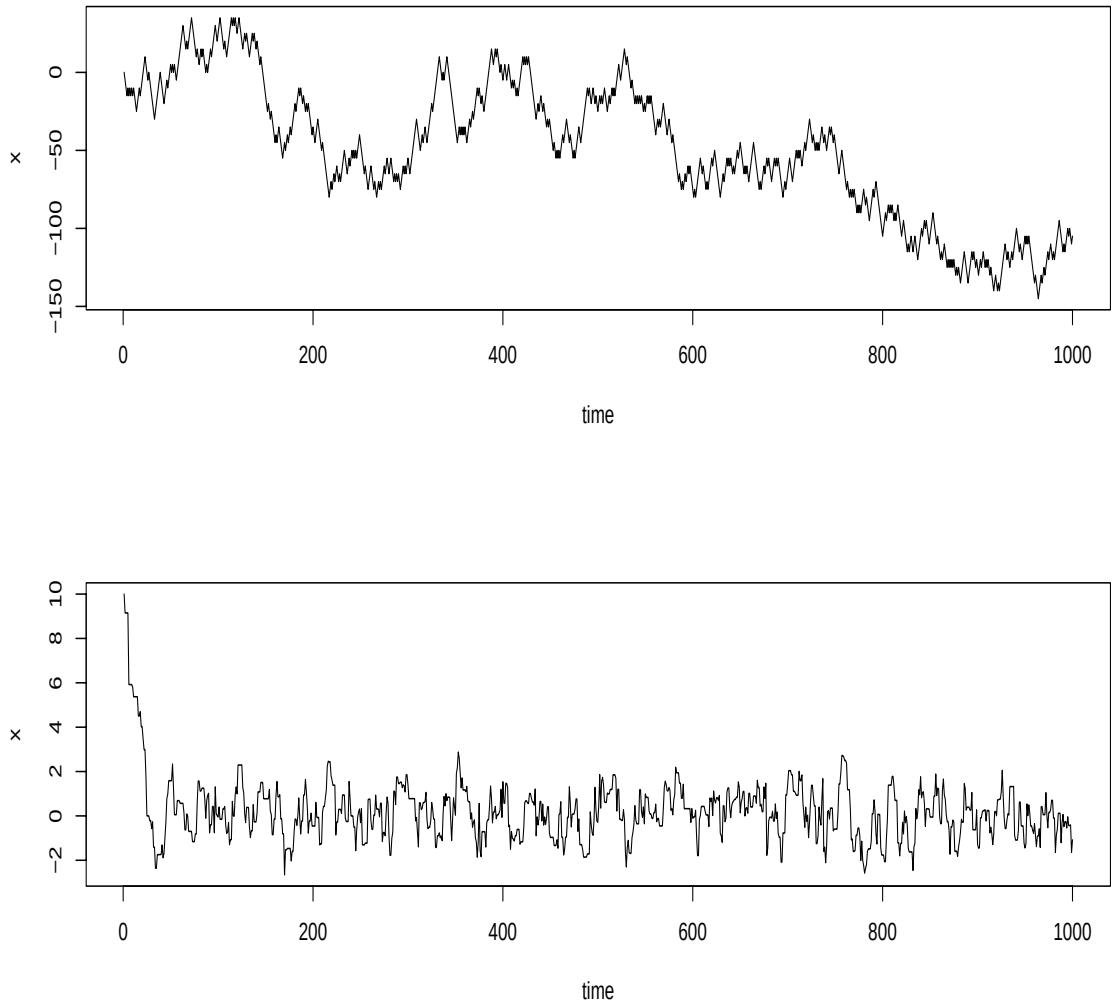


FIGURE 24.2. The first chain does not settle down into an equilibrium. The second does.

Flip a coin with $\mathbb{P}(Heads) = p$ *and* $\mathbb{P}(Tails) = q = 1 - p$. *If it is heads, take one step to the right. If it is tails, take one step to the left. If you hit one of the endpoints, stay there. The transition matrix is*

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ q & 0 & p & 0 & \cdots & 0 & 0 \\ 0 & q & 0 & p & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad \blacksquare$$

Example 24.8 *Suppose the state space is* $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$. *Then* $X_1, X_2, \ldots$ *represents the weather for a sequence of days. The weather today clearly depends on yesterday's weather. It might also depend on the weather two days ago but as a first approximation we might assume that the dependence is only one day back. In that case the weather is a Markov chain and a typical transition matrix might be*

	Sunny	*Cloudy*
Sunny	*0.4*	*0.6*
Cloudy	*0.8*	*0.2*

For example, if it is sunny today, there is a 60 per cent chance it will be cloudy tommorrow. ■

Let

$$p_{ij}(n) = \mathbb{P}(X_{m+n} = j | X_m = i) \tag{24.4}$$

be the probability of of going from state i to state j in n steps. Let $\mathbf{P}_n$ be the matrix whose (i, j) element is $p_{ij}(n)$. These are called the **n-step transition probabilities.**

Theorem **24.9 (The Chapman-Kolmogorov equations.)** *The n-step probabilities satisfy*

$$p_{ij}(m + n) = \sum_k p_{ik}(m) p_{kj}(n). \tag{24.5}$$

PROOF. Recall that, in general,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y|X = x).$$

This fact is true in the more general form

$$\mathbb{P}(X = x, Y = y|Z = z) = \mathbb{P}(X = x|Z = z)\mathbb{P}(Y = y|X = x, Z = z).$$

Also, recall the law of total probability:

$$\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y).$$

Using these facts and the Markov property we have

$$\begin{aligned} p_{ij}(m+n) &= \mathbb{P}(X_{m+n} = j|X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j, X_m = k|X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j|X_m = k, X_0 = i)\mathbb{P}(X_m = k|X_0 = i) \\ &= \sum_k \mathbb{P}(X_{m+n} = j|X_m = k)\mathbb{P}(X_m = k|X_0 = i) \\ &= \sum_k p_{ik}(m)p_{kj}(n). \quad \blacksquare \end{aligned}$$

Look closely at equation (24.5). This is nothing more than the equation for matrix multiplication. Hence we have shown that

$$\mathbf{P}_{m+n} = \mathbf{P}_m\mathbf{P}_n. \tag{24.6}$$

By definition, $\mathbf{P}_1 = \mathbf{P}$. Using the above theorem, $\mathbf{P}_2 = \mathbf{P}_{1+1} = \mathbf{P}_1\mathbf{P}_1 = \mathbf{PP} = \mathbf{P}^2$. Continuing this way, we see that

$$\mathbf{P}_n = \mathbf{P}^n \equiv \underbrace{\mathbf{P} \times \mathbf{P} \times \cdots \times \mathbf{P}}_{\text{multiply the matrix n times}}. \tag{24.7}$$

Let $\mu_n = (\mu_n(1), \ldots, \mu_n(N))$ be a row vector where

$$\mu_n(i) = \mathbb{P}(X_n = i) \tag{24.8}$$

is the marginal probability that the chain is in state i at time n. In particular, μ_0 is called the **initial distribution.** To simulate a Markov chain, all you need to know is μ_0 and $\mathbf{P}$. The simulation would look like this:

Step 1: Draw $X_0 \sim \mu_0$. Thus, $\mathbb{P}(X_0 = i) = \mu_0(i)$.
Step 2: Suppose the outcome of step 1 is i. Draw $X_1 \sim \mathbf{P}$. In other words, $\mathbb{P}(X_1 = j|X_0 = i) = p_{ij}$.
Step 3: Suppose the outcome of step 2 is j. Draw $X_2 \sim \mathbf{P}$. In other words, $\mathbb{P}(X_2 = k|X_1 = j) = p_{jk}$.
And so on.

It might be difficult to understand the meaning of μ_n. Imaging simulating the chain many times. Collect all the outcomes at time n from all the chains. This histogram would look approximately like μ_n. A consequence of theorem 24.9 is the following.

Lemma 24.10 *The marginal probabilities are given by*

$$\mu_n = \mu_0 \mathbf{P}^n.$$

Proof.

$$\begin{aligned} \mu_n(j) &= \mathbb{P}(X_n = j) \\ &= \sum_i \mathbb{P}(X_n = j|X_0 = i)P(X_0 = i) \\ &= \sum_i \mu_0(i)p_{ij}(n) = \mu_0 \mathbf{P}^n. \quad \blacksquare \end{aligned}$$

Summary

1. Transition matrix: $\mathbf{P}(i,j) = \mathbb{P}(X_{n+1} = j | X_n = i)$.
2. n-step matrix: $\mathbf{P}_n(i,j) = \mathbb{P}(X_{n+m} = j | X_m = i)$.
3. $\mathbf{P}_n = \mathbf{P}^n$.
4. Marginal: $\mu_n(i) = \mathbb{P}(X_n = i)$.
5. $\mu_n = \mu_0 \mathbf{P}^n$.

STATES. The states of a Markov chain can be classified according to various properties.

Definition 24.11 *We say that i* **reaches** *j (or j is* **accessible** *from i) if $p_{ij}(n) > 0$ for some n, and we write $i \to j$. If $i \to j$ and $j \to i$ then we write $i \leftrightarrow j$ and we say that i and j* **communicate.**

Theorem **24.12** *The communication relation satisfies the following properties:*

1. $i \leftrightarrow i$.
2. *If $i \leftrightarrow j$ then $j \leftrightarrow i$.*
3. *If $i \leftrightarrow j$ and $j \leftrightarrow k$ then $i \leftrightarrow k$.*
4. *The set of states $\mathcal{X}$ can be written as a disjoint union of* **classes** *$\mathcal{X} = \mathcal{X}_1 \bigcup \mathcal{X}_2 \bigcup \cdots$ where two states i and j communicate with each other if and only if they are in the same class.*

If all states communicate with each other then the chain is called **irreducible**. A set of states is **closed** if, once you enter that set of states you never leave. A closed set consisting of a single state is called an **absorbing state.**

Example 24.13 *Let* $\mathcal{X} = \{1, 2, 3, 4\}$ *and*

$$\mathbf{P} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The classes are $\{1, 2\}, \{3\}$ *and* $\{4\}$. *State 4 is an absorbing state.* ■

Suppose we start a chain in state i. Will the chain ever return to state i? If so, that state is called persistent or recurrent.

Definition 24.14 *State* i *is* **recurrent** *or* **persistent** *if*

$$\mathbb{P}(X_n = i \text{ for some } n \geq 1 \mid X_0 = i) = 1.$$

Otherwise, state i *is* **transient**.

Theorem **24.15** *A state* i *is recurrent if and only if*

$$\sum_n p_{ii}(n) = \infty. \tag{24.9}$$

A state i *is transient if and only if*

$$\sum_n p_{ii}(n) < \infty. \tag{24.10}$$

Proof. Define

$$I_n = \begin{cases} 1 & \text{if } X_n = i \\ 0 & \text{if } X_n \neq i. \end{cases}$$

The number of times that the chain is in state i is $Y = \sum_{n=0}^{\infty} I_n$. The mean of Y, given that the chain starts in state i, is

$$\begin{aligned}\mathbb{E}(Y|X_0 = i) &= \sum_{n=0}^{\infty} \mathbb{E}(I_n|X_0 = i) \\ &= \sum_{n=0}^{\infty} \mathbb{P}(X_n = i|X_0 = i) \\ &= \sum_{n=0}^{\infty} p_{ii}(n).\end{aligned}$$

Define $a_i = \mathbb{P}(X_n = i \text{ for some } n \geq 1 \mid X_0 = i)$. If i is recurrent, $a_i = 1$. Thus, the chain will eventually return to i. Once it does return to i, we argue again that since $a_i = 1$, the chain will return to state i again. By repeating this argument, we conclude that $\mathbb{E}(Y|X_0 = i) = \infty$. If i is transient, then $a_i < 1$. When the chain is in state i, there is a probability $1 - a_i > 0$ that it will never return to state i. Thus, the probability that the chain is in state i exactly n times is $a_i^{n-1}(1 - a_i)$. This is a geometric distribution which has finite mean. ■

Theorem **24.16** *Facts about recurrence.*

1. *If state i is recurrent and $i \leftrightarrow j$ then j is recurrent.*
2. *If state i is transient and $i \leftrightarrow j$ then j is transient.*
3. *A finite Markov chain must have at least one recurrent state.*
4. *The states of a finite, irreducible Markov chain are all recurrent.*

Theorem **24.17 (Decomposition Theorem.)** *The state space $\mathcal{X}$ can be written as the disjoint union*

$$\mathcal{X} = \mathcal{X}_T \bigcup \mathcal{X}_1 \bigcup \mathcal{X}_2 \cdots$$

where $\mathcal{X}_T$ are the transient states and each $\mathcal{X}_i$ is a closed, irreducible set of recurrent states.

CONVERGENCE OF MARKOV CHAINS. To discuss the convregence of chains, we need a few more definitions. Suppose that $X_0 = i$. Define the **recurrence time**

$$T_{ij} = \min\{n > 0 : \ X_n = j\} \tag{24.11}$$

assuming X_n ever returns to state i, otherwise define $T_{ij} = \infty$. The **mean recurrence time** of a recurrent state i is

$$m_i = E(T_{ii}) = \sum_n n f_{ii}(n) \tag{24.12}$$

where

$$f_{ij}(n) = P(X_1 \neq j, X_2 \neq j, \ldots, X_{n-1} \neq j, X_n = j | X_0 = i).$$

A recurrent state is **null** if $m_i = \infty$ otherwise it is called **non-null** or **positive.**

Lemma 24.18 *If a state is null and recurrent, then $p_{ii}^n \to 0$.*

Lemma 24.19 *In a finite state Markov chain, all recurrent states are positive.*

Consider a three state chain with transition matrix

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Suppose we start the chain in state 1. Then we will be in state 3 at times 3, 6, 9, ... This is an example of a periodic chain. Formally, the **period** of state i is d if $p_{ii}(n) = 0$ whenever n is not divisible by d and d is the largest integer with this property. Thus, $d = \gcd\{n : \ p_{ii}(n) > 0\}$ where gcd means "greater common divisor." State i is **periodic** if $d(i) > 1$ and **aperiodic** if $d(i) = 1$. A state with period 1 is called **aperiodic**.

