Lecture Notes on *Advanced Statistical Theory*[1]

Ryan Martin
Department of Statistics
North Carolina State University
www4.stat.ncsu.edu/~rmartin

January 3, 2017

[1]These notes were written to supplement the lectures for the Stat 511 course given by the author at the University of Illinois Chicago. The accompanying textbook for the course is Keener's *Theoretical Statistics*, Springer, 2010, and is referred to frequently though out the notes. The author makes no guarantees that these notes are free of typos or other, more serious errors. Feel free to contact the author if you have any questions or comments.

Contents

Chapter 1

Introduction and Preparations

1.1 Introduction

Stat 511 is a first course in advanced statistical theory. This first chapter is intended to set the stage for the material that is the core of the course. In particular, these notes define the notation we shall use throughout, and also set the conceptual and mathematical level we will be working at. Naturally, both the conceptual and mathematical level will be higher than in an intermediate course, such as Stat 411 at UIC.

On the mathematical side, real analysis and, in particular, measure theory, is very important in probability and statistics. Indeed, measure theory is the foundation on which modern probability is built and, by the close connection between probability and statistics, it is natural that measure theory also permeates the statistics literature. Measure theory itself can be very abstract and difficult. I am not an expert in measure theory, and I don't expect you to be an expert either. But, in general, to read and understand research papers in statistical theory, one should at least be familiar with the basic terminology and results of measure theory. My presentation here is meant to introduce you to these basics, so that we have a working measure-theoretic vocabulary moving forward to our main focus in the course. Keener (2010), the course textbook, also takes a similar approach to its measure theory presentation. Besides measure theory, I will also give some brief introduction to group theory and convex sets/functions. The remainder of this first set of notes concerns the transitions from measure theory to probability and from probability to statistics.

On the conceptual side, besides being able to apply theory to particular examples, I hope to communicate *why* such theory was developed; that is, not only do I want you be familiar with results and techniques, but I hope you can understand the motivation behind these developments. Along these lines, in this chapter, I will discuss the basic ingredients of a statistical inference problem, along with some discussion about statistical reasoning, addressing the fundamental question: *how to reason from sample to population?* Surprisingly, there's no fully satisfactory answer to this question.

1.2 Mathematical preliminaries

1.2.1 Measure and integration

Measure theory is the foundation on which modern probability theory is built. All statisticians should, at least, be familiar with the terminology and the key results (e.g., Lebesgue's dominated convergence theorem). The presentation below is based on material in Lehmann and Casella (1998); similar things are presented in Keener (2010).

A *measure* is a generalization of the concept of length, area, volume, etc. More specifically, a measure μ is a non-negative set-function, i.e., μ assigns a non-negative number to subsets A of an abstract set $\mathbb{X}$, and this number is denoted by $\mu(A)$. Similar to lengths, μ is assumed to be *additive*:

$$\mu(A \cup B) = \mu(A) + \mu(B), \quad \text{for each disjoint } A \text{ and } B.$$

This extends by induction to any finite set $A_1, \ldots, A_n$ of disjoint sets. But a stronger assumption is *σ-additivity*:

$$\mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i), \quad \text{for all disjoint } A_1, A_2, \ldots.$$

Note that finite additivity does not imply σ-additivity. All of the (probability) measures we're familiar with are σ-additive. But there are some peculiar measures which are finitely additive but not σ-additive. The classical example of this is the following.

Example 1.1. Take $\mathbb{X} = \{1, 2, \ldots\}$ and define a measure μ as

$$\mu(A) = \begin{cases} 0 & \text{if } A \text{ is finite} \\ 1 & \text{if } A \text{ is co-finite,} \end{cases}$$

where a set A is "co-finite" if it's the complement of a finite set. It is easy to see that μ is additive. Taking a disjoint sequence $A_i = \{i\}$ we find that $\mu(\bigcup_{i=1}^{\infty} A_i) = \mu(\mathbb{X}) = 1$ but $\sum_{i=1}^{\infty} \mu(A_i) = \sum_{i=1}^{\infty} 0 = 0$. Therefore, μ is not σ-additive.

In general, a measure μ cannot be defined for all subsets $A \subseteq \mathbb{X}$. But the class of subsets on which the measure can be defined is, in general, a *σ-algebra*, or *σ-field*.

Definition 1.1. A σ-algebra $\mathcal{A}$ is a collection of subsets of $\mathbb{X}$ such that:

- $\mathbb{X}$ is in $\mathcal{A}$;
- If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$;
- and if $A_1, A_2, \ldots \in \mathcal{A}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

The sets $A \in \mathcal{A}$ are said to be *measurable*. We refer to $(\mathbb{X}, \mathcal{A})$ as a measurable space and, if μ is a measure defined on $(\mathbb{X}, \mathcal{A})$, then $(\mathbb{X}, \mathcal{A}, \mu)$ is a measure space.

A measure μ is *finite* if $\mu(\mathbb{X})$ is a finite number. Probability measures (see Section 1.3.1) are special finite measures where $\mu(\mathbb{X}) = 1$. A measure μ is said to be σ-finite if there exists a sequence of sets $\{A_i\} \subset \mathcal{A}$ such that $\bigcup_{i=1}^{\infty} A_i = \mathbb{X}$ and $\mu(A_i) < \infty$ for each i.

Example 1.2. Let $\mathbb{X}$ be a countable set and $\mathcal{A}$ the class of *all* subsets of $\mathbb{X}$; then clearly $\mathcal{A}$ is a σ-algebra. Define μ according to the rule

$$\mu(A) = \text{number of points in } A, \quad A \in \mathcal{A}.$$

Then μ is a σ-finite measure which is refered to as *counting measure*.

Example 1.3. Let $\mathbb{X}$ be a subset of d-dimensional Euclidean space $\mathbb{R}^d$. Take $\mathcal{A}$ to be the smallest σ-algebra that contains the collection of open rectangles

$$A = \{(x_1, \ldots, x_d) : a_i < x_i < b_i, i = 1, \ldots, d\}, \quad a_i < b_i.$$

Then $\mathcal{A}$ is the Borel σ-algebra on $\mathbb{X}$, which contains all open and closed sets in $\mathbb{X}$; but there are subsets of $\mathbb{X}$ that do not belong to $\mathcal{A}$! Then the (unique) measure μ, defined by

$$\mu(A) = \prod_{i=1}^{d} (b_i - a_i), \quad \text{for rectangles } A \in \mathcal{A}$$

is called *Lebesgue measure*, and it's σ-finite.

Next we consider integration of a real-valued function f with respect to a measure μ on $(\mathbb{X}, \mathcal{A})$. This more general definition of integral satisfies most of the familiar properties from calculus, such as linearity, monotonicity, etc. But the calculus integral is defined only for a class of functions which is generally too small for our applications.

The class of functions of interest are those which are *measurable*. In particular, a real-valued function f is measurable if and only if, for every real number a, the set $\{x : f(x) \leq a\}$ is in $\mathcal{A}$. If A is a measurable set, then the indicator function $I_A(x)$, which equals 1 when $x \in A$ and 0 otherwise, is measurable. More generally, a simple function

$$s(x) = \sum_{k=1}^{K} a_k I_{A_k}(x),$$

is measurable provided that $A_1, \ldots, A_K \in \mathcal{A}$. Continuous f are also usually measurable; indeed, if f is continuous, then $\{x : f(x) > a\}$ is open, so if the σ-algebra contains all open sets, as the Borel σ-algebra does, then f is measurable.

The integral of a non-negative simple function s with respect to μ is defined as

$$\int s \, d\mu = \sum_{k=1}^{K} a_i \mu(A_k). \tag{1.1}$$

Take a non-decreasing sequence of non-negative simple functions $\{s_n\}$ and define

$$f(x) = \lim_{n\to\infty} s_n(x). \tag{1.2}$$

It can be shown that f defined in (1.2) is measurable. Then the integral of f with respect to μ is defined as

$$\int f\, d\mu = \lim_{n\to\infty} \int s_n\, d\mu,$$

the limit of the simple function integrals. It turns out that the left-hand side does not depend on the particular sequence $\{s_n\}$, so it's unique. In fact, an equivalent definition for the integral of a non-negative f is

$$\int f\, d\mu = \sup_{0 \le s \le f,\ \text{simple}} \int s\, d\mu. \tag{1.3}$$

For a general measurable function f which may take negative values, define

$$f^+(x) = \max\{f(x), 0\} \quad \text{and} \quad f^-(x) = -\min\{f(x), 0\}.$$

Both the positive part f^+ and the negative part f^- are non-negative, and $f = f^+ - f^-$. The integral of f with respect to μ is defined as

$$\int f\, d\mu = \int f^+\, d\mu - \int f^-\, d\mu,$$

where the two integrals on the right-hand side are defined through (1.3). In general, a measurable function f is said to be μ-integrable, or just integrable, if $\int f^+\, d\mu$ and $\int f^-\, d\mu$ are both finite.

Example 1.4 (Counting measure)**.** If $\mathbb{X} = \{x_1, x_2, \ldots\}$ and μ is counting measure, then

$$\int f\, d\mu = \sum_{i=1}^{\infty} f(x_i).$$

Example 1.5 (Lebesgue measure)**.** If $\mathbb{X}$ is a Euclidean space and μ is Lebesgue measure, then $\int f\, d\mu$ exists and is equal to the usual Riemann integral of f from calculus whenever the latter exists. But the Lebesgue integral exists for f which are not Riemann integrable.

Next we list some important results from analysis, related to integrals. The first two have to do with interchange of limits[1] and integration, which is often important in statistical problems. The first is relatively weak, but is used in the proof of the second.

[1]Recall the notions of "lim sup" and "lim inf" from analysis. For example, if x_n is a sequence of real numbers, then $\limsup_{n\to\infty} x_n = \inf_n \sup_{k\ge n} x_k$ and, intuitively, this is the largest accumulation point of the sequence; similarly, $\liminf_{n\to\infty} x_n$ is the smallest accumulation point, and if the largest and smallest accumulations points are equal, then the sequence converges and the common accumulation point is the limit. Also, if f_n is a sequence of real-valued functions, then we can define $\limsup f_n$ and $\liminf f_n$ by applying the previous definitions pointwise.

Theorem 1.1 (Fatou's lemma)**.** *Given $\{f_n\}$, non-negative and measurable,*

$$\int \left(\liminf_{n\to\infty} f_n\right) d\mu \le \liminf_{n\to\infty} \int f_n \, d\mu.$$

The opposite inequality holds for lim sup*, provided that $|f_n| \le g$ for integrable g.*

Theorem 1.2 (Dominated convergence)**.** *Given measurable $\{f_n\}$, suppose that*

$$f(x) = \lim_{n\to\infty} f_n(x) \quad \mu\text{-almost everywhere},$$

and $|f_n(x)| \le g(x)$ for all n, for all x, and for some integrable function g. Then f_n and f are integrable, and

$$\int f \, d\mu = \lim_{n\to\infty} \int f_n \, d\mu.$$

Proof. First, by definition of f as the pointwise limit of f_n, we have that $|f_n - f| \le |f_n| + |f| \le 2g$ and $\limsup_n |f_n - f| = 0$. From Exercise 8, we get

$$\left| \int f_n \, d\mu - \int f \, d\mu \right| = \left| \int (f_n - f) \, d\mu \right| \le \int |f_n - f| \, d\mu$$

and, for the upper bound, by the "reverse Fatou's lemma," we have

$$\limsup_n \int |f_n - f| \, d\mu \le \int \limsup_n |f_n - f| \, d\mu = 0.$$

Therefore, $\int f_n \, d\mu \to \int f \, d\mu$, which completes the proof. □

Note, the phrase "μ-almost everywhere" used in the theorem means that the property holds everywhere except on a μ-null set, i.e., a set N with $\mu(N) = 0$. These sets of measure zero are sets which are "small" in a measure-theoretic sense, as opposed to meager first-category sets which are small is a topological sense. Roughly, sets of measure zero can be ignored in integration and certain kinds of limits, but one should always be careful.

The next theorem is useful for bounding integrals of products of two functions. You may be familiar with this name from other courses, such as linear algebra—it turns out actually, that certain collections of integrable functions act very much the same as vectors in a finite-dimensional vector space.

Theorem 1.3 (Cauchy–Schwarz inequality)**.** *If f and g are measurable, then*

$$\left(\int fg \, d\mu \right)^2 \le \int f^2 \, d\mu \cdot \int g^2 \, d\mu.$$

Proof. If either f^2 or g^2 is not integrable, then the inequality is trivial; so assume that both f^2 and g^2 are integrable. Take any λ; then $\int (f + \lambda g)^2 \, d\mu \ge 0$. In particular,

$$\underbrace{\int g^2 \, d\mu}_{a} \cdot \lambda^2 + \underbrace{2 \int fg \, d\mu}_{b} \cdot \lambda + \underbrace{\int g^2 \, d\mu}_{c} \ge 0 \quad \forall\, \lambda.$$

In other words, the quadratic (in λ) can have at most one real root. Using the quadratic formula,

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

it is clear that the only way there can be fewer than two real roots is if $b^2 - 4ac$ is ≤ 0. Using the definitions of a, b, and c we find that

$$4\Big(\int fg\, d\mu\Big)^2 - 4\int f^2\, d\mu \cdot \int g^2\, d\mu \leq 0,$$

and from this the result follows immediately. A different proof, based on Jensen's inequality, is given in Example 1.8. □

The next result defines "double-integrals" and shows that, under certain conditions, the order of integration does not matter. Fudging a little bit on the details, for two measure spaces $(\mathbb{X}, \mathcal{A}, \mu)$ and $(\mathbb{Y}, \mathcal{B}, \nu)$, define the product space

$$(\mathbb{X} \times \mathbb{Y}, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu),$$

where $\mathbb{X} \times \mathbb{Y}$ is usual set of ordered pairs (x, y), $\mathcal{A} \otimes \mathcal{B}$ is the smallest σ-algebra that contains all the sets $A \times B$ for $A \in \mathcal{A}$ and $B \in \mathcal{B}$, and $\mu \times \nu$ is the product measure defined as

$$(\mu \times \nu)(A \times B) = \mu(A)\nu(B).$$

This concept is important for us because independent probability distributions induce a product measure. Fubini's theorem is a powerful result that allows certain integrals over the product to be done one dimension at a time.

Theorem 1.4 (Fubini)**.** *Let $f(x, y)$ be a non-negative measurable function on $\mathbb{X} \times \mathbb{Y}$. Then*

$$\int_{\mathbb{X}} \Big[\int_{\mathbb{Y}} f(x, y)\, d\nu(y)\Big]\, d\mu(x) = \int_{\mathbb{Y}} \Big[\int_{\mathbb{X}} f(x, y)\, d\mu(x)\Big]\, d\nu(y). \tag{1.4}$$

The common value above is the double integral, written $\int_{\mathbb{X}\times\mathbb{Y}} f\, d(\mu \times \nu)$.

Our last result has something to do with constructing new measures from old. It also allows us to generalize the familiar notion of probability densities which, in turn, will make our lives easier when discussing the general statistical inference problem. Suppose f is a non-negative[2] measurable function. Then

$$\nu(A) = \int_A f\, d\mu \tag{1.5}$$

defines a new measure ν on $(\mathbb{X}, \mathcal{A})$. An important property is that $\mu(A) = 0$ implies $\nu(A) = 0$; the terminology is that ν is *absolutely continuous with respect to* μ, or ν is *dominated* by μ, written $\nu \ll \mu$. But it turns out that, if $\nu \ll \mu$, then there exists f such that (1.5) holds. This is the famous Radon–Nikodym theorem.

[2] f can take negative values, but then the measure is a *signed measure*.

Theorem 1.5 (Radon–Nikodym)**.** *Suppose $\nu \ll \mu$. Then there exists a non-negative μ-integrable function f, unique modulo μ-null sets, such that* (1.5) *holds. The function f, often written as $f = d\nu/d\mu$ is the Radon–Nikodym derivative of ν with respect to μ.*

We'll see later that, in statistical problems, the Radon–Nikodym derivative is the familiar density or, perhaps, a likelihood ratio. The Radon–Nikodym theorem also formalizes the idea of change-of-variables in integration. For example, suppose that μ and ν are σ-finite measures defined on $\mathbb{X}$, such that $\nu \ll \mu$, so that there exists a unique Radon–Nikodym derivative $f = d\nu/d\mu$. Then, for a ν-integrable function φ, we have

$$\int \varphi \, d\nu = \int \varphi \, f \, d\mu;$$

symbolically this makes sense: $d\nu = (d\nu/d\mu)\, d\mu$.

1.2.2 Basic group theory

An important mathematical object is that of a *group*, a set of elements together with a certain operation having a particular structure. Our particular interest (Section 1.3.6) is in groups of transformations and how they interact with probability distributions. Here we set some very basic terminology and understanding of groups. A course on abstract algebra would cover these concepts, and much more.

Definition 1.2. A *group* is a set $\mathscr{G}$ together with a binary operation $\cdot$, such that:

- (closure) for each $g_1, g_2 \in \mathscr{G}$, $g_1 \cdot g_2 \in \mathscr{G}$;
- (identity) there exists $e \in \mathscr{G}$ such that $e \cdot g = g$ for all $g \in \mathscr{G}$;
- (inverse) for each $g \in \mathscr{G}$, there exists $g^{-1} \in \mathscr{G}$ such that $g^{-1} \cdot g = e$;
- (associative) for each $g_1, g_2, g_3 \in \mathscr{G}$, $g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3$.

The element e is called the *identity*, and the element g^{-1} is called the *inverse* of g. The group $\mathscr{G}$ is called *abelian*, or *commutative*, if $g_1 \cdot g_2 = g_2 \cdot g_1$ for all $g_1, g_2 \in \mathscr{G}$.

Some basic examples of groups include $(\mathbb{Z}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{R} \setminus \{0\}, \times)$; the latter requires that the origin be removed since 0 has no multiplicative inverse. These three groups are abelian. The general linear group of dimension m, consisting of all $m \times m$ non-singular matrices, is a group under matrix multiplication; this is not an abelian group. Some simple properties of groups are given in Exercise 10.

We are primarily interested in groups of transformations. Let $\mathbb{X}$ be a space (e.g., a sample space) and consider a collection $\mathscr{G}$ of functions g, mapping $\mathbb{X}$ to itself. Consider the operation $\circ$ of function composition. The identity element e is the function $e(x) = x$ for all $x \in \mathbb{X}$. If we require that $(\mathscr{G}, \circ)$ be a group with identity e, then each $g \in \mathscr{G}$ is a one-to-one function. To see this, take any $g \in \mathscr{G}$ and take $x_1, x_2 \in \mathbb{X}$ such that $g(x_1) = g(x_2)$. Left composition by g^{-1} gives $e(x_1) = e(x_2)$ and, consequently, $x_1 = x_2$; therefore, g is one-to-one. Some examples of groups of transformations are:

- For $\mathbb{X} = \mathbb{R}^m$, define the map $g_c(x) = x + c$, a shift of the vector x by a vector c. Then $\mathscr{G} = \{g_c : c \in \mathbb{R}^m\}$ is an abelian group of transformations.
- For $\mathbb{X} = \mathbb{R}^m$, define the map $g_c(x) = cx$, a rescaling of the vector x by a constant c. Then $\mathscr{G} = \{g_c : c > 0\}$ is an abelian group of transformations.
- For $\mathbb{X} = \mathbb{R}^m$, let $g_{a,b}(x) = ax + b1_m$, a combination of the shift and scaling of x. Then $\mathscr{G} = \{g_{a,b} : a > 0, b \in \mathbb{R}\}$ is a group of transformations; not abelian.
- For $\mathbb{X} = \mathbb{R}^m$, let $g_A(x) = Ax$, where $A \in GL(m)$. Then $\mathscr{G} = \{g_A : A \in GL(m)\}$ is a group of transformations; not abelian.
- Let $\mathbb{X} = \{1, 2, \ldots, m\}$ and define $g_\pi(x) = (x_{\pi(1)}, \ldots, x_{\pi(m)})$, where π is a permutation of the indices. Then $\mathscr{G} = \{g_\pi : \text{permutations } \pi\}$ is a group of transformations; not abelian.

In the literature on groups of transformations, it is typical to write gx instead of $g(x)$. For a given group of transformations $\mathscr{G}$ on $\mathbb{X}$, there are some other classes of functions which are of interest. A function α, mapping $\mathbb{X}$ to itself, is called *invariant* (with respect to $\mathscr{G}$) if $\alpha(gx) = \alpha(x)$ for all $x \in \mathbb{X}$ and all $g \in \mathscr{G}$. A function β, mapping $\mathbb{X}$ to itself, is *equivariant* (with respect to $\mathscr{G}$) if $\beta(gx) = g\beta(x)$ for all $x \in \mathbb{X}$ and all $g \in \mathscr{G}$. The idea is that α is not sensitive to changes induced by mapping $x \mapsto gx$ for $g \in \mathscr{G}$, and β doesn't care whether g is applied before or after. Next is a simple but important example.

Example 1.6. Let $\mathbb{X} = \mathbb{R}^m$ and define maps $g_c(x) = x + c1_m$, the location shifts. The function $\beta(x) = \bar{x}1_m$ is equivariant with respect to $\mathscr{G}$, where $\bar{x}$ is the average of the entries of x. The function $\alpha(x) = x - \bar{x}1_m$ is invariant with respect to $\mathscr{G}$.

A slightly different notion of invariance with respect to a group of transformations, in a context relevant to probability and statistics, will be considered in Section 1.3.6.

1.2.3 Convex sets and functions

There is a special property that functions can have which will we will occasionally take advantage of later on. This property is called *convexity*. Throughout this section, unless otherwise stated, take $f(x)$ to be a real-valued function defined over a p-dimensional Euclidean space $\mathbb{X}$. The function f is said to be convex on $\mathbb{X}$ if, for any $x, y \in \mathbb{X}$ and any $\alpha \in [0, 1]$, the following inequality holds:

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

For the case $p = 1$, this property is easy to visualize. Examples of convex (univariate) functions include e^x, $-\log x$, x^r for $r > 1$.

In the case where f is twice differentiable, there is an alternative characterization of convexity. This is something that's covered in most intermediate calculus courses.

Proposition 1.1. *A twice-differentiable function f, defined on p-dimensional space, is convex if and only if*

$$\nabla^2 f(x) = \left(\left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)\right)_{i,j=1,\ldots,p},$$

the matrix of second derivatives, is positive semi-definite for each x.

Convexity is important in optimization problems (maximum likelihood, least squares, etc) as it relates to existence and uniqueness of global optima. For example, if the criterion (loss) function to be minimized is convex and a local minimum exists, then convexity guarantees that it's a global minimum.

"Convex" can be used as an adjective for sets, not just functions. A set C, in a linear space, is convex if, for any points x and y in C, the convex combination $ax + (1-a)y$, for $a \in [0,1]$, is also a point in C. In other words, a convex set C contains line segments connecting all pairs of points in C. Examples of convex sets are interval of numbers, circles in the plane, and balls/ellipses in higher dimensions. There is a connection between convex sets and convex functions: if f is a convex real-valued function, then, for any real t, the set $C_t = \{x : f(x) \leq t\}$ is convex (see Exercise 15). There will be some applications of convex sets in the later chapters.[3]

1.3 Probability

1.3.1 Measure-theoretic formulation

It turns out the mathematical probability is just a special case of the measure theory stuff presented above. Our probabilities are finite measures, our random variables are measurable functions, our expected values are Lebesgue integrals.

Start with an essentially arbitrary measurable space $(\Omega, \mathcal{F})$, and introduce a probability measure P; that is $\mathsf{P}(\Omega) = 1$. Then $(\Omega, \mathcal{F}, \mathsf{P})$ is called a probability space. The idea is that Ω contains all possible outcomes of the random experiment. Consider, for example, the heights example above in Section 1.4.1. Suppose we plan to sample a single UIC student at random from the population of students. Then Ω consists of all students, and exactly one of these students will be the one that's observed. The measure P will encode the underlying sampling scheme. But in this example, it's not the particular student chosen that's of interest: we want to know the student's height, which is a measurement or characteristic of the sampled student. How do we account for this?

A random variable X is nothing but a measurable function from Ω to another space $\mathbb{X}$. It's important to understand that X, as a mapping, is not random; instead, X is a function of a randomly chosen element ω in Ω. So when we are discussing probabilities that X satisfies such and such properties, we're actually thinking about the probability (or measure) the set

[3] e.g., the parameter space for natural exponential families is convex; Anderson's lemma, which is used to prove minimaxity in normal mean problems, among other things, involves convex sets; etc.

of ω's for which $X(\omega)$ satisfies the particular property. To make this more precise we write

$$\mathsf{P}(X \in A) = \mathsf{P}\{\omega : X(\omega) \in A\} = \mathsf{P}X^{-1}(A).$$

To simplify notation, etc, we will often ignore the underlying probability space, and work simply with the probability measure $\mathsf{P}_X(\cdot) = \mathsf{P}X^{-1}(\cdot)$. This is what we're familiar with from basic probability and statistics; the statement $X \sim \mathsf{N}(0,1)$ means simply that the probability measure induced on $\mathbb{R}$ by the mapping X is a standard normal distribution. When there is no possibility of confusion, we will drop the "X" subscript and simply write P for P_X.

When P_X, a measure on the X-space $\mathbb{X}$, is dominated by a σ-finite measure μ, the Radon–Nikodym theorem says there is a density $d\mathsf{P}_X/d\mu = p_X$, and

$$\mathsf{P}_X(A) = \int_A p_X \, d\mu.$$

This is the familiar case we're used to; when μ is counting measure, p_X is a probability mass function and, when μ is Lebesgue measure, p_X is a probability density function. One of the benefits of the measure-theoretic formulation is that we do not have to handle these two important cases separately.

Let φ be a real-valued measurable function defined on $\mathbb{X}$. Then the expected value of $\varphi(X)$ is

$$\mathsf{E}_X\{\varphi(X)\} = \int_{\mathbb{X}} \varphi(x) \, d\mathsf{P}_X(x) = \int_{\mathbb{X}} \varphi(x) p_X(x) \, d\mu(x),$$

the latter expression holding only when $\mathsf{P}_X \ll \mu$ for a σ-finite measure μ on $\mathbb{X}$. The usual properties of expected value (e.g., linearity) hold in this more general case; the same tools we use in measure theory to study properties of integrals of measurable functions are useful for deriving such things.

In these notes, it will be assumed you are familiar with all the basic probability calculations defined and used in basic probability and statistics courses, such as Stat 401 and Stat 411 at UIC. In particular, you are expected to know the common distributions (e.g., normal, binomial, Poisson, gamma, uniform, etc) and how to calculate expectations for these and other distributions. Moreover, I will assume you are familiar with some basic operations involving random vectors (e.g., covariance matrices) and some simple linear algebra stuff. Keener (2010), Sections 1.7 and 1.8, introduces these concepts and notations.

In probability and statistics, product spaces are especially important. The reason, as we eluded to before, is that independence of random variables is connected with product spaces and, in particular, product measures. If $X_1, \ldots, X_n$ are iid P_X, then their joint distribution is the product measure

$$\mathsf{P}_{X_1} \times \mathsf{P}_{X_2} \times \cdots \times \mathsf{P}_{X_n} = \mathsf{P}_X \times \mathsf{P}_X \cdots \times \mathsf{P}_X = \mathsf{P}_X^n.$$

The first term holds with only "independence;" the second requires "identically distributed;" the last term is just a short-hand notation for the middle term.

When we talk about convergence theorems, such as the law of large numbers, we say something like: for an infinite sequence of random variables $X_1, X_2, \ldots$ some event happens

with probability 1. But what is the measure being referenced here? In the iid case, it turns out that it's an *infinite product measure*, written as P_X^∞. We'll have more to say about this when the time comes.

1.3.2 Conditional distributions

Conditional distributions in general are rather abstract. When the random variables in question are discrete (μ = counting measure), however, things are quite simple; the reason is that events where the value of the random variable is fixed have positive probability, so the ordinary conditional probability formula involving ratios can be applied. When one or more of the random variables in question are not discrete, then more care must be taken.