Lemma 24.20 *If state i has period d and $i \leftrightarrow j$ then j has period d.*

Definition 24.21 *A state is* **ergodic** *if it is recurrent, non-null and aperiodic. A chain is ergodic if all its states are ergodic.*

Let $\pi = (\pi_i : \ i \in \mathcal{X})$ be a vector of non-negative numbers that sum to one. Thus π can be thought of as a probability mass function.

Definition 24.22 *We say that π is a* **stationary** *(or* **invariant***) distribution if $\pi = \pi \mathbf{P}$.*

Here is the intuition. Draw X_0 from distribution π and suppose that π is a stationary distribution. Now draw X_1 according to the transition probability of the chain. The distribition of X_1 is then $\mu_1 = \mu_0 \mathbf{P} = \pi \mathbf{P} = \pi$. The distribution of X_2 is $\pi \mathbf{P}^2 = (\pi \mathbf{P})\mathbf{P} = \pi \mathbf{P} = \pi$. Continuing this way, we see that the distribution of X_n is $\pi \mathbf{P}^n = \pi$. In other words:

If at any time the chain has distribution π then it will continue to have distribution π forever.

Definition 24.23 *We say that a chain has* **limiting distribution** *if*

$$\mathbf{P}^n \to \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix}$$

for some π. In other words, $\pi_j = \lim_{n\to\infty} \mathbf{P}_{ij}^n$ exists and is independent of i.

Here is the main theorem.

Theorem **24.24** *An irreducible, ergodic Markov chain has a unique stationary distribution π. The limiting distrubution exists and is equal to π. If g is any bounded function, then, with probability 1,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} g(X_n) \to \mathbb{E}_\pi(g) \equiv \sum_j g(j)\pi_j. \tag{24.13}$$

The last statement of the theorem, is the law of large numbers for Markov chains. It says that sample averages converge to their expectations. Finally, there is a special condition which will be useful later. We say that π satisfies **detailed balance** if

$$\pi_i p_{ij} = p_{ji}\pi_j. \tag{24.14}$$

Detailed balance guarantees that π is a stationary distribution.

Theorem **24.25** *If π satisfies detailed balance then π is a stationary distribution.*

Proof. We need to show that $\pi\mathbf{P} = \pi$. The j^{th} element of $\pi\mathbf{P}$ is $\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$. ■

The importance of detailed balance will become clear when we discuss Markov chain Monte Carlo methods.

Warning! Just because a chain has a stationary distribution does not mean it converges.

Example 24.26 *Let*

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Let $\pi = (1/3, 1/3, 1/3)$. *Then* $\pi P = \pi$ *so* π *is a stationary distribution. If the chain is started with the distribution* π *it will stay in that distribution. Imagine simulating many chains and checking the marginal distribution at each time* n. *It will always be the uniform distribution* π. *But this chain does not have a limit. It continues to cycle around forever.* ■

Examples of Markov Chains

Example 24.27 (Random Walk) . *Let* $\mathcal{X} = \{\ldots, -2, -1, 0, 1, 2, \ldots, \}$ *and suppose that* $p_{i,i+1} = p$, $p_{i,i-1} = q = 1 - p$. *All states communicate hence either all the states are recurrent or all are transient. To see which, suppose we start at* $X_0 = 0$. *Note that*

$$p_{00}(2n) = \binom{2n}{n} p^n q^n \tag{24.15}$$

since, the only way to get back to 0 is to have n *heads (steps to the right) and* n *tails (steps to the left). We can approximate this expression using Stirling's formula which says that*

$$n! \sim n^n \sqrt{n} e^{-n} \sqrt{2\pi}.$$

Inserting this approximation into (24.15) shows that

$$p_{00}(2n) \sim \frac{(4pq)^n}{\sqrt{n\pi}}.$$

It is easy to check that $\sum_n p_{00}(n) < \infty$ *if and only if* $\sum_n p_{00}(2n) < \infty$. *Moreover,* $\sum_n p_{00}(2n) = \infty$ *if and only if* $p = q = 1/2$. *By Theorem (24.15), the chain is recurrent if* $p = 1/2$ *otherwise it is transient.* ■

Example 24.28 *Let* $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$. *Let*

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Then $C_1 = \{1, 2\}$ *and* $C_2 = \{5, 6\}$ *are irreducible closed sets. States 3 and 4 are transient because the of the path* $3 \to 4 \to 6$ *and once you hit state 6 you cannot return to 3 or 4. Since* $p_{ii}(1) > 0$, *all the states are aperiodic. In summary, 3 and 4 are transient while 1,2,5 and 6 are ergodic.* ■

Example 24.29 (Hardy-Weinberg.) *Here is a famous example from genetics. Suppose a gene can be type A or type a. There are three types of people (called genotypes): AA, Aa and aa. Let* (p, q, r) *denote the fraction of people of each genotype. We assume that everyone contributes one of their two copies of the gene at random to their children. We also assume that mates are selected at random. The latter is not realistic however, it is often reasonable to assume that you do not choose your mate based on whether they are AA, Aa or aa. (This would be false if the gene was for eye color and if people chose mates based on eye color.) Imagine if we pooled eveyone's genes together. The proportion*

of A genes is $P = p + (q/2)$ and the proportion of a genes is $Q = r + (q/2)$. A child is AA with probability P^2, aA with probability $2PQ$ and aa with probability Q^2. Thus, the fraction of A genes in this generation is

$$P^2 + PQ = \left(p + \frac{q}{2}\right)^2 + \left(p + \frac{q}{2}\right)\left(r + \frac{q}{2}\right).$$

However, $r = 1 - p - q$. Substitute this in the above equation and you get $P^2 + PQ = P$. A similar calculation shows that fraction of a genes is Q. We have shown that the proportion of type A and type a is P and Q and this remains stable after the first generation. The proportion of people of type AA, Aa, aa is thus $(P^2, 2PQ, Q^2)$ from the second generation and on. This is called the Hardy-Weinberg law.

Assume everyone has exactly one child. Now consider a fixed person and let X_n be the genotype of their n^{th} descendent. This is a Markov chain with state space $\mathcal{X} = \{AA, Aa, aa\}$. Some basic calculations will show you that the transition matrix is

$$\begin{bmatrix} P & Q & 0 \\ \frac{P}{2} & \frac{P+Q}{2} & \frac{Q}{2} \\ 0 & P & Q \end{bmatrix}.$$

The stationary distribution is $\pi = (P^2, 2PQ, Q^2)$. ■

INFERENCE FOR MARKOV CHAINS. Consider a chain with finite state space $\mathcal{X} = \{1, 2, \ldots, N\}$. Suppose we observe n observations $X_1, \ldots, X_n$ from this chain. The unknown parameters of a Markov chain are the initial probabilities $\mu_0 = (\mu_0(1), \mu_0(2), \ldots,)$ and the elements of the transition matrix $\mathbf{P}$. Each row of $\mathbf{P}$ is a multinomial distribution. So we are essentially estimating N distributions (plus the initial probabilities). Let n_{ij} be the observed number of transitions from state i to state j. The likelihood function is

$$\mathcal{L}(\mu_0, \mathbf{P}) = \mu_0(x_0) \prod_{r=1}^{n} p_{X_{r-1}, X_r} = \mu_0(x_0) \prod_{i=1}^{N} \prod_{j=1}^{N} p_{ij}^{n_{ij}}.$$

There is only one observation on μ_0 so we can't estimate that. Rather, we focus on estimating $\mathbf{P}$. The MLE is obtained by maximizing $\mathcal{L}(\mu_0, \mathbf{P})$ subject to the constraint that the elements are non-negative and the rows sum to 1. The solution is

$$\widehat{p}_{ij} = \frac{n_{ij}}{n_i}$$

where $n_i = \sum_{j=1}^{N} n_{ij}$. Here we are assuming that $n_i > 0$. If note, then we set $\widehat{p}_{ij} = 0$ by convention.

Theorem **24.30 (Consistency and Asymptotic Normality of the MLE .)** *Assume that the chain is ergodic. Let $\widehat{p}_{ij}(n)$ denote the mle after n observations. Then $\widehat{p}_{ij}(n) \xrightarrow{\text{P}} p_{ij}$. Also,*

$$\left[\sqrt{N_i(n)}(\widehat{p}_{ij} - p_{ij})\right] \rightsquigarrow N(0, \Sigma)$$

where the left hand side is a matrix, $N_i(n) = \sum_{r=1}^{n} I(X_r = i)$ and

$$\Sigma_{ij,k\ell} = \begin{cases} p_{ij}(1 - p_{ij}) & (i,j) = (k,\ell) \\ -p_{ij}p_{i\ell} & i = k, j \neq \ell \\ 0 & \text{otherwise.} \end{cases}$$

24.3 Poisson Processes

One of the most studied and useful stochastic processes is the Poisson process. It arises when we count occurences of events over time, for example, traffic accidents, radioactive decay, arrival of email messages etc. As the name suggests, the Poisson process is intimately related to the Poisson distribution. Let's first review the Poisson distribution.

Recall that X has a Poisson distribution with parameter λ – written $X \sim \text{Poisson}(\lambda)$ – if

$$\mathbb{P}(X = x) \equiv p(x;\ \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Also recall that $\mathbb{E}(X) = \lambda$ and $\mathbb{V}(X) = \lambda$. If $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\nu)$ and $X \amalg Y$, then $X + Y \sim \text{Poisson}(\lambda + \nu)$. Finally, if $N \sim \text{Poisson}(\lambda)$ and $Y|N = n \sim \text{Binomial}(n, p)$, then the marginal distribution of Y is $Y \sim \text{Poisson}(\lambda p)$.

Now we describe the Poisson process. Imaging you are at your computer. Each time a new email message arrives you record the time. Let X_t be the number of messages you have received up to and including time t. Then, $\{X_t : t \in [0, \infty)\}$ is a stochastic process with state space $\mathcal{X} = \{0, 1, 2, \ldots\}$. A process of this form is called a **counting process**. A Poisson process is a counting process that satisfies certain conditions. In what follows, we will sometimes write $X(t)$ instead of X_t. Also, we need the following notation. Write $f(h) = o(h)$ if $f(h)/h \to 0$ as $h \to 0$. This means that $f(h)$ is smaller than h when h is close to 0. For example, $h^2 = o(h)$.

Definition 24.31 *A* **Poisson process** *is a stochastic process* $\{X_t : t \in [0,\infty)\}$ *with state space* $\mathcal{X} = \{0,1,2,\ldots\}$ *such that*

1. $X(0) = 0$.
2. *For any* $0 = t_0 < t_1 < t_2 < \cdots < t_n$, *the increments*

$$X(t_1) - X(t_0),\ X(t_2) - X(t_1),\ \cdots, X(t_n) - X(t_{n-1})$$

are independent.

3. *There is a function* $\lambda(t)$ *such that*

$$\mathbb{P}(X(t+h) - X(t) = 1) = \lambda(t)h + o(h) \quad (24.16)$$
$$\mathbb{P}(X(t+h) - X(t) \geq 2) = o(h). \quad (24.17)$$

We call $\lambda(t)$ *the* **intensity function**.

The last condition means that the probability of an event in $[t, t+h]$ is approximately $h\lambda(t)$ while the probability of more than one event is small.

Theorem **24.32** *If* X_t *is a Poisson process with intensity function* $\lambda(t)$, *then*

$$X(s+t) - X(s) \sim \text{Poisson}(m(s+t) - m(s))$$

where

$$m(t) = \int_0^t \lambda(s)\, ds.$$

In particular, $X(t) \sim \text{Poisson}(m(t))$. *Hence,* $\mathbb{E}(X(t)) = m(t)$ *and* $\mathbb{V}(X(t)) = m(t)$.

Definition 24.33 *A Poisson process with intensity function $\lambda(t) \equiv \lambda$ for some $\lambda > 0$ is called a* **homogeneous Poisson process** *with rate λ. In this case,*

$$X(t) \sim \text{Poisson}(\lambda t).$$

Let $X(t)$ be a homogeneous Poisson process with rate λ. Let W_n be the time at which the n^{th} event occurs and set $W_0 = 0$. The random variables $W_0, W_1, \ldots,$ are called **waiting times**. Let $S_n = W_{n+1} - W_n$. Then $S_0, S_1, \ldots,$ are called **sojourn times** or **interarraival times**.

Theorem **24.34** *The sojourn times $S_0, S_1, \ldots$ are* IID *random variables. Their distribution is exponential with mean $1/\lambda$, that is, they have density*

$$f(s) = \lambda e^{-\lambda s}, \quad s \geq 0.$$

The waiting time $W_n \sim \text{Gamma}(n, 1/\lambda)$ i.e. it has density

$$f(w) = \frac{1}{\Gamma(n)} \lambda^n w^{n-1} e^{-\lambda t}.$$

Hence, $\mathbb{E}(W_n) = n/\lambda$ and $\mathbb{V}(W_n) = n/\lambda^2$.

Proof. First, we have

$$\mathbb{P}(S_1 > t) = \mathbb{P}(X(t) = 0) = e^{-\lambda t}$$

with shows that the CDF for S_1 is $1 - e^{-\lambda t}$. This shows the result for S_1. Now,

$$\begin{aligned} \mathbb{P}(S_2 > t | S_1 = s) &= \mathbb{P}(\text{no events in } (s, s+t] | S_1 = s) \\ &= \mathbb{P}(\text{no events in } (s, s+t]) \quad \text{(increments are independent)} \\ &= e^{-\lambda t}. \end{aligned}$$

Hence, S_2 has an exponential distribution and is independent of S_1. The result follows by repeating the argument. The result for W_n follows since a sum of exponentials has a Gamma distribution. ■

Example 24.35 *Figures 24.3 and 24.4 show requests to a WWW server in Calgary.*[1] *Assuming that this is a homogeneous Poisson process,* $N \equiv X(T) \sim \text{Poisson}(\lambda T)$. *The likelihood is*

$$\mathcal{L}(\lambda) \propto e^{-\lambda T}(\lambda T)^N$$

which is maxmimized at

$$\widehat{\lambda} = \frac{N}{T} = 48.0077$$

in units per minute.