Suppose random variables X and Y have a joint distribution with density function $p_{X,Y}(x, y)$, with respect to some dominating (product) measure $\mu \times \nu$. Then the corresponding marginal distributions have densities with respect to μ and ν, respectively, given by

$$p_X(x) = \int p_{X,Y}(x, y)\, d\nu(y) \quad \text{and} \quad p_Y(y) = \int p_{X,Y}(x, y)\, d\mu(x).$$

Moreover, the conditional distribution of Y, given $X = x$, also has a density with respect to ν, and is given by the ratio

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x, y)}{p_X(x)}.$$

As a function of x, for given y, this is clearly μ-measurable since the joint and marginal densities are measurable. Also, for a given x, $p_{Y|X}(y \mid x)$ defines a probability measure Q_x, called the *conditional distribution of Y, given $X = x$*, through the integral

$$\mathsf{Q}_x(B) = \int_B p_{Y|X}(y \mid x)\, d\nu(y).$$

That is, $p_{Y|X}(y \mid x)$ is the Radon–Nikodym derivative of the conditional distribution Q_x with respect to ν. This is the familiar setup and, from here, we can define conditional probabilities and expectations as usual. That is,

$$\mathsf{P}(Y \in B \mid X = x) = \int_B p_{Y|X}(y \mid x)\, d\nu(y).$$

Here I use the more standard notation for conditional probability. The law of total probability then allows us to write

$$\mathsf{P}(Y \in B) = \int \mathsf{P}(Y \in B \mid X = x) p_X(x)\, d\mu(x),$$

in other words, marginal probabilities for Y may be obtained by taking expectation of the conditional probabilities. More generally, for any ν-integrable function φ, we may write the conditional expectation

$$\mathsf{E}\{\varphi(Y) \mid X = x\} = \int \varphi(y) p_{Y|X}(y|x)\, d\nu(y).$$

We may evaluate the above expectation for any x, so we actually have defined a (μ-measurable) function, say, $g(x) = \mathsf{E}(Y \mid X = x)$; here I took $\varphi(y) = y$ for simplicity. Now, $g(X)$ is a random variable, to be denoted by $\mathsf{E}(Y \mid X)$, and we can ask about its mean, variance, etc. The corresponding version of the law of total probability for conditional expectations is

$$\mathsf{E}(Y) = \mathsf{E}\{\mathsf{E}(Y \mid X)\}. \tag{1.6}$$

This formula is called *smoothing* in Keener (2010) but I would probably call it a law of iterated expectation. This is actually a very powerful result that can simplify lots of calculations; Keener (2010) uses this a lot. There are versions of iterated expectation for higher moments, e.g.,

$$\mathsf{V}(Y) = \mathsf{V}\{\mathsf{E}(Y \mid X)\} + \mathsf{E}\{\mathsf{V}(Y \mid X)\}, \tag{1.7}$$

$$\mathsf{C}(X, Y) = \mathsf{E}\{\mathsf{C}(X, Y \mid Z)\} + \mathsf{C}\{\mathsf{E}(X \mid Z), \mathsf{E}(Y \mid Z)\}, \tag{1.8}$$

where $\mathsf{V}(Y \mid X)$ is the conditional variance, i.e., the variance of Y relative to its conditional distribution and, similarly, $\mathsf{C}(X, Y \mid Z)$ is the conditional covariance of X and Y.

Three somewhat technical remarks:

- We have defined the conditional distribution Q_x through its corresponding conditional density, but this is not always appropriate. There are real cases where the most general definition of conditional distribution (Keener 2010, Sec. 6.2) is required, e.g., in the proof of the Neyman–Fisher factorization theorem and of Bayes's theorem.
- What makes the "familiar" case above relatively simple is the assumption that the joint distribution of (X, Y) is dominated by a product measure $\mu \times \nu$. There are many real situations where this assumption holds, but there are also important cases where it does not. One of the most common types of examples where the joint distribution does not have a density with respect to a product measure is when Y is a function of X, i.e., $Y = g(X)$. To see what happens in such a case, let's try to develop some intuition. Since Y is determined by X, the "joint distribution" of (X, Y)—whatever that means here—is determined by the marginal distribution of X. Therefore, a reasonable guess as to what the "joint density" for (X, Y) looks like is

  $$\text{“}p_{X,Y}(x, y)\text{”} = p_X(x) \cdot I_{A_y}(x), \quad \text{where} \quad A_y = \{x : g(x) = y\}.$$

  The reason for the quotation marks in the above explanation is that I have not been careful about the dominating measure. Following this intuition, the corresponding "conditional density" looks like

  $$\text{“}p_{X|Y}(x \mid y)\text{”} = \frac{p_X(x) I_{A_y}(x)}{p_Y(y)}. \tag{1.9}$$

  It turns out that this intuition can all be made rigorous, but it is not worth going through the details here; the interested reader is encouraged to check out Appendix B.3.3 in Schervish (1995), in particular, Corollary B.55.

- Conditional distributions are not unique: the conditional density can be redefined arbitrarily on a set of ν-measure zero, without affecting the integral that defines Q_x. We will not dwell on this point here, but students should be aware of the subtleties of conditional distributions; the wikipedia page[4] on the *Borel paradox* gives a clear explanation of these difficulties, along with references, e.g., to Jaynes (2003), Chapter 15.

As a final word about conditional distributions, it is worth mentioning that these are particularly useful in the specification of complex models. Indeed, it can be difficult to specify a meaningful joint distribution for a collection of random variables in a given application. However, it is often possible to write down a series of conditional distributions that, together, specify a meaningful joint distribution. That is, we can simplify the modeling step by working with several lower-dimensional conditional distributions. Keener (2010, Sec. 6.3) discusses this in some detail. This is particularly useful for specifying prior distributions for unknown parameters in a Bayesian analysis; see Chapter 4 and Keener (2010, Sec. 15.1).

1.3.3 Jensen's inequality

Convex sets and functions appear quite frequently in statistics and probability applications, so it can help to see the some applications. The first result, relating the expectation of a convex function to the function of the expectation, should be familiar.

Theorem 1.6 (Jensen's inequality)**.** *Suppose φ is a convex function on an open interval $\mathbb{X} \subseteq \mathbb{R}$, and X is a random variable taking values in $\mathbb{X}$. Then*

$$\varphi[\mathsf{E}(X)] \leq \mathsf{E}[\varphi(X)].$$

If φ is strictly convex, then equality holds if and only if X is constant.

Proof. First, take x_0 to be any fixed point in $\mathbb{X}$. Then there exists a linear function $\ell(x) = c(x - x_0) + \varphi(x_0)$, through the point $(x_0, \varphi(x_0))$, such that $\ell(x) \leq \varphi(x)$ for all x. To prove our claim, take $x_0 = \mathsf{E}(X)$, and note that

$$\varphi(X) \geq c[X - \mathsf{E}(X)] + \varphi[\mathsf{E}(X)].$$

Taking expectations on both sides gives the result. □

Jensen's inequality can be used to confirm: $\mathsf{E}(1/X) \geq 1/\mathsf{E}(X)$, $\mathsf{E}(X^2) \geq \mathsf{E}(X)^2$, and $\mathsf{E}[\log X] \leq \log \mathsf{E}(X)$. An interesting consequence is the following.

Example 1.7 (Kullback–Leibler divergence)**.** Let f and g be two probability density functions dominated by a σ-finite measure μ. The Kullback–Leibler divergence of g from f is defined as

$$\mathsf{E}_f\{\log[f(X)/g(X)]\} = \int \log(f/g) f \, d\mu.$$

[4] `https://en.wikipedia.org/wiki/BorelKolmogorov_paradox`

It follows from Jensen's inequality that

$$\begin{aligned}\mathsf{E}_f\{\log[f(X)/g(X)]\} &= -\mathsf{E}_f\{\log[g(X)/f(X)]\} \\ &\geq -\log \mathsf{E}_f[g(X)/f(X)] \\ &= -\log \int (g/f) f \, d\mu = 0.\end{aligned}$$

That is, the Kullback–Leibler divergence is non-negative for all f and g. Moreover, it equals zero if and only if $f = g$ (μ-almost everywhere). Therefore, the Kullback–Leibler divergence acts like a distance measure between to density functions. While it's not a metric in a mathematical sense[5], it has a lot of statistical applications. See Exercise 23.

Example 1.8 (Another proof of Cauchy–Schwartz)**.** Recall that f^2 and g^2 are μ-measurable functions. If $\int g^2 \, d\mu$ is infinite, then there is nothing to prove, so suppose otherwise. Then $p = g^2 / \int g^2 \, d\mu$ is a probability density on $\mathbb{X}$. Moreover,

$$\Big(\frac{\int fg \, d\mu}{\int g^2 \, d\mu}\Big)^2 = \Big(\int (f/g) p \, d\mu\Big)^2 \leq \int (f/g)^2 p \, d\mu = \frac{\int f^2 \, d\mu}{\int g^2 \, d\mu},$$

where the inequality follows from Theorem 1.6. Rearranging terms one gets

$$\Big(\int fg \, d\mu\Big)^2 \leq \int f^2 \, d\mu \cdot \int g^2 \, d\mu,$$

which is the desired result.

Another application of convexity and Jensen's inequality will come up in the decision-theoretic context to be discussed later. In particular, when the loss function is convex, it will follow from Jensen's inequality that randomized decision rules are inadmissible and, hence, can be ignored.

1.3.4 A concentration inequality

We know that sample means of iid random variables, for large sample sizes, will "concentrate" around the population mean. A concentration inequality gives a bound on the probability that the sample mean is outside a neighborhood of the population mean. *Chebyshev's inequality* (Exercise 26) is one example of a concentration inequality and, often, these tools are the key to proving limit theorems and even some finite-sample results in statistics and machine learning.

Here we prove a famous but relatively simple concentration inequality for sums of independent bounded random variables. By "bounded random variables" we mean X_i such that $\mathsf{P}(a_i \leq X_i \leq b_i) = 1$. For one thing, boundedness implies existence of moment generating functions. We start with a simple result for one bounded random variable with mean zero; the proof uses some properties of convex functions. Portions of what follows are based on notes prepared by Prof. Larry Wasserman.[6]

[5] it's not symmetric and does not satisfy the triangle inequality

[6] `http://www.stat.cmu.edu/~larry/=stat705/Lecture2.pdf`

Lemma 1.1. *Let X be a random variable with mean zero, bounded within the interval $[a, b]$. Then the moment generating function $M_X(t) = \mathsf{E}(e^{tX})$ satisfies*

$$M_X(t) \leq e^{t^2(b-a)^2/8}.$$

Proof. Write $X = Wa + (1 - W)b$, where $W = (X - a)/(b - a)$. The function $z \mapsto e^{tz}$ is convex, so we get

$$e^{tX} \leq We^{ta} + (1 - W)e^{tb}.$$

Taking expectation, using the fact that $\mathsf{E}(X) = 0$, gives

$$M_X(t) \leq -\frac{a}{b-a}e^{ta} + \frac{b}{b-a}e^{tb}.$$

The right-hand side can be rewritten as $e^{h(\zeta)}$, where

$$\zeta = t(b - a) > 0, \quad h(z) = -cz + \log(1 - c + ce^z), \quad c = -a/(b - a) \in (0, 1).$$

Obviously, $h(0) = 0$; similarly, $h'(z) = -c + ce^z/(1 - c + ce^z)$, so $h'(0) = 0$. Also,

$$h''(z) = \frac{c(1-c)e^z}{(1-c+ce^z)^2}, \quad h'''(z) = \frac{c(1-c)e^z(1-c-ce^z)}{(1-c+ce^z)^3}.$$

It is easy to verify that $h'''(z) = 0$ iff $z = \log(\frac{1-c}{c})$. Plugging this z value in to h'' gives $1/4$, and this is the global maximum. Therefore, $h''(z) \leq 1/4$ for all $z > 0$. Now, for some $z_0 \in (0, \zeta)$, there is a second-order Taylor approximation of $h(\zeta)$ around 0:

$$h(\zeta) = h(0) + h'(0)\zeta + h''(z_0)\frac{\zeta^2}{2} \leq \frac{\zeta^2}{8} = \frac{t^2(b-a)^2}{8}.$$

Plug this bound in to get $M_X(t) \leq e^{h(\zeta)} \leq e^{t^2(b-a)^2/8}$. □

Lemma 1.2 (Chernoff)**.** *For any random variable X, $\mathsf{P}(X > \varepsilon) \leq \inf_{t>0} e^{-t\varepsilon}\mathsf{E}(e^{tX})$.*

Proof. See Exercise 27. □

Now we are ready for the main result, Hoeffding's inequality. The proof combines the results in the two previous lemmas.

Theorem 1.7 (Hoeffding's inequality)**.** *Let $Y_1, Y_2, \ldots$ be independent random variables, with $\mathsf{P}(a \leq Y_i \leq b) = 1$ and mean μ. Then*

$$\mathsf{P}(|\bar{Y}_n - \mu| > \varepsilon) \leq 2e^{-2n\varepsilon^2/(b-a)^2}.$$

Proof. We can take $\mu = 0$, without loss of generality, by working with $X_i = Y_i - \mu$. Of course, X_i is still bounded, and the length of the bounding interval is still $b - a$. Write

$$\mathsf{P}(|\bar{X}_n| > \varepsilon) = \mathsf{P}(\bar{X}_n > \varepsilon) + \mathsf{P}(-\bar{X}_n > \varepsilon).$$

Start with the first term on the right-hand side. Using Lemma 1.2,

$$\mathsf{P}(\bar{X}_n > \varepsilon) = \mathsf{P}(X_1 + \cdots + X_n > n\varepsilon) \leq \inf_{t>0} e^{-tn\varepsilon} M_X(t)^n,$$

where $M_X(t)$ is the moment generating function of X_1. By Lemma 1.1, we have

$$\mathsf{P}(\bar{X}_n > \varepsilon) \leq \inf_{t>0} e^{-tn\varepsilon} e^{nt^2(b-a)^2/8}.$$

The minimizer, over $t > 0$, of the right-hand side is $t = 4\varepsilon/(b-a)^2$, so we get

$$\mathsf{P}(\bar{X}_n > \varepsilon) \leq e^{-2n\varepsilon^2/(b-a)^2}.$$

To complete the proof, apply the same argument to $\mathsf{P}(-\bar{X}_n > \varepsilon)$, obtain the same bound as above, then sum the two bounds together. □

There are lots of other kinds of concentration inequalities, most are more general than Hoeffding's inequality above. Exercise 29 walks you through a concentration inequality for normal random variables and a corresponding strong law. Modern work on concentration inequalities deals with more advanced kinds of random quantities, e.g., random functions or stochastic processes. The next subsection gives a special case of such a result.

1.3.5 The "fundamental theorem of statistics"

Consider the problem where $X_1, \ldots, X_n$ are iid with common distribution function F on the real line; for simplicity, lets assume throughout that F is everywhere continuous. Of course, if we knew F, then, at least in principle, we know everything about the distribution of the random variables. It should also be clear, at least intuitively, that, if n is large, then we would have seen "all the possible values" of a random variable $X \sim F$, in their relative frequencies, and so it should be possible to learn F from a long enough sequence of data. The result below, called the Glivenko–Cantelli theorem or, by some, the *fundamental theorem of statistics*, demonstrates that our intuition is correct.

First we need a definition. Given $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, we want to construct an estimator $\hat{F}_n$ of F. A natural choice is the "empirical distribution function:"

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i), \quad x \in \mathbb{R},$$

that is, $\hat{F}_n(x)$ is just the proportion of the sample with values not exceeding x. It is a simple consequence from Hoeffding's inequality above (paired with the Borel–Cantelli lemma) that $\hat{F}_n(x)$ converges almost surely to $F(x)$ for each x. The Glivenko–Cantelli theorem says that $\hat{F}_n$ converges to F not just pointwise, but *uniformly*.

Theorem 1.8 (Glivenko–Cantelli)**.** *Given $X_1, \ldots, X_n \overset{iid}{\sim} F$, where F is everywhere continuous on $\mathbb{R}$, let $\hat{F}_n$ be the empirical distribution function as defined above. Set*

$$\|\hat{F}_n - F\|_\infty := \sup_x |\hat{F}_n(x) - F(x)|.$$

Then $\|\hat{F}_n - F\|_\infty$ converges to zero almost surely.

Proof. Our goal is to show that, for any $\varepsilon > 0$,

$$\limsup_n \sup_x |\hat{F}_n(x) - F(x)| \leq \varepsilon, \quad \text{almost surely.}$$

To start, given (arbitrary) $\varepsilon > 0$, let $-\infty = t_1 < t_2 < \cdots < t_J = \infty$ be a partition of $\mathbb{R}$ such that

$$F(t_{j+1}^-) - F(t_j) \leq \varepsilon, \quad j = 1, \ldots, J-1.$$

Exercise 30 demonstrates the existence of such a partition. Then, for any x, there exists j such that $t_j \leq x < t_{j+1}$ and, by monotonicity,

$$\hat{F}_n(t_j) \leq \hat{F}_n(x) \leq \hat{F}_n(t_{j+1}^-) \quad \text{and} \quad F(t_j) \leq F(x) \leq F(t_{j+1}^-).$$

This implies that

$$\hat{F}_n(t_j) - F(t_{j+1}^-) \leq \hat{F}_n(x) - F(x) \leq \hat{F}_n(t_{j+1}^-) - F(t_j).$$

By adding-then-subtracting appropriate quantities on the upper- and lower-bounds, we get

$$\begin{aligned}
\hat{F}_n(x) - F(x) &\geq \hat{F}_n(t_j) - F(t_j) + F(t_j) - F(t_{j+1}^-) \\
\hat{F}_n(x) - F(x) &\leq \hat{F}_n(t_{j+1}^-) - F(t_{j+1}^-) + F(t_{j+1}^-) - F(t_j).
\end{aligned}$$

By the way the partition was defined, we have

$$\hat{F}_n(t_j) - F(t_j) - \varepsilon \leq \hat{F}_n(x) - F(x) \leq \hat{F}_n(t_{j+1}^-) - F(t_{j+1}^-) + \varepsilon.$$

If we apply the law of large numbers for each of the finitely many j, then the upper and lower bounds converge to $\pm\varepsilon/2$, uniformly in x, which completes the proof. □

Even stronger convergence results for the empirical distribution function are known. In particular, the Dvoretzky et al. (1956) show that

$$\mathsf{P}\big(\|\hat{F}_n - F\|_\infty > \varepsilon\big) \leq 2e^{-2n\varepsilon^2},$$

which implies that the rate of convergence is $n^{-1/2}$, i.e., $\|\hat{F}_n - F\|_\infty = O_P(n^{-1/2})$.

What are the implications of this result for statistics, i.e., why is it called the "fundamental theorem of statistics"? It means that any quantity which can be expressed in terms of the distribution function F can be estimated from data. In most cases, the "parameter" of interest (see below) is a function(al) of the distribution function F. For example, the mean of a distribution can be expressed as $\theta = \theta(F) = \int x \, dF(x)$, the median is $\theta = \theta(F) = F^{-1}(0.5)$, etc. The Glivenko–Cantelli theorem says that any $\theta(F)$ can be estimated with $\theta(\hat{F}_n)$ and, moreover, one can expect that these plug-in estimators will have good properties. As we see in Section 1.3.6 below, the distribution will be indexed by a parameter θ of interest, i.e., we write F_θ instead of F and $\theta(F)$. Glivenko–Cantelli insures that one can learn about F_θ from the sample; in order to be able to learn about the parameter of interest, we require that θ be *identifiable*, i.e., that $\theta \mapsto F_\theta$ be a one-to-one function.

It is worth emphasizing that pointwise convergence, $\hat{F}_n(x) \to F(x)$ for each x, is an automatic consequence of the law of large numbers (which, for bounded random variables, is a consequence of Hoeffding). The effort that is required here is to strengthen the conclusion from pointwise to uniform convergence. This turns out to be a general problem—converting pointwise convergence to uniform convergence—and there is considerable, and very technical, work on the subject. A nice introduction is given by van der Vaart (1998, Chap. 19), and more general versions of the Glivenko–Cantelli theorem are given, along with extensions (e.g., "Donsker theorems"), and an introduction to the tools needed to prove such theorems.

1.3.6 Parametric families of distributions

As we will discuss in Section 1.4.1, in a statistical problem, there is not just one probability measure in question, but a whole family of measures P_θ indexed[7] by a parameter $\theta \in \Theta$. You're already familiar with this setup; $X_1, \ldots, X_n$ iid $\mathsf{N}(\theta, 1)$ is one common example. A very important and broad class of distributions is the *exponential family*. That is, for a given dominating measure μ, an exponential family has density function (Radon–Nikodym derivative with respect to μ) of the form

$$p_\theta(x) = e^{\langle \eta(\theta), T(x) \rangle + A(\theta)} h(x),$$

where $\eta(\theta)$, $T(x)$, $A(\theta)$, and $h(x)$ are some functions, and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. You should be familiar with these distributions from a previous course, such as Stat 411 at UIC. We will discuss exponential families in some detail later.

In this section we will consider another special family of probability measures which are characterized by a "base measure" and a group of transformations. We begin with an important special case.

Example 1.9. Let P_0 be a probability measure with symmetric density p_0 with respect to Lebesgue measure on $\mathbb{R}$. Symmetry implies that the median is 0; if the expected value exists, then it equals 0 too. For $X \sim \mathsf{P}_0$, define $X' = X + \theta$ for some real number θ. Then the distribution of X' is $\mathsf{P}_\theta(A) := \mathsf{P}_0(X + \theta \in A)$. Doing this for all θ generates the family $\{\mathsf{P}_\theta : \theta \in \mathbb{R}\}$. The normal family $\mathsf{N}(\theta, 1)$ is a special case.

The family of distributions in Example 1.9 are generated by a single distribution, centered at 0, and a collection of "location shifts." There are four key properties of these location shifts: first, shifting by zero doesn't change anything; second, the result of any two consecutive location shifts can be achieved by a single location shift; third, the order in which location shifts are made is irrelevant; fourth, for any given location, there is a shift that takes the location back to 0. It turns out that these two properties characterize what's called a *group* of transformations, discussed in Section 1.2.2. Keener (2010, Ch. 10) has a few details about group transformation models, and Eaton (1989) is a thorough introduction.

[7] Note that the subscript in P_θ serves a different purpose than the subscript P_X described in Section 1.3.1.

To generalize the location shift example, start with a fixed probability measure P on $(\mathbb{X}, \mathcal{A})$. Now introduce a group $\mathscr{G}$ of transformations on $\mathbb{X}$, and take $\mathsf{P}_e = \mathsf{P}$; here the subscript "$e$" refers to the group identity e. Then define the family $\{\mathsf{P}_g : g \in \mathscr{G}\}$ as

$$\mathsf{P}_g(A) = \mathsf{P}_e(g^{-1}A), \quad A \in \mathcal{A}.$$

That is, $\mathsf{P}_g(A)$ is the probability, under $X \sim \mathsf{P}_e$, that gX lands in A. In the case where P_e has a density p_e with respect to Lebesgue measure, we have

$$p_g(x) = p_e(g^{-1}x) \Big| \frac{dg^{-1}x}{dx} \Big|,$$

which is just the usual change-of-variable formula from introductory probability; of course, the above formula assumes that each $g \in \mathscr{G}$ is differentiable.

The explanation in the previous paragraph concerns the construction of a family of distributions which is, in a certain sense, invariant with respect to $\mathscr{G}$. In many case, like the normal example above, there is already a family $\mathbb{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ of distributions on $\mathbb{X}$, indexed by Θ. If $\mathscr{G}$ is a group of transformations on $\mathbb{X}$, then one could ask if the family is invariant with respect to $\mathscr{G}$. That is, if $X \sim \mathsf{P}_\theta$, is it possible that there is no $\theta' \in \Theta$ such that $gX \sim \mathsf{P}_{\theta'}$? In short, is $\mathscr{G}\mathbb{P} = \mathbb{P}$? There are some distribution families and some groups of transformations for which this holds (Exercise 11).

1.4 Conceptual preliminaries

1.4.1 Ingredients of a statistical inference problem

Statistics in general is concerned with the collection and analysis of data. The data-collection step is an important one, but this will not be considered here—we will assume the data is given, and concern ourselves only with how these data should be analyzed. In our case, the general statistical problem we will face consists of data X, possibly vector-valued, taking values in $\mathbb{X}$, and a model that describes the mechanism that produced this data. For example, if $X = (X_1, \ldots, X_n)$ is a vector consisting of the recorded heights of n UIC students, then the model might say that these individuals were sampled completely at random from the entire population of UIC students, and that heights of students in the population are normally distributed. In short, we would write something like $X_1, \ldots, X_n$ are iid $\mathsf{N}(\mu, \sigma^2)$; here "iid" stands for independent and identically distributed. There would be nothing to analyze if the population in question were completely known. In the heights example, it shall be assumed that at least one of μ and σ^2 are unknown, and we want to use the observed data X to learn something about these unknown quantities. So, in some sense, the population in question is actually just a class/family of distributions—in the heights example this is the collection of all (univariate) normal distributions. More generally, we shall specify a parametric family $\{\mathsf{P}_\theta : \theta \in \Theta\}$, discussed in Section 1.3.6, as the *model* for the observable data X; in other words, $X \sim \mathsf{P}_\theta$ for some $\theta \in \Theta$, though the specific θ that corresponds to the observed $X = x$ is unknown. The statistician's charge then is to learn something about the true θ from the

observations. What it means to "learn something" is not so easy to explain; I will attempt to clarify this in the next section.

The data and model should be familiar ingredients in the statistical inference problem. There is an important but less-familiar piece of the statistical inference problem—the *loss function*—that is not given much attention in introductory inference courses. To facilitate this discussion, consider the problem of trying to estimate θ based on data $X \sim \mathsf{P}_\theta$. The loss function L records how much I lose by guessing that θ equals any particular a in Θ. In other words, $(\theta, a) \mapsto L(\theta, a)$ is just a real-valued function defined on $\Theta \times \Theta$. In introductory courses, one usually takes $L(\theta, a) = (a - \theta)^2$, the so-called squared error loss, without explanation. In this course, we will consider more general loss functions, in more general inference problems, particularly when we discuss *decision theory*.

To summarize, the statistical inference problem consists of data X taking values in a sample space Θ and a family of probability distributions $\{\mathsf{P}_\theta : \theta \in \Theta\}$. In some cases, we will need to consider the loss function $L(\cdot, \cdot)$, and in other cases there will be a known probability distribution Π sitting on the parameter space Θ, representing some prior knowledge about the unknown parameter, which we will need to incorporate somehow. In any case, the goal is to identify the particular P_θ which produced the observed X.

1.4.2 Reasoning from sample to population

It is generally believed that statistics and probability are closely related. While this claim is true in some sense, the connection is not immediate or obvious. Surely, the general sampling model "$X \sim \mathsf{P}_\theta$" is a probabilistic statement. For example, if $X \sim \mathsf{N}(\theta, 1)$ with θ known, then we can compute $\mathsf{P}_\theta(X \leq c) = \Phi(c - \theta)$ for any c, where Φ is the standard normal distribution function. Similar calculations can be made for other distributions depending on a known θ. But this exercise is to make probabilistic statements about a yet-to-be-observed value of a random variable X with parameter θ known. That is, probability is designed to describe uncertainty about a sample to be taken from a fixed and known population. The statistics problem, on the other hand, is one where the sample is given but some characteristic of the population is unknown. This is basically the opposite of the probability problem and, in this light, it seems very hard. Moreover, it is not clear how to use probability, or even if it should be used at all.

A crucial issue is that it is not clear how to interpret probability statements about $X \sim \mathsf{P}_\theta$ after X is observed.[8] An illustration of this idea is in the context of p-values for hypothesis testing. If the p-value is small, then the observed value is an "outlier" with respect to the hypothesized distribution. It is typical to interpret such an outcome as evidence against the hypothesis, but this is a *choice* the statistician is making—there is no basis for handling the problem in this way, mathematical or otherwise. The point here is that the sampling model on its own is insufficient for statistical inference, something more is needed.

[8]Students likely would have encountered this difficulty in their first exposure to Bayes's formula, where a conditional probability is reversed and there is an attempt to use probability to explain uncertainty about the outcome of an experiment that has already been performed but not yet observed.