In the preceding example we might want to check if the assumption that the data follow a Poisson process is reasonable. To proceed, let us first note that if a Poisson process is observed on $[0, T]$ then, given $N = X(T)$, the event times $W_1, \ldots, W_N$ are a sample from a Uniform$(0, T)$ distribution. To see intutively why this is true, note that the probability of finding an event in a small interval $[t, t + h]$ is proportional to h. But this doesn't depend on t. Since the probability is constant over t it must be uniform.

Figure 24.5 shows a histogram and a some density estimates. The density does not appear to be uniform. To test this formally, let is divide into 4 equal length intervals I_1, I_2, I_3, I_4. Let p_i be the probability of a point being in I_i. The null hypothesis is that $p_1 = p_2 = p_3 = p_4$. We can test this hypothesis using either a likelihood ratio test or a χ^2 test. The latter is

$$\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

[1] See http://ita.ee.lbl.gov/html/contrib/Calgary-HTTP.html for more information.

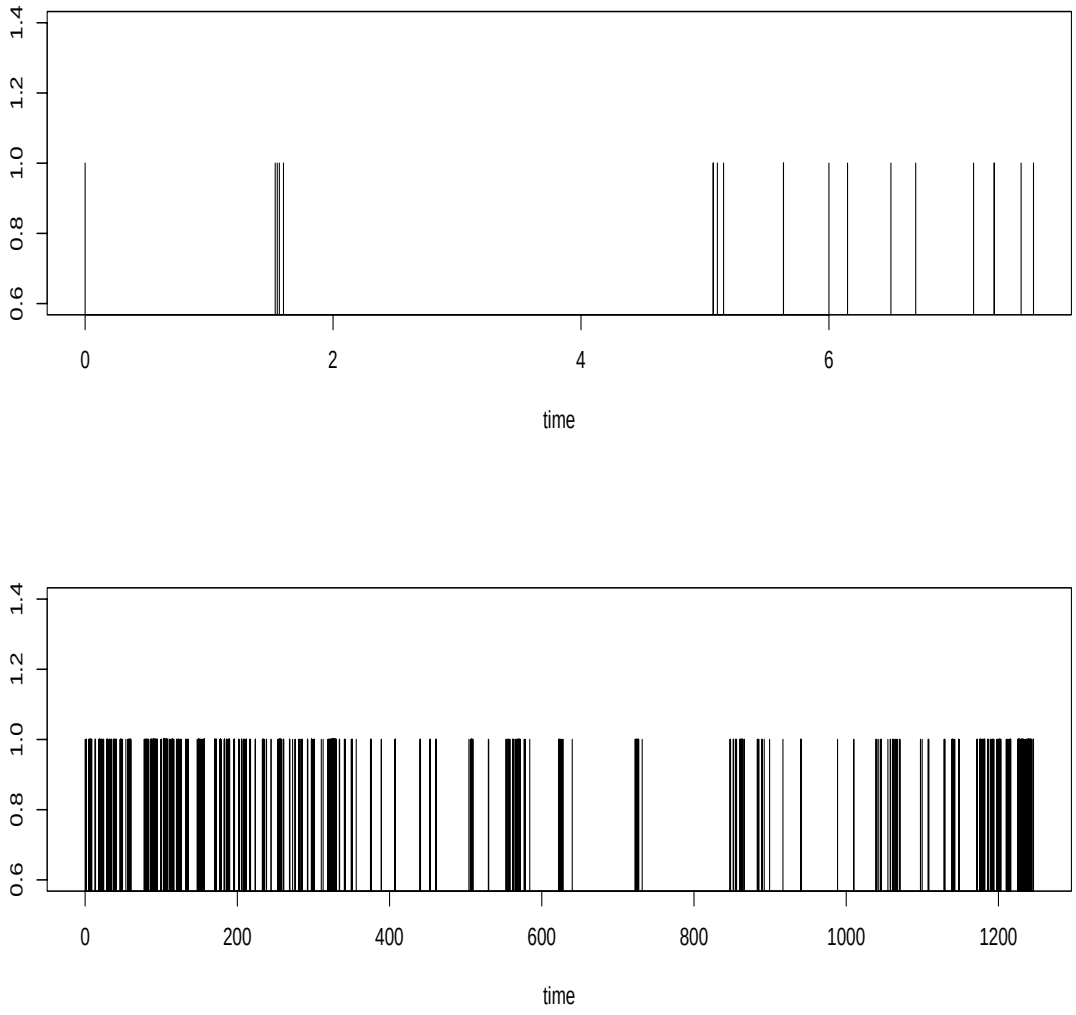


FIGURE 24.3. Hits on a web server. First plot is first 20 events. Second plot is several hours worth of data.

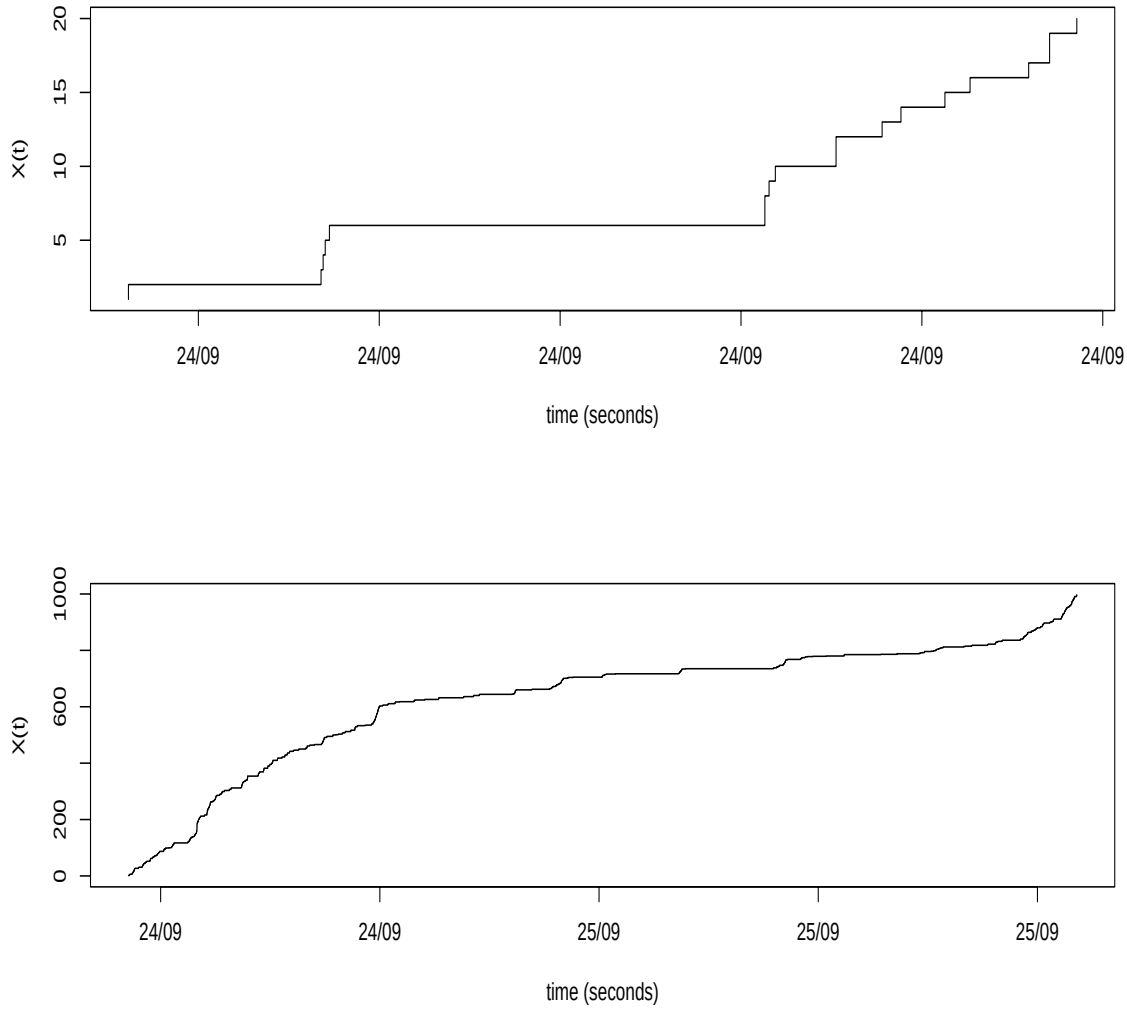


FIGURE 24.4. $X(t)$ for hits on a web server. First plot is first 20 events. Second plot is several hours worth of data.

where O_i is the number of observations in I_i and $E_i = n/4$ is the expected number under the null. This yields $\chi^2 = 252$ which a p-value of 0. Strong evidence againt the null.

24.4 Hidden Markov Models

24.5 Bibliographic Remarks

This is standard material and there are many good references. In this Chapter, I followed the following references quite closely: *Probability and Random Processes* by Grimmett and Stirzaker (1982), *An Introduction to Stochastic Modeling* by Taylor and Karlin (1998), *Stochastic Modeling of Scientific Data* by Guttorp (1995) and *Probability Models for Computer Science* by Ross (2002). Some of the exercises are from these texts.

24.6 Exercises

1. Let $X_0, X_1, \ldots$ be a Markov chain with states $\{0, 1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{bmatrix} 0.1 & 0.2 & 0.7 \\ 0.9 & 0.1 & 0.0 \\ 0.1 & 0.8 & 0.1 \end{bmatrix}$$

   Assume that $\mu_0 = (0.3, 0.4, 0.3)$. Find $\mathbb{P}(X_0 = 0, X_1 = 1, X_2 = 2)$ and $\mathbb{P}(X_0 = 0, X_1 = 1, X_2 = 1)$.

2. Let $Y_1, Y_2, \ldots$ be a sequence of iid observations such that $\mathbb{P}(Y = 0) = 0.1$, $\mathbb{P}(Y = 1) = 0.3$, $\mathbb{P}(Y = 2) = 0.2$, $\mathbb{P}(Y = 3) = 0.4$. Let $X_0 = 0$ and let

$$X_n = \max\{Y_1, \ldots, Y_n\}.$$

   Show that $X_0, X_1, \ldots$ is a Markov chain and find the transition matrix.

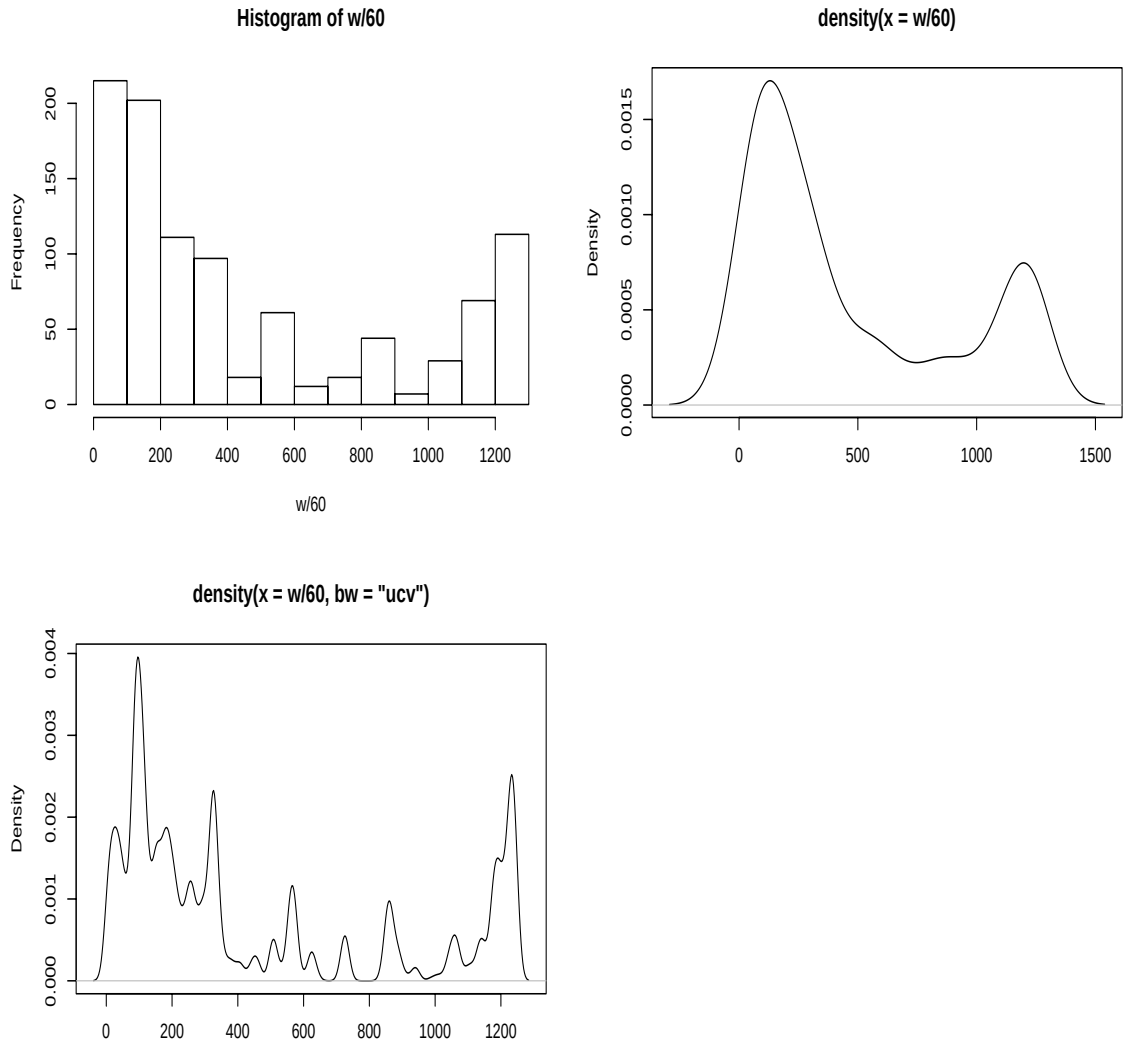


FIGURE 24.5. Density estimate of event times.

3. Consider a two state Markov chain with states $\mathcal{X} = \{1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$$

where $0 < a < 1$ and $0 < b < 1$. Prove that

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{bmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{bmatrix}.$$

4. Consider the chain from question 3 and set $a = .1$ and $b = .3$. Simulate the chain. Let

$$\widehat{p}_n(1) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = 1)$$

$$\widehat{p}_n(2) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = 2)$$

be the proportion of times the chain is in state 1 and state 2. Plot $\widehat{p}_n(1)$ and $\widehat{p}_n(2)$ versus n and verify that they converge to the values predicted from the answer in the previous question.

5. An important Markov chain is the **branching process** which is used in biology, genetics, nuclear physics and many other fields. Suppose that an animal has Y children. Let $p_k = P(Y = k)$. Hence, $p_k \geq 0$ for all k and $\sum_{k=0}^{\infty} p_k = 1$. Assume each animal has the same lifespan and that they produce offspring according to the distribution p_k. Let X_n be the number of animals in the n^{th} generation. Let $Y_1^{(n)}, \ldots, Y_{X_n}^{(n)}$ be the offspring produced in the n^{th} generation. Note that

$$X_{n+1} = Y_1^{(n)} + \cdots + Y_{X_n}^{(n)}.$$

Let $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \mathbb{V}(Y)$. Assume throughout this question that $X_0 = 1$. Let $M(n) = \mathbb{E}(X_n)$ and $V(n) = \mathbb{V}(X_n)$.

(a) Show that $M(n+1) = \mu M(n)$ and $V(n+1) = \sigma^2 M(n) + \mu^2 V(n)$.

(b) Show that $M(n) = \mu^n$ and that $V(n) = \sigma^2 \mu^{n-1}(1 + \mu + \cdots + \mu^{n-1})$.

(c) What happens to the variance if $\mu > 1$? What happens to the variance if $\mu = 1$? What happens to the variance if $\mu < 1$?

(d) The population goes extinct if $X_n = 0$ for some n. Let us thus define the extinction time N by

$$N = \min\{n : \ X_n = 0\}.$$

Let $F(n) = P(N \leq n)$ be the cdf of the random variable N. Show that

$$F(n) = \sum_{k=0}^{\infty} p_k (F(n-1))^k, \quad n = 1, 2, \ldots$$

Hint: Note that the event $\{N \leq n\}$ is the same as event $\{X_n = 0\}$. Thus, $\mathbb{P}(\{N \leq n\}) = \mathbb{P}(\{X_n = 0\})$. Let k be the number of offspring of the original parent. The population becomes extinct at time n if and only if each of the k sub-populations generated from the k offspring goes extinct in $n-1$ generations.

(e) Suppose that $p_0 = 1/4$, $p_1 = 1/2$, $p_2 = 1/4$. Use the formula from (5d) to compute the cdf $F(n)$.

6. Let
$$\mathbf{P} = \begin{bmatrix} 0.40 & 0.50 & 0.10 \\ 0.05 & 0.70 & 0.25 \\ 0.05 & 0.50 & 0.45 \end{bmatrix}$$
Find the stationary distribution π.

7. Show that if i is a recurrent state and $i \leftrightarrow j$, then j is a recurrent state.

8. Let
$$\mathbf{P} = \begin{bmatrix} \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$
Which states are transient? Which states are recurrent?

9. Let
$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
Show that $\pi = (1/2, 1/2)$ is a stationary distribution. Does this chain converge? Why/why not?

10. Let $0 < p < 1$ and $q = 1 - p$. Let
$$\mathbf{P} = \begin{bmatrix} q & p & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 \\ q & 0 & 0 & p & 0 \\ q & 0 & 0 & 0 & p \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$
Find the limiting distribution of the chain.

11. Let $X(t)$ be an inhomogeneous Poisson process with intensity function $\lambda(t) > 0$. Let $\Lambda(t) = \int_0^t \lambda(u)du$. Define $Y(s) = X(t)$ where $s = \Lambda(t)$. Show that $Y(s)$ is a homogeneous Poisson process with intensity $\lambda = 1$.

12. Let $X(t)$ be a Poisson process with intensity λ. Find the conditional distribution of $X(t)$ given that $X(t+s) = n$.

13. Let $X(t)$ be a Poisson process with intensity λ. Find the probability that $X(t)$ is odd, i.e. $\mathbb{P}(X(t) = 1, 3, 5, \ldots)$.

14. Suppose that people logging in to the University computer system is described by a Poisson process $X(t)$ with intensity λ. Assume that a person stays logged in for some random time with cdf G. Assume these times are all independent. Let $Y(t)$ be the number of people on the system at time t. Find the distribution of $Y(t)$.

15. Let $X(t)$ be a Poisson process with intensity λ. Let $W_1, W_2, \ldots,$ be the waiting times. Let f be an arbitrary function. Show that
$$\mathbb{E}\left(\sum_{i=1}^{X(t)} f(W_i)\right) = \lambda \int_0^t f(w)dw.$$

16. A two dimensional Poisson point process is a process of random points on the plane such that (i) for any set A, the number of points falling in A is Poisson with mean $\lambda\mu(A)$ where $\mu(A)$ is the area of A, (ii) the number of events in nonoverlapping regions is independent. Consider an arbitrary point x_0 in the plane. Let X denote the distance from x_0 to the nearest random point. Show that
$$\mathbb{P}(X > t) = e^{-\lambda\pi t^2}$$
and
$$\mathbb{E}(X) = \frac{1}{2\sqrt{\lambda}}.$$

25
Simulation Methods

In this we will see that by generating data in a clever way, we can solve a number of problems such as integrating or maximizing a complicated function.[1] For integration, we will study three methods: (i) basic Monte Carlo integration, (ii) importance sampling and (iii) Markov chain Monte Carlo (MCMC).

25.1 Bayesian Inference Revisited

Simulation methods are especially useful in Bayesian inference so let us briefly review the main ideas in Bayesian inference. Given a prior $f(\theta)$ and data $X^n = (X_1, \ldots, X_n)$ the posterior density is

$$f(\theta|X^n) = \frac{\mathcal{L}(\theta)f(\theta)}{\int \mathcal{L}(u)f(u)du}$$

[1] My main source for this chapter is *Monte Carlo Statistical Methods* by C. Robert and G. Casella.

where $\mathcal{L}(\theta)$ is the likelihood function. The posterior mean is

$$\overline{\theta} = \int \theta f(\theta|X^n)d\theta = \frac{\int \theta \mathcal{L}(\theta)f(\theta)d\theta}{\int \mathcal{L}(\theta)f(\theta)d\theta}.$$

If $\theta = (\theta_1, \ldots, \theta_k)$ is multidimensional, then we might be interested in the posterior for one of the components, θ_1, say. This marginal posterior density is

$$f(\theta_1|X^n) = \int \int \cdots \int f(\theta_1, \ldots, \theta_k|X^n)d\theta_2 \cdots d\theta_k$$

which involves high dimensional integration.

You can see that integrals play a big role in Bayesian inference. When θ is high dimensional, it may not be feasible to calculate these integrals analytically. Simulation methods will often be very helpful.

25.2 Basic Monte Carlo Integration

Suppose you want to evaluate the integral $I = \int_a^b h(x)dx$ for some function h. If h is an "easy" function like a polynomial or triginometric function then we can do the integral in closed form. In practice, h can be very complicated and there may be no known closed form expression for I. There are many numerical techniques for evaluating I such as Simpson's rule, the trapezoidal rule, Guassian quadrature and so on. In some cases these techniques work very well. But other times they may not work so well. In particular, it is hard to extend them to higher dimensions. Monte Carlo integration is another approach to evaluating I which is notable for its simplicity, generality and scalability.

Let us begin by writing

$$I = \int_a^b h(x)dx = \int_a^b h(x)(b-a)\frac{1}{b-a}dx = \int_a^b w(x)f(x)dx$$

where $w(x) = h(x)(b-a)$ and $f(x) = 1/(b-a)$. Notice that f is the density for a uniform random variable over (a, b). Hence,

$$I = E_f(w(X))$$

where $X \sim \text{Unif}(a, b)$.

Suppose we generate $X_1, \ldots, X_N \sim \text{Unif}(a, b)$ where N is large. By the law of large numbers

$$\widehat{I} \equiv \frac{1}{N}\sum_{i=1}^{N} w(X_i) \xrightarrow{p} E(w(X)) = I.$$

This is the basic **Monte Carlo integration method**. We can also compute the standard error of the estimate

$$\widehat{\text{se}} = \frac{s}{\sqrt{N}}$$

where

$$s^2 = \frac{\sum_i (Y_i - \widehat{I})^2}{N-1}$$

where $Y_i = w(X_i)$. A $1-\alpha$ confidence interval for I is $\widehat{I} \pm z_{\alpha/2}\widehat{se}$. We can take N as large as we want and hence make the length of the confidence inteval very small.

Example 25.1 *Let's try this on an example where we know the true answer. Let $h(x) = x^3$. Hence, $I = \int_0^1 x^3 dx = 1/4$. Based on $N = 10,000$ I got $\widehat{I} = .2481101$ with a standard error of .0028. Not bad at all.*

A simple generalization of the basic method is to consider integrals of the form

$$I = \int h(x)f(x)dx$$

where $f(x)$ is a probability density function. Taking f to be a Unif (a,b) gives us the special case above. The only difference is

that now we draw $X_1, \ldots, X_N \sim f$ and take

$$\widehat{I} \equiv \frac{1}{N} \sum_{i=1}^{N} h(X_i)$$

as before.

Example 25.2 *Let*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

be the standard Normal pdf. Suppose we want to compute the cdf at some point x:

$$I = \int_{-\infty}^{x} f(s)ds = \Phi(x).$$

Of course, we could look this up in a table or use the pnorm command in R. But suppose you don't have access to either. Write

$$I = \int h(s)f(s)ds$$

where

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

Now we generate $X_1, \ldots, X_N \sim N(0,1)$ and set

$$\widehat{I} = \frac{1}{N} \sum_i h(X_i) = \frac{\text{number of observations } \leq x}{N}.$$

For example, with $x = 2$, the true answer is $\Phi(2) = .9772$ and the Monte Carlo estimate with $N = 10,000$ yields .9751. Using $N = 100,000$ I get .9771.

Example 25.3 (Two Binomials) *Let $X \sim \text{Binomial}(n, p_1)$ and $Y \sim \text{Binomial}(m, p_2)$. We would like to estimate $\delta = p_2 - p_1$. The mle is $\widehat{\delta} = \widehat{p}_2 - \widehat{p}_1 = (Y/m) - (X/n)$. We can get the standard error $\widehat{se}$ using the delta method (remember?) which yields*

$$\widehat{se} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$$

and then construct a 95 per cent confidence interval $\widehat{\delta} \pm 2\,\widehat{se}$. *Now consider a Bayesian analysis. Suppose we use the prior* $f(p_1, p_2) = f(p_1)f(p_2) = 1$, *that is, a flat prior on* (p_1, p_2). *The posterior is*

$$f(p_1, p_2|X, Y) \propto p_1^X (1-p_1)^{n-X}\, p_2^Y (1-p_2)^{m-Y}.$$

The posterior mean of δ *is*

$$\overline{\delta} = \int_0^1 \int_0^1 \delta(p_1, p_2) f(p_1, p_2|X, Y) = \int_0^1 \int_0^1 (p_2 - p_1) f(p_1, p_2|X, Y).$$

If we want the posterior density of δ *we can first get the posterior cdf*

$$F(c|X, Y) = P(\delta \leq c|X, Y) = \int_A f(p_1, p_2|X, Y)$$

where $A = \{(p_1, p_2) : \; p_2 - p_1 \leq c\}$. *The density can then be obtained by differentiating* F.

To avoid all these integrals, let's use simulation. Note that $f(p_1, p_2|X, Y) = f(p_1|X)f(p_2|Y)$ *which implies that* p_1 *and* p_2 *are independent under the posterior distribution. Also, we see that* $p_1|X \sim \text{Beta}(X+1, n-X+1)$ *and* $p_2|Y \sim \text{Beta}(Y+1, m-Y+1)$. *Hence, we can simulate* $(P_1^{(1)}, P_2^{(1)}), \ldots, (P_1^{(N)}, P_2^{(N)})$ *from the posterior by drawing*

$$\begin{aligned} P_1^{(i)} &\sim \text{Beta}(X+1, n-X+1) \\ P_2^{(i)} &\sim \text{Beta}(Y+1, m-Y+1) \end{aligned}$$

for $i = 1, \ldots, N$. *Now let* $\delta^{(i)} = P_2^{(i)} - P_1^{(i)}$. *Then,*

$$\overline{\delta} \approx \frac{1}{N} \sum_i \delta^{(i)}.$$

We can also get a 95 per cent posterior interval for δ *by sorting the simulated values, and finding the* .025 *and* .975 *quantile. The posterior density* $f(\delta|X, Y)$ *can be obtained by applying density*

estimation techniques to $\delta^{(1)}, \ldots, \delta^{(N)}$ *or, simply by plotting a histogram.*

For example, suppose that $n = m = 10$, $X = 8$ *and* $Y = 6$. *The R code is:*

```
x <- 8; y <- 6
n <- 10; m <- 10
B <- 10000
p1 <- rbeta(B,x+1,n-x+1)
p2 <- rbeta(B,y+1,m-y+1)
delta <- p2 - p1
print(mean(delta))
left <- quantile(delta,.025)
right <- quantile(delta,.975)
print(c(left,right))
```

I found that $\overline{\delta}$ *and a 95 per cent posterior interval of (-0.52,0.20). The posterior density can be estimated from a histogram of the simulated values as shown in Figure 25.1.*