To further illustrate this point, consider Fisher's *fiducial argument* for statistical inference. Suppose data X and parameter θ are both scalars, and let $F_\theta(x)$ be the distribution function. Take any $p \in [0,1]$ and assume that the equation $p = F_\theta(x)$ can be solved uniquely for x, given θ, and for θ, given x. That is, there exists $x_p(\theta)$ and $\theta_p(x)$ such that

$$p = F_\theta(x_p(\theta)) = F_{\theta_p(x)}(x), \quad \forall\ (x, \theta).$$

If the sampling model is "monotone" in the sense that, for all (p, x, θ),

$$x_p(\theta) \geq x \iff \theta_p(x) \leq \theta,$$

then it is an easy calculation to show that

$$p = \mathsf{P}_\theta\{X \leq x_p(\theta)\} = \mathsf{P}_\theta\{\theta_p(X) \leq \theta\}.$$

Fisher's idea was to take the latter expression and give it an interpretation *after* $X = x$ is observed. That is, he *defined*

$$\text{"}\mathsf{P}\{\theta \geq \theta_p(x)\}\text{"} = p, \quad \forall\ p \in [0,1], \quad x \text{ is observed } X.$$

The collection $\{\theta_p(x) : p \in [0,1]\}$ defines the quantiles of a distribution and, therefore, a distribution itself. Fisher called this the *fiducial distribution* and he made the controversial claim that he had carried out the Bayesian task of getting a sort of "posterior distribution" for the parameter without a prior distribution or by invoking Bayes's theorem; see Zabell (1992) for more on this. Our goal here is not to discuss the validity of Fisher's claims, but simply to point out that Fisher's construction of a fiducial distribution, albeit intuitive, requires a sort of "leap of faith"—in fact, the word *fiducial* actually means "based on belief or faith." Therefore, the fiducial argument is not a mathematical derivation of a solution to the statistical inference problem based on the sampling model alone.

For the most part, we will avoid philosophical concerns in this course, but students should be aware that (i) statistical inference is hard, and (ii) there is no widely agreed upon setup. The issue is that the statistical inference problem is ill-posed, from a mathematical point of view, so one cannot deduce, from first principles, a "correct answer." (For this reason, no one can say that one approach is "right" or better than another approach; the little poem in Figure 1.1 is relevant here.) Fisher thought very carefully about such things and, although his fiducial argument is not fully satisfactory, he was on the right track. The fiducial argument was, at its core, meant to facilitate

> *the conversion of information in the observed data into a meaningful summary of the evidence supporting the truthfulness of various hypotheses related to the parameter of interest.*

This is my *definition of statistical inference*; no textbooks (that I know of) give a formal definition, so they neither agree nor disagree with my definition. Following this idea, there have been attempts to extend/improve upon Fisher's original argument, including generalized

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.
The First approached the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to ball:
"God bless me!" but the Elephant
Is very like a wall!"
The Second, feeling of the tusk,
Cried "Ho! what have we here
So very round and smooth and sharp?
To me 'tis mighty clear
This wonder of an Elephant
Is very like a spear!"

The Third approached the animal,
And happening to take
The squirming trunk within his hands,
Thus boldly up and spake:
"I see," quoth he, "the Elephant
is very like a snake!"
The Fourth reached out an eager hand,
And felt about the knee.
"What most this wondrous beast is like
Is mighty plain," quoth he;
"Tis clear enough the Elephant
is very like a tree!"
The Fifth, who chanced to touch the ear,
Said: "E'en the blindest man
Can tell what this resembles most;
Deny the fact who can
This marvel of an Elephant
is very like a fan!"

The Sixth no sooner had begun
About the beast to grope,
Than, seizing within his scope,
"I see," quoth he, "the Elephant
is very like a rope!"
And so these men of Indostan
Disputed loud and long.
Each in his own opinion
Exceeding stiff and strong
Though each was partly in the right
And all were in the wrong!
Moral: So oft in theologic wars,
The disputants, I ween,
Rail on utter ignorance
Of what each other mean,
And prate about an Elephant
Not one of them has seen!

Figure 1.1: *Three Blind Men and the Elephant*, John Godfrey Saxe, 1880.

fiducial inference (Hannig 2009), structural inference (Fraser 1968), and Dempster–Shafer theory (Dempster 2008; Shafer 1976). An important point that is missing in these existing approaches is a statement of what makes their summaries "meaningful." The new *inferential model* framework (Martin and Liu 2013, 2015b) makes clear what "meaningful" means, but I will not go into this point here; see Martin and Liu (2016). Although these alternative approaches discussed above, which are neither frequentist nor Bayesian, have yet to reach the mainstream, the developments are promising and I am hopeful.

1.5 Exercises

1. Keener, Problem 1.1, page 17.
2. Show that if $A_1, A_2, \ldots$ are members of a σ-algebra $\mathcal{A}$, then so is $\bigcap_{i=1}^{\infty} A_i$.
3. Keener, Problem 1.6, page 17.
4. For $A_1, A_2, \ldots \in \mathcal{A}$, define

$$\limsup A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{x\colon\ x \text{ is in } A_n \text{ for infinitely many } n\}.$$

   Show that $\limsup A_n$ is also in $\mathcal{A}$.
5. Prove the *Borel–Cantelli Lemma*: If μ is a finite measure (i.e., $\mu(\mathbb{X}) < \infty$) and $\sum_{n=1}^{\infty} \mu(A_n) < \infty$, then $\mu(\limsup A_n) = 0$.
6. Keener, Problem 1.8, page 17.

7. Prove that if f and g are measurable functions, then so are $f+g$ and $f \vee g = \max\{f, g\}$. [Hint: For proving $f + g$ is measurable, note that if $f(x) \leq a - g(x)$ then there is a rational number r that sits between $f(x)$ and $a - g(x)$.]

8. Show that if f is μ-integrable, then $|\int f \, d\mu| \leq \int |f| \, d\mu$. [Hint: Write $|f|$ in terms of f^+ and f^-.]

9. (a) Use Fubini's theorem to show that, for a non-negative random variable X, with distribution function F, we have $\mathsf{E}(X) = \int_0^\infty \{1 - F(x)\} \, dx$.

   (b) Use this result to derive the mean of an exponential distribution with scale parameter θ.

10. Let $(\mathscr{G}, \cdot)$ be a group. Show that:

    (a) $g \cdot e = g$ for all $g \in \mathscr{G}$;

    (b) $g \cdot g^{-1} = g$ for all $g \in \mathscr{G}$;

    (c) the identity e is unique;

    (d) for each g, the inverse g^{-1} is unique;

    (e) for each g, $(g^{-1})^{-1} = g$.

11. Let $\mathbb{P} = \{\mathsf{N}(0, \theta) : \theta > 0\}$. Show that $\mathbb{P}$ is invariant with respect to the group $\mathscr{G} = \{g_a(x) = ax : a > 0\}$.

12. Suppose φ is convex on (a, b) and ψ is convex and nondecreasing on the range of φ. Prove that $\psi \circ \varphi$ is convex on (a, b), where $\circ$ denotes function composition.

13. Suppose that $\varphi_1, \ldots, \varphi_n$ are convex functions, and $a_1, \ldots, a_n$ are positive constants. Prove that $\varphi(x) = \sum_{i=1}^n a_i \varphi_i(x)$ is convex.

14. Let $\{C_t : t \in T\}$ be a collection of convex sets. Prove that $\bigcap_{t \in T} C_t$ is also convex.

15. Let f be a convex real-valued function and, for any real t, define $C_t = \{x : f(x) \leq t\}$. Prove that C_t is convex.

16. Keener, Problem 1.26, page 21.

17. Keener, Problems 1.36 and 1.37, page 22.

18. Let $X \sim \mathsf{N}(\mu, \sigma^2)$ and, given $X = x$, $Y \sim \mathsf{N}(x, \tau^2)$. Find the conditional distribution of X, given $Y = y$.

19. Prove the conditional expectation formulas (1.6), (1.7), and (1.8).

20. Let X be a random variable.

    (a) If $\mathsf{E}(X^2) < \infty$, find c to minimize $\mathsf{E}\{(X - c)^2\}$.

(b) If $\mathsf{E}|X| < \infty$, find c to minimize $\mathsf{E}|X - c|$.

21. Keener, Problem 1.17, page 19. [Hint: Consider the *probability generating function* $g(t) = \mathsf{E}(t^X)$ and, in particular, $g(1)$ and $g(-1)$.]

22. *A "reverse" Jensen's inequality.* Let X be a bounded random variable, i.e., $\mathsf{P}(X \in [a, b]) = 1$. If f is an increasing function, then $\mathsf{E}f(X) \le f(\mathsf{E}(X) + d)$, where $d = b - a$.

23. Let f and g be density functions corresponding to $\mathsf{N}(\theta, 1)$ and $\mathsf{N}(\mu, 1)$, respectively. Compute the Kullback–Leibler divergence $K(f, g)$.

24. *Markov's inequality.*

    (a) Let X be a positive random variable with mean $\mathsf{E}(X)$. Show that

    $$\mathsf{P}(X > \varepsilon) \le \varepsilon^{-1}\mathsf{E}(X), \quad \forall\, \varepsilon > 0.$$

    (b) Consider a measure space $(\mathbb{X}, \mathcal{A}, \mu)$, where $\mu(\mathbb{X}) < \infty$, and a μ-integrable function f. State and prove a general measure-theoretic version of Markov's inequality.

25. *A "reverse" Markov inequality.* For a random variable X, taking values in $(0, 1)$, prove that

    $$\mathsf{P}(X > a) \ge \frac{\mathsf{E}(X) - a}{1 - a}, \quad a \in (0, 1).$$

26. *Chebyshev's inequality.* Let X be a random variable with mean μ and variance σ^2. Use Markov's inequality to show that

    $$\mathsf{P}(|X - \mu| > \varepsilon) \le \varepsilon^{-2}\sigma^2, \quad \forall\, \varepsilon > 0.$$

27. Prove Chernoff's bound, Lemma 1.2.

28. (a) Specialize Hoeffding's inequality (Theorem 1.7) to the case where $X_1, \ldots, X_n$ are iid $\mathsf{Ber}(\mu)$ random variables.

    (b) Given η, small, find $\varepsilon = \varepsilon(n, \eta)$ such that $\mathsf{P}(|\bar{X}_n - \mu| \le \varepsilon) \ge 1 - \eta$.

29. (a) Let $Z \sim \mathsf{N}(0, 1)$. Show that $\mathsf{P}(|Z| > \varepsilon) \le \varepsilon^{-1}e^{-\varepsilon^2/2}$.

    (b) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$, and $\bar{X}_n$ the sample mean. Give a bound like the one above for $\mathsf{P}(|\bar{X}_n - \mu| > \varepsilon)$.

    (c) Use your inequality in (b), with the Borel–Cantelli lemma in Exercise 5, to prove the following strong law of large numbers for normals: If $X_1, X_2, \ldots$ are iid $\mathsf{N}(\mu, \sigma^2)$, then $\bar{X}_n \to \mu$ almost surely.

30. Let F be a distribution function on real line, and $\varepsilon > 0$ a fixed number. Let $t_1 = -\infty$ and, for $j > 1$, define

    $$t_{j+1} = \sup\{t : F(t) \le F(t_j) + \varepsilon\}.$$

(a) Show that this sequence is finite and defines the partition used in the proof of Theorem 1.8.

(b) How many points t_j are needed for the partition?

31. Show that if X is distributed according to a scale family, then $Y = \log X$ is distributed according to a location family.

32. Let X be a positive random variable, and consider the family $\mathbb{P}$ of distributions generated by X and transformations $\mathscr{G} = \{g_{b,c} : b > 0, c > 0\}$ given by

$$g_{b,c}(x) = bx^{1/c}.$$

(a) Show $\mathscr{G}$ is a group under function composition.

(b) If X has a unit-rate exponential distribution, then show that the family $\mathbb{P}$ generated by $\{g_{b,c}\}$ is the Weibull family, with density

$$(c/b)(x/b)^{c-1} \exp\{-(x/b)^c\}, \quad x > 0.$$

33. Recall the standard $100(1-\alpha)\%$ confidence interval for a normal mean θ with known variance $\sigma^2 = 1$, i.e., $\bar{X} \pm z_{1-\alpha/2}\sigma n^{-1/2}$, where $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$.

(a) When we say that the coverage probability is $1 - \alpha$, what do we mean?

(b) Explain how this interval is to be interpreted after data is observed.

34. A fundamental concept in frequentist statistical theory is *sampling distributions*. For an observable sample $X_1, \ldots, X_n$ from a distribution depending on some parameter θ, let $T = T(X_1, \ldots, X_n)$ be some statistic.

(a) What do we mean by the sampling distribution of T?

(b) Explain how the sampling distribution is used to reason towards statistical inference. If it helps, you can use an example to explain.

Chapter 2

Exponential Families, Sufficiency, and Information

2.1 Introduction

In statistics, sufficiency and information are fundamental concepts, no matter what approach one adopts—Bayesian, frequentist, or other. The basic idea is that, for a given statistical model $\{\mathsf{P}_\theta : \theta \in \Theta\}$, indexed by a (finite-dimensional) parameter space Θ, there are functions of the observable data $X = (X_1, \ldots, X_n)$ which contain all the available information in X concerning the unknown parameter θ. Such functions are called *sufficient statistics*, and the idea is that it generally suffices for, say, point estimation, to restrict attention to functions of sufficient statistics. The notion of "information" introduced above is meant to be informal; however, for rigor, we must formulate a precise notion of information for a statistical problem. In particular, we shall focus on the Fisher information, obviously due to R. A. Fisher. The key result is that the Fisher information for θ in a $T = T(X)$ function of the observable data X is no more than the Fisher information for θ in X itself, and the two measures of information are equal if and only if T is a sufficient statistic.

The definition of sufficiency is not helpful for finding a sufficient statistic in a given problem. Fortunately, the Neyman–Fisher factorization theorem makes this task quite easy. The idea is that, with some simple algebra on the likelihood function, a sufficient statistic can be readily obtained. Sufficient statistics are not unique, however. Therefore, there is some interest in trying to find the "best" sufficient statistic in a given problem. This best sufficient statistic is called *minimal*, and we discuss some techniques for finding the minimal sufficient statistic. It can happen, however, that even a minimal sufficient statistic T provides an inefficient reduction of the data X—either the dimension of T is larger than that of θ, or there is redundant information in T. In such cases, it makes sense to consider conditioning on an *ancillary* statistic, a sort of complement to a sufficient statistic that contains no information about θ. There are special cases where the minimal sufficient statistic T is *complete*. This means that T contains no redundant information about θ, so conditioning on an ancillary statistic is unnecessary (Basu's theorem).

We begin this chapter with an introduction to exponential families. This is a broad class that contains almost all distributions encountered in an intermediate statistics course like Stat 411 at UIC. Roughly speaking, what makes exponential families so useful is that their corresponding likelihood functions are nice that lots of tricks can be done. For example, the regularity conditions needed for, say, asymptotic normality of the maximum likelihood estimators or the Cramer–Rao inequality, hold for (regular) exponential families. A key result is Theorem 2.1 below, which is a nice application of the Lebesgue dominated convergence theorem. Of particular importance here is that, essentially, only (regular) exponential families admit a suitable dimension reduction via sufficiency (Theorem 2.4). Details about exponential families, including everything presented here, are discussed in the (technical) monograph by Brown (1986).

2.2 Exponential families of distributions

We discussed previously the general concept of a parametric family of probability measures, with some detail to a special case with a structure induced by a group of transformations. In this section we discuss a very important class of distributions that contains many of the common statistical models, such as binomial, Poisson, normal, etc.

Definition 2.1. A collection of probability measures $\{\mathsf{P}_\theta : \theta \in \Theta\}$ on $(\mathbb{X}, \mathcal{A})$, each dominated by a σ-finite measure μ, is called an *exponential family* if the Radon–Nikodym derivatives $p_\theta(x) = (d\mathsf{P}_\theta/d\mu)(x)$ satisfy

$$p_\theta(x) = h(x)e^{\langle \eta(\theta), T(x)\rangle - A(\theta)} \tag{2.1}$$

for some functions h, A, η, and T, where

$$\eta(\theta) = (\eta_1(\theta), \ldots, \eta_d(\theta))^\top \quad \text{and} \quad T(x) = (T_1(x), \ldots, T_d(x))^\top.$$

Here $\langle x, y\rangle$ denotes the usual Euclidean inner product between d-vectors, $\langle x, y\rangle = \sum_{i=1}^d x_i y_i$.

There is a subtle point here concerning the support of these distributions. Formally, the *support* of a distribution is the set $\{x : p_\theta(x) > 0\}$. For some families, the support does not depend on the parameter θ, but for others it does. The normal family $\mathsf{N}(\theta, 1)$ does not have support that depends on θ, but the uniform family $\mathsf{Unif}(0, \theta)$ does. The point I want to make here is that the definition (2.1) implicitly excludes the possibility that the support depends on θ, because it says that x and θ directly interact in the density formula *only* through the inner-product $\langle \eta(\theta), T(x)\rangle$. From here on, when discussing exponential families, the support does not depend on the parameter.

When convenient we may write $a(\theta) = e^{-A(\theta)}$ and write (2.1) as $p_\theta(x) = a(\theta)h(x)e^{\langle \eta(\theta), T(x)\rangle}$. When considering expectations with respect to an exponential family distribution, we might occasionally absorb the $h(x)$ term into $d\mu(x)$; see, e.g., Theorem 2.1.

Example 2.1. Suppose $X \sim \mathsf{Pois}(\theta)$. The Poisson distribution is dominated by counting measure on $\mathbb{X} = \{0, 1, \ldots\}$, with density

$$p_\theta(x) = \frac{e^{-\theta}\theta^x}{x!} = \frac{1}{x!} \cdot e^{x \log \theta - \theta}, \quad x = 0, 1, \ldots$$

The right-hand side is of the form (2.1), so Poisson belongs to the exponential family.

Example 2.2. The following distributions are exponential family members: $\mathsf{N}(\theta, 1)$, $\mathsf{N}(\theta_1, \theta_2^2)$, $\mathsf{Exp}(\theta)$, $\mathsf{Gam}(\theta_1, \theta_2)$, $\mathsf{Beta}(\theta_1, \theta_2)$, $\mathsf{Bin}(n, \theta)$, and $\mathsf{Geo}(\theta)$. Common examples of distributions which are not members of an exponential family include $\mathsf{Cau}(\theta, 1)$, and $\mathsf{Unif}(0, \theta)$.

There are a number of nice statistical properties of exponential families, specifically related to existence of sufficient statistics and, later on, to existence of minimum variance unbiased estimates. The following mathematical properties of exponential families will be useful in proving these statistical results.

Proposition 2.1. *Consider an exponential family with μ-densities*

$$p_\theta(x) = a(\theta)h(x)e^{\langle \theta, T(x)\rangle}. \tag{2.2}$$

The set $\Theta = \left\{\theta : \int h(x)e^{\langle \theta, T(x)\rangle}\, d\mu(x) < \infty\right\}$ is convex.

Proof. Exercise 1. □

In Proposition 2.1, the parameter θ is called the *natural parameter*, and Θ the corresponding natural parameter space. We have stated this in terms of the notation θ, but the basic idea is to start with (2.1) and take $\eta(\theta)$ as the parameter; that is, just do a reparametrization. The result states that the natural parameter space is a "nice" set.

The next theorem is useful for a number of calculations, in particular, for calculating moments in exponential families or Fisher information. The proof is a nice application of the dominated convergence theorem presented earlier.

Theorem 2.1. *Let $X \in \mathbb{X} \subseteq \mathbb{R}^d$ have density $p_\theta(x) = a(\theta)e^{\langle \theta, x\rangle}$ with respect to μ.*[1] *Let $\varphi : \mathbb{X} \to \mathbb{R}$ be μ-measurable and set $\Theta_\varphi = \{\theta : \int |\varphi(x)|e^{\langle \theta, x\rangle}\, d\mu(x) < \infty\}$. Then for θ in the interior of Θ_φ,*

$$m(\theta) := \int \varphi(x)e^{\langle \theta, x\rangle}\, d\mu(x)$$

is continuous and has continuous derivatives of all order. Moreover, the derivative can be taken under the integral sign; i.e., for $i = 1, \ldots, d$,

$$\frac{\partial m(\theta)}{\partial \theta_i} = \int x_i \varphi(x)e^{\langle \theta, x\rangle}\, d\mu(x).$$

Proof. Consider the case of one-dimensional θ; the general d-dimensional case is similar. Choose a fixed $\theta \in \Theta_\varphi$. For a suitable $\varepsilon > 0$, define $d_n(x) = (e^{\varepsilon x/n} - 1)/(\varepsilon/n)$, so that

$$\frac{m(\theta + \varepsilon/n) - m(\theta)}{\varepsilon/n} = \int \varphi(x)e^{\theta x} d_n(x)\, d\mu(x).$$

It is clear that $d_n(x) \to d(x) = x$ for each x, where $d(x)$ is the derivative of $z \mapsto e^{xz}$ at $z = 0$. Write $f_n(x) = \varphi(x)e^{\theta x}d_n(x)$, so that $f_n(x) \to x\varphi(x)e^{\theta x}$ as $n \to \infty$ for all x. It remains to

[1] Here I have re-expressed the dominating measure, i.e., $d\mu(x) \leftarrow h(x)\, d\mu(x)$.

show that the μ-integral of f_n converges to the μ-integral of f. To show this, we employ the dominated convergence theorem.

Note the following inequalities for the exponential function:

$$|e^z - 1| \le |z|e^{|z|} \quad \text{and} \quad |z| \le e^{|z|}, \quad \forall\, z \in \mathbb{R}. \tag{2.3}$$

With these inequalities, we may write

$$|f_n(x)| \le |\varphi(x)|e^{\theta x}\varepsilon^{-1}e^{2\varepsilon|x|} \le |\varphi(x)|e^{\theta x}\varepsilon^{-1}\big(e^{2\varepsilon x} + e^{-2\varepsilon x}\big).$$

If we choose ε such that $\theta \pm 2\varepsilon$ belong to Θ_φ, then the upper bound, say, $g(x)$, is μ-integrable. Therefore, the dominated convergence theorem says

$$\frac{dm(\theta)}{d\theta} = \lim_{n\to\infty} \int f_n(x)\, d\mu(x) = \int f(x)\, d\mu(x) = \int x\varphi(x)e^{\theta x}\, d\mu(x).$$

That is, the derivative of the integral is the integral of the derivative, as was to be shown. To show that one can take more derivatives, and that these further derivatives can be evaluated by taking derivative inside the integral, can be checked by repeating the above argument. □

A couple remarks about the above result are worth making here.

- The proof requires that Θ_φ has a non-empty interior—this ensures that it is possible to find an open interval/box, depending on ε, centered at the given θ that fits inside Θ_φ. Of course, Θ_φ having non-empty interior implies that Θ itself has a non-empty interior. This latter assumption is often included in the list of "regularity conditions" in classical definitions of exponential families; see, e.g., Hogg et al. (2012).
- The strategy of the proof given above extends directly for higher derivatives. However, in applying that strategy, say, for the second derivative, it appears, at first look, that one needs to adjust the function $\varphi(x)$ to $\varphi_1(x) = x\varphi(x)$, which requires that one pick a θ in the interior of Θ_{φ_1}. It turns out, however, that no such change is required. To see this, we need to show that, in the notation of the above proof, for any $\theta \in \Theta_\varphi$,

  $$\int |x|^k|\varphi(x)|e^{\theta x}\, d\mu(x) < \infty, \quad \text{any } k = 1, 2, \ldots,.$$

  The intuition is that polynomials are dominated by exponentials, and, to see this, note that, for any $\delta > 0$,

  $$e^{\delta|x|} = \sum_{s=0}^{\infty} \frac{(\delta|x|)^s}{s!} \implies |x|^k \ll \frac{k!}{\delta^k}e^{\delta|x|}.$$

  Therefore,

  $$|x|^k|\varphi(x)|e^{\theta x} < k!\,\delta^{-k}|\varphi(x)|e^{\theta x}(e^{\delta x} + e^{-\delta x}),$$

  and, for a sufficiently small δ, which does not depend on k, the upper bound is μ-integrable. So, application of the argument above to higher derivatives does not require changing the function φ or the set Θ_φ.

The same result holds for general exponential families, not just for those in the natural or canonical form. The message is simply that for nice enough functions φ, the expected value of $\varphi(X)$ is a very nice function of the parameter θ. As an application of Theorem 2.1 we have the following.

Corollary 2.1. *Suppose that $X = (X_1, \ldots, X_d)$ has an exponential family density of the form $p_\theta(x) = a(\theta)e^{\langle \theta, x\rangle}$. Then for $i, j = 1, \ldots, d$,*

$$\mathsf{E}_\theta(X_i) = -\frac{\partial}{\partial \theta_i} \log a(\theta) \quad \textit{and} \quad \mathsf{C}_\theta(X_i, X_j) = -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log a(\theta).$$

Higher-order moments of X can be found similarly.

Proof. Start with the identity $\int a(\theta)e^{\langle x,\theta\rangle}\, d\mu(x) = 1$ for all θ. Now differentiate both sides with respect to θ as many times as needed, move derivative under the integral sign, and solve for the appropriate moment. □

This can be written in perhaps a more familiar form by recognizing that $-\log a(\theta) = A(\theta)$. In this case, for example, $\mathsf{E}_\theta(X_i) = (\partial/\partial\theta_i)A(\theta)$. The general formulae for means and variances in the exponential family are given in Exercise 3. An alternative derivation of the above result can be found by noticing that exponential families admit a moment-generating function, and it's given by

$$M_\theta(u) = e^{A(\theta+u)-A(\theta)} = a(\theta)/a(\theta+u). \tag{2.4}$$

2.3 Sufficient statistics

2.3.1 Definition and the factorization theorem

A statistic is simply a measurable function of the data; that is, if $T : \mathbb{X} \to \mathbb{T}$ is measurable, then $T(X)$ is a *statistic*. But not all statistics will be useful for the statistical inference problem. It is the goal of this section to begin understanding what kind of mappings T are worth using. The definition involves general conditional distributions.

Definition 2.2. Suppose $X \sim \mathsf{P}_\theta$. Then the statistic $T = T(X)$, mapping $(\mathbb{X}, \mathcal{A})$ to $(\mathbb{T}, \mathcal{B})$, is sufficient for $\{\mathsf{P}_\theta : \theta \in \Theta\}$ if the conditional distribution of X, given $T = t$, is independent of θ. More precisely, suppose there exists a map $K : \mathcal{A} \times \mathbb{T} \to [0, 1]$, independent of θ, such that $K(\cdot, t)$ is a probability measure on $(\mathbb{X}, \mathcal{A})$ for each t and $K(A, \cdot)$ is a measurable function for each $A \in \mathcal{A}$, with

$$\mathsf{P}_\theta(X \in A, T \in B) = \int_B K(A, t)\, d\mathsf{P}_\theta^T(t), \quad \forall\, A \in \mathcal{A}, B \in \mathcal{B}.$$

Here, P_θ^T stands for the marginal distribution of T. Then T is sufficient for θ.

The key to the above definition is that the conditional probability of X, given $T = t$, characterized by the kernel $K(\cdot, t)$, does not depend on θ. So, for example, if φ is some integrable function, then

$$\mathsf{E}_\theta\{\varphi(X) \mid T = t\} = \int \varphi(x) dK(x, t)$$

actually does not depend on θ. Taking this argument one step further, one finds that sufficiency implies that knowing the value of T is sufficient to generate new data X', say, with the same probabilistic properties as X.

The measure-theoretic difficulties that arise with the conditional distributions in the continuous case makes identifying sufficient statistics via the definition difficult in these cases; there's a slicker way to do it, which we discuss shortly. However, for discrete problems, where conditioning is very straightforward, the definition is just fine.

Example 2.3. Suppose $X_1, \ldots, X_n$ are iid $\mathsf{Ber}(\theta)$. Let $T(x) = \sum_{i=1}^n x_i$. Then

$$\mathsf{P}_\theta(X = x \mid T(X) = t) = \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \binom{n}{t}^{-1}.$$

This is independent of θ, so $T(X)$ is sufficient for θ. Here $K(\cdot, t)$ is just a uniform distribution over all those n-tuples of 0's and 1's that consist of exactly t 1's.