Example 25.4 (Dose Response) *Suppose we conduct an experiment by giving rats one of ten possible doses of a drug, denoted by* $x_1 < x_2 < \ldots < x_{10}$. *For each dose level* x_i *we use* n *rats and we observe* Y_i, *the number that survive. Thus we have ten independent binomials* $Y_i \sim \text{Binomial}(n, p_i)$. *Suppose we know from biological considerations that higher doses should have higher probability of death. Thus,* $p_1 \leq p_2 \leq \cdots \leq p_{10}$. *Suppose we want to estimate the dose at which the animals have a 50 per cent chance of dying. This is called the LD50. Formally,* $\delta = x_j$ *where*

$$j = \min\{i : \ p_i \geq .50\}.$$

Notice that δ *is implicitly a (complicated) function of* $p_1, \ldots, p_{10}$ *so we can write* $\delta = g(p_1, \ldots, p_{10})$ *for some* g. *This just means that if we know* $(p_1, \ldots, p_{10})$ *then we can find* δ. *The posterior*

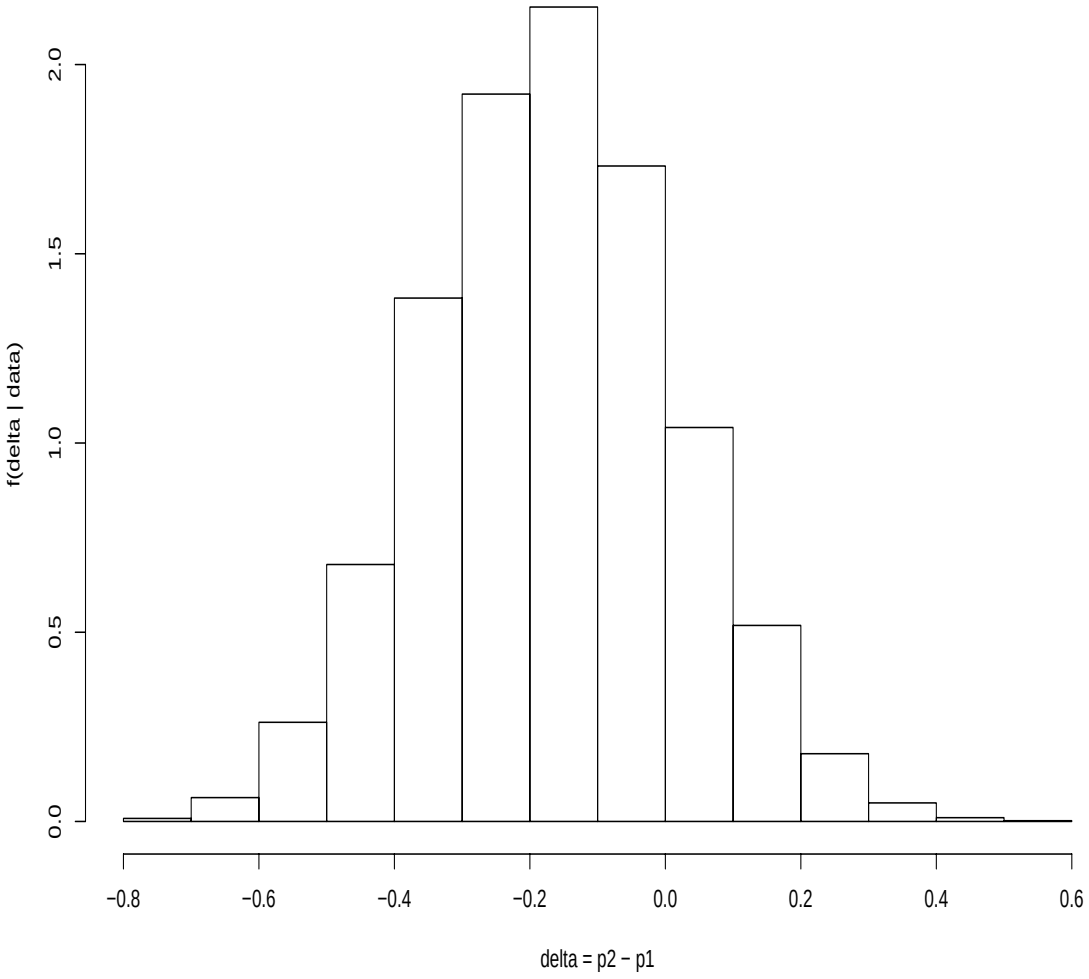


FIGURE 25.1. Posterior of δ from simulation.

mean of δ *is*

$$\int\int\cdots\int_A g(p_1,\ldots,p_{10})f(p_1,\ldots,p_{10}|Y_1,\ldots,Y_{10})dp_1dp_2\ldots dp_{10}.$$

The integral is over the region

$$A=\{(p_1,\ldots,p_{10}):\ p_1\leq\cdots\leq p_{10}\}.$$

Similarly, the posterior cdf of δ *is*

$$\begin{aligned}F(c|Y_1,\ldots,Y_{10}) &= P(\delta\leq c|Y_1,\ldots,Y_{10})\\ &= \int\int\cdots\int_B f(p_1,\ldots,p_{10}|Y_1,\ldots,Y_{10})dp_1dp_2\ldots dp_{10}\end{aligned}$$

where

$$B=A\bigcap\{(p_1,\ldots,p_{10}):\ g(p_1,\ldots,p_{10})\leq c\}.$$

We would need to do a 10 dimensional integral over a resticted region A. Instead, let's use simulation.

Let us take a flat prior truncated over A. Except for the truncation, each P_i *has once again a Beta distribution. To draw from the posterior we do the following steps:*

(1) Draw $P_i\sim\text{Beta}(Y_i+1,n-Y_i+1), i=1,\ldots,10.$

(2) If $P_1\leq P_2\leq\cdots\leq P_{10}$ *keep this draw. Otherwise, throw it away and draw again until you get one you can keep.*

(3) Let $\delta=x_j$ *where*

$$j=\min\{i:\ P_i>.50\}.$$

We repeat this N times to get $\delta^{(1)},\ldots,\delta^{(N)}$ *and take*

$$E(\delta|Y_1,\ldots,Y_{10})\approx\frac{1}{N}\sum_i\delta^{(i)}.$$

δ is a discrete variable. We can estimate its probability mass function by

$$P(\delta = x_j | Y_1, \ldots, Y_{10}) \approx \frac{1}{N} \sum_{i=1}^{N} I(\delta^{(i)} = j).$$

For example, suppose that $x = (1, 2, \ldots, 10)$. Figure 25.2 shows some data with $n = 15$ animals at each dose. The R code is a bit tricky.

```
## assume x and y are given
n <- rep(15,10)
B <- 10000
p <- matrix(0,B,10)
check <- rep(0,B)
for(i in 1:10){
     p[,i] <- rbeta(B, y[i]+1, n[i]-y[i]+1)
     }
for(i in 1:B){
     check[i] <- min(diff(p[i,]))
     }
good <- (1:B)[check >= 0]
p <- p[good,]
B <- nrow(p)
delta <- rep(0,B)
for(i in 1:B){
     temp <- cumsum(p[i,])
     j <- (1:10)[temp > .5]
     j <- min(j)
     delta[i] <- x[j]
     }
print(mean(delta))
left <- quantile(delta,.025)
right <- quantile(delta,.975)
print(c(left,right))
```

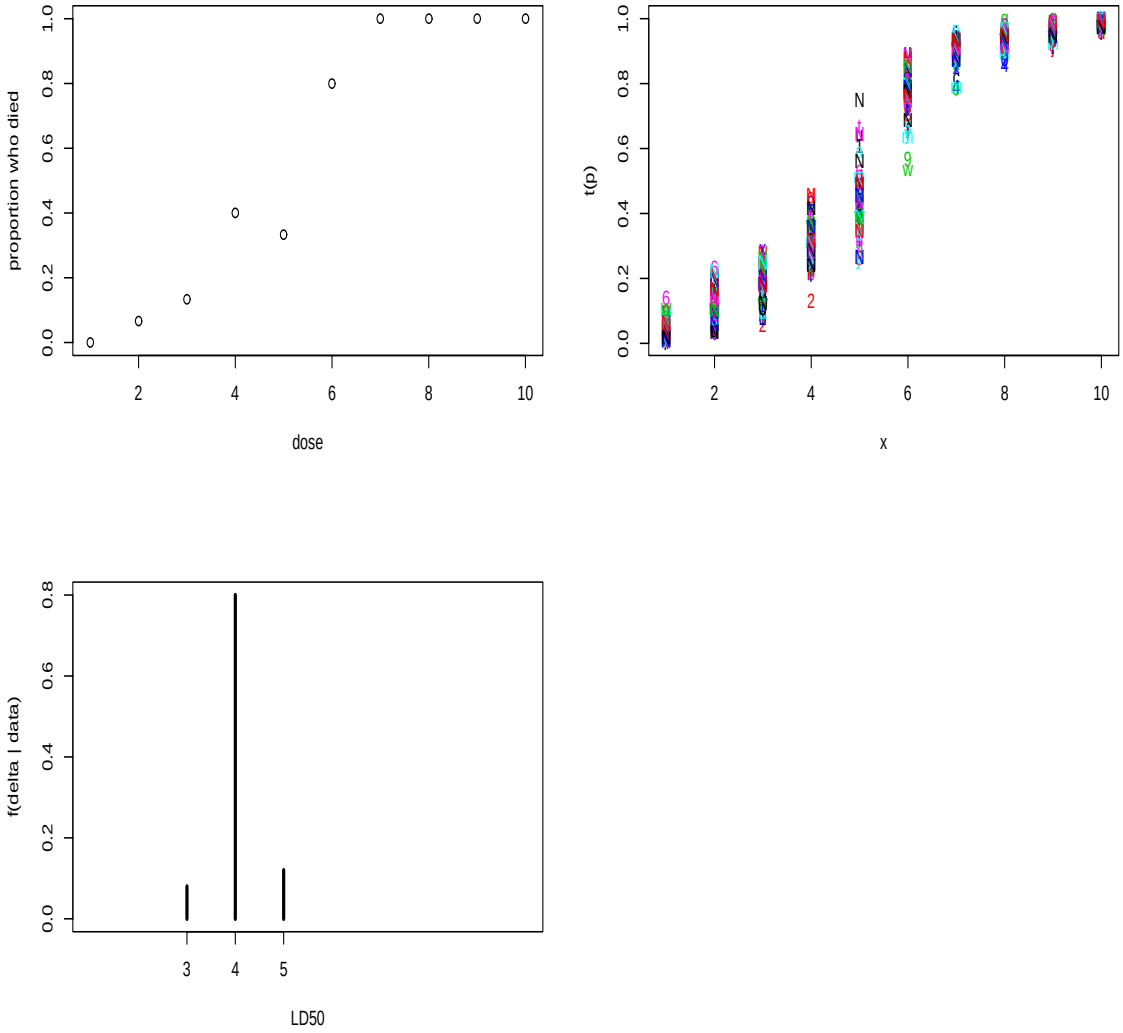


FIGURE 25.2. Dose repoonse data.

The posterior draws for $p_1, \ldots, p_{10}$ are shown in the second panel in the figure. We find that that $\overline{\delta} = 4.04$ with a 95 per cent interval of (3,5). The third panel shows the posterior for δ which is necessarily a discrete distribution.

25.3 Importance Sampling

Consider the integral $I = \int h(x)f(x)dx$ where f is a probability density. The basic Monte Carlo method involves sampling from f. However, there are cases where we may not know how to sample from f. For example, in Bayesian inference, the posterior density density is is obtained by multiplying the likelihood $\mathcal{L}(\theta)$ times the prior $f(\theta)$. There is no guarantee that $f(\theta|x)$ will be a known distribution like a Normal or Gamma or whatever.

Importance sampling is a generalization of basic Monte Carlo which overcomes this problem. Let g be a probability density that we know how to simulate from. Then

$$I = \int h(x)f(x)dx = \int \frac{h(x)f(x)}{g(x)}g(x)dx = E_g(Y)$$

where $Y = h(X)f(X)/g(X)$ and the expectation $E_g(Y)$ is with respect to g. We can simulate $X_1, \ldots, X_N \sim g$ and estimate I by

$$\widehat{I} = \frac{1}{N}\sum_i Y_i = \frac{1}{N}\sum_i \frac{h(X_i)f(X_i)}{g(X_i)}.$$

This is called **importance sampling.** By the law of large numbers, $\widehat{I} \xrightarrow{p} I$. However, there is a catch. It's possible that $\widehat{I}$ might have an infinite standard error. To see why, recall that I is the mean of $w(x) = h(x)f(x)/g(x)$. The second moment of this quantity is

$$E_g(w^2(X)) = \int \left(\frac{h(x)f(x)}{g(x)}\right)^2 g(x)dx = \int \frac{h^2(x)f^2(x)}{g(x)}dx.$$

If g has thinner tails than f, then this integral might be infinite. To avoid this, a basic rule in importance sampling is to sample from a density g with thicker tails than f. Also, suppose that $g(x)$ is small over some set A where $f(x)$ is large. Again, the ratio of f/g could be large leading to a large variance. This implies that we should choose g to be similar in shape to f. In summary, a good choice for an importance sampling density g should be similar to f but with thicker tails. In fact, we can say what the optimal choice of g is.

Theorem **25.5** *The choice of g that minimizes the variance of $\widehat{I}$ is*

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(s)|f(s)ds}.$$

PROOF. The variance of $w = fh/g$ is

$$\begin{aligned} E_g(w^2) - (E(w^2))^2 &= \int w^2(x)g(x)dx - \left(\int w(x)g(x)dx\right)^2 \\ &= \int \frac{h^2(x)f^2(x)}{g^2(x)}g(x)dx - \left(\int \frac{h(x)f(x)}{g(x)}g(x)dx\right)^2 \\ &= \int \frac{h^2(x)f^2(x)}{g^2(x)}g(x)dx - \left(\int h(x)f(x)dx\right)^2. \end{aligned}$$

The second integral does not depend on g so we only need to minimize the first integral. Now, from Jensen's inequality we have

$$E_g(W^2) \geq (E_g(|W|))^2 = \left(\int |h(x)|f(x)dx\right)^2.$$

This establishes a lower bound on $E_g(W^2)$. However, $E_{g^*}(W^2)$ equals this lower bound which proves the claim. □

This theorem is interesting but it is only of theoretical interest. If we did not know how to sample from f then it is unlikely

that we could sample from $|h(x)|f(x)/\int |h(s)|f(s)ds$. In practice, we simply try to find a thick tailed distribution g which is similar to $f|h|$.

Example 25.6 (Tail Probability) *Let's estimate* $I = P(Z > 3) = .0013$ *where* $Z \sim N(0,1)$. *Write* $I = \int h(x)f(x)dx$ *where* $f(x)$ *is the standard Normal density and* $h(x) = 1$ *if* $x > 3$ *and 0 otherwise. The basic Monte Carlo estimator is* $\widehat{I} = N^{-1}\sum_i h(X_i)$. *Using* $N = 100$ *we find (from simulating many times) that* $E(\widehat{I}) = .0015$ *and* $Var(\widehat{I}) = .0039$. *Notice that most observations are wasted in the sense that most are not near the right tail. We will estimate this with importance sampling taking* g *to be a Normal(4,1) density. We draw values from* g *and the estimate is now* $\widehat{I} = N^{-1}\sum_i f(X_i)h(X_i)/g(X_i)$. *In this case we find that* $E(\widehat{I}) = .0011$ *and* $Var(\widehat{I}) = .0002$. *We have reduced the standard deviation by a factor of 20.*

Example 25.7 (Measurement Model With Outliers) *Suppose we have measurements* $X_1, \ldots, X_n$ *of some physical quantity* θ. *We might model this as*

$$X_i = \theta + \epsilon_i.$$

If we assume that $\epsilon_i \sim N(0,1)$ *then* $X_i \sim N(\theta_i, 1)$. *However, when taking measurements, it is often the case that we get the occasional wild obvervation, or outlier. This suggests that a Normal might be a poor model since Normals have thin tails which implies that extreme observations are rare. One way to improve the model is to use a density for* ϵ_i *with a thicker tail, for example, a t-distribution with* ν *degrees of freedom which has the form*

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

Smaller values of ν *correspond to thicker tails. For the sake of illustration we will take* $\nu = 3$. *Suppose we observe* n $X_i = \theta + \epsilon_i$, $i = 1, \ldots, n$ *where* ϵ_i *has a t distribution with* $\nu = 3$. *We will*

take a flat prior on θ. The likelihood is $\mathcal{L}(\theta) = \prod_{i=1}^n t(X_i - \theta)$ and the posterior mean of θ is

$$\overline{\theta} = \frac{\int \theta \mathcal{L}(\theta) d\theta}{\int \mathcal{L}(\theta) d\theta}.$$

We can estimate the top and bottom integral using importance sampling. We draw $\theta_1, \ldots, \theta_N \sim g$ and then

$$\overline{\theta} \approx \frac{\frac{1}{N}\sum_{j=1}^N \frac{\theta_j \mathcal{L}(\theta_j)}{g(\theta_j)}}{\frac{1}{N}\sum_{j=1}^N \frac{\mathcal{L}(\theta_j)}{g(\theta_j)}}.$$

To illustrate the idea, I drew $n = 2$ observations. The posterior mean (ccomputed numerically) is -0.54. Using a Normal importance sampler g yields an estimate of -0.74. Using a Cauchy (t-distribution with 1 degree of freedom) importance sampler yields an estimate of -0.53.

25.4 MCMC Part I: The Metropolis-Hastings Algorithm

Consider again the problem of estimating the integral $I = \int h(x) f(x) dx$. In this chapter we introduce Markov chain Monte Carlo (MCMC) methods. The idea is to constuct a Markov chain $X_1, X_2, \ldots,$ whose stationary distribution is f. Under certain conditions it will then follow that

$$\frac{1}{N}\sum_{i=1}^N h(X_i) \xrightarrow{p} E_f(h(X)) = I.$$

This works because there is a law of large numbers for Markov chains called **the ergodic theorem.**

The **Metropolis-Hastings** algorithm is a specific MCMC method that works as follows. Let $q(y|x)$ be an arbitrary, friendly distribution (i.e. we know how to sample from $q(y|x)$). The conditional density $q(y|x)$ is called the **proposal distribution.**

The Metropolis-Hastings algorithm creates a sequence of observations $X_0, X_1, \ldots$, as follows.

Metropolis-Hastings Algorithm. Choose X_0 arbitrarily. Suppose we have generated $X_0, X_1, \ldots, X_i$. To generate X_{i+1} do the following:

(1) Generate a **proposal** or **candidate** value $Y \sim q(y|X_i)$.

(2) Evaluate $r \equiv r(X_i, Y)$ where

$$r(x, y) = \min\left\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)},\ 1\right\}.$$

(3) Set

$$X_{i+1} = \begin{cases} Y & \text{with probability } r \\ X_i & \text{with probability } 1 - r. \end{cases}$$

REMARK 1: A simple way to execute step (3) is to generate $U \sim (0, 1)$. If $U < r$ set $X_{i+1} = Y$ otherwise set $X_{i+1} = X_i$.

REMARK 2: A common choice for $q(y|x)$ is $N(x, b^2)$ for some $b > 0$. This means that the proposal is draw from a Normal, centered at the current value.

REMARK 3: If the proposal density q is symmetric, $q(y|x) = q(x|y)$, then r simplifies to

$$r = \min\left\{\frac{f(Y)}{f(X_i)},\ 1\right\}.$$

The Normal proposal distribution mentioned in Remark 2 is an example of a symmetric proposal density.

By construction, $X_0, X_1, \ldots$ is a Markov chain. But why does this Markov chain have f as its stationary distribution? Before we explain why, let us first do an example.

Example 25.8 *The Cauchy distribution has density*

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

Our goal is to simulate a Markov chain whose stationary distribution is f. As suggested in the remark above, we take $q(y|x)$ to be a $N(x, b^2)$. So in this case,

$$r(x,y) = \min\left\{\frac{f(y)}{f(x)},\ 1\right\} = \min\left\{\frac{1+x^2}{1+y^2},\ 1\right\}.$$

So the algorithm is to draw $Y \sim N(X_i, b^2)$ and set

$$X_{i+1} = \begin{cases} Y & \text{with probability } r(X_i, Y) \\ X_i & \text{with probability } 1 - r(X_i, Y). \end{cases}$$

The R code is very simple:

```
metrop <- function(N,b){
     x <- rep(0,N)
     for(i in 2:N){
          y <- rnorm(1,x[i-1],b)
          r <- (1+x^2)/(1+y^2)
          u <- runif(1)
          if(u < r)x[i] <- y else x[i] <- x[i-1]
          }
     return(x)
     }
```

The simulator requires a choice of b. Figure 25.3 shows three chains of length $N = 1000$ using $b = .1$, $b = 1$ and $b = 10$. The histograms of the simulated values are shown in Figure 25.4. Setting $b = .1$ forces the chain to take small steps. As a result, the chain doesn't "explore" much of the sample space. The histogram

from the sample does not approximate the true density very well. Setting $b = 10$ causes the proposals to often be far in the tails, making r small and hence we reject the proposal and keep the chain at its current position. The result is that the chain "gets stuck" at the same place quite often. Again, this means that the histogram from the sample does not approximate the true density very well. The middle choice avoids these extremes and results in a Markov chain sample that better represents the density sooner. In summary, there are are tuning parameters and the efficiency of the chain depends on these parameters. We'll discus this in more detail later.

If the sample from the Markov chain starts to "look like" the target distribution f quickly, then we say that the chain is "mixing well." Constructing a chain that mixes well is somewhat of an art.

WHY IT WORKS. Recall from the previous chapter that a distribution π satisfies **detailed balance** for a Markov chain if

$$p_{ij}\pi_i = p_{ji}\pi_j.$$

We then showed that if π satisfies detailed balance, then it is a stationary distribution for the chain.

Because we are now dealing with continuous state Markov chains, we will change notation a little and write $p(x, y)$ for the probability of making a transition from x to y. Also, let's use $f(x)$ instead of π for a distribution. In this new notation, f is a stationary distribution if $f(x) = \int f(y)p(y, x)dy$ and detailed balance holds for f if

$$f(x)p(x, y) = f(y)p(y, x).$$

Detailed balance implies that f is a stationary distribution since, if detailed balance holds, then

$$\int f(y)p(y, x)dy = \int f(x)p(x, y)dy = f(x)\int p(x, y)dy = f(x)$$

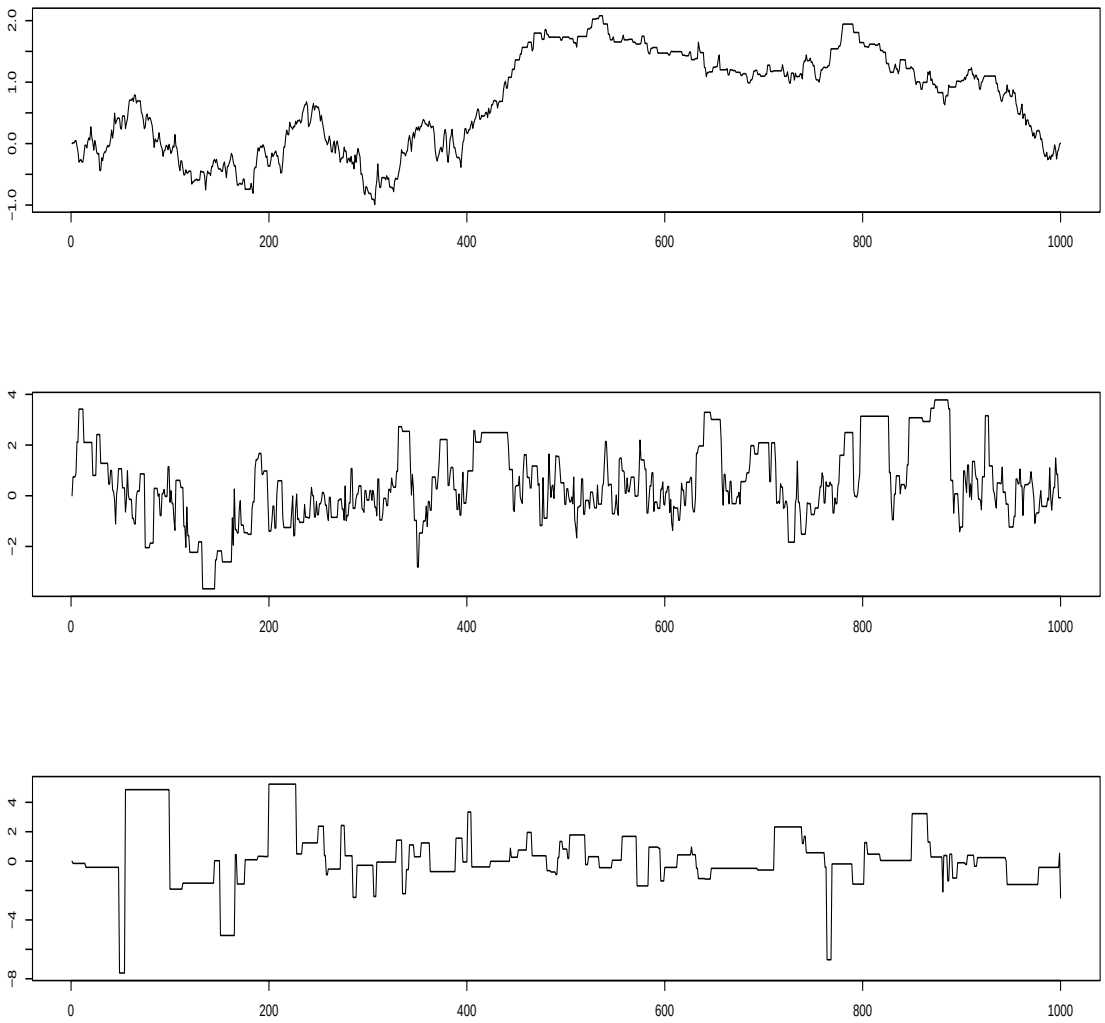


FIGURE 25.3. Three Metropolis chains corresponding to $b = .1$, $b = 1$, $b = 10$.

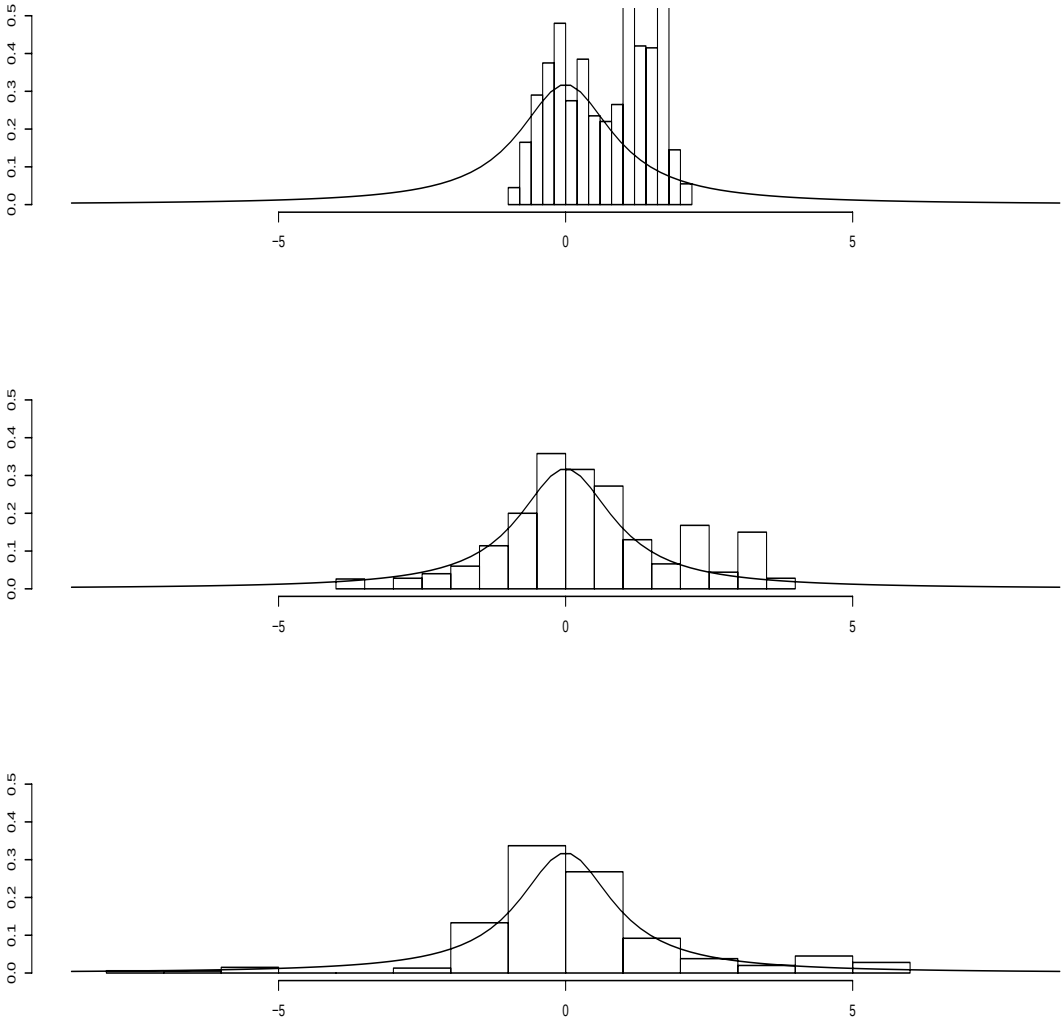


FIGURE 25.4. Histograms from the three chains. The true density is also plotted.

which shows that $f(x) = \int f(y)p(y,x)dy$ as required. Our goal is to show that f satisfies detailed balance which will imply that f is a stationary distribution for the chain.