Example 2.4. Suppose $X_1, \ldots, X_n$ are iid $\mathsf{Pois}(\theta)$. Take $T(x) = \sum_{i=1}^n x_i$. Then

$$\begin{aligned}\mathsf{P}_\theta(X_1 = x_1 \mid T(X) = t) &= \frac{e^{-\theta}\theta^{x_1}/x_1! \cdot e^{-(n-1)\theta}[(n-1)\theta]^{t-x_1}/(t-x_1)!}{e^{-n\theta}(n\theta)^t/t!} \\ &= \binom{t}{x_1} (1/n)^{x_1} (1-1/n)^{t-x_1}.\end{aligned}$$

That is, the conditional distribution of X_1, given $T = t$, is $\mathsf{Bin}(t, 1/n)$. This holds for all X_i's, not just X_1, and, in fact, the conditional distribution of the X vector, given $T = t$ is is multinomial with size t and weights $1/n$. This distribution is independent of θ, hence $T(X)$ is sufficient for θ.

Example 2.5. Let $X_1, \ldots, X_n$ be an iid sample from a distribution P_θ. Define the order statistics $(X_{(1)}, \ldots, X_{(n)})$, the ordered list of data points. Given the order, there are $n!$ possible values of X, and they all have the same probability. Since this is independent of P_θ, the order statistics must be sufficient for θ. Note, however, that sufficiency may fail without the iid assumption.

In the case where the family $\{\mathsf{P}_\theta : \theta \in \Theta\}$ consists of distributions dominated by a common σ-finite measure μ, there is a very convenient tool to identify a sufficient statistic.

Theorem 2.2 (Neyman–Fisher Factorization Theorem)**.** *Let $\{\mathsf{P}_\theta : \theta \in \Theta\}$ be dominated by a common σ-finite measure μ, with densities $p_\theta = d\mathsf{P}_\theta/d\mu$. Then $T(X)$ is sufficient for θ if and only if there exists non-negative functions h and g_θ such that*

$$p_\theta(x) = g_\theta[T(x)]h(x) \quad \textit{all } \theta, \ \mu\textit{-almost all } x. \tag{2.5}$$

Proof. For a detailed proof, paying close attention to the measure-theoretic concerns about conditions, see Keener (2010), Sec. 6.4. To provide some intuitive understanding of why the factorization (2.5) involving $T(x)$ is related to sufficiency of T, I will provide a rough sketch of the "if" part of the proof. Recall that $T = T(X)$ is a function of X, so the joint distribution of (X, T) is not dominated by a product measure, so some care is needed in defining joint and conditional densities. In particular, according to (1.9), the conditional density of X given $T = t$ is

$$\text{“}p_\theta(x \mid t)\text{”} = \frac{p_\theta(x) I_{T(x)=t}}{m_\theta(t)},$$

where $m_\theta(t)$ is the marginal density of T. (Here, as in Chapter 1, the quotation marks are meant to warn the reader that I am not being careful here about dominating measures.) We assume that (2.5) holds, so we have, first, that

$$m_\theta(t) = \int g_\theta[T(x)] h(x) I_{T(x)=t} \text{“}d\mu(x)\text{”} = g_\theta(t) k(t),$$

where $k(t)$ does not depend on θ, and, second, that

$$\text{“}p_\theta(x \mid t)\text{”} = \frac{g_\theta[T(x)] h(x) I_{T(x)=t}}{g_\theta(t) k(t)}.$$

Since the part involving $g_\theta(t)$ cancels in the top and bottom, we conclude that the conditional distribution of X, given $T = t$, does not depend on θ, hence T is sufficient. □

This theorem allows us to easily identify sufficient statistics, simply by algebraic manipulations of the joint density/likelihood function.

Example 2.6. Suppose that $X = (X_1, \ldots, X_n)$ consists of iid $\mathsf{Unif}(0, \theta)$ samples. Then the joint distribution can be written as

$$p_\theta(x) = \prod_{i=1}^n \theta^{-1} I_{(0,\theta)}(x_i) = \theta^{-n} I_{(0,\theta)}(\max x_i).$$

Since $p_\theta(x)$ depends on θ only through $T(x) = \max x_i$, it follows from Theorem 2.2 that $T(X) = \max X_i$ is sufficient for θ.

Example 2.7. Suppose $X = (X_1, \ldots, X_n)$ consists of iid $\mathsf{N}(\mu, \sigma^2)$ samples. The joint density is

$$\frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{n}{2\sigma^2} \mu^2 \Big\}.$$

Therefore, by Theorem 2.2, $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient for (μ, σ^2). Equivalently, $T'(X) = (\bar{X}, s^2(X))$ is also sufficient.

2.3.2 Minimal sufficient statistics

It should be clear that sufficient statistics are not unique; in fact, in Example 2.7, two sufficient statistics were identified, and the order statistics are sufficient too as usual. More generally, if T is sufficient, then so is $\psi(T)$ for any one-to-one mapping ψ. That being said, it's desirable to find the sufficient statistic which is "smallest" in some sense. This *minimal sufficient* statistic $T = T_{\min}$ is so that, for any other sufficient statistic U, there exists a mapping h such that $T = h(U)$. A powerful technique for finding minimal sufficient statistics is described by the following theorem.

Theorem 2.3. *Suppose, for each $\theta \in \Theta$, P_θ has a density $p_\theta(x) = g_\theta[T(x)]h(x)$ wrt μ. If $p_\theta(x) = cp_\theta(y)$, for some $c = c(x, y)$, implies $T(x) = T(y)$, then T is minimal sufficient.*

Proof. See Keener (2010), page 47. □

Example 2.8. Suppose $X = (X_1, \ldots, X_n)$ is an iid sample with common density

$$p_\theta(x) = h(x)e^{\sum_{j=1}^d \eta_j(\theta)T_j(x) - A(\theta)}. \tag{2.6}$$

Then $T(X) = [T_1(X), \ldots, T_d(X)]$, with $T_j(X) = \sum_{i=1}^n T_j(X_i)$, is sufficient. To see that T is minimal sufficient (under some condition), we shall apply Theorem 2.3. Take x and y such that $p_\theta(x) = cp_\theta(y)$ for some $c = c(x, y)$. Then this implies $\langle \eta(\theta), T(x)\rangle = \langle \eta(\theta), T(y)\rangle + c'$ for some $c' = c'(x, y)$. Take two points θ_0 and θ_1 in Θ and, by subtraction, we get

$$\langle \eta(\theta_0) - \eta(\theta_1), T(x)\rangle = \langle \eta(\theta_0) - \eta(\theta_1), T(y)\rangle$$

which implies

$$\langle \eta(\theta_0) - \eta(\theta_1), T(x) - T(y)\rangle = 0,$$

i.e., $T(x) - T(y)$ and $\eta(\theta_0) - \eta(\theta_1)$ are orthogonal. Since θ_0 and θ_1 are arbitrary, this implies $T(x) - T(y)$ must be orthogonal to the linear space spanned by $\mathcal{S} = \{\eta(\theta_0) - \eta(\theta_1) : \theta_0, \theta_1 \in \Theta\}$. If $\mathcal{S}$ spans the whole $\mathbb{R}^d$ (see explanation below), then this implies $T(x) = T(y)$ and, hence, T is minimal sufficient by Theorem 2.3.

The condition in the previous example—that the space $\mathcal{S}$ spans the entire space—is enough to prove that the natural sufficient statistic, $T(X)$, in the exponential family is minimal sufficient. But, this condition alone is not entirely satisfactory. First, it is not an obvious thing to check and, second, desirable properties, such as asymptotic normality of maximum likelihood estimators, require even more regularity. For this reason, we often impose a stronger condition, one that implies, in particular, that the space $\mathcal{S}$ above spans the entire space. To make this formal, first say that an exponential family of the form (2.6) is *full rank* if $\eta(\Theta)$ has non-empty interior and $[T_1(x), \ldots, T_d(x)]$ do not satisfy a linear constraint for μ-almost all x. You will recognize these as further regularity conditions classically imposed on exponential families.[2] If $\eta(\Theta)$ has non-empty interior, then so does

[2] See, for example, Chapter 3 of my Stat 411 notes.

$\{\eta(\theta_0) - \eta(\theta_1) : \theta_0, \theta_1 \in \Theta\}$, which implies that the span $\mathcal{S}$ in the previous example fills the whole space. In fact, if Θ contains an open set, and η is a continuous one-to-one map, then $\eta(\Theta)$ also contains an open set.

It is a beneficial exercise to consider how a collection of vectors containing an open set would imply that its span fills the entire space. For this, consider an open set in two dimensions. Take a vector $v = (v_1, v_2)$ in this open set. That the set is open means that there exists a sufficiently small $\varepsilon > 0$ such that $\tilde{v} = (v_1 + \varepsilon, v_2 + \varepsilon)$ also resides in the set. The claim is that, as long as $v_1 \neq v_2$, the pair of vectors $(v, \tilde{v})$ are linearly independent. From linear algebra, there is a test for linear independence based on the determinant of the matrix obtained by stacking the set of vectors in question. In this case,

$$\det \begin{pmatrix} v_1 & v_1 + \varepsilon \\ v_2 & v_2 + \varepsilon \end{pmatrix} = v_1 v_2 + v_1 \varepsilon - v_1 v_2 - v_2 \varepsilon = \varepsilon(v_1 - v_2).$$

Of course, if $v_1 \neq v_2$, then this determinant cannot be zero, hence $(v, \tilde{v})$ are linearly independent. Finally, a pair of linearly independent vectors in two dimensions is a basis, and hence its span fills the space.

Most exponential families we know are full rank: e.g., normal, binomial, Poisson, gamma, etc. The classical example of a non-full rank exponential family is $\mathsf{N}(\theta, \theta^2)$, which is one of those so-called *curved* exponential families (Keener 2010, Chap. 5). The name "curved" comes from the fact that the natural parameter space is a curve or, more generally, a set whose effective dimension is smaller than the actual dimension. In this case, the natural parameter $\eta(\theta)$ is given by

$$\eta_1(\theta) = 1/\theta \quad \text{and} \quad \eta_2(\theta) = -1/2\theta^2.$$

Since $\eta_2 = -\eta_1^2/2$, it is clear that the natural parameter space $\eta(\Theta)$ looks like an upside-down parabola. Since the one-dimensional subset of the two-dimensional space cannot contain an open set, we conclude that this curved exponential family cannot be full rank. However, the natural sufficient statistic $T = (T_1, T_2) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is still minimal sufficient. To see this, we need to check that the set $\mathcal{S}$ defined in Example 2.8 spans $\mathbb{R}^2$. Take two pairs of points (x_1, y_1) and (x_2, y_2) and consider the two vectors of differences:

$$v_j = \left(x_j - y_j, \tfrac{1}{2}(y_j^2 - x_j^2)\right)^\top, \quad j = 1, 2.$$

Stick the two vectors into a matrix, i.e.,

$$\begin{pmatrix} x_1 - y_1 & x_2 - y_2 \\ \frac{1}{2}(y_1^2 - x_1^2) & \frac{1}{2}(y_2^2 - x_2^2) \end{pmatrix}.$$

If, for example, we take $x_1 = 1$ and $y_1 = -1$, then the determinant of the matrix equals $y_2^2 - x_2^2$. So, in the case $x_1 = 1$ and $y_1 = -1$, as long as $x_2 \neq \pm y_2$, the determinant is non-zero, the vectors are linearly independent, and the span fills the space. Therefore, the natural sufficient statistic T above is minimal sufficient, even though the exponential family is not full rank.

The reduction of dimension via sufficiency greatly simplifies things. Therefore, it is interesting to ask in what problem is such a substantial reduction of dimension possible. It turns out that it is essentially only the exponential family case. As a last bit of terminology, say that a distribution family admits a *continuous k-dimensional sufficient statistic* U if the factorization (2.5) holds for all x (not almost all x) and if $U(x) = [U_1(x), \ldots, U_k(x)]$ is continuous in x. The following theorem is in Lehmann and Casella (1998, Chap. 1.6)

Theorem 2.4 (Characterization of Sufficiency)**.** *Suppose $X_1, \ldots, X_n$ are real-valued and iid from a distribution with continuous density $f_\theta(x)$ with respect to Lebesgue measure, supported on an interval $\mathbb{X}$ that does not depend on θ. Denote the joint density by*

$$p_\theta(x) = f_\theta(x_1) \cdots f_\theta(x_n),$$

and assume there is a continuous k-dimensional sufficient statistic. Then

- *if $k = 1$, then* (2.6) *holds for some h, η_1 and A.*
- *if $k > 1$ and if $f_\theta(x_i)$ have continuous partial derivatives with respect to x_i, then* (2.6) *holds for some $d \leq k$.*

This theorem says that, among those smooth absolutely continuous families with fixed support, essentially the only ones that admit a continuous sufficient statistic are the exponential families. Note that the theorem says nothing about those irregular problems where the support depends on θ; indeed, the family $\mathsf{Unif}(0, \theta)$ admits a one-dimensional sufficient statistic for all n.

2.3.3 Ancillary and complete statistics

There are sufficient statistics of varying dimensions; e.g., if $X_1, \ldots, X_n$ are iid $\mathsf{N}(\theta, 1)$, then the order statistics and the mean $\bar{X}$ are both sufficient. The ability of a sufficient statistic to admit a signficant reduction seems related to the amount of ancillary information it contains. A statistic $U(X)$ is said to be ancillary if its distribution is independent of θ. Ancillary statistics themselves contain no information about θ, but even minimal sufficient statistics can contain ancillary information. For example, in Exercise 8, the minimal sufficient statistic is not complete.

Just because ancillary statistics contain no information about θ doesn't mean they're not useful. A common, but not universally accepted/used, suggestion is to perform analysis conditional on the values of ancillary statistics. Conditioning on something that carries no information about θ causes no logical difficulties, and Fisher argued that conditioning on ancillary statistics gives a clever way to give the resulting inference more meaning for the problem at hand. Intuitively, it restricts the sample space to a "relevant subset"—the set where $U(X) = u_{\text{obs}}$—getting closer to inference conditioned on the observed X. This idea is often used when, e.g., a maximum likelihood estimate, is not minimal sufficient. We will discuss this more in Section 2.5.

A statistic T is *complete* if $\mathsf{E}_\theta\{f(T)\} = 0$ for all θ implies $f = 0$ almost everywhere. In other words, there are no non-constant functions of T which are ancillary. Alternatively,

a complete sufficient statistic is one that contains exactly all the information about θ in X; that is, it contains no redundant information about θ since every feature $f(T)$ of T has information about θ. To see how this relates to the formal definition, note that no non-zero functions of T are ancillary. Complete sufficient statistics are especially effective at reducing the data; in fact, complete sufficient statistics are minimal.

Theorem 2.5. *If T is complete and sufficient, then T is also minimal sufficient.*

Proof. Let T' be a minimal sufficient statistic. By minimality, we have $T' = f(T)$ for some function f. Write $g(T') = \mathsf{E}_\theta(T \mid T')$, which does not depend on θ by sufficiency of T'. Moreover, by iterated expectation, $\mathsf{E}_\theta g(T') = \mathsf{E}_\theta(T)$. Therefore, $\mathsf{E}_\theta\{T - g(T')\} = 0$ for all θ and, since $T - g(T') = T - g(f(T))$ is a function of T, completeness implies $T = g(T')$ almost everywhere. Since $T = g(T')$ and $T' = f(T)$ we see that T and T' are equivalent up to one-to-one transformations; hence, T is also minimal sufficient (Exercise 7). □

Because of the strength of a complete sufficient statistic, it is helpful to be able to identify cases when one exists. Not surprisingly, exponential families admit a complete sufficient statistic.

Theorem 2.6. *If X is distributed as a full rank d-dimensional exponential family with density* (2.6)*, then $[T_1(X), \ldots, T_d(X)]$ is complete.*

Proof. This is just a sketch in a simple case; for the detailed proof in the general case, see Brown (1986, Theorem 2.12). Take the one-dimensional case with $p_\theta(x) = e^{\theta x - A(\theta)}$ and dominating measure μ. Then $T(x) = x$. Let $f(x)$ be some integrable function with $\mathsf{E}_\theta\{f(X)\} = 0$ for all θ. Writing out the integral form of expectation gives

$$\int f(x) e^{\theta x} \, d\mu(x) = 0 \quad \forall\, \theta.$$

The integral is essentially the Laplace transform of f. The Laplace transform of the zero function is constant equal to zero and, since Laplace transforms are (μ-a.e.) unique, it follows that f must equal the zero function (μ-a.e.). Therefore, X is complete. □

Example 2.9. Theorem 2.6 shows that $T(X) = \sum_{i=1}^n X_i$ is complete when $X_1, \ldots, X_n$ is an iid sample from $\mathsf{Ber}(\theta)$, $\mathsf{Pois}(\theta)$, and $\mathsf{N}(\theta, 1)$.

Example 2.10. Let $X_1, \ldots, X_n$ be an iid sample from $\mathsf{N}(\theta, \theta^2)$. It was shown above that $T = (T_1, T_2) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is minimal sufficient. However, it is not complete. To see this, consider the function $f(t_1, t_2) = t_1^2 - \frac{n+1}{2} t_2$. Then

$$\begin{aligned} \mathsf{E}_\theta\{f(T_1, T_2)\} &= \mathsf{E}_\theta(T_1^2) - \tfrac{n+1}{2} \mathsf{E}_\theta(T_2) \\ &= n\theta^2 + (n\theta)^2 - \tfrac{n+1}{2} \cdot 2n\theta^2 \\ &= 0 \quad \forall\, \theta. \end{aligned}$$

Since this not-exactly-zero function of $T = (T_1, T_2)$ has mean equal zero, the statistic T is not complete. This is not a contradiction to Theorem 2.6 because this curved exponential

family is not full rank, i.e., the natural parameter space is a one-dimensional curve in the two-dimensional plane and, therefore, does not contain an open set. An example similar to this one is considered in Exercise 8.

Example 2.11. Let $X_1, \ldots, X_n$ be iid $\mathsf{Unif}(0, \theta)$. The claim is that $T(X) = X_{(n)}$ is complete. A straightforward calculation shows that the density of T is

$$p_\theta(t) = nt^{n-1}/\theta^n, \quad 0 < t < n.$$

Suppose that $\mathsf{E}_\theta\{f(T)\} = 0$ for all θ. Then we have

$$\int_0^\theta t^{n-1} f^+(t)\, dt = \int_0^\theta t^{n-1} f^-(t)\, dt \quad \forall\, \theta > 0.$$

Since this holds for integration ranges $[0, \theta]$ for all θ, it must hold for all intervals $[a, b]$. The set of all intervals generates the Borel σ-algebra, so, in fact,

$$\int_A t^{n-1} f^+(t)\, dt = \int_A t^{n-1} f^-(t)\, dt \quad \text{for all Borel sets } A.$$

Therefore, f must be zero almost everywhere, so T is complete.

According to Basu's theorem, it's pointless to condition on ancillaries (as described briefly at the beginning of this section) in cases where the sufficient statistic is complete.

Theorem 2.7 (Basu)**.** *If T is a complete sufficient statistic for $\{\mathsf{P}_\theta : \theta \in \Theta\}$, then any ancillary statistic U is independent of T.*

Proof. Since U is ancillary, the probability $p_A = \mathsf{P}_\theta(U \in A)$ does not depend on θ for any A. Define the conditional distribution $\pi_A(t) = \mathsf{P}_\theta(U \in A \mid T = t)$; by iterated expectation, $\mathsf{E}_\theta\{\pi_A(T)\} = p_A$ for all A and all θ. Therefore, by completeness, $\pi_A(t) = p_A$ for almost all t. Since the conditional distribution $\pi_A(t)$ for U, given $T = t$ does not depend on t, the two must be independent. □

Example 2.12. Basu's theorem can be used to show that the mean and variance of an independent sample from $\mathsf{N}(\mu, \sigma^2)$ are independent. Suppose first that σ^2 is known to be equal to 1. We know that the sample mean $\bar{X}$ is complete and sufficient for μ, and also that the sample variance $s^2(X) = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is ancillary. Therefore, Basu's theorem says $\bar{X}$ and $s^2(X)$ are independent. But this was for $\sigma^2 = 1$—how to extend to the unknown σ^2 case? The point is that the general σ^2 case corresponds to a simple scale transformation of the data, which clearly cannot alter the correlation structure between $\bar{X}$ and $s^2(X)$. Therefore, $\bar{X}$ and $s^2(X)$ are independent for all (μ, σ^2).

Example 2.13. Suppose $X_1, \ldots, X_n$ is an iid sample from $\mathsf{N}(0, 1)$, and let $\bar{X}$ and M denote the sample mean and sample median, respectively. The goal is to calculate the covariance between $\bar{X}$ and M. Introduce a mean parameter ξ; in so doing, we find that $\bar{X}$ is complete and sufficient, while $\bar{X} - M$ is ancillary. Then Basu's theorem says $\bar{X}$ and $\bar{X} - M$ are independent, and hence:

$$0 = \mathsf{C}(\bar{X}, \bar{X} - M) = \mathsf{V}(\bar{X}) - \mathsf{C}(\bar{X}, M) \quad \Longrightarrow \quad \mathsf{C}(\bar{X}, M) = n^{-1}.$$

It is common in stat theory courses and textbooks to give the impression that Basu's theorem is just a trick for doing certain calculations, like in the two examples above. However, the real contribution of Basu's theorem is that point mentioned above about conditioning on ancillary statistics. More on this in Section 2.5

2.4 Fisher information

2.4.1 Definition

We colloquially understand that a sufficient statistic contains all the information in $X_1, \ldots, X_n$ concerning the parameter of interest θ. The concept of "Fisher information" will make this more precise. Some of the material here comes from Chapter 2.3 in Schervish (1995).

Definition 2.3. Suppose θ is d-dimensional and $p_\theta(x)$ is the density of X with respect to μ. Then the following are the *FI regularity conditions*.

I. $\partial p_\theta(x)/\partial\theta_i$ exists μ-a.e. for each i.

II. $\int p_\theta(x)\, d\mu(x)$ can be differentiated under the integral sign.

III. The support of p_θ is the same for all θ.

Definition 2.4 (Score; Fisher information)**.** Assume the FI regularity conditions. The score vector is defined as $\partial \log p_\theta(X)/\partial\theta_i$ for $i = 1, \ldots, d$. The Fisher information $I_X(\theta)$ is the covariance matrix of the score vector; i.e.,

$$I_X(\theta)_{ij} = \mathsf{C}_\theta\Big(\frac{\partial \log p_\theta(X)}{\partial\theta_i}, \frac{\partial \log p_\theta(X)}{\partial\theta_j}\Big). \tag{2.7}$$

It turns out that the expected value of the score is zero. In this case, if we can differentiate twice under the integral sign (as we can in exponential families; cf. Theorem 2.1), then there is an alternative formula for the Fisher information:

$$I_X(\theta)_{ij} = -\mathsf{E}_\theta\Big\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\Big\}.$$

If $X_1, \ldots, X_n$ are iid from a distribution that satisfies the FI regularity conditions, then it's fairly easy to show that $I_{X_1,\ldots,X_n}(\theta) = nI_{X_1}(\theta)$; see Exercise 14. That is, information is accumulated as more data comes in, which makes sense if it's meant to measure the information in a data set. If data are not iid, then the information still increases, but at a rate no faster than that for the iid case.

It is worth mentioning that the FI regularity conditions are not necessary here. In particular, that $\theta \mapsto p_\theta(x)$ be differentiable for all x is far too strong. Fisher information can be defined under the much less strict condition of *differentiability in quadratic mean* (van der Vaart 1998, Chap. 6). See Chapter 6.

2.4.2 Sufficiency and information

The next result helps the interpretation of sufficient statistics containing all available information relevant to θ.

Theorem 2.8. *Assume the FI regularity conditions. Suppose θ is d-dimensional and P_θ is dominated by μ. If $T = g(X)$ is a statistic, then $I_X(\theta) - I_T(\theta)$ is positive semidefinite. The matrix is all zeros if and only if T is sufficient.*

Proof. Since T is a function of X, the joint distribution is determined by the marginal distribution of X. In particular, by (1.9),

$$p_\theta^{X|T}(x \mid t) = \begin{cases} p_\theta^X(x)/p_\theta^T(t) & \text{if } T(x) = t \\ 0 & \text{otherwise.} \end{cases}$$

Here we use p_θ for all densities, and the superscripts indicate the distribution. Therefore,

$$p_\theta^X(x) = p_\theta^{X,T}(x,t) = p_\theta^T(t) p_\theta^{X|T}(x \mid t), \quad \text{if } T(x) = t.$$

Taking logs gives

$$\frac{\partial \log p_\theta^X(X)}{\partial \theta_i} = \frac{\partial \log p_\theta^T(T)}{\partial \theta_i} + \frac{\partial \log p_\theta^{X|T}(X \mid T)}{\partial \theta_i} \quad \forall\, \theta. \tag{2.8}$$

We will show that the two terms on the right-hand side are uncorrelated, and that the last term is zero if and only if T is sufficient. Using the iterated expectation,

$$\begin{aligned} \mathsf{C}_\theta\Big(\frac{\partial \log p_\theta^T(T)}{\partial \theta_i}, \frac{\partial \log p_\theta^{X|T}(X|T)}{\partial \theta_i}\Big) &= \mathsf{E}_\theta\Big\{\frac{\partial \log p_\theta^T(T)}{\partial \theta_i} \frac{\partial \log p_\theta^{X|T}(X|T)}{\partial \theta_i}\Big\} \\ &= \mathsf{E}_\theta\Big\{\frac{\partial \log p_\theta^T(T)}{\partial \theta_i} \mathsf{E}_\theta\Big(\frac{\partial \log p_\theta^{X|T}(X|T)}{\partial \theta_i} \,\Big|\, T\Big)\Big\}. \end{aligned}$$

We claim that the inner conditional expectation is zero with P_θ^T-probability 1. To show this, note first that

$$1 = \int p_\theta^{X|T}(x \mid t)\, d\mu(x) \implies 0 = \frac{\partial}{\partial \theta_i} \int p_\theta^{X|T}(x \mid t)\, d\mu(x),$$

for all t outside a P_θ^T-null set. If we can interchange the latter derivative and conditional expectation, then we're done. Since

$$\int p_\theta^{X|T}(x \mid t)\, d\mu(x) = \frac{1}{p_\theta^T(t)} \int_{\{x:T(x)=t\}} p_\theta^X(x)\, d\mu(x),$$

taking derivative with respect to θ_i and simplifying gives

$$\frac{1}{p_\theta^T(t)} \frac{\partial}{\partial \theta_i} \int_{\{x:T(x)=t\}} p_\theta^X(x)\, d\mu(x) - \frac{\partial}{\partial \theta_i} \log p_\theta^T(t). \tag{2.9}$$

The restriction on the range of integration does not prevent us from interchanging derivative and p_θ^X integral[3] (as in FI), we get

$$\begin{aligned}
&\frac{\partial}{\partial\theta_i}\int_{\{x:T(x)=t\}} p_\theta^X(x)\,d\mu(x) \\
&\qquad = \int_{\{x:T(x)=t\}} \Big[\frac{\partial}{\partial\theta_i}\log p_\theta^T(t) + \frac{\partial}{\partial\theta_i}\log p_\theta^{X|T}(x|t)\Big] p_\theta^X(x)\,d\mu(x) \\
&\qquad = p_\theta^T(t)\frac{\partial}{\partial\theta_i}\log p_\theta^T(t) + p_\theta^T(t)\int \frac{\partial}{\partial\theta_i}\log p_\theta^{X|T}(x|t) p_\theta^{X|T}(x|t)\,d\mu(x).
\end{aligned}$$

This latter calculation shows that (2.9) simplifies to

$$\int \frac{\partial}{\partial\theta_i}\log p_\theta^{X|T}(x|t) p_\theta^{X|T}(x|t)\,d\mu(x).$$

Therefore,

$$\frac{\partial}{\partial\theta_i}\int p_\theta^{X|T}(x\mid t)\,d\mu(x) = \int \frac{\partial}{\partial\theta_i}\log p_\theta^{X|T}(x|t) p_\theta^{X|T}(x|t)\,d\mu(x),$$

and since we can interchange derivative and integral with respect to the conditional distribution, we get that the conditional expectation of the conditional score function is zero (with P_θ^T-probability 1). This, in turn, shows that the two terms on the right-hand side in (2.8) are uncorrelated. Then the covariance matrix of the sum on the right-hand side of (2.8) is the sum of the respective covariance matrices. Therefore,

$$I_X(\theta) = I_T(\theta) + \mathsf{E}_\theta\{I_{X|T}(\theta)\},$$

and it's clear that $I_X(\theta) - I_T(\theta)$ is positive semidefinite. The matrix $\mathsf{E}_\theta\{I_{X|T}(\theta)\}$ is all zeros if and only if the conditional score $\partial\log p_{X|T,\theta}(X|T)/\partial\theta_i$ is constant (must be zero, right?) in θ or, in other words, T is sufficient. □

This formalizes the claim made in the introduction that sufficient statistics T preserve all the information about θ in data X. That is, in the one-dimensional case, we have $I_{T(X)} \le I_X$ with equality if and only if T is sufficient.

2.4.3 Cramer–Rao inequality

We have seen that Fisher information provides a justification for the claim that sufficient statistics contain all the relevant information in a sample. However, the Fisher information plays an even deeper role in statistical inference, in particular, it is involved in many optimality results that provide a baseline for comparison among estimators, tests, etc. Below is a familiar but important result, which states that, under some conditions, the variance of an estimator cannot be less than a bound involving the Fisher information.

[3] In fact, it can be shown that the derivative of $\int g(x)p_\theta^X(x)\,d\mu(x)$ with respect to each coordinate in θ can be taken under the integral, for any bounded function g; this follows from the dominated convergence theorem. For the present case, take $g(x)$ to be the appropriate indicator function.

Theorem 2.9 (Cramer–Rao)**.** *For simplicity, take θ to be a scalar, and assume that p_θ satisfies the FI regularity conditions. Let $X_1, \dots, X_n \overset{iid}{\sim} p_\theta$ and let $T = T(X_1, \dots, X_n)$ be a real-valued statistic with $\mathsf{E}_\theta(T) = g(\theta)$. Then*

$$\mathsf{V}_\theta(T) \geq \{g'(\theta)\}^2 \{nI(\theta)\}^{-1}.$$

Proof. By Exercise 15, the covariance between T and the score function is $g'(\theta)$. Since the variance of the score is $nI(\theta)$, the Cauchy–Schwartz inequality gives

$$g'(\theta)^2 \leq \mathsf{V}_\theta(T)\{nI(\theta)\}.$$

Solve for $\mathsf{V}_\theta(T)$ to get the desired result. □

One application of the Cramer–Rao inequality is in experimental design. In those problems, one has control on certain inputs and the goal is to select those inputs in such a way that the estimator has, say, as small a variance as possible. In such cases, the strategy is to chose those inputs in such a way that the Fisher information is maximized,[4] which has some intuition driven by Cramer–Rao, i.e., the variance is small if the Fisher information is big.

2.4.4 Other measures of information

Is Fisher information the only measure of information? Technically, the answer is NO, there are other measures, but they don't get much attention. The reason is that the Fisher information is the "right" choice provided that the model satisfies the FI regularity conditions. Since most models (e.g., regular exponential families) satisfies these, there is not so much reason to look beyond Fisher information. However, there are models which do not satisfy the regularity conditions, such as $\mathsf{Unif}(0, \theta)$. In such cases, the Fisher information is not defined, so it obviously cannot be used. The question is if there is some other kind of information, one that reduces to the Fisher information when it exists, but is more versatile in the sense that it can be defined when Fisher information cannot.

Towards extending the Fisher information, it helps to understand where the Fisher information comes from. Recall the Kullback–Leibler divergence from Chapter 1, which (roughly) measures the distance between two models. Consider here two models with density functions p_θ and $p_{\theta+\varepsilon}$, with respect to the same dominating measure μ, where the latter means a small change in the parameter. Kullback (1997, Sec. 1.6) shows that the Kullback–Leibler divergence of $p_{\theta+\varepsilon}$ from p_θ is approximately quadratic in ε, in particular, $K(p_\theta, p_{\theta+\varepsilon}) \approx \varepsilon^\top I(\theta)\varepsilon$, as $\varepsilon \to 0$, where $I(\theta)$ is the Fisher information matrix from before; see Exercise 18. Then the key idea to generalizing Fisher information matrix is to recognize that, outside the regular cases, the Kullback–Leibler divergence or, better, the squared *Hellinger distance*, defined as

$$h^2(\theta, \theta') = \int (p_\theta^{1/2} - p_{\theta'}^{1/2})^2 \, d\mu,$$

[4]Since $I(\theta)$ is a generally matrix, it is not clear what it means to "maximize" it; usually this is formulated in terms of maximizing a certain functional of $I(\theta)$, such as its determinant.

is not quadratic in ε anymore. But this same expansion can be carried out and the coefficient defines a suitable "Hellinger information." For example, consider the $\mathsf{Unif}(0,\theta)$ case. The Hellinger divergence is

$$h^2(\theta+\varepsilon,\theta) = \int \Big[\frac{1}{\sqrt{\theta+\varepsilon}} I_{(0,\theta+\varepsilon)}(x) - \frac{1}{\sqrt{\theta}} I_{(0,\theta)}(x)\Big]^2 \, dx = \cdots = \frac{\varepsilon}{\theta} + o(\varepsilon). \tag{2.10}$$

This has a linear instead of quadratic approximation, a result of the non-regularity of the $\mathsf{Unif}(0,\theta)$ distribution. But a "Hellinger information" for $\mathsf{Unif}(0,\theta)$ can be defined as θ^{-1}. There are versions of the Cramer–Rao bound for the Hellinger information too, but I will not present this here.[5]

2.5 Conditioning

Here we discuss some interesting examples in which the classical frequentist approach gives strange answers. These examples shall be used to motivate conditioning in inference.

Example 2.14. Suppose X_1 and X_2 are iid with distribution P_θ satisfying

$$\mathsf{P}_\theta(X = \theta - 1) = \mathsf{P}_\theta(X = \theta + 1) = 0.5, \quad \theta \in \mathbb{R}.$$

The goal is to construct a confidence interval for the unknown θ. Consider

$$C = \begin{cases} \{\bar{X}\} & \text{if } X_1 \neq X_2 \\ \{X_1 - 1\} & \text{if } X_1 = X_2. \end{cases}$$

To be clear, in either case, C is a singleton. It can be shown that C has confidence 75%. But let's look at this procedure more carefully. From the structure of the problem, if $X_1 \neq X_2$, then one observation is $\theta - 1$ and the other is $\theta + 1$. In this case, $\bar{X}$ is exactly equal to θ so, *given* $X_1 \neq X_2$, C is guaranteed to be correct. On the other hand, if $X_1 = X_2$, then C is $\{\theta\}$ with probability 0.5 and $\{\theta - 2\}$ with probability 0.5 so, *given* $X_1 = X_2$, C is correct with probability 0.5. Putting this together, C has confidence 100% when $X_1 \neq X_2$ and 50% when $X_1 = X_2$. So on average the confidence is 75% but since, for a given problem, we know which case we're in, wouldn't it make sense to report the *conditional confidence* of either 100% or 50%?

Example 2.15. Suppose data X can take values in $\{1,2,3\}$ and $\theta \in \{0,1\}$. The probability distribution for X for each θ is described by the following table.

x	1	2	3
$p_0(x)$	0.0050	0.0050	0.99
$p_1(x)$	0.0051	0.9849	0.01

[5] I am currently working on this Hellinger information thing with applications to optimal experimental design, and the papers are in preparation.

The most powerful level $\alpha = 0.01$ test of $H_0 : \theta = 0$ versus $H_1 : \theta = 1$ is based on the likelihood ratio $p_0(x)/p_1(x)$ for the given $X = x$. It can be shown that this test has power 0.99, which suggests that there is a lot of confidence in the decision made based on the observed x. But is this the case? If $X = 1$ is observed, then the likelihood ratio is $0.005/0.0051 \approx 1$. In general, a likelihood ratio close to 1 does not give strong preference to either H_0 or H_1, so measuring our certainty about the procedures choice using the "global" measure of power might be misleading.

Example 2.16. Consider the following experiment: flip a fair coin and, if the coin lands on Heads, then take $X \sim \mathsf{N}(\theta, 1)$; otherwise, take $X \sim \mathsf{N}(\theta, 99)$. Suppose that the outcome of the coin flip is *known*. The goal is to use X to estimate θ. What distribution should we use to construct a confidence interval, say? The marginal variance of X is $(1 + 99)/2 = 50$. However, this seems like a poor explanation of the actual error in X as an estimator of θ, since we actually know whether X is sampled from $\mathsf{N}(\theta, 1)$ or $\mathsf{N}(\theta, 99)$. Then the question is, why not use the "conditional" variance, given the outcome of the coin flip? This is an intuitively natural thing to do, but this is *not* what frequentism says to do.

Example 2.17. Let (X_{1i}, X_{2i}), $i = 1, \ldots, n$ be an iid bivariate sample from a distribution with density $p_\theta(x_1, x_2) = e^{-\theta x_1 - x_2/\theta}$, where x_1, x_2, and θ are all positive. It can be shown that the minimal sufficient statistic is $T = (T_1, T_2)$, where $T_j = \sum_{i=1}^n X_{ji}$, $j = 1, 2$. Note that the minimal sufficient statistic is two-dimensional while the parameter is only one-dimensional. For estimating θ, a reasonable choice is $\hat{\theta} = \{T_2/T_1\}^{1/2}$, the maximum likelihood estimator. However, this is *not* a minimal sufficient statistic, so we have to choose whether we should condition or not. An ancillary statistic to condition on is $A = \{T_1 T_2\}^{1/2}$. As discussed in Ghosh et al. (2010), the unconditional Fisher information in T and in $\hat{\theta}$, respectively, are

$$I_T(\theta) = \frac{2n}{\theta^2} \quad \text{and} \quad I_{\hat{\theta}}(\theta) = \frac{2n}{\theta^2} \frac{2n}{2n+1};$$

of course, as expected, $I_T(\theta) > I_{\hat{\theta}}(\theta)$. The conditional Fisher information, however, is

$$I_{\hat{\theta}|A}(\theta) = I_T(\theta) \frac{K_1(2A)}{K_0(2A)}, \tag{2.11}$$

where K_0 and K_1 are Bessel functions. A plot of the ratio—call it $r(A)$—on the right-hand side above, as a function of $A = a$, is shown in Figure 2.1. When A is large, $r(A)$ is near 1 so $I_{\hat{\theta}|A}(\theta) \approx I_T(\theta)$. However, if A is not large, then the conditional information can be much larger, and since larger information is "better," we can see that there is an advantage to conditioning in this case.

The foregoing examples are meant to shed light on the drawbacks of pure frequentism. At least in some examples, there is clearly a reason to consider conditioning on something: sometimes it's clear what to condition on (Example 2.16) and other times it's not (Example 2.17). Conditional inference is when sampling distributions are based on conditional distributions of estimators given the observed value of an ancillary statistic; e.g., in Example 2.14, $|X_1 - X_2|$ is an ancillary statistic. When the estimator is a complete sufficient

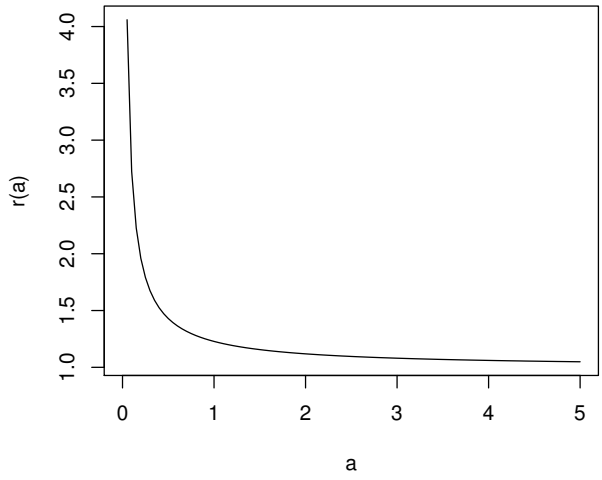


Figure 2.1: Plot of the ratio $r(a)$ on the right-hand of (2.11) as a function of $A = a$, the ancillary statistic.

statistic, then Basu's theorem says there is no need to condition. But in problems where the estimator is not a complete sufficient statistic (Example 2.17), there is a need to condition. There are extensive discussions in the literature about conditional inference, for example, Fraser (2004) and Ghosh et al. (2010); Berger (2014) gives a more recent discussion. Despite the benefits of conditional inference, this stuff has not really permeated applied statistics; this is due to some additional technicalities, both in finding an appropriate ancillary to condition on, and in actually doing the conditioning. A nice applied look at conditional inference and related topics is in Brazzale et al. (2007). A different look at conditioning and related issues can be found in Martin and Liu (2014, 2015a).

2.6 Discussion

2.6.1 Generalized linear models

An important application of exponential family distributions are the so-called *generalized linear models*, or GLMs for short. These models generalize the usual linear models (e.g., regression and analysis of variance) presented in introductory statistics methodology courses. This topic typically does not appear in a course such as this—there's no mention of GLMs in Keener (2010)—and I think the reason for its omission is that the details of the theory can be understood in the simpler exponential family setup discussed in Section 2.2, leaving the specific modeling and computational details to other courses/texts. However, I think it's important for students to get some minimal exposure to this application of exponential family models in this theoretical course, if only so that they know of the existence of such things so they can read more on their own or take more specialized courses. Here I give

a very brief explanation of GLMs with a few examples. A comprehensive introduction to GLMs is given in McCullagh and Nelder (1983).

Consider a problem with two variables: Y is called the *response variable* and X the *predictor variable* or *covariate*, where X is, say, d-dimensional. The usual linear model states that, given $X = x$, the mean of the response variable Y is a linear function of x, i.e., $\mathsf{E}(Y \mid X = x) = x^\top \beta$, where β is a d-dimensional parameter of coefficients. If we have independent samples, i.e., $\{(X_i, Y_i) : i = 1, \ldots, n\}$, then the model states that, given $X_i = x_i$, Y_i are independent with mean $\mu_i = x_i^\top \beta$, $i = 1, \ldots, n$. A key point is that β is the same for each i; moreover, this is one of the most common independent but *not iid* models students will see. The least-squares method can be used to estimate β based on the observations, and this solution has many desirable properties that we will not dig into here.

A question to consider is this: is this kind of linear model always appropriate? That is, should the mean of the response variable distribution (conditioned on the predictor X) be expressed as a linear function of the predictor? As an example, consider the case where Y is Poisson or Bernoulli distributed. In both cases, the mean of the distribution has a constraint—in $(0, \infty)$ in one case and in $(0, 1)$ in the other—so a linear function, which has no constraints, i.e., can take values in $(-\infty, \infty)$, may not be appropriate. A GLM can handle this by not extending too far beyond the comfort of a linear model.

Suppose that response variables $Y_1, \ldots, Y_n$ are independent with densities

$$p_{\theta_i}(y_i) = h(y_i) e^{\eta(\theta_i) y_i - A(\theta_i)}, \quad i = 1, \ldots, n,$$

which is of the exponential family form in Section 2.2, but with a different parameter θ_i for each data point. Assume that there is some common structure, in particular, that the mean $\mu_i = \mathsf{E}_{\theta_i}(Y_i)$ satisfies the condition $g(\mu_i) = x_i^\top \beta$ for some smooth one-to-one function g, called the *link function*. When the link function is such that $\eta(\theta_i) = x_i^\top \beta$, it is called the *canonical link*. The result of this construction is a general way to introduce an effectively linear model connecting the response variable Y_i to the predictor variable X_i but avoid the shortcomings of an actual linear model.

As a quick example, consider the case where $Y_i \sim \mathsf{Pois}(\theta_i)$, $i = 1, \ldots, n$, independent. It is easy to check that Poisson is an exponential family model and $\eta(\theta_i) = \log \theta_i$. Since θ_i is also the mean of Y_i, if we want to construct a Poisson GLM with canonical link, then $g(u) = \log u$, so that

$$\theta_i = e^{x_i^\top \beta} \iff \log \theta_i = x_i^\top \beta.$$

The latter formula explains why this Poisson GLM is often called a log-linear model. Another example, for the Bernoulli model, is given in Exercise 20.

2.6.2 A bit more about conditioning

In stat theory courses and books, sufficiency is treated as a critically important aspect of statistical inference. Here I want to make a case that there is nothing really special about sufficient statistics, provided that some appropriate conditioning is done. The message here is that conditioning is a more fundamental concept than sufficiency.

I'll make this case using a simple example. Let X_1, X_2 be iid $\mathsf{N}(\theta, 1)$. A reasonable estimator of θ is $\bar{X} = (X_1 + X_2)/2$, a sufficient statistic, with sampling distribution is $\mathsf{N}(\theta, 1/2)$. Consider, on the other hand, the estimator $\hat{\theta} = X_1$, which is not a sufficient statistic. Classical considerations would suggest that inference based on X_1 is worse than that based on $\bar{X}$. However, consider the conditional sampling distribution of X_1, given $X_2 - X_1$. It is easy to check that

$$X_1 \mid (X_2 - X_1) \sim \mathsf{N}(\theta + \tfrac{X_2 - X_1}{2}, 1/2),$$

and, for example, confidence intervals based on this conditional sampling distribution are the same as those based on the marginal sampling distribution of the sufficient statistic $\bar{X}$. So, in this problem, one could argue that there is really nothing special about the sufficient statistic $\bar{X}$, since one can get effectively the same sampling distribution using some other non-sufficient statistic, provided that proper conditioning is performed. The result here is more general (see Exercise 22), though continuity seems to be important.

Sufficiency, when it gives something meaningful, can be convenient, since conditioning isn't needed, which saves some effort. However, there are cases where sufficiency provides no improvement. For example, in the Student-t location problem with known degrees of freedom, the full data is the minimal sufficient statistic. However, one can easily get a reasonable (location equivariant) estimator, such as the sample mean, and condition on the the maximal invariant, an ancillary statistic. The point is that conditioning works when sufficiency doesn't, and even when sufficiency does work, conditioning can be just as good. So, I would argue that conditioning is more fundamental than sufficiency.

2.7 Exercises

1. Hölder's inequality is a generalization of the Cauchy–Schwartz inequality.

   > Let $1 \leq p, q \leq \infty$ be numbers with $1/p + 1/q = 1$. Let f and g be functions such that f^p and g^q are μ-integrable. Then
   >
   > $$\int |fg| \, d\mu \leq \Big(\int |f|^p \, d\mu \Big)^{1/p} \Big(\int |g|^q \, d\mu \Big)^{1/q}.$$
   >
   > Cauchy–Schwartz corresponds to $p = q = 2$.

   Use Hölder's inequality to prove Proposition 2.1.

2. Prove the inequalities in (2.3).

3. Suppose X has an exponential family distribution with density

   $$p_\theta(x) = h(x) e^{\eta(\theta) T(x) - A(\theta)}.$$

   Derive the mean and variance formulas

   $$\mathsf{E}_\theta[T(X)] = \frac{A'(\theta)}{\eta'(\theta)}, \quad \mathsf{V}_\theta[T(X)] = \frac{A''(\theta)}{[\eta'(\theta)]^2} - \frac{\eta''(\theta) A'(\theta)}{[\eta'(\theta)]^3}.$$

4. Prove (2.4), a formula for the exponential family moment-generating function.

5. A discrete random variable with pmf

$$p_\theta(x) = a(x)\theta^x/C(\theta), \quad x \in \{0, 1, \ldots\};\ a(\theta) \geq 0;\ \theta > 0$$

has a *power series distribution*.

   (a) Show that the power series distribution is an exponential family.

   (b) Show that binomial and Poisson are special cases of power series distributions.

6. (a) Prove *Stein's identity*.[6] For $X \sim \mathsf{N}(\mu, \sigma^2)$, let φ be a differentiable function with $\mathsf{E}_\theta|\varphi'(X)| < \infty$. Then

$$\mathsf{E}[\varphi(X)(X - \mu)] = \sigma^2\mathsf{E}[\varphi'(X)].$$

   [Hint: Without loss of generality, assume $\mu = 0$ and $\sigma = 1$. Use integration-by-parts. You'll need to show that $\varphi(x)e^{-x^2/2} \to 0$ as $x \to \pm\infty$. There is also an approach that uses Fubini's theorem.]

   (b) Let $X \sim \mathsf{N}(\mu, \sigma^2)$. Use Stein's identity to find the first four moments, $\mathsf{E}(X^k)$, $k = 1, 2, 3, 4$, of X. [Hint: For $\mathsf{E}(X^k)$ use $\varphi(x) = x^{k-1}$.]

7. Argue that a one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic.

8. Suppose $X_1, \ldots, X_n$ are iid $\mathsf{N}(\theta, \theta^2)$.

   (a) Show that $\mathsf{N}(\theta, \theta^2)$ has an exponential family form.

   (b) Find the minimal sufficient statistic for θ.

   (c) Show that your minimal sufficient statistic is not complete.

9. The inverse Gaussian family, denoted by $\mathsf{IG}(\lambda, \mu)$ has density function

$$(\lambda/2\pi)^{1/2}\exp\{(\lambda\mu)^{1/2}\}x^{-3/2}\exp\{-(\lambda x^{-1} + \mu x)/2\}, \quad x > 0;\ \lambda, \mu > 0.$$

   (a) Show that $\mathsf{IG}(\lambda, \mu)$ is an exponential family.

   (b) Show that $\mathsf{IG}(\lambda, \mu)$ is invariant with respect to the group of scale transformations, i.e., $\mathscr{G} = \{g_c(x) = cx : c > 0\}$.[7]

   (c) Let $T_1(X) = n^{-1}\sum_{i=1}^n X_i$ and $T_2(X) = \sum_{i=1}^n (1/X_i - 1/T_1(X))$. Show that (T_1, T_2) is complete and sufficient.

[6] This is just a special case; a similar result holds for all exponential families.

[7] The inverse Gaussian, together with normal and gamma, are the only three distributions which are both exponential families and group families.

(d) Show that $T_1 \sim \mathsf{IG}(n\lambda, n\mu)$.[8]

10. Suppose that pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid from a bivariate normal distribution, where $\mathsf{E}(X_1) = \mathsf{E}(Y_1) = 0$, $\mathsf{V}(X_1) = \mathsf{V}(Y_1) = 1$, and $\mathsf{E}(X_1 Y_1) = \theta$. Here $\theta \in (-1, 1)$ is the correlation between X and Y.

    (a) Find a (two-dimensional) minimal sufficient statistic for θ.

    (b) Prove that the minimal sufficient statistic is not complete.

    (c) Let $Z_1 = \sum_{i=1}^n X_i^2$ and $Z_2 = \sum_{i=1}^n Y_i^2$. Show that both Z_1 and Z_2 are ancillary, but not (Z_1, Z_2).

11. This exercise describes an alternative approach to finding minimal sufficient statistics. It is related to that given in Theorem 2.3.

    (a) Prove the following theorem:

    > *Consider a finite family of distributions with densities $p_0, p_1, \ldots, p_K$, all having the same support. Then*
    >
    > $$T(X) = \Big(\frac{p_1(X)}{p_0(X)}, \frac{p_2(X)}{p_0(X)}, \ldots, \frac{p_K(X)}{p_0(X)}\Big)$$
    >
    > *is minimal sufficient.*

    (b) Prove the following theorem: Let $\mathbb{P}$ be a parametric family of distributions with common support, and $\mathbb{P}_0$ a subset of $\mathbb{P}$. If T is minimal sufficient for $\mathbb{P}_0$ and sufficient for $\mathbb{P}$, then it's minimal sufficient for $\mathbb{P}$.

    (c) Use the previous two results to prove that, for the $\mathsf{Pois}(\theta)$ family, $T = \sum_{i=1}^n X_i$ is a minimal sufficient statistic. [Hint: Pick a two-element subset $\mathbb{P}_0 = \{p_0 = \mathsf{Pois}(\theta_0), p_1 = \mathsf{Pois}(\theta_1)\}$ of $\mathbb{P} = \{\mathsf{Pois}(\theta) : \theta > 0\}$.]

12. (a) Consider a location family with densities $p_\theta(x) = p(x - \theta)$, $\theta \in \mathbb{R}$. For $X \sim p_\theta$, show that the Fisher information for θ is

    $$I_X(\theta) = \int_{-\infty}^{\infty} [p'(x)]^2 / p(x)\, dx,$$

    which is independent of θ.

    (b) Consider a scale family $p_\theta(x) = p(x/\theta)/\theta$, $\theta > 0$. For $X \sim p_\theta$, show that the Fisher information for θ is

    $$I_X(\theta) = \frac{1}{\theta^2} \int \Big[\frac{x p'(x)}{p(x)} + 1\Big]^2 p(x)\, dx.$$

13. For each case, find the Fisher information based on a single observation X.

[8] It can also be shown that $T_2 \sim (1/\lambda)\mathsf{ChiSq}_{n-1}$, but the proof is more difficult.

(a) $\mathsf{Ber}(\theta)$.

(b) $\mathsf{Pois}(\theta)$.

(c) $\mathsf{Cau}(\theta, 1)$.

(d) $\mathsf{N}(0, \theta)$, where $\theta > 0$ denotes the variance.

14. For iid $X_1, \ldots, X_n$, show that $I_{X_1,\ldots,X_n}(\theta) = nI_{X_1}(\theta)$.

15. Let p_θ be a density that satisfies the FI regularity conditions, and let $T = T(X_1, \ldots, X_n)$ have $\mathsf{E}_\theta(T) = g(\theta)$. Show that $\mathsf{C}_\theta(T, U_\theta) = g'(\theta)$, where $U_\theta = \sum_{i=1}^n \frac{\partial}{\partial\theta} \log p_\theta(X_i)$ is the score function.

16. Suppose the Fisher information in X about θ is $I_X(\theta)$, where θ is a scalar. Let ξ be a (scalar-valued) smooth one-to-one reparametrization of θ, and write $\tilde{I}_X(\xi)$ for the Fisher information in X about ξ. Show that $\tilde{I}_X(\xi) = (d\theta/d\xi)^2 I_X(\theta)$. Generalize to the case of vector θ and ξ.

17. Let $X \sim \mathsf{N}_n(\theta, \Sigma)$ be a single n-dimensional normal sample; here, the covariance matrix Σ is known but the vector θ is unknown.

    (a) Find the Fisher information matrix $I_X(\theta)$. [Hint: You can verify this directly via the formulas, or generalize the result in Exercise 12(a).]

    (b) Suppose that $\theta = D\xi$, where D is a $n \times p$ matrix of rank p, where $p < n$, and ξ is an unknown $p \times 1$ vector. Here D is the "design matrix." Use the result in Exercise 16 to find the Fisher information $\tilde{I}_X(\xi)$ in X about ξ.

    (The information matrix in part (b) depends on the design matrix D, and the theory of optimal designs seeks to choose D to make $\tilde{I}_X(\xi)$, as "large as possible." Of course, the Fisher information here is a matrix, so one must define what it means for a matrix to be large, but the intuition is perfectly clear.)

18. Let $\{p_\theta : \theta \in \Theta\}$ be a class of μ-densities that satisfy the FI regularity conditions. By interchanging differentiation and integration, work out a quadratic Taylor approximation to the function $\eta \mapsto K(p_\theta, p_\eta)$, for η near θ, where K is the Kullback–Leibler divergence from Chapter 1. You should see the Fisher information emerge in the quadratic approximation.

19. Fill in the missing details in (2.10).

20. Let $Y_i \sim \mathsf{Ber}(\theta_i)$, $i = 1, \ldots, n$, independent.

    (a) Show that the Bernoulli model is an exponential family with $\eta(\theta) = \log \frac{\theta}{1-\theta}$.

    (b) Find the canonical link and write down the formula for θ_i in terms of a predictor variable x_i and a parameter β like in Section 2.6.1.