Consider two points x and y. Either

$$f(x)q(y|x) < f(y)q(x|y) \quad \text{or} \quad f(x)q(y|x) > f(y)q(x|y).$$

We will ignore ties which we can do in the continuous setting. Without loss of generality, assume that $f(x)q(y|x) > f(y)q(x|y)$. This implies that

$$r(x,y) = \frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}$$

and that $r(y,x) = 1$. Now $p(x,y)$ is the probability of jumping from x to y. This requires two things: (i) the proposal distribution must generate y and (ii) you must accept y. Thus,

$$p(x,y) = q(y|x)r(x,y) = q(y|x)\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)} = \frac{f(y)}{f(x)}q(x|y).$$

Therefore,

$$f(x)p(x,y) = f(y)q(x|y). \tag{25.1}$$

On the other hand, $p(y,x)$ is the probability of jumping from y to x. This requires two things: (i) the proposal distribution must generate x and (ii) you must accept x. This occurs with probability $p(y,x) = q(x|y)r(y,x) = q(x|y)$. Hence,

$$f(y)p(y,x) = f(y)q(x|y). \tag{25.2}$$

Comparing (25.1) and (25.2), we see that we have shown that detailed balance holds.

25.5 MCMC Part II: Different Flavors

There are different types of MCMC algorithm. Here we will consider a few of the most popular versions.

RANDOM-WALK-METROPOLIS-HASTINGS. In the previous section we considered drawing a proposal Y of the form

$$Y = X_i + \epsilon_i$$

where ϵ_i comes from some distribution with density g. In other words, $q(y|x) = g(y - x)$. We saw that in this case,

$$r(x, y) = \min\left\{1, \frac{f(y)}{f(x)}\right\}.$$

This is called a **random-walk-Metropolis-Hastings** method. The reason for the name is that, if we did not do the accept-reject step, we would be simulating a random walk. The most common choice for g is a $N(0, b^2)$. The hard part is choosng b so that the chain mixes well. A good rule of thumb is: choose b so that you accept the proposals about 50 per cent of the time.

Warning: This method doesn't make sense unless X takes values on the whole real line. If X is restricted to some interval then it is best to transform X, say, $Y = m(X)$ say, where Y takes values on the whole real line. For example, if $X \in (0, \infty)$ then you might take $Y = \log X$ and then simulate the distribution for Y instead of X.

INDEPENDENCE-METROPOLIS-HASTINGS. This is like an importance-sampling version of MCMC. In this we draw the proposal from a fixed distribution g. Generally, g is chosen to be an approximation to f. The acceptance probability becomes

$$r(x, y) = \min\left\{1, \frac{f(y)}{f(x)}\frac{g(x)}{g(y)}\right\}.$$

GIBBS SAMPLING. The two previous methods can be easily adapted, in principle, to work in higher dimensions. In practice,

tuning the chains to make them mix well is hard. Gibbs sampling is a way to turn a high-dimensional problem into several one dimensional problems.

Here's how it works for a bivariate problem. Suppose that (X, Y) has density $f_{X,Y}(x, y)$. First, suppose that it is possible to simulate from the conditional distributions $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. Let (X_0, Y_0) be starting values. Asume we have drawn $(X_0, Y_0), \ldots, (X_n, Y_n)$. Then the Gibbs sampling algorithm for getting (X_{n+1}, Y_{n+1}) is:

The $n+1$ step in Gibbs sampling:

$$\begin{aligned} X_{n+1} &\sim f_{X|Y}(x|Y_n) \\ Y_{n+1} &\sim f_{Y|X}(y|X_{n+1}) \end{aligned}$$

This generalizes in the obvious way to higher dimensions.

Example 25.9 (Normal Hierarchical Model) *Gibbs sampling is very useful for a class of models called* **hierarchical models**. *Here is a simple case. Suppose we draw a sample of k cities. From each city we draw n_i people and observe how many people Y_i have a disease. Thus, $Y_i \sim \text{Binomial}(n_i, p_i)$. We are allowing for different disease rates in different cities. We can also think of the $p_i's$ as random draws from some distribution F. We can write this model in the following way:*

$$\begin{aligned} P_i &\sim F \\ Y_i|P_i = p_i &\sim \text{Binomial}(n_i, p_i). \end{aligned}$$

We are interested in estimating the $p_i's$ and the overall disease rate $\int p dF(p)$.

To proceed, it will simplify matters if we make some transformations that allow us to use some Normal approxmations. Let

$\widehat{p}_i = Y_i/n_i$. Recall that $\widehat{p}_i \approx N(p_i, s_i)$ where $s_i = \sqrt{\widehat{p}_i(1-\widehat{p}_i)/n_i}$. Let $\psi_i = \log(p_i/(1-p_i))$ and define $Z_i \equiv \widehat{\psi}_i = \log(\widehat{p}_i/(1-\widehat{p}_i))$. By the delta method,

$$\widehat{\psi}_i \approx N(\psi_i, \sigma_i^2)$$

where $\sigma_i^2 = 1/(n\widehat{p}_i(1-\widehat{p}_i))$. Experience shows that the Normal approximation for ψ is more accurate than the Normal approximation for p so we shall work with ψ. We shall treat σ_i as known. Furthermore, we shall take the distribution of the ψ_i's to be Normal. The hierarchical model is now

$$\begin{aligned} \psi_i &\sim N(\mu, \tau^2) \\ Z_i|\psi_i &\sim N(\psi_i, \sigma_i^2). \end{aligned}$$

As yet another simplification we take $\tau = 1$. The unknown parameter are $\theta = (\mu, \psi_1, \ldots, \psi_k)$. The likelihood function is

$$\begin{aligned} \mathcal{L}(\theta) &\propto \prod_i f(\psi_i|\mu) \prod_i f(Z_i|\psi) \\ &\propto \prod_i \exp\left\{-\frac{1}{2}(\psi_i - \mu)^2\right\} \exp\left\{-\frac{1}{2\sigma_i^2}(Z_i - \psi_i)^2\right\}. \end{aligned}$$

If we use the prior $f(\mu) \propto 1$ then the posterior is proportional to the likelihood. To use Gibbs sampling, we need to find the conditional distribution of each parameter conditional on all the others. Let us begin by finding $f(\mu|\text{rest})$ where "rest" refers to all the other variables. We can throw away any terms that don't involve μ. Thus,

$$\begin{aligned} f(\mu|\text{rest}) &\propto \prod_i \exp\left\{-\frac{1}{2}(\psi_i - \mu)^2\right\} \\ &\propto \exp\left\{-\frac{k}{2}(\mu - b)^2\right\} \end{aligned}$$

where

$$b = \frac{1}{k}\sum_i \psi_i.$$

Hence we see that $\mu|\text{rest} \sim N(b, 1/k)$. *Next we will find* $f(\psi|\text{rest})$. *Again, we can throw away any terms not involving* ψ_i *leaving us with*

$$\begin{aligned} f(\psi_i|\text{rest}) &\propto \exp\left\{-\frac{1}{2}(\psi_i-\mu)^2\right\}\exp\left\{-\frac{1}{2\sigma_i^2}(Z_i-\psi_i)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2d_i^2}(\psi_i-e_i)^2\right\} \end{aligned}$$

where

$$e_i = \frac{\frac{Z_i}{\sigma_i^2}+\mu}{1+\frac{1}{\sigma_i^2}} \quad \text{and} \quad d_i^2 = \frac{1}{1+\frac{1}{\sigma_i^2}}$$

and so $\psi_i|\text{rest} \sim N(e_i, d_i^2)$. *The Gibbs sampling algorithm then involves iterating the following steps* N *times:*

$$\begin{aligned} \text{draw } \mu &\sim N(b, v^2) \\ \text{draw } \psi_1 &\sim N(e_1, d_1^2) \\ &\vdots \quad\quad \vdots \\ \text{draw } \psi_k &\sim N(e_k, d_k^2). \end{aligned}$$

It is understood that at each step, the most recently drawn version of each variable is used.

Let's now consider a numerical example. Suppose that there are $k = 20$ *cities and we sample* $n = 20$ *people from each city. The R function for this problem is:*

```
gibbs.fun <- function(y,n,N){
     k <- length(y)
     p.hat <- y/n
     Z <- log(p.hat/(1-p.hat))
     sigma <- sqrt(1/(n*p.hat*(1-p.hat)))
     v <- sqrt(1/sum(1/sigma^2))
     mu <- rep(0,N)
     psi <- matrix(0,N,k)
     for(i in 2:N){
```

```
        ### draw mu given rest
        b <- v^2*sum(psi[i-1,]/sigma^2)
        mu[i] <- rnorm(1,b,v)

        ### draw psi given rest
        e <- (Z + mu[i]/sigma^2)/(1 + (1/sigma^2))
        d <- sqrt(1/(1 + (1/sigma^2)))
        psi[i,] <- rnorm(k,e,d)
        }
    list(mu=mu,psi=psi)
    }
```

After running the chain, we can convert each ψ_i back into p_i by way of $p_i = e^{\psi_i}/(1+e^{\psi_i})$. The raw proportions are shown in Figure 25.6. Figure 25.5 shows "trace plots" of the Markov chain for p_1 and μ. (We should also look at the plots for $p_2, \ldots, p_{20}$.) Figure 25.6 shows the posterior for μ based on the simulated values. The second panel of Figure 25.6 shows the raw proportions and the Bayes estimates. Note that the Bayes estimates are "shrunk" together. The parameter τ controls the amount of shrinkage. We set $\tau = 1$ but in practice, we should treat τ as another unknown parameter and let the data determine how much shrinkage is needed.

So far we assumed that we know how to draw samples from the conditionals $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. If we don't know how, we can still use the Gibbs sampling algorithm by drawing each observation using a Metropolis-Hastings step. Let q be a proposal distribution for x and let $\widetilde{q}$ be a proposal distribution for y. When we do a Metropolis step for X we treat Y as fixed. Similarly, when we do a Metropolis step for Y we treat X as fixed. Here are the steps:

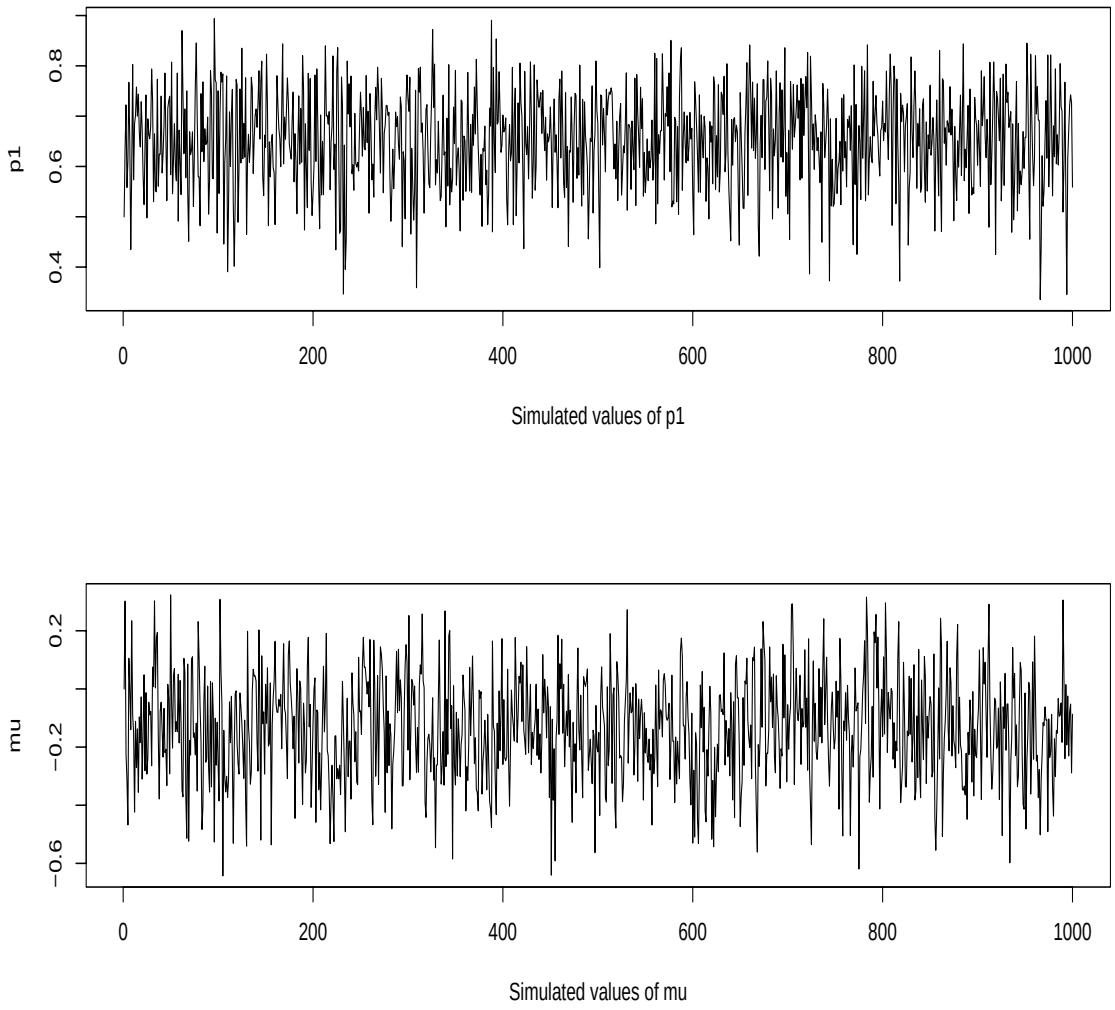


FIGURE 25.5. Simulated values of p_1 and μ.