(c) Look up the "logistic distribution" (e.g., on `wikipedia`) to see why they call this Bernoulli GLM with canonical link *logistic regression*.

21. Let X_1, X_2 be iid $\mathsf{Unif}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$.

    (a) Show that $A = (X_2 - X_1)/2$ is an ancillary statistic.

    (b) Find the distribution of $\bar{X}$, given $A = a$.

    (c) Compare $\mathsf{V}(\bar{X})$ and $\mathsf{V}(\bar{X} \mid A = a)$.

    (See Example 2.2 in Fraser (2004) for a different illustration in this example: there he shows that "optimal" unconditional confidence intervals for θ are junk, whereas the the conditional confidence interval is very reasonable.)

22. Let X_1, X_2 be iid exponential with mean θ.

    (a) Find the distribution of $\bar{X} = (X_1 + X_2)/2$.

    (b) Find the distribution of X_1, given X_2/X_1.

    (c) Compare confidence intervals obtained from the distributions in (a) and (b).

Chapter 3

Likelihood and Likelihood-based Methods

3.1 Introduction

Likelihood is surely one of the most important concepts in statistical theory. We have seen the role it plays in sufficiency, through the factorization theorem. But, more importantly, the likelihood function establishes a preference among the possible parameter values given data $X = x$. That is, a parameter values θ_1 with larger likelihood is better than parameter value θ_2 with smaller likelihood, in the sense that the model P_{θ_1} provides a better fit to the observed data than P_{θ_2}. This leads naturally to procedures for inference which select, as a point estimator, the parameter value that makes the likelihood the largest, or rejects a null hypothesis if the hypothesized value has likelihood too small. The likelihood function is also of considerable importance in Bayesian analysis, as we'll see later.

Estimators and hypothesis tests based on the likelihood function have some general desirable properties, in particular, there are widely applicable large-sample approximations of the relevant sampling distributions. A main focus of this chapter is the mostly rigorous derivation of these important results. There is also a very brief introduction to some advanced likelihood theory, including higher-order approximations and the use of pseudo-likelihoods, namely, profile and marginal likelihoods, when nuisance parameters are present. A few remarks about computation relevant in likelihood-based contexts is given in Section 3.7. The last section gives brief historical discussion of likelihood and a controversial result, due to Birnbaum, on what is called the *likelihood principle*.

3.2 Likelihood function

Consider a class of probability models $\{\mathsf{P}_\theta : \theta \in \Theta\}$, defined on the measurable space $(\mathbb{X}, \mathcal{A})$, absolutely continuous with respect to a dominating σ-finite measure μ. In this case, for each θ, the Radon–Nikodym derivative $(d\mathsf{P}_\theta/d\mu)(x)$ is the usual probability density function for the observable X, written as $p_\theta(x)$. For fixed θ, we know that $p_\theta(x)$ characterizes the

sampling distribution of X as well as that of any statistic $T = T(X)$. But how do we use/interpret $p_\theta(x)$ as a function of θ for fixed x? This is a special function with its own name—the *likelihood function*.

Definition 3.1. Given $X = x$, the likelihood function is $L(\theta) = p_\theta(x)$.

The intuition behind the choice of name is that a θ for which $L(\theta)$ is large is "more likely" to be true value compared to a θ' for which $L(\theta')$ is small. The name "likelihood" was coined by Fisher (1973):

> What has now appeared is that the mathematical concept of probability is ... inadequate to express our mental confidence or indifference in making ... inferences, and that the mathematical quantity which usually appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. To distinguish it from probability, I have used the term "likelihood" to designate this quantity; since both words "likelihood" and "probability" are loosely used in common speech to cover both kinds of relationship.

Fisher's point is that $L(\theta)$ is a measure of how *plausible* θ is, but that this measure of plausibility is different from our usual understanding of probability; see Aldrich (1997) for more on Fisher and likelihood. While we understand the probability (density) $p_\theta(x)$, for fixed θ, as a pre-experimental summary of our uncertainty about where X will fall, the likelihood $L(\theta) = p_\theta(x)$, for fixed x, gives a post-experimental summary of how likely it is that model P_θ produced the observed $X = x$. In other words, the likelihood function provides a ranking of the possible parameter values—those θ with greater likelihood are better, in that they fit the data better, than those θ with smaller likelihood. Therefore, only the shape of the likelihood function is relevant, not the scale.

The likelihood function is useful across all approaches to statistics. We've already seen some uses of the likelihood function. In particular, the factorization theorem states that the (shape of the) likelihood function depends on the observed data $X = x$ only through the sufficient statistic. The next section discusses some standard and some not-so-standard uses of the likelihood.

3.3 Likelihood-based methods and first-order theory

3.3.1 Maximum likelihood estimation

Probably the most familiar use of the likelihood function is maximum likelihood estimation. Given a class of potential models P_θ indexed by Θ, a subset of $\mathbb{R}^d$, we observe $X = x$ and we'd like to know which model is the most likely to have produced this x. This defines an optimization problem, and the result, namely

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta), \tag{3.1}$$

is the *maximum likelihood estimate* (MLE) of θ. Naturally, $\mathsf{P}_{\hat{\theta}}$ is then considered the most likely model, that is, among the class $\{\mathsf{P}_\theta : \theta \in \Theta\}$, the model $\mathsf{P}_{\hat{\theta}}$ provides the best fit to the observed $X = x$. In terms of "ranking" intuition, $\hat{\theta}$ is ranked the highest.

When the likelihood function is smooth, the optimization problem can be posed as a root-finding problem. That is, the MLE $\hat{\theta}$ can be viewed as a solution to the equation

$$\nabla \ell(\theta) = 0, \tag{3.2}$$

where ∇ denotes the gradient operator, $\ell = \log L$ is the log-likelihood, and the right-hand side is a d-vector of zeroes. Equation (3.2) is called the *likelihood equation*. Some remarks about solving the likelihood equation are given in Section 3.7. Our focus here will be on studying the theoretical large-sample properties of solutions $\hat{\theta}$ of the (3.2).

A first desirable large-sample property is consistency, which suggests that, if n is large, then $\hat{\theta} = \hat{\theta}_n$ will be close to θ with high P_θ-probability. More formally, we say that an estimator $\hat{\theta}_n$, not necessarily the MLE, is consistent if $\hat{\theta}_n \to \theta$ in P_θ-probability,[1] i.e.,

$$\lim_{n\to\infty} \mathsf{P}_\theta\{\|\hat{\theta}_n - \theta\| > \varepsilon\} = 0, \quad \forall\, \varepsilon > 0.$$

Here $\|\cdot\|$ is a suitable norm on the space Θ, e.g., the Euclidean norm. The definition can be strengthened to require that $\hat{\theta}_n \to \theta$ with P_θ-probability 1, though this is harder to prove.

We will talk in more detail in Chapter 6 about some general kinds of consistency results, which will contain consistency of the MLE as a special case. Suffice it say, under suitable conditions, there exists a consistent sequence of solutions to the likelihood equation (3.2); if, for example, that solution is unique, then a consistency result for the MLE obtains.

A more useful large-sample property is one that describes the limiting distribution. This (i) gives an exact characterization of the rate of convergence, and (ii) allows for the construction of asymptotically exact statistical procedures. Though it is possible to get non-normal limits, all "standard" problems that admit a limiting distribution have a normal limit. From previous experience, we know that MLEs typically have an asymptotic normality property. Here is one version of such a theorem, similar to Theorem 9.14 in Keener (2010), with conditions given in C1–C4 below. Condition C3 is the most difficult to check, but it does hold for regular exponential families; see Exercise 11. We focus on the one-dimensional case, but the exact same theorem, with obvious modifications, holds for d-dimensional θ.

C1. The support of P_θ does not depend on θ.

C2. For each x in the support, $f_x(\theta) := \log p_\theta(x)$ is three times differentiable with respect to θ in an interval $(\theta^\star - \delta, \theta^\star + \delta)$; moreover, $\mathsf{E}_{\theta^\star}|f'_X(\theta^\star)|$ and $\mathsf{E}_{\theta^\star}|f''_X(\theta^\star)|$ are finite and there exists a function $M(x)$ such that

$$\sup_{\theta\in(\theta^\star-\delta,\theta^\star+\delta)} |f'''_x(\theta)| \le M(x) \quad \text{and} \quad \mathsf{E}_{\theta^\star}[M(X)] < \infty. \tag{3.3}$$

[1] In this kind of notation, the subscript θ on P_θ is what indicates that θ is the true value.

C3. Expectation with respect to $\mathsf{P}_{\theta^\star}$ and differentiation at $\theta^\star$ can be interchanged, which implies that the score function has mean zero and that the Fisher information exists and can be evaluated using either of the two familiar formulas.

C4. The Fisher information at $\theta^\star$ is positive.

Theorem 3.1. *Suppose $X_1, \dots, X_n$ are iid P_θ, where $\theta \in \Theta \subseteq \mathbb{R}$. Assume C1–C4, and let $\hat{\theta}_n$ be a consistent sequence of solutions to* (3.2)*. Then, for any interior point $\theta^\star$,*

$$n^{1/2}(\hat{\theta}_n - \theta^\star) \to \mathsf{N}(0, I(\theta^\star)^{-1}), \quad \textit{in distribution under } \mathsf{P}_{\theta^\star}.$$

Proof. Let $\ell_n(\theta) = n^{-1} \log L_n(\theta)$ be scaled log-likelihood. Since $\theta^\star$ is an interior point, there exists an open neighborhood A of $\theta^\star$ contained in Θ. From consistency of $\hat{\theta}_n$, the event $\{\hat{\theta}_n \in A\}$ has $\mathsf{P}_{\theta^\star}$-probability converging to 1. Therefore, it suffices [Exercise 7(c)] to consider the behavior of $\hat{\theta}_n$ only when it is in A where the log-likelihood is well-behaved, in particular, $\ell_n'(\hat{\theta}_n) = 0$. Next, take a second-order Taylor approximation of $\ell_n'(\hat{\theta}_n)$ around $\theta^\star$:

$$0 = \ell_n'(\theta^\star) + \ell_n''(\theta^\star)(\hat{\theta}_n - \theta^\star) + \tfrac{1}{2}\ell_n'''(\tilde{\theta}_n)(\hat{\theta}_n - \theta^\star)^2, \quad \text{for } \hat{\theta}_n \text{ near } \theta^\star,$$

where $\tilde{\theta}_n$ is between $\hat{\theta}_n$ and $\theta^\star$. After a bit of simple algebra, we get

$$n^{1/2}(\hat{\theta}_n - \theta^\star) = -\frac{n^{1/2}\ell_n'(\theta^\star)}{\ell_n''(\theta^\star) + \frac{1}{2}\ell_n'''(\tilde{\theta}_n)(\hat{\theta}_n - \theta^\star)}, \quad \text{for } \hat{\theta}_n \text{ near } \theta^\star.$$

So, it remains to show that the right-hand side above has the stated asymptotically normal distribution. Let's look at the numerator and denominator separately.

Numerator. The numerator can be written as

$$n^{1/2}\ell_n'(\theta^\star) = n^{1/2} \cdot \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta} \log p_\theta(X_i)\Big|_{\theta=\theta^\star}.$$

The summands are iid with mean zero and variance $I(\theta^\star)$, by our assumptions about interchanging derivatives and integrals. Therefore, the standard Central Limit Theorem says that $n^{1/2}\ell_n'(\theta^\star) \to \mathsf{N}(0, I(\theta^\star))$ in distribution.

Denominator. The first term in the denominator converges in $\mathsf{P}_{\theta^\star}$-probability to $-I(\theta^\star)$ by the usual law of large numbers. It remains to show that the second term in the denominator is negligible. For this, note that by (3.3),

$$|\ell_n'''(\tilde{\theta}_n)| \le \frac{1}{n}\sum_{i=1}^n M(X_i), \quad \text{for } \hat{\theta}_n \text{ close to } \theta^\star.$$

The ordinary law of large numbers, again, says that the upper bound converges to $\mathsf{E}_{\theta^\star}[M(X_1)]$, which is finite. Consequently, $\ell_n'''(\tilde{\theta}_n)$ is bounded in probability, and since $\hat{\theta}_n - \theta^\star \to 0$ in $\mathsf{P}_{\theta^\star}$-probability by assumption, we may conclude that

$$(\hat{\theta}_n - \theta^\star)\ell_n'''(\tilde{\theta}_n) \to 0 \quad \text{in } \mathsf{P}_{\theta^\star}\text{-probability}.$$

It then follows from Slutsky's Theorem [Exercise 7(b)] that

$$-\frac{n^{1/2}\ell_n'(\theta^\star)}{\ell_n''(\theta^\star)+\frac{1}{2}(\hat{\theta}_n-\theta^\star)\ell_n'''(\tilde{\theta}_n)} \to \frac{\mathsf{N}(0,I(\theta^\star))}{-I(\theta^\star)} = \mathsf{N}(0,I(\theta^\star)^{-1}), \quad \text{in distribution,}$$

which is the desired result. □

The take-away message here is that, under certain conditions, if n is large, then the MLE $\hat{\theta}$ has sampling distribution close to $\mathsf{N}(\theta,[nI(\theta)]^{-1})$ under P_θ. To apply this result, e.g., to construct an asymptotically approximate confidence interval, one needs to replace $I(\theta)$ with a quantity that does not depend on the unknown parameter. Standard choices are the *expected* Fisher information $I(\hat{\theta}_n)$ and the *observed* Fisher information $-\ell_n''(\hat{\theta}_n)$; see Exercise 19 and Efron and Hinkley (1978). The latter is often preferred, for it has some desirable "conditioning" properties.

With asymptotic normality of the MLE, it is possible to derive the asymptotic distribution of any smooth function of the MLE. This is the well-known *delta theorem*, which you're invited to prove in Exercise 8. The delta theorem is actually more general, showing how to create new central limit theorems from existing ones; that is, the Delta Theorem is not specific to MLEs, etc. The delta theorem also offers an alternative—called *variance stabilizing transformations* (see Exercise 10)—to the plug-in rules discussed above to eliminate θ from the variance in the asymptotic normal approximation.

It is possible to drop the requirement that the likelihood be three times differentiable if one assumes that the second derivative exists and has a certain Lipschitz property:

> $\log p_\theta(x)$ is twice differentiable at $\theta^\star$, and there exists a function $g_r(x,\theta)$ such that, for each interior point $\theta^\star$,
>
> $$\sup_{\theta:|\theta-\theta^\star|\le r}\left|\frac{\partial^2}{\partial\theta^2}\log p_\theta(x)-\frac{\partial^2}{\partial\theta^2}\log p_{\theta^\star}(x)\right| \le g_r(x,\theta^\star), \tag{3.4}$$
>
> with $\lim_{r\to 0}\mathsf{E}_\theta\{g_r(X,\theta)\}=0$ for each θ.

With this assumption, the same asymptotic normality result holds. See Exercise 12. Interestingly, it is possible to get asymptotic normality under a much weaker condition, namely, *differentiable in quadratic mean*, which assumes less than differentiability of $\theta\mapsto p_\theta(x)$, but the details are a bit more technical; see Chapter 6.

3.3.2 Likelihood ratio tests

For two competing hypotheses H_0 and H_1 about the parameter θ, the likelihood ratio is often used to make a comparison. For example, for $H_0:\theta=\theta_0$ versus $H_1:\theta=\theta_1$, the likelihood ratio is $L(\theta_0)/L(\theta_1)$, and large (resp. small) values of this ratio indicate that the data x favors H_0 (resp. H_1). A more difficult and somewhat more general problem is $H_0:\theta\in\Theta_0$

versus $H_1 : \theta \notin \Theta_0$, where Θ_0 is a subset of Θ. In this case, one can define the likelihood ratio as

$$T_n = T_n(X, \Theta_0) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)}. \tag{3.5}$$

The interpretation of this likelihood ratio is the same as before, i.e., if the ratio is small, then data lends little evidence to the null hypothesis. For practical purposes, we need to know what it means for the ratio to be "small;" this means we need the *null distribution* of T_n, i.e., the distribution of T_n under P_θ, when $\theta \in \Theta_0$.

For $\Theta \subseteq \mathbb{R}^d$, consider the testing problem $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$, where Θ_0 is a subset of Θ that specifies the values $\theta_{01}, \ldots, \theta_{0m}$ of $\theta_1, \ldots, \theta_m$, where m is a fixed integer between 1 and d. The following result, known as *Wilks's Theorem*, gives conditions under which the null distribution of $W_n = -2 \log T_n$ is asymptotically of a convenient form.

Theorem 3.2. *Suppose the conditions of Theorem 3.1 hold. Under the setup described in the previous paragraph, $W_n \to \mathsf{ChiSq}(m)$ in distribution, under P_θ with $\theta \in \Theta_0$.*

Proof. We focus here only on the case $d = m = 1$.[2] That is, $\Theta_0 = \{\theta_0\}$ is a singleton, and we want to know the limiting distribution of W_n under P_{θ_0}. Clearly,

$$W_n = -2\ell_n(\theta_0) + 2\ell_n(\hat{\theta}_n),$$

where $\hat{\theta}_n$ is the MLE and ℓ_n is the log-likelihood. By the assumed continuity of the log-likelihood, do a two-term Taylor approximation of $\ell_n(\theta_0)$ around $\hat{\theta}_n$:

$$\ell_n(\theta_0) = \ell_n(\hat{\theta}_n) + \ell_n'(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n) + \frac{\ell_n''(\tilde{\theta}_n)}{2}(\theta_0 - \hat{\theta}_n)^2,$$

where $\tilde{\theta}_n$ is between θ_0 and $\hat{\theta}_n$. Since $\ell_n'(\hat{\theta}_n) = 0$, we get

$$W_n = -\ell_n''(\tilde{\theta}_n)(\theta_0 - \hat{\theta}_n)^2 = -\frac{\ell_n''(\tilde{\theta}_n)}{n}\{n^{1/2}(\hat{\theta}_n - \theta_0)\}^2.$$

From Theorem 3.1, we have that $n^{1/2}(\hat{\theta}_n - \theta_0) \to \mathsf{N}(0, I(\theta_0)^{-1})$ in distribution, as $n \to \infty$. Also, in the proof of that theorem, we showed that $n^{-1}\ell_n''(\tilde{\theta}_n) \to -I(\theta_0)$ under P_{θ_0} for any consistent $\tilde{\theta}_n$. Indeed, we can write

$$\ell_n''(\tilde{\theta}_n) = \ell_n''(\theta_0) + \ell_n''(\tilde{\theta}_n) - \ell_n''(\theta_0),$$

and we have that

$$|\ell_n''(\tilde{\theta}_n) - \ell_n''(\theta_0)| \leq \frac{1}{n}\sum_{i=1}^n \Big| \frac{\partial^2}{\partial\theta^2} \log p_\theta(X_i)\big|_{\theta=\tilde{\theta}_n} - \frac{\partial^2}{\partial\theta^2} \log p_\theta(X_i)\big|_{\theta=\theta_0} \Big|.$$

[2]The proof of the general case is lengthy, so I won't reproduce it here. See Section 17.2 in Keener (2010) or Theorem 7.125 in Schervish (1995) for details. There are also some more general results with algebraic conditions on the null parameter space Θ_0 presented in Drton et al. (2009).

Using Condition C2, the upper bound is bounded by $n^{-1}\sum_{i=1}^{n} M(X_i)\cdot|\tilde{\theta}_n-\theta_0|$, which goes to zero in probability under P_{θ_0} since $\tilde{\theta}_n$ is consistent. Therefore, $\ell_n''(\tilde{\theta}_n)$ has the same limiting behavior as $\ell_n''(\theta_0)$. Finally, by Slutsky, we get

$$W_n \to I(\theta_0)\mathsf{N}(0, I(\theta_0)^{-1})^2 \equiv \mathsf{N}(0,1)^2 \equiv \mathsf{ChiSq}(1). \qquad \square$$

Wilks's theorem facilitates construction of an approximate size-α test of H_0 when n is large, i.e., by rejecting H_0 iff W_n is more than $\chi^2_{m,1-\alpha}$, the $100(1-\alpha)$ percentile of the $\mathsf{ChiSq}(m)$ distribution. The advantage of Wilks' theorem appears in cases where the exact sampling distribution of W_n is intractable, so that an exact (analytical) size-α test is not available. Monte Carlo can often be used to find a test (see Section 3.7), but Wilks's theorem gives a good answer and only requires use of a simple chi-square table. One can also use the Wilks's theorem result to obtain approximation confidence regions. Let $W_n(\theta_0) = -2\log T_n(X;\theta_0)$, where θ_0 is a fixed generic value of the full d-dimensional parameter θ, i.e., $H_0 : \theta = \theta_0$. Then an approximate $100(1-\alpha)\%$ confidence region for θ is $\{\theta_0 : W_n(\theta_0) \le \chi^2_{m,1-\alpha}\}$. An interesting and often overlooked aspect of Wilks's theorem is that the asymptotic null distribution does not depend on the true values of those parameters unspecified under the null. For example, in a gamma distribution problem with the goal of testing if the shape is equal to some specified value, the null distribution of W_n does not depend on the true value of the scale.

3.4 Cautions concerning the first-order theory

One might be tempted to conclude that the desirable properties of the likelihood-based methods presented in the previous section are universal, i.e., that maximum likelihood estimators will "always work." Moreover, based on the form of the asymptotic variance of the MLE and its similarity to the Cramer–Rao lower bound in Chapter 2, it is tempting to conclude that the MLE is asymptotically efficient.[3] However, both of these conclusions are technically *false* in general. Indeed, there are examples where

- the MLE is not unique (Exercise 14) or even does not exist (Exercise 15);
- the MLE "works" (in the sense of consistency), but the conditions of the theory are not met so asymptotic normality fails (Exercise 16); and
- the MLE is *not even consistent!*

Non-uniqueness or non-existence of the MLE are roadblocks to practical implementation but, for some reason, aren't viewed as much of a concern from a theoretical point of view. The case where the MLE works but is not asymptotically normal is also not really a problem, provided that one recognizes the non-regular nature of the problem and makes the necessary

[3]More on efficiency in Chapter 6; a recommended read about this history of likelihood and these theoretical developments is Stigler (2007).

adjustments.[4] The most concerning of these points is inconsistency of the MLE. Since consistency is a rather weak property, inconsistency of the MLE means that its performance is poor and can give very misleading results. The most famous example of inconsistency of the MLE, due to Neyman and Scott (1948), is given next.

Example 3.1 (Neyman and Scott 1948)**.** Let X_{ij} be independent normal random variables, $X_{ij} \sim \mathsf{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, n$ and $j = 1, 2$; the case of two j levels is the simplest, but the result holds for any fixed number of levels. The point here is that X_{i1} and X_{i2} have the same mean μ_i, but there are possibly n different means. The full parameter is $\theta = (\mu_1, \ldots, \mu_n, \sigma^2)$, which is of dimension $n + 1$. It is easy to check that the MLEs are given by

$$\hat{\mu}_i = \tfrac{1}{2}(X_{i1} + X_{i2}), \quad i = 1, \ldots, n$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{i1} - X_{i2})^2.$$

A routine argument (Exercise 17) shows that, as $n \to \infty$, $\hat{\sigma}^2 \to \frac{1}{2}\sigma^2 \neq \sigma^2$ in probability, so that the MLE of σ^2 is inconsistent!

The issue here that is causing inconsistency is that the dimension of the nuisance parameter, the means $\mu_1, \ldots, \mu_n$, is increasing with n. In general, when the dimension of the parameter depends on n, consistency of the MLE will be a concern (see Exercise 20) so one should be careful. Modification of the basic maximum likelihood approach can fix this, see Section 3.6. More generally, these shortcomings of the standard MLE provide motivation for the popular modifications, i.e., shrinkage, penalization, etc.

The fact that maximum likelihood is not necessarily a reliable strategy in general may be surprising to hear. Lucian Le Cam, in a paper from 1960, wrote[5]

> The author is firmly convinced that a recourse to maximum likelihood is justifiable only when one is dealing with families of distributions that are extremely regular. The cases in which ML estimates are readily obtainable and have been proved to have good properties are extremely restricted.

and, later, in his 1986 book, wrote

> The terms "likelihood" and "maximum likelihood" seem to have been introduced by RA Fisher who seems also to be responsible for a great deal of the propaganda on the merits of the maximum likelihood method... In view of Fisher's vast influence, it is perhaps not surprising that the presumed superiority of the method is still for many an article of faith promoted with religious fervor. This state of affairs remains, in spite of a long accumulation of evidence to the effect that maximum likelihood estimates are often useless or grossly misleading.

[4] It turns out that, as far as I know, the theory for "non-regular" problems is not fully understood, i.e., it seems like there are sets of examples where the theory is known but no general results. This is related to the ideas presented in Section 2.4.4.

[5] I got this quote from Section 11 in van der Vaart 2002, a recommended read.

Le Cam's criticisms of Fisher and likelihood-based methods are based on the following point: when likelihood methods work, there are other methods which are just as good, and, when likelihood methods fail, there are other methods that do work. In this light, Le Cam's views are not debatable. However, I think that likelihood is an important object and likelihood-based methods can be useful, if used responsibly. The main point is that it's dangerous to just assume that the likelihood-based methods will work how you expect them to.

3.5 Alternatives to the first-order theory

3.5.1 Bootstrap

Bootstrap is designed to get a sampling distribution for an estimator, which can be used for constructing tests and confidence regions, based on only a single data set. This is a very popular tool, likely due to its simplicity. The first paper on bootstrap is Efron (1979) and Chapter 29 in DasGupta (2008) is a nice summary of the literature up to that point. There are now lots of sophisticated bootstrap techniques and results, but here I will give only the simplest setup, to keep the concepts clear.

Recall that if we have an estimator $\hat{\theta}_n$, this is based on just one data set. The sampling distribution of $\hat{\theta}_n$, which is relevant to the construction of confidence intervals and tests, is based on lots of samples and lots of values of $\hat{\theta}_n$, so obviously it is not available to us. The asymptotic theory in the previous section are concerned with providing a simple approximation of that unknown sampling distribution. Bootstrap, instead, tries to produce an approximate sampling distribution numerically by resampling from the available data.

Let P_θ denote the distribution of an observable X. We observe iid copies $X_1, \ldots, X_n$ from P_θ, and produce an estimator $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$. To learn the sampling distribution of $\hat{\theta}_n$, we would need lots of copies of the sample $(X_1, \ldots, X_n)$. The basic idea behind the bootstrap is to sample, with replacement, from the one available data set. Write such a resample as $X_i^\star$, $i = 1, \ldots, n$; in this case, it is possible that there are ties because sampling is with replacement. Based on $(X_1^\star, \ldots, X_n^\star)$, compute

$$\hat{\theta}_n^\star = \hat{\theta}(X_1^\star, \ldots, X_n^\star).$$

Repeat this process B times, yielding a *bootstrap sample* of B values of $\hat{\theta}_n$, which I will write as $\hat{\theta}_{n,1}^\star, \ldots, \hat{\theta}_{n,B}^\star$. Then the claim is that the distribution of this bootstrap sample is a good approximation of the actual sampling distribution of $\hat{\theta}_n$. For example, we can get a crude 90% confidence interval for θ by taking the 5th and 95th percentiles of the bootstrap sample.

This is very easy to do, since the resampling can be done quickly with a computer. The question is if it works. The basis for the claim that the bootstrap distribution is a good approximation of the sampling distribution is an asymptotic one. Roughly, the claim is that the bootstrap distribution and the sampling distribution merge as $n \to \infty$. To understand this at an intuitive level, recall the fundamental theorem of statistics, namely, that the empirical distribution function converges uniformly almost surely to the true distribution function. Then the resampling step is simply iid sampling from the empirical distribution.

So, if the empirical distribution is close to the true distribution, then sampling from the former should be equivalent to sampling from the latter, hence the claim. Interestingly, some bootstrap methods, perhaps more sophisticated than what is presented here, have "automatic" higher-order accuracy properties (e.g., DiCiccio and Romano 1995).

Despite the simplicity of bootstrap, and the wide range of applications where it works, it is not a tool that works universally. That is, there are known cases where bootstrap fails, so one cannot use it blindly. There are tools available for correcting bootstrap when it fails, but these modifications are not intuitive. See the discussion in DasGupta (2008, Chap. 29) and the references therein for more details on bootstrap.