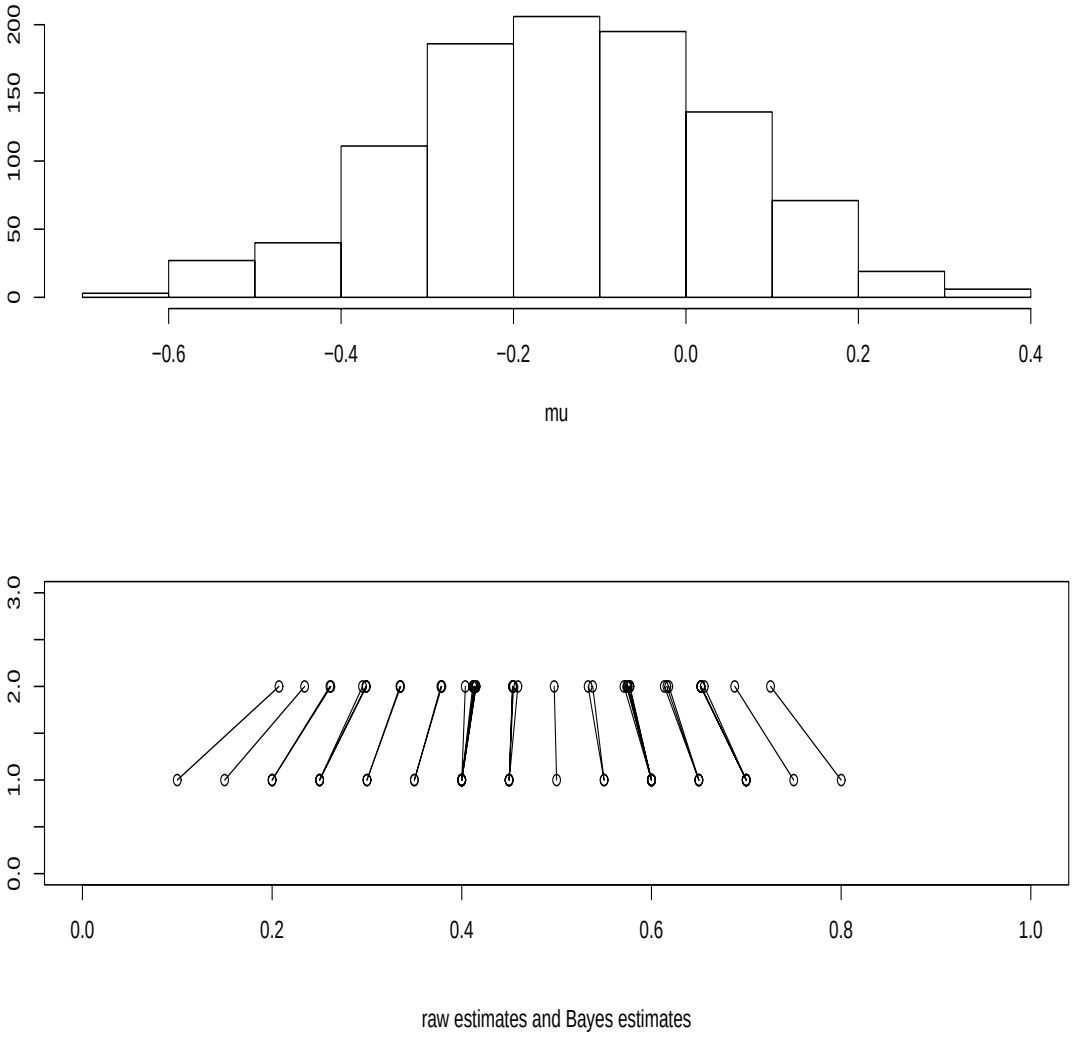


FIGURE 25.6. Top panel: posterior histogram of μ. Lower panel: raw proptions and the Bayes posterior estimates.

Metropolis within Gibbs:

(1a) Draw a proposal $Z \sim q(z|X_n)$.

(1b) Evaluate

$$r = \min\left\{\frac{f(Z,Y_n)}{f(X_n,Y_n)}\frac{q(X_n|Z)}{q(Z|X_n)},\ 1\right\}.$$

(1c) Set

$$X_{n+1} = \begin{cases} Z & \text{with probability } r \\ X_n & \text{with probability } 1-r. \end{cases}$$

(2a) Draw a proposal $Z \sim \widetilde{q}(z|Y_n)$.

(2b) Evaluate

$$r = \min\left\{\frac{f(X_{n+1},Z)}{f(X_{n+1},Y_n)}\frac{\widetilde{q}(Y_n|Z)}{\widetilde{q}(Z|Y_n)},\ 1\right\}.$$

(2c) Set

$$Y_{n+1} = \begin{cases} Z & \text{with probability } r \\ Y_n & \text{with probability } 1-r. \end{cases}$$

Again, this generalizes to more than two dimensions.

Chapter 8

Fundamental Concepts in Inference

Many inferential problems can be identified as being one of three types: estimation, confidence sets or hypothesis testing.

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 f, or a prediction for a future value Y of some random variable.

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

A confidence set C_n is a set that contains a quantity of interest θ with some prescribed probability $1 - \alpha$. For example, we might want to construct an interval $C_n = (a, b)$ that traps an unknown parameter θ 95 per cent of the time.

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.

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

Example 8.1 (Coin Flipping.) *Suppose* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ *denote* n *independent coin flips. A reasonable point estimate of* p *is* $\widehat{p} = n^{-1}\sum_{i=1}^{p} X_i$. *A* $1-\alpha$ *per cent confidence interval for* p *is* $C_n = (\widehat{p}_n - \epsilon_n, \widehat{p}_n + \epsilon_n)$ *where* $\epsilon_n^2 = \{2n\}^{-1}\log(2/\alpha)$. *To test the null hypothesis that the coin is fair, that is,* $p = 1/2$, *we can reject the null hypothesis when* $|\widehat{p}_n - \frac{1}{2}|/\sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}$ *is larger than 2. The justification for these procedures will be the subject of the next few chapters.*

8.1 Point Estimation

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

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

We define

$$\mathsf{bias}\,(\widehat{\theta}_n) = \mathbb{E}_\theta(\widehat{\theta}_n) - \theta \tag{8.1}$$

to be the bias of $\widehat{\theta}_n$. We say that $\widehat{\theta}_n$ is **unbiased** if $\mathbb{E}(\widehat{\theta}_n) = \theta$. Unbiasedness used to receive much attention but these days it is not considered very important; many of the estimators we will use are biased. A point estimator $\widehat{\theta}_n$ of a parameter θ is **consistent** if $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$. Consistency is a reasonable requirement for estimators. 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:

$$\mathsf{se} = \mathsf{se}\,(\widehat{\theta}_n) = \sqrt{\mathbb{V}(\widehat{\theta}_n)}. \tag{8.2}$$

Often, it is not possible to compute the standard error but usually we can estimate the standard error. The estimated standard error is denoted by $\widehat{\mathsf{se}}$.

Example 8.2 *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

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

Recall 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 density for θ.

Theorem **8.3** *The* MSE *can be written as*

$$\mathsf{MSE} = \mathsf{bias}\,(\widehat{\theta}_n)^2 + \mathbb{V}_\theta(\widehat{\theta}_n). \tag{8.3}$$

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

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

Theorem **8.4** *If* bias $\to 0$ *and* se $\to 0$ *as* $n \to \infty$ *then* $\widehat{\theta}_n$ *is consistent, that is,* $\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta$.

PROOF. If bias $\to 0$ and se $\to 0$ then, by Theorem 8.3, MSE $\to 0$. It follows that $\widehat{\theta}_n \xrightarrow{\mathrm{qm}} \theta$. (Recall definition 6.3.) The result follows from part (b) of Theorem 6.4. $\blacksquare$

Example 8.5 *Returning to the coin flipping example, we have that* $\mathbb{E}_p(\widehat{p}_n) = p$ *so that* bias $= p - p = 0$ *and* se $= \sqrt{p(1-p)/n} \to 0$. *Hence,* $\widehat{p}_n \xrightarrow{\mathrm{P}} p$, *that is,* $\widehat{p}_n$ *is a consistent estimator.* $\blacksquare$

Consider n data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ from a regression model

$$Y_i = f(X_i) + \epsilon_i$$

where $\mathbb{E}(\epsilon_i) = 0$ and f is an unknown regresion function. An estimate of f is

$$\widehat{f}_n(x) = \text{average of all } Y_i \text{ such that } |X_i - x| < \delta$$

for some $\delta > 0$. Notice that for every x we have an unknown parameter $f(x)$ and an estimator $\widehat{f}_n(x)$. The **integrated mean squared error** IMSE is defined by

$$\begin{aligned} \mathsf{IMSE} &= \int \mathsf{MSE}(\widehat{f}_n(x))^2 dx \\ &= \int \left(\mathsf{bias}\,(\widehat{f}_n(x))^2 + \mathbb{V}(\widehat{f}_n(x))^2 \right) dx \\ &= \int \mathsf{bias}\,(\widehat{f}_n(x))^2 dx + \int \mathbb{V}(\widehat{f}_n(x))^2 dx \\ &= \text{integrated bias}^2 + \text{integrated variance}. \end{aligned} \tag{8.4}$$

Small values of δ cause small bias but large variance. Large values of δ cause large bias but small variance. This leads to the **bias-variance tradeoff** that we will discuss in detail when we cover nonparametric regression.

In many cases, a point estimate has an asymptoptic Normal distribution, that is,

$$\widehat{\theta}_n \approx N(\theta, \mathsf{se}^2). \tag{8.5}$$

This is a fact we will often make use of.

8.2 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

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta. \tag{8.6}$$

In words, (a, b) traps θ with probability $1 - \alpha$. We call $1 - \alpha$ the **coverage** of the confidence interval. Commonly, people use 95 per cent confidence

intervals which corresponds to choosing $\alpha = 0.05$. Note: C_n **is random and** θ **is fixed!** If θ is a vector then we use a confidence set (such 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 per cent 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 per cent confidence interval for a parameter θ_1. On day 2, you collect new data and construct a 95 per cent confidence interval for an unrelated parameter θ_2. On day 3, you collect new data and construct a 95 per cent 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 per cent your intervals will trap the true parameter value. There is no need to introduce the idea of repeating the same experiment over and over.

Example 8.6 *Every day, newspapers report opinion polls. For example, they might say that "83 per cent of the population favor arming pilots with guns." Usually, you will see a statement like "this poll is accurate to within 4 points 95 per cent of the time." They are saying that* 83 ± 4 *is a 95 per cent 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 everyday for the rest of your life, 95 per cent 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.*

Later, 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 C_n, given the

data, is 95 per cent." However, these Bayesian intervals refer to degree-of-belief probabilities. These Bayesian intervals will not, in general, trap the parameter 95 per cent of the time.

Example 8.7 *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 (5.4) 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 (8.5) holds, that is, $\widehat{\theta}_n \approx N(\theta, \widehat{\mathsf{se}}^2)$. In this case we can construct (approximate) confidence intervals as follows.

Theorem 8.8 (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*

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}}\,,\ \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}\,). \tag{8.7}$$

Then

$$\mathbb{P}_\theta(\theta \in C_n) \to 1 - \alpha. \tag{8.8}$$

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 per cent confidence intervals, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96 \approx 2$ leading to the approximate 95 per cent confidence interval $\widehat{\theta}_n \pm 2\widehat{\mathsf{se}}$. We will discuss the construction of confidence intervals in more generality in the rest of the book.

Example 8.9 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ *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,* $\widehat{p}_n \approx N(p, \widehat{se}^2)$. *Therefore, an approximate* $1-\alpha$ *confidence interval is*

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

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

8.3 Hypothesis Testing

Your colleague has written an algorithm for simulating flips of a fair coin. You are suspicious of your colleague's programming skills. Let $X_1, \ldots, X_n$ denote n simulated coin flips and let $p = \mathbb{P}(X = 1)$. Let H_0 denote the hypothesis that the simulator is correct and let H_1 denote the hypothesis that the simulator is not fair. H_0 is called the **null hypothesis** and H_1 is called the **alternative hypothesis**. We can write the hypotheses as

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

If we get 100 simulated flips and all 100 are 1, we might reasonably take this as evidence against H_0. This is an example of **hypothesis testing**. Later we will see that common practice is to reject H_0 when

$$\frac{|\widehat{p}_n - \frac{1}{2}|}{\sqrt{\frac{1}{4n}}} > 2.$$

We will discuss this and other hypothesis tests in detail in a later chapter.

8.4 Technical Appendix

Our definition of confidence interval requires that $\mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha$ for all $\theta \in \Theta$. An **pointwise asymptotic** confidence interval requires that

$\liminf_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n) \geq 1-\alpha$ for all $\theta \in \Theta$. An **uniform asymptotic** confidence interval requres 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. In general, it might not be a uniform asymptotic confidence interval.