3.5.2 Monte Carlo and plausibility functions

The focus on asymptotic theory is arguably driven by tradition—when statistics was first being developed, there were no computers available, so only asymptotic analytical approximations were possible. Technology has changed dramatically since then, so an interesting question to ask is if we need asymptotic approximations anymore. That is, why not just crunch everything out exactly using the readily available computing power? I am not suggesting that asymptotic are not useful, but it is important to keep in mind what asymptotics are really for, namely, to help simplify calculations that are too difficult to carry out exactly.

Martin (2015) worked out an approach that is based on getting a Monte Carlo estimate (see Section 3.7.2) of the distribution function of the likelihood ratio statistic; the challenge is that Monte Carlo generally needs be run for lots of parameter values. It can be proved that this approach, which defines a "plausibility function" provides exact inference, without asymptotics. That is, the 95% "plausibility intervals" have coverage probability exactly equal to 0.95. Then the question is, if one has the resources to carry out these computations (which are not bad in simple problems), then why not do so and avoid any asymptotic approximations? That paper shows some examples to demonstrate the efficiency of the method, but there are theoretical and computational problems to be worked out. I should also mention that, although the aforementioned paper doesn't really say anything about this, the approach presented there has some connections to the *inferential model* (IM) framework (e.g., Martin and Liu 2013, 2015a,c). The point is that this approach defines an IM without using a complete specification of the sampling model, which simplifies things a lot, but apparently without sacrificing too much on efficiency; see Martin (2016).

3.6 On advanced likelihood theory

3.6.1 Overview

One goal of advanced likelihood theory is to handle problems with nuisance parameters. For example, suppose that θ splits into a pair of sub-vectors $\theta = (\psi, \lambda)$, where ψ is the parameter of interest and λ is a nuisance parameter, i.e., a parameter that is unknown but not of interest. We would like to make inference on ψ in the presence of unknown λ. This

is a challenging problem because the likelihood is a function of both ψ and λ, and it is not immediately clear how to get rid of λ. For example, if the model is sufficiently regular, then $\hat{\theta}$ is asymptotically normal. One could proceed to make inference on ψ by grabbing $\hat{\psi}$ from $\hat{\theta}$ and the corresponding block from the asymptotic covariance matrix. However, that covariance matrix will generally depend on all of θ, so the question is if plugging in $\hat{\lambda}$ into that covariance matrix is a sufficient way to eliminate λ—I don't think so.

A second goal of advanced likelihood theory is to get more accurate approximations than what is obtained by asymptotic normality of the MLE or Wilks's theorem. The basic tool that drives the proofs of these two results is a two-term Taylor approximation of the log-likelihood. If we take higher-order approximation, effectively handling the remainder terms with more care, then we can often obtain sharper asymptotic approximations. This will lead to more accurate tests and/or confidence regions. The details of these higher-order approximations are beyond our scope, so I only give one example of this. Good/readable references include Young and Smith (2005, Chap. 9) and Brazzale et al. (2007, Chap. 2); a nice overview of this kind of advanced asymptotics is given in Reid (2003).

3.6.2 "Modified" likelihood

Write $\theta = (\psi, \lambda)$, where ψ is the parameter of interest and λ is an unknown nuisance parameter. How to construct a likelihood for ψ alone? There are basically three possible techniques; the first of which you would have seen tastes of in a first course on statistical theory, in the context of likelihood ratio tests.

Profile likelihood

Suppose, for the moment, that ψ was known and only λ was unknown. In which case, we could find the MLE for λ, given this known value of ψ, which we denote by $\hat{\lambda}_\psi$. This can be done for any value of ψ, so we may write

$$L_p(\psi) = L(\psi, \hat{\lambda}_\psi),$$

called the profile likelihood of ψ. This function can, generally, be treated like a genuine likelihood function. For example, Wilks's theorem shows that -2 times the log profile likelihood ratio is asymptotically chi-square under a given H_0. Some non-asymptotic use of the profile likelihood ratio is discussed in Martin (2015). However, a word of caution is in order: in the Neyman–Scott example above, the MLE $\hat{\sigma}^2$ is the "profile maximum likelihood estimator" and is inconsistent!

Marginal and conditional likelihood

Much of the material in this section is taken from Section 2.4 in Boos and Stefanski (2013). For data X, let (S, T) be a one-to-one function of X; of course, the function is not allowed to depend on the parameter $\theta = (\psi, \lambda)$. So, in terms of likelihoods, with some abuse of notation, we can write $p_\theta(X) = p_\theta(S, T)$. Of course, the right-hand side, which is a joint

density for T and S, can be factored into a product of a marginal density and a conditional density. Suppose that either

$$p_\theta(T, S) = p_\theta(T \mid S) p_\psi(S) \tag{3.6}$$
$$p_\theta(T, S) = p_\psi(T \mid S) p_\theta(S). \tag{3.7}$$

We consider the two cases separately.

- In the former case, Equation (3.6), the marginal distribution of S depends only on the interest parameter, so we can take as a "modified" or "pseudo" likelihood

  $$L_m(\psi) = p_\psi(S).$$

  This is called a *marginal likelihood* since it is based on the marginal distribution a function $S = S(X)$. Note, however, that this is not a real likelihood because some information relevant to ψ, contained in the conditional part $p_\theta(T \mid S)$, has been thrown out; the hope, however, is that the elimination of the nuisance parameter λ through marginalization will lead to advantages that outweigh the loss of information.
- In the latter case, Equation (3.7), the conditional distribution of T given S does not depend on λ, so we can take as a modified or pseudo likelihood

  $$L_c(\psi) = p_\psi(T \mid S).$$

  Again, this is not a real likelihood since some information has been thrown out, but the elimination of the nuisance parameter has value.

In the Neyman–Scott example, consider the transformation $X = (X_{ij})$ to

$$S_i = 2^{-1/2}(X_{i1} - X_{i2}) \quad \text{and} \quad T_i = 2^{-1/2}(X_{i1} + X_{i2}), \quad i = 1, \ldots, n.$$

Then the following properties are easy to check:

- the marginal distribution of $S = (S_1, \ldots, S_n)$ does not depend on $(\mu_1, \ldots, \mu_2)$;
- $S_1, \ldots, S_n$ are iid $\mathsf{N}(0, \sigma^2)$;
- and S and T are independent.

Therefore, conditioning has no effect so the marginal/conditional likelihood for the interest parameter σ^2, based on S only, is

$$L_m(\sigma^2) \propto (\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n S_i^2}.$$

It is straightforward to check that the marginal/conditional maximum likelihood estimator of σ^2 is $\tilde{\sigma}^2 = (2/n)\sum_{i=1}^n S_i^2 = 2\hat{\sigma}^2$. Since $\hat{\sigma}^2 \to \frac{1}{2}\sigma^2$, it is clear that $\tilde{\sigma}^2$ is consistent.

The obvious challenge in implementing a marginal or conditional likelihood approach is finding the appropriate statistics (S, T). Unfortunately, there really aren't' any general

strategies for finding this, experience is the only guide. Probably the only semi-general strategy that one can apply to obtain a conditional likelihood is the following, that applies for certain exponential family models. Suppose that the density function for X is of the form

$$p_\theta(x) \propto \exp\{\langle \psi, T(x)\rangle + \langle \lambda, S(x)\rangle - A(\psi, \lambda)\}.$$

Then it is pretty easy to see that the conditional distribution of $T = T(X)$, given $S = S(X)$, will be of exponential family form and will not depend on the nuisance parameter λ. Boos and Stefanski (2013), Section 2.4.6, give a really nice application of this strategy in the context of logistic regression; see, also, Section 5 of Bølviken and Skovlund (1996).

3.6.3 Asymptotic expansions

The two primary technical tools are the *Edgeworth* and *saddlepoint* expansions. Both of these are based on cumulant generating functions—log of the moment generating function. The idea is to approximate the density $p_n(s)$ of the normalized sum $S_n = (\sum_{i=1}^n X_i - n\mu)/\sqrt{n\sigma^2}$, where $X_1, \ldots, X_n$ are iid with mean μ and variance σ^2. The central limit theorem says $p_n(s) \to \varphi(s)$, the standard normal density, as $n \to \infty$. These expansions are designed to get a more accurate approximation of $p_n(s)$ for finite n. We will not discuss the general details here—see, e.g., DasGupta (2008)—just an application.

For a given sample, suppose the minimal sufficient statistic S for θ can be expressed as (T, A), where T is the MLE and A is an ancillary statistic. Since A is ancillary, it is natural to base inference on the sampling distribution of T, given $A = a$, where a is the observed value of A. Despite this being a natural thing to do, it is generally not so easy to compute this conditional distribution. The $p^\star$ *formula*, based on the asymptotic expansions above, provides a very good approximation to this conditional distribution.

Write the log-likelihood function as $\ell(\theta; t, a)$ and the observed Fisher information matrix (Exercise 19) $J(\theta; t, a)$. The $p^\star$ formula is then

$$p^\star_\theta(t \mid a) = c(\theta, a) |\det\{J(t; t, a)\}|^{1/2} e^{\ell(\theta; t, a) - \ell(t; t, a)},$$

where $c(\theta, a)$ is a normalizing constant that does not depend on t. Then the claimed approximation result is, for any t, as $n \to \infty$,

$$p_\theta(t \mid a) = p^\star_\theta(t \mid a)\{1 + O(n^{-1})\};$$

that is, the exact conditional density $p_\theta(t \mid a)$ of T, given $A = a$, equals $p^\star_\theta(t \mid a)$ modulo an error that vanishes at the rate of n^{-1}. This comes from a saddlepoint expansion which comes with general approximation bounds. For some problems, including group transformation problems, the $p^\star$ formula is exact.

There are also very nice approximations of the distribution function of a statistic, e.g., the $r^\star$ approximation explained in Reid (2003). This, unfortunately, is a bit too technical for us to consider here. But there are other tricks to improve asymptotic approximations, such as the Bartlett correction (Exercise 21), which are relatively easy to use.

3.7 A bit about computation

3.7.1 Optimization

Newton's method is a simple and powerful tool for doing optimization or, more precisely, root finding. You should be familiar with this method from a calculus course. The idea is based on the fact that, locally, any differentiable function can be suitably approximated by a linear function. This linear function is then used to define a recursive procedure that will, under suitable conditions, eventually find the desired solution.

Recall the likelihood equation (3.2). Then the MLE is a solution to this equation, i.e., a root of the gradient of the log-likelihood function. Assume that the gradient $\nabla\ell(\theta)$ is also differentiable, and let $D(\theta)$ denote that matrix of derivatives, i.e., $D(\theta)_{ij} = (\partial^2/\partial\theta_i\partial\theta_j)\ell(\theta)$. Assume $D(\theta)$ is non-singular for all θ. The idea behind Newton's method is as follows. Pick some guess , say $\theta^{(0)}$ of the MLE $\hat{\theta}$. Now approximate $\nabla\ell(\theta)$ by a linear function:

$$\nabla\ell(\theta) = \nabla\ell(\theta^{(0)}) + D(\theta^{(0)})(\theta - \theta^{(0)}) + \text{error}.$$

Ignore the error, solve for θ, and call the solution $\theta^{(1)}$:

$$\theta^{(1)} = \theta^{(0)} - D(\theta^{(0)})^{-1}\nabla\ell(\theta^{(0)}).$$

If $\theta^{(0)}$ is close to the solution of the likelihood equation, then so will $\theta^{(1)}$ (draw a picture!). The idea is to iterate this process until the solutions converge. So the method is to pick a "reasonable" starting value $\theta^{(0)}$ and, at iteration $t \geq 0$ set

$$\theta^{(t+1)} = \theta^{(t)} - D(\theta^{(t)})^{-1}\nabla\ell(\theta^{(t)}).$$

Then stop the algorithm when t is large and/or $\|\theta^{(t+1)} - \theta^{(t)}\|$ is small.

There are lots of tools available for doing optimization, the Newton method described above is just one simple approach. Fortunately, there are good implementations of these methods already available in the standard software. For example, the routine `optim` in R is a very powerful and simple-to-use tool for generic optimization. For problems that have a certain form, specifically, problems that can be written in a "latent variable" form, there is a very clever tool called the EM algorithm (e.g. Dempster et al. 1977) for maximizing the likelihood. Section 9.6 in Keener (2010) gives some description of this method.

An interesting and unexpected result is that sometimes optimization can be used to do integration. The technical result I'm referring to is the *Laplace Approximation*, and some further comments on this will be made in Chapter 4 on Bayesian methods.

3.7.2 Monte Carlo integration

In hypothesis testing, suppose $H_0 : \theta = \theta_0$, i.e., the null gives a complete specification of the parameter. In this case, it is straightforward to derive exact tests using Monte Carlo.

That is, sample lots of data sets $(X_1, \ldots, X_n) \overset{\text{iid}}{\sim} \mathsf{P}_{\theta_0}$ and, for each data set, compute the corresponding W_n, or any other test statistic for that matter. A size-α test says

$$\text{Reject } H_0 \text{ iff } W_n > k_\alpha,$$

where k_α depends on the null distribution of W_n. Now choose k_α as the $100(1-\alpha)$ percentile from the Monte Carlo sample of W_n's. This is easy to do; see Exercise 23.

The challenge is trying to do something similar when the null hypothesis leaves some components of θ unspecified. In this case, it is not clear what distribution you should sample from in the Monte Carlo step. For example, if $\theta = (\theta_1, \theta_2)$ and the null hypothesis specifies the value $\theta_{1,0}$ of θ_1, then what value of θ_2 should be used to simulate? In general, the null distribution of W_n in this case, at least for finite n, would depend on the particular value of θ_2, so this is an important question. In some cases, one can show that the distribution of W_n does not depend on the unspecified parameter (see Exercise 24), but this can be difficult to do. Martin (2015) has some relevant comments on this.

3.8 Discussion

Likelihood, likelihood-based methods, and the desirable large-sample results presented above, have been important to the development of statistical practice. It is an interesting question, however, if the likelihood function is fundamental to statistics. We understand that, in a certain sense, the likelihood function contains all the relevant information in the data concerning the unknown parameter. But, the statistical methods describe above (e.g., the likelihood ratio test) use information beyond what is contained in the likelihood function. In particular, the sampling distribution of the relevant statistic is needed to choose the test's cutoff. That the results of a statistical procedure depend on things other than the observed data is somewhat controversial.

To fully buy in to the claim that the likelihood contains all relevant information in data about the parameter, one must be willing to say that any two data sets that produce the same likelihood function (up to proportionality constant) should produce the same results concerning the parameter of interest. Such a position has a name—the *likelihood principle*. Those classical methods that depend on sampling distributions, including the likelihood ratio tests discussed above, violate the likelihood principle. Besides the likelihood principle, there are two other statistical principles that have taken hold:

- *Sufficiency principle:* Any two data sets that admit the same (minimal) sufficient statistics should lead to the same conclusions about θ.
- *Conditionality principle:* If two different experiments are to be considered, and the choice between the two is random, and the randomization does not depend on θ, then conclusions about θ should be based only the experiment actually performed.

These two principles are difficult to argue with and, historically, this have been mostly accepted as reasonable principles. (I am, of course, glossing over some non-trivial details so

that I can paraphrase the main point.) There is a famous result of Birnbaum (1962), arguably the most controversial in all of statistics, that says that statistical inference that follows the sufficiency and conditionality principles must also follow the likelihood principle. This suggests that those statistical methods based on sampling distributions (e.g., the likelihood ratio tests), which violate the likelihood principle, must also violate either the sufficiency or conditionality principles. To summarize: Birnbaum's result implies that frequentist methods are "illogical" in this specific sense.

Birnbaum's result has had some significant effect, in particular, to the development and acceptance of Bayesian methods. The fact is, the only known statistical approach which satisfies the likelihood principle is the Bayesian approach (with a subjective prior). For example, Jimmie Savage, in his discussion of Birnbaum's paper, writes

> I, myself, came to take ... Bayesian statistics ... seriously only through recognition of the likelihood principle.

That is, if not for Birnbaum's result on the likelihood principle, Savage never would have taken the Bayesian approach seriously. So, in this way, Birnbaum's result is indirectly responsible for much of the developments in Bayesian statistics.

As you are reading this, you should be feeling a bit uncomfortable: *those classical methods taught in Stat 101 courses everywhere violate some logical principles!?* There has always been doubt about the validity of Birnbaum's claim and, recently, it has been shown that it is actually *false*! See Evans (2013) and Mayo (2014). Besides relieving statistics of the constraints of Birnbaum's claim, these developments open the door to some new ideas about the foundations of statistics; see Martin and Liu (2014).

3.9 Exercises

1. Let $X_1, \ldots, X_n$ be iid $\mathsf{N}(\mu, \sigma^2)$. Find the MLE of (μ, σ^2).

2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Gamma}(\theta, 1)$, where the shape parameter $\theta > 0$ is unknown.

   (a) There is no closed-form expression for the MLE $\hat{\theta}$. Write R code to find the MLE numerically; try to do it without using any built-in optimization routines.

   (b) Simulate 1000 data sets (with $\theta = 7$ and $n = 10$) and, for each, calculate the MLE. Summarize your simulation with a histogram that describes the sampling distribution of the MLE.

   (c) Does your sampling distribution appear to be approximately normal?

3. Maximum likelihood estimators have a desirable *invariance* property. That is, if $\hat{\theta}$ is the MLE of θ, and $\eta = g(\theta)$ is some transformation, then the MLE of η is $\hat{\eta} = g(\hat{\theta})$. Explain the intuition behind this fact.

4. For iid data $X_1, \ldots, X_n$, let $\ell_n(\theta)$ be the log-likelihood.

(a) Use Jensen's inequality and the law of large numbers to argue that, for any non-zero a, $\ell_n(\theta^\star + a) - \ell_n(\theta^\star) < 0$ for all large n with $\mathsf{P}_{\theta^\star}$-probability 1.

(b) Use this to argue for consistency of a sequence of solutions of the likelihood equation. [Hint: Fix $\varepsilon > 0$ and apply (a) with $a = \varepsilon$ and $a = -\varepsilon$.]

5. Suppose that the density p_θ of X satisfies the conditions of the factorization theorem from Chapter 2. Use Theorem 2.3 to show that if the MLE $\hat{\theta} = \hat{\theta}(X)$ is unique and sufficient, then it is also minimal sufficient.

6. Prove the following version of the *Continuous Mapping Theorem*:

> Let $X, \{X_n : n \geq 1\}$ be random variables taking values in a metric space $(\mathbb{X}, |\cdot|)$, such as $\mathbb{R}^d$. Let g be an everywhere continuous function that maps $\mathbb{X}$ to another metric space $\mathbb{X}'$. Show that if $X_n \to X$ in probability, then $g(X_n) \to g(X)$ in probability.

Remarks: (i) g does not need to be everywhere continuous, it is enough that X is in the continuity set with probability 1, and (ii) the same result holds if you replace convergence in probability with convergence in distribution or convergence with probability 1.

7. (a) Consider two sequences of random variables, X_n and Y_n, such that $X_n \to X$ and $Y_n \to c$, both in distribution, where X is a random variable and c is a constant. Prove that $(X_n, Y_n) \to (X, c)$ in distribution. [Hint: By definition, a sequence of random variables X_n converges in distribution to X iff $\mathsf{E}f(X_n) \to \mathsf{E}f(X)$ for all bounded and continuous functions f.]

(b) Use (a) and Continuous Mapping Theorem to prove *Slutsky's Theorem*:

> If $A_n \to a$ and $B_n \to b$, both in probability, and $X_n \to X$ in distribution, then $A_n + B_n X_n \to a + bX$ in distribution.

[Hints: (i) Convergence in distribution and convergence in probability are equivalent when the limit is a constant, and (ii) to apply part (a), think of (A_n, B_n) as one sequence, with constant limit (a, b).]

(c) As a special case of Slutsky's Theorem, show that if $Y_n \to Y$ in distribution and B_n is an event with $\mathsf{P}(B_n) \to 1$, then $Y_n I_{B_n} + Z_n I_{B_n^c} \to Y$ for any sequence Z_n of random variables.

8. Prove the *Delta Theorem*:

> For random variables T_n, assume that $n^{1/2}(T_n - \theta) \to \mathsf{N}(0, v(\theta))$ in distribution, where $v(\theta)$ is the asymptotic variance. Let $g(\cdot)$ be a function differentiable at θ, with $g'(\theta) \neq 0$. Then $n^{1/2}\{g(T_n) - g(\theta)\} \to \mathsf{N}(0, v_g(\theta))$, in distribution, where $v_g(\theta) = [g'(\theta)]^2 v(\theta)$.

9. The delta theorem in the previous exercise assumes that g' exists and is non-zero at θ. What happens when $g'(\theta) = 0$?

   Theorem. Assume $g'(\theta) = 0$, g'' is continuous, and $g''(\theta) \neq 0$. Then there exists a sequence of constants c_n and a function $h(\cdot)$ such that

   $$c_n h(\theta)[g(\hat{\theta}_n) - g(\theta)] \to \mathsf{ChiSq}(1) \quad \text{in distribution, } n \to \infty.$$

   (a) Prove the theorem and identify the particular c_n and $h(\theta^\star)$. [Hint: Quadratic Taylor approximation of $g(\hat{\theta}_n)$ at θ and continuous mapping theorem.]

   (b) Give an example of an iid model, an estimator $\hat{\theta}_n$, a true θ, and a function g such that the chi-square approximation above is exact. [Hint: $\mathsf{N}(0,1)^2 = \mathsf{ChiSq}(1)$.]

10. Let $\hat{\theta}_n$ be a sequence of estimators such that $\sqrt{n}(\hat{\theta}_n - \theta) \to \mathsf{N}(0, v(\theta))$ in distribution, where $v(\theta) > 0$ is the (asymptotic) variance function.

    (a) A *variance stabilizing transformation* is a function g such that the asymptotic variance of $g(\hat{\theta}_n)$ does not depend on θ. Use the delta theorem to find the condition required for a function g to be variance stabilizing.

    (b) Let λ be a fixed real number, and suppose that $v(\theta) = \theta^{2(1-\lambda)}$, $\theta > 0$. Use the sufficient condition you derived in Part (a) to show that

    $$g(\theta) = \begin{cases} (\theta^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\ \log \theta & \text{if } \lambda = 0, \end{cases}$$

    is variance stabilizing. (This is the *Box–Cox transformation*.)

    (c) Suppose that $X_1, \ldots, X_n$ are iid with density $p_\theta(x) = (1/\theta)e^{-x/\theta}$, $x > 0$, $\theta > 0$. The maximum likelihood estimator $\hat{\theta}_n$ is asymptotically normal, with variance function $v(\theta)$ of the form in Part (b). Find the corresponding λ and g.

    (d) In the context of Part (c), use the asymptotic distribution of $g(\hat{\theta}_n)$ to find an asymptotically correct $100(1-\alpha)\%$ confidence interval for θ.

11. Consider an exponential family density $p_\theta(x) = h(x)e^{\eta(\theta)T(x) - A(\theta)}$. Under what conditions can you find a function $M(x)$ satisfying (3.3)?

12. Show that (3.4) implies existence of a function $M(x)$ satisfying (3.3).

13. Let $X_1, \ldots, X_n$ be iid exponential observations with unknown mean θ.

    (a) Find the exact size-α likelihood ratio test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

    (b) Find the approximate size-α test based on Wilks's theorem.

    (c) Plot the power functions of the two tests above and compare.

14. *Non-uniqueness of the MLE.* Let $X_1, \ldots, X_n$ be iid with density $p_\theta(x) = 2e^{-|x-\theta|}$, $x \in \mathbb{R}$, $\theta \in \mathbb{R}$. This is called a shifted Laplace (or double-exponential) distribution.

    (a) Argue that the MLE of θ is not unique.

    (b) To verify this, take $\theta = 0$, simulate $n = 10$ observations from the Laplace distribution, plot the likelihood, and identify the flat peak. [Hint: To simulate from the standard Laplace distribution, simulate a standard exponential and then flip a fair coin to decide if the sign should be positive or negative.]

15. *Non-existence of MLE.* Consider a mixture of two normal distributions, i.e.,

$$\pi \mathsf{N}(\mu_1, \sigma_1^2) + (1 - \pi)\mathsf{N}(\mu_2, \sigma_2^2).$$

    Suppose $X_1, \ldots, X_n$ are iid from the above mixture. Argue that the MLE of $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \pi)$ does not exist.[6] [Hint: What happens to the likelihood, as a function of σ_1 if $\mu_1 = X_1$, say?]

16. Let $X_1, \ldots, X_n$ be iid $\mathsf{Unif}(0, \theta)$.

    (a) Show that the MLE is $\hat{\theta} = \max X_i$.

    (b) Explain why the MLE cannot be asymptotically normal and why this is not a counter-example to the theory presented in Section 3.3.

    (c) Show that $n(\theta - \hat{\theta})$ converges is distribution to $\mathsf{Exp}(\theta)$.

17. Refer to Example 3.1, the Neyman–Scott problem.

    (a) Derive the stated MLE $\hat{\sigma}^2$ for σ^2.

    (b) Show that $\hat{\sigma}^2$ is inconsistent.

18. Suppose $X_1, \ldots, X_n$ are independent, with $X_i \sim \mathsf{N}(\theta_i, 1)$, $i = 1, \ldots, n$. In vector notation, we may write $X \sim \mathsf{N}_n(\theta, I_n)$, where $X = (X_1, \ldots, X_n)^\top$ is the observable and $\theta = (\theta_1, \ldots, \theta_n)^\top$ is the unknown mean vector.

    (a) Use Exercise 3 to find the MLE of $\psi = \|\theta\|^2$, the squared length of θ.

    (b) Show that the MLE of ψ is biased.

    (c) Does the bias above disappear as $n \to \infty$? Explain what's going on.

19. For iid data, the asymptotic variance of the MLE $\hat{\theta}$ is $[nI(\theta)]^{-1}$ and, for constructing confidence intervals, one needs an estimate of $nI(\theta)$. A reasonable choice is $nI(\hat{\theta})$, but

[6] Mixture models can be very difficult creatures, and often serve as good counterexamples for properties, like existence of MLEs, that hold for "regular" problems. But despite difficulties, mixture models are actually very useful in both theory and applications.

Table 3.1: The data of Rubin (1981) on SAT coaching experiments.

School (i)	Treatment Effect (X_i)	Standard Error (σ_i)
1	28.39	14.9
2	7.94	10.2
3	−2.75	16.3
4	6.82	11.0
5	−0.64	9.4
6	0.63	11.4
7	18.01	10.4
8	12.16	17.6

there is something else that is often easier to get and works at least as good. For scalar θ and log-likelihood $\ell(\theta)$, define the *observed Fisher information* $J(\theta)$ as

$$J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta^2}.$$

There is a multi-parameter version as well, where $J(\theta)$ is the matrix of negative second partial derivatives of the log-likelihood. The claim is that $J(\hat{\theta})$ is a good estimate of $nI(\theta)$ compared to $nI(\hat{\theta})$. Suppose we have samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid from a bivariate distribution with density $p_\theta(x, y) = e^{-x/\theta - \theta y}$, $x, y > 0$, $\theta > 0$. In this case, $\hat{\theta}$, $nI(\hat{\theta})$, and $J(\hat{\theta})$ can be found analytically. Use simulations to compare the sampling distributions of $[nI(\hat{\theta})]^{1/2}(\hat{\theta} - \theta)$ and $[J(\hat{\theta})]^{1/2}(\hat{\theta} - \theta)$.

20. Suppose $X_1, \ldots, X_n$ are independent, with $X_i \sim \mathsf{N}(\lambda, \sigma_i^2 + \psi)$, where $\sigma_1, \ldots, \sigma_n$ are known, but $\theta = (\psi, \lambda)$ is unknown. Here $\psi \geq 0$ is the parameter of interest and λ is a nuisance parameter. For the data in Table 3.1, taken from the SAT coaching study in Rubin (1981), find and plot the profile likelihood function for ψ.

21. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$ and consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

    (a) Show that the likelihood ratio statistic $W = W(\mu_0)$ can be expressed as $W = n \log\{1 + T^2/(n-1)\}$, where $T = n^{1/2}(\bar{X} - \mu_0)/S$ is the usual t-statistic.

    (b) Show that $\mathsf{E}(W) = 1 + bn^{-1} + O(n^{-2})$, where $b = 3/2$. [Hint: Find (e.g., on `wikipedia`) formulas for even moments of Student-t random variables.]

    (c) Define $W_b = W/(1 + bn^{-1})$; this is called the *Bartlett corrected* likelihood ratio statistic. Compare, using simulations, the accuracy of the $\mathsf{ChiSq}(1)$ approximations for W and W_b for relatively small values of n; you may take $\mu = \mu_0 = 0$ and $\sigma = 1$.

22. *One-step estimation* is a method by which a consistent estimator is updated, via a single iteration of Newton's method in Section 3.7, to get an asymptotically efficient

estimator. That is, let $\hat{\theta}_0$ be a consistent estimator of θ, and define the one-step version: $\hat{\theta}_1 = \hat{\theta}_0 - D(\hat{\theta}_0)^{-1}\nabla\ell(\hat{\theta}_0)$, where $D(\theta)$ is the matrix of second derivatives of $\ell(\theta)$. It can be shown that $\hat{\theta}_1$ is asymptotically efficient, like the MLE. As a simple example of this, suppose $X_1, \ldots, X_n$ are iid $\mathsf{N}(\theta, 1)$. Take $\hat{\theta}_0$ to be the sample median, which is consistent but not asymptotically efficient. Find the one-step version $\hat{\theta}_1$ and argue for its efficiency.

23. In the setup of Exercise 2, consider testing $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$. Use Monte Carlo to find the cutoff k_α for the size-α likelihood ratio test. Take $n = 10$, $\alpha = 0.05$, and produce a Monte Carlo sample of size $M = 5000$. Compare your cutoff k_α with that based on the large-sample chi-square approximation.

24. Consider a general location-scale model, i.e., where X has density $\sigma^{-1}p(\sigma^{-1}(x-\mu))$, where $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$, and p is a density on $\mathbb{R}$. Given data $X_1, \ldots, X_n$ iid from this model, suppose the goal is to test $H_0 : \sigma = \sigma_0$ versus $H_1 : \sigma \neq \sigma_0$. Show/argue that the (exact, not asymptotic) null distribution of the likelihood ratio statistic does not depend on μ.

25. The mathematical formulation of Birnbaum's theorem on the likelihood principle and Evans's clarification thereof involves *equivalence relations*. An equivalence relation $\sim$ on $\mathbb{X}$ is a binary relation that satisfies:

    Reflexivity: $x \sim x$ for all $x \in \mathbb{X}$.
    Symmetry: If $x \sim y$, then $y \sim x$.
    Transitivity: If $x \sim y$ and $y \sim z$, then $x \sim z$.

    (a) One of the most common examples of equivalence relations in statistics is equality of μ-measurable functions up to sets of μ-measure zero. That is, write $f \sim g$ if $f = g$ μ-almost everywhere. Prove that $\sim$ is an equivalence relation on the set of all μ-measurable functions.

    (b) Another example of equivalence relations appears in group theory. Consider a group $\mathscr{G}$ of transformations $g : \mathbb{X} \to \mathbb{X}$. Write $x \sim y$ if there exists $g \in \mathscr{G}$ such that $y = gx$. Prove that $\sim$ is an equivalence relation on $\mathbb{X}$.

    (c) Let $\mathbb{X}$ be a set equipped with an equivalence relation $\sim$. Given $x \in \mathbb{X}$, define the equivalence class $E_x = \{y \in \mathbb{X} : y \sim x\}$, the set of all things equivalent to x. For two x, y in show that E_x and E_y are either disjoint or exactly equal. That is, $\sim$ induces a partition $\{E_x : x \in \mathbb{X}\}$ of $\mathbb{X}$ into equivalence classes.

Chapter 4

Bayesian Inference

4.1 Introduction

The classical frequentist approach to statistics is one that Stat 511 students are familiar with. That is, for a given procedure—estimator, test, confidence interval, etc—the frequentist is interested in the performance of that procedure in terms of repeated sampling. For example, the quality of a test is measured by its power function, which is nothing but the limiting proportion of times the test rejects the null hypothesis when sampling from a distribution contained in the alternative hypothesis. This is fine, but it's important to understand the limitations of such considerations. In particular, the power function for a test provides no comfort when a fixed set of data is available and you want to measure uncertainty about the truthfulness of the null hypothesis. So, there is some reason to look for a different approach, one that might allow you to report a sort of *probability* that the null hypothesis is true, given the observed data. A Bayesian approach makes this possible, but we need to look at the problem from a very different perspective.

By the time students reach Stat 511, they surely know something about the Bayesian approach. For example, I'm sure everyone knows that, in the Bayesian context, the unknown parameter is treated as a random variable, with a prior distribution, and Bayes's theorem is used to produce a posterior distribution. But it is natural to ask why the parameter, which is some fixed but unknown quantity, should be treated as random. For example, it seems foolish to assume that the mean income in Cook County is selected at random, right? This is a subtle but important point. The justification for the Bayesian approach is based on the following sort of "axiom:"

Uncertainties can only be described with probability.

This means that, for anything we don't know—e.g., the parameter θ in a statistical problem—the only logical way to describe our beliefs is with probability. This is what sets the Bayesian approach apart from the classical approach. In the latter, θ is assumed fixed but unknown. But what does it mean for θ to be "unknown?" Do we really know nothing about it, do we not know how to summarize what knowledge we have, or are we uneasy using this knowledge?

It seems unrealistic that we actually know *nothing* about θ. For example, if θ is the mean income in Cook county, we know that θ is positive and less than \$1 billion; we'd also believe that $\theta \in (\$40K, \$60K)$ is more likely that $\theta \in (\$200K, \$220K)$. If, for each event concerning θ, we assign some numerical score that represents our uncertainty, and those scores satisfy certain consistency properties,[1] then we have effectively assigned a probability distribution on the parameter space. This is the thing called the *prior distribution*.

What is particular interesting about this argument is that there is no notion of repeated sampling, etc, that we are used to seeing in a basic probability course. That is, this prior distribution is simply a description of one's own uncertainty, and need not have anything to do with *chance*, per se. This is not as foreign as it might originally seem. For example, suppose you and several friends have been invited to a party next saturday. When your friend asks if you will attend, you might respond with something like "there's a 50–50 chance that I'll go." Although not on the scale of probabilities, this has such an interpretation. The same thing goes for weather reports, e.g., "there's a 30% chance of rain tomorrow." Note that these events are different from the kind of experiments that can be repeated over and over, like rolling a die, and yet probabilities can be defined. Fortunately, these subjective probabilities can be manipulated just like ordinary frequency-based probability.

So, in the statistics problem, we are uncertain about the parameter θ. We then describe our uncertainty with (subjective) probabilities. That is, we assign probabilities to events like $\{\theta > 7\}$, $\{-0.33 \leq \theta < 0.98\}$, etc, which describes the prior distribution Π for θ. This is effectively the same as assuming that the unknown parameter itself is a random variable with a specified distribution. It is a common misconception to say that Bayesian analysis assumes the parameter is a random variable. On the contrary, a Bayesian starts by assigning probabilities to all things which are uncertain; that this happens to be equivalent to taking θ to be a random variable is just a consequence.

Having some basic understanding of the logic behind the Bayesian approach, in the rest of this chapter we will investigate some of the specifics of Bayesian analysis. Here we will not get into any philosophical discussions about Bayesian versus non-Bayesian approaches, but there have been such discussions for many years.[2] Here I will describe first, the Bayes model and how the prior is updated to a posterior distribution via Bayes theorem. Then we will discuss how this posterior distribution is used for inference and give some examples. Next I will attempt to describe several motivations for a Bayesian analysis. If one elects to use a Bayesian analysis, then perhaps the most important question is how to choose the prior. There are a number of now fairly standard methods, which I will briefly describe.

[1] The consistency properties are quite reasonable, e.g., if one event is a subset of another, then the former cannot have a greater score than the latter, but constructing such a system of scoring from scratch is not so easy; usually it's done by assuming a particular probability model.

[2] Nowadays, most people understand that both Bayes and non-Bayes approaches have their advantages and disadvantages, i.e., neither is clearly better than the other. So the discussion is more about making the best use of the tools available in a given problem.

4.2 Bayesian analysis

4.2.1 Basic setup of a Bayesian inference problem

Just as before, start with a sample (measurable) space $(\mathbb{X}, \mathcal{A})$ which is equipped with a family of probability distributions $\mathsf{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$. Suppose also that there exists a σ-finite measure μ such that $\mathsf{P}_\theta \ll \mu$ for all θ, so that we have Radon–Nikodym derivatives (densities) $p_\theta(x) = (d\mathsf{P}_\theta/d\mu)(x)$ with respect to μ. The difference is that some probability distribution Π on Θ is also available from somewhere. We call Π the *prior distribution*. To help keep track of what's random and what's fixed, I will use the notation Θ for a random variable distributed according to Π, and θ for the observed values. There shouldn't be any confusion in using the notation Θ for both the parameter space and the random variable version of the parameter.

The Bayesian setup assumes the following hierarchical model:

$$\Theta \sim \Pi \quad \text{and} \quad X \mid (\Theta = \theta) \sim p_\theta(x). \tag{4.1}$$

The goal is to take the information from the observed $X = x$ and update the prior information about the "parameter" Θ. This is accomplished quite generally via Bayes' theorem. But before seeing the technical stuff, it helps to understand the reasoning behind this particular choice. If uncertainty about θ is described by the (subjective) probability distribution Π, then the uncertainty about θ *after seeing data* x should be described by the conditional distribution Π_x, the *posterior* distribution of Θ given $X = x$. We'll discuss how this posterior distribution is used for inference shortly.

4.2.2 Bayes's theorem

We are all familiar with Bayes's theorem from an introductory probability course. In simple presentations, the theorem provides a formula for the probability $\mathsf{P}(A \mid B)$ in terms of the opposite conditional probability $\mathsf{P}(B \mid A)$ and the marginal probabilities $\mathsf{P}(A)$ and $\mathsf{P}(B)$. Here we give a very general measure-theoretic version of this result.

Theorem 4.1 (Bayes's Theorem)**.** *Under the setup described above, let Π_x denote the conditional distribution of Θ given $X = x$. Then $\Pi_x \ll \Pi$ for P_Π-almost all x, where $\mathsf{P}_\Pi = \int \mathsf{P}_\theta \, d\Pi(\theta)$ is the marginal distribution of X from model (4.1). Also, the Radon–Nikodym derivative of Π_x with respect to Π is*

$$\frac{d\Pi_x}{d\Pi}(\theta) = \frac{p_\theta(x)}{p_\Pi(x)},$$

for those x such that the marginal density $p_\Pi(x) = (d\mathsf{P}_\Pi/d\mu)(x)$ is neither 0 nor ∞. Since the set of all x such that $p_\Pi(x) \in \{0, \infty\}$ is a P_Π-null set, the Radon–Nikodym derivative can be defined arbitrarily for such x.

Proof. This proof comes from Schervish (1995, p. 16–17). Define

$$C_0 = \{x : p_\Pi(x) = 0\} \quad \text{and} \quad C_\infty = \{x : p_\Pi(x) = \infty\}.$$

Since $\mathsf{P}_\Pi(A) = \int_A p_\Pi(x)\, d\mu(x)$, it follows that

$$\begin{aligned} \mathsf{P}_\Pi(C_0) &= \int_{C_0} p_\Pi(x)\, d\mu(x) = 0 \\ \mathsf{P}_\Pi(C_\infty) &= \int_{C_\infty} p_\Pi(x)\, d\mu(x) = \int_{C_\infty} \infty\, d\mu(x). \end{aligned}$$

The last integral will equal ∞ if $\mu(C_\infty) > 0$; but since the last quantity cannot equal ∞ (it's a probability), it must be that $\mu(C_\infty) = 0$ and, hence, $\mathsf{P}_\Pi(C_\infty) = 0$. This proves the last statement in the theorem about the denominator.

To prove the main claim, recall that the posterior Π_x must satisfy

$$\mathsf{P}(\Theta \in B,\, X \in A) = \int_A \Pi_x(B)\, d\mathsf{P}_\Pi(x), \quad \text{all } A,\, B \text{ measurable.} \tag{4.2}$$

Joint distributions are symmetric (i.e., "conditional times marginal" can go in both directions), so the left-hand side (LHS) of (4.2) can also be written as

$$\text{LHS} = \int_B \int_A p_\theta(x)\, d\mu(x)\, d\Pi(\theta) = \int_A \Big[\int_B p_\theta(x)\, d\Pi(\theta) \Big]\, d\mu(x),$$

where the second equality follows from Fubini's theorem. Since we are forcing the left- and right-hand sides to be equal, i.e., LHS = RHS, we must have that

$$\text{RHS} = \int_A \Big[\int_B p_\theta(x)\, d\Pi(\theta) \Big]\, d\mu(x).$$

But, RHS can also be written as

$$\text{RHS} = \int_A \Big[\Pi_x(B) \int_\Theta p_\theta(x)\, d\Pi(\theta) \Big]\, d\mu(x).$$

Since both expressions for RHS must be equal for all A and B, we must have

$$\Pi_x(B) \int_\Theta p_\theta(x)\, d\Pi(\theta) = \int_B p_\theta(x)\, d\Pi(\theta), \quad \text{for } \mathsf{P}_\Pi\text{-almost all } x.$$

Solving for $\Pi_x(B)$, we see that $\Pi_x \ll \Pi$ and the formula for the posterior density (Radon–Nikodym derivative) follows too. □

In the case where the prior Π has a density with respect to some measure ν, then we get the more familiar form of the Bayes posterior update.

Corollary 4.1. *Suppose that $\Pi \ll \nu$ with Radon–Nikodym derivative π. Then the posterior distribution Π_x is also absolutely continuous with respect to ν, and its density, call it π_x, is given by*

$$\pi_x(\theta) = \frac{p_\theta(x)\pi(\theta)}{p_\Pi(x)} \propto p_\theta(x)\pi(\theta).$$

Proof. Follows from the basic chain-rule property of the Radon–Nikodym derivatives; see Exercise 1. □

The take-away message is that given prior distribution Π and likelihood $p_\theta(x)$ one can construct a posterior distribution Π_x and, in the case that Π has a density, the posterior also has a density and it's proportional to the prior density times the likelihood.

4.2.3 Inference

The posterior distribution is all that's needed for inference on θ, that is, once the posterior is available, we can use it to calculate all kinds of things. For example, a typical point estimate for θ is the posterior mean or mode. The posterior mean is defined as

$$\hat{\theta}_{\text{mean}} = \mathsf{E}(\Theta \mid X = x) = \int \theta \, d\Pi_x(\theta),$$

and, in the case where Π_x has a density π_x, the posterior mode[3] is defined as

$$\hat{\theta}_{\text{mode}} = \arg\max_\theta \pi_x(\theta),$$

which is similar to the maximum likelihood estimate. There are more formal notions of Bayes estimators (or, more generally, Bayes rules) which we'll encounter a bit later.

For set estimation, a Bayesian uses what's called a *credible set*. A $100(1-\alpha)\%$ credible set is a set $C \subset \Theta$ such that $\Pi_x(C) = 1 - \alpha$. There are a variety of ways to construct such a set. For a real-valued parameter θ, such a set can be found as

$$C = \{\theta : \theta \text{ is between the } \alpha/2 \text{ and } 1-\alpha/2 \text{ quantiles of } \Pi_x\}.$$

Alternatively, and more generally, if the posterior Π_x has a density π_x, then a *highest posterior density* region can be used. That is,

$$C = \{\theta : \pi_x(\theta) \geq c_\alpha\},$$

where c_α is chosen such that $\Pi_x(C) = 1 - \alpha$. The key point here is that, unlike a frequentist confidence interval, a credible interval *does not* necessarily have the property that the coverage probability of C equals $1 - \alpha$.

[3] Also called the *MAP estimator*, for "maximum *a posteriori*."

Hypothesis testing is similar. A hypothesis about θ defines a subset H of the parameter space, and its prior probability is $\Pi(H)$. According to Bayes's theorem, the posterior probability of H is

$$\Pi_x(H) = \frac{\int_H p_\theta(x)\, d\Pi(\theta)}{p_\Pi(x)}.$$

Then the Bayesian will "reject" H if this posterior probability is too small. One thing to notice is that, in this setup, if the prior probability of H is zero, then so is the posterior probability. Therefore, a Bayesian must do some different things when $\Pi(H) = 0$. These different things are related to Bayes factors and model selection, but we will not discuss this anymore here.

4.2.4 Marginalization

In our discussion of likelihood, we considered the problem where θ was a vector but only a feature or a component of θ is of interest. For concreteness, suppose that $\theta = (\psi, \lambda)$, where both ψ and λ are unknown but only ψ is of interest. In that case, some modification to the usual likelihood function was required, e.g., a profile likelihood obtained by maximizing the likelihood over λ, pointwise in ψ. Though this modification is conceptually simple, the profile likelihood has some shortcomings, e.g., it does not have any built-in adjustment for the uncertainty in λ.

In the Bayesian setting, marginalization is straightforward. From probability theory, you know that, given a joint density $p(x, y)$ for a random vector (X, Y), the marginal density for X can be obtained by integration, i.e., $p(x) = \int p(x, y)\, dy$. In our present context, the posterior distribution for $\Theta = (\Psi, \Lambda)$ is a joint density, and the marginal posterior for Ψ can be obtained by integrating the joint posterior density $\pi_x(\psi, \lambda)$ over λ: $\pi_x(\psi) = \int \pi_x(\psi, \lambda)\, d\lambda$. This approach is arguably simpler than in the non-Bayesian setting where some modification to the likelihood must be invented and then implemented. For the Bayesian, the rules of probability say how to handle marginalization.

In many practical situations, it is not possible to do the required integration by hand, so some kind of numerical method is needed. In some cases, numerical integration (e.g., via Riemann sums, trapezoid/Simpson's rule, etc, or via the function `integrate` in R) can be used for this purpose. Generally, the joint posterior itself is intractable so some kind of simulations are needed for posterior inference. In this case, marginalization is particularly simple. Suppose a sample $\{(\Psi^{(m)}, \Lambda^{(m)}),\, m = 1, \ldots, M\}$ from the posterior distribution $\pi_x(\psi, \lambda)$ is available. Then a sample from the marginal posterior for Ψ is obtained by ignoring the Λ portion of the posterior sample, i.e., construct a credible interval for ψ based on $\Psi^{(1)}, \ldots, \Psi^{(M)}$.

An extreme version of the marginalization problem is that of *prediction*, i.e., where the quantity of interest is X_{n+1}, the next observation. For the prediction problem, the Bayesian will integrate the conditional density $p_\theta(x)$ with respect to the posterior distribution of Θ, given $(X_1, \ldots, X_n)$, to get the so-called *predictive distribution* of X_{n+1}, given $(X_1, \ldots, X_n)$. Exercises 6 and 7 invite you to explore the prediction problem further.

4.3 Some examples

Example 4.1. Suppose $X_1, \ldots, X_n$ are iid $\mathsf{N}(\theta, \sigma^2)$ where σ is known. In addition, assume that θ has a $\mathsf{N}(\omega, \tau^2)$ prior distribution for fixed ω and τ. In particular, prior beliefs suggest that θ would be located somewhere near ω but being above or below ω is equally likely. Since $\bar{X}$ is a sufficient statistic, the posterior will depend on $(X_1, \ldots, X_n)$ only through the mean $\bar{X}$ (why?). In this case, the posterior $\theta \mid (\bar{X} = \bar{x})$ is normal, with

$$\text{mean} = \frac{n\tau^2}{\sigma^2 + n\tau^2}\,\bar{x} + \frac{\sigma^2}{\sigma^2 + n\tau^2}\,\omega \quad \text{and} \quad \text{var} = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}.$$

Since the posterior has a closed-form expression, finding posterior moments, credible intervals, or testing hypotheses is easy. For example, a point estimate for θ is $\hat{\theta} = \text{mean}$, where mean is as displayed above.

Example 4.2. Suppose $X_1, \ldots, X_n$ is an independent sample from a $\mathsf{Pois}(\theta)$ population. Consider a $\mathsf{Gamma}(a, b)$ prior distribution for θ. Multiplying prior and likelihood gives

$$p_\theta(x)\pi(\theta) = \text{const} \times e^{-n\theta}\theta^{n\bar{x}}\theta^{a-1}e^{-\theta/b} = \text{const} \times \theta^{n\bar{x}+a-1}e^{-(n+1/b)\theta}.$$

It is clear that, after normalization, the posterior density must be that of a $\mathsf{Gamma}(n\bar{x} + a, [n + 1/b]^{-1})$ distribution. Again, the fact that the posterior has a closed-form expression makes things easy. For example, suppose $a = 5$ and $b = 2$; furthermore, suppose in a sample os size $n = 10$ the observed mean is $\bar{x} = 7$. Then the posterior distribution for θ is $\mathsf{Gamma}(75, 0.095)$ and a 95% credible interval for θ is

$$\texttt{qgamma(c(0.025, 0.975), shape=75, scale=0.095)} = (5.60, 8.23).$$

An important observation is that, in the two previous examples, the posterior distribution is within the same family as the prior. These are special cases of a more general concept of *conjugate priors*. Specifically, a class of distributions makes up a conjugate class if, for any prior Π in the class, the posterior distribution Π_x is also a member of the class. In all such problems, analysis of the posterior is straightforward—just like in the previous two examples. But one may question how realistic is a conjugate prior when its only justification is that it allows for simple calculations.

Next is an example with a non-standard model and a non-conjugate prior. This one illustrates a numerical method—Markov chain Monte Carlo, or MCMC—often used by Bayesians to compute the posterior when there is no nice closed-form expression; see Keener (2010, Ch. 15).

Example 4.3. Suppose $X_1, \ldots, X_n$ are iid samples with density

$$p_\theta(x) = \frac{1 - \cos(x - \theta)}{2\pi}, \quad 0 \le x \le 2\pi,$$

where θ is an unknown parameter in $[-\pi, \pi]$. We take a $\mathsf{Unif}(-\pi, \pi)$ prior distribution for Θ. This time, the posterior does not have a nice form, though its features can be found

via numerical integration. In this case, we use Markov chain Monte Carlo, or MCMC, to simulate from the posterior and estimate various quantities. We shall employ the Metropolis–Hastings algorithm which uses a uniform random walk proposal distribution, with window width $a = 0.5$; R code is given in Figure 4.1. A histogram of the 10000 samples from Π_x are shown in Figure 4.2(a), along with a plot of the true posterior density, and it is clear that the sampling procedure is doing the right thing; panel (b) shows a trace plot to help assess whether the Markov chain has "converged."

Example 4.4. Let $X_1, \ldots, X_n$ be iid $\mathsf{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. As a prior, consider a "distribution" with density $\pi(\mu, \sigma^2) \propto 1/\sigma^2$. This kind of prior is called *improper* because the integral over (μ, σ^2) is not finite. The meaningfulness of such a prior is questionable,[4] but one can still formally apply Bayes theorem to get a posterior distribution. While this might seem strange, the use of improper priors is quite standard; see Section 4.5.4. Here the goal is simply to work out the marginal posterior distribution for μ.

First, the normal likelihood can be written as

$$L(\mu, \sigma^2) = (1/\sigma^2)^{n/2} e^{-D/\sigma^2},$$

where $D = \frac{1}{2}\{(n-1)s^2 + n(\mu - \bar{x})^2\}$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$. With prior $\pi(\mu, \sigma^2) \propto 1/\sigma^2$, the posterior density satisfies

$$\pi_x(\mu, \sigma^2) \propto (1/\sigma^2)^{n/2+1} e^{-D/\sigma^2}.$$

The right-hand side above is proportional to the density of a known distribution, namely, the normal inverse gamma distribution,[5] but this is not particularly important.

For the marginal posterior distribution of μ, we need to integrate out σ^2 from the posterior $\pi_x(\mu, \sigma^2)$. The key here is that the right-hand side in the above display is, as a function σ^2, proportional to an inverse gamma density which has form

$$\frac{b^a}{\Gamma(a)} \Big(\frac{1}{x}\Big)^{a+1} e^{-b/x}.$$

Therefore, if we integrate over σ^2 in the above display, then we get

$$\int_0^\infty (1/\sigma^2)^{n/2+1} e^{-D/\sigma^2} \, d\sigma^2 = \frac{\Gamma(n/2)}{D^{n/2}},$$

and the marginal posterior density must satisfy

$$\pi_x(\mu) \propto \frac{\Gamma(n/2)}{D^{n/2}} \propto \Big(\frac{1}{(n-1)s^2 + n(\mu - \bar{x})^2}\Big)^{n/2}.$$

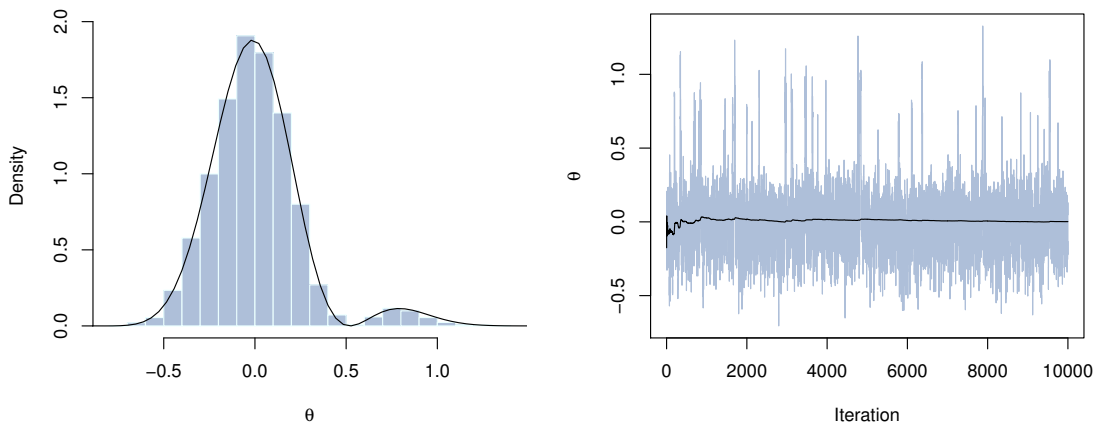
The expression on the right-hand side above, as a function of μ, is proportional to a location-scale transformation of Student-t density with $n-1$ degrees of freedom; that is, given x, the distribution of $n^{1/2}(\mu - \bar{x})/s$ is $\mathsf{t}(n-1)$.

[4] Probably the best way to interpret an improper prior is as a sort of weight attached to each parameter point, in this case, (μ, σ^2). For the prior in this example, those pairs (μ, σ^2) with small σ^2 are given more prior weight.

[5] `http://en.wikipedia.org/wiki/Normal-inverse-gamma_distribution`

```
mh <- function(x0, f, dprop, rprop, N, B) {

  x <- matrix(NA, N + B, length(x0))
  fx <- rep(NA, N + B)
  x[1,] <- x0
  fx[1] <- f(x0)
  ct <- 0
  for(i in 2:(N + B)) {

    u <- rprop(x[i-1,])
    fu <- f(u)
    r <- log(fu) + log(dprop(x[i-1,], u)) - log(fx[i-1]) - log(dprop(u, x[i-1,]))
    R <- min(exp(r), 1)
    if(runif(1) <= R) {

      ct <- ct + 1
      x[i,] <- u
      fx[i] <- fu

    } else {

      x[i,] <- x[i-1,]
      fx[i] <- fx[i-1]

    }

  }
  out <- list(x=x[-(1:B),], fx=fx[-(1:B)], rate=ct / (N + B))
  return(out)

}

X <- c(3.91, 4.85, 2.28, 4.06, 3.70, 4.04, 5.46, 3.53, 2.28, 1.96, 2.53,
       3.88, 2.22, 3.47, 4.82, 2.46, 2.99, 2.54, 0.52, 2.50)
lik <- function(theta) {

  o <- drop(exp(apply(log(1 - cos(outer(X, theta, "-"))), 2, sum)))
  ind <- (theta <= pi) & (theta >= -pi)
  o <- o * ind
  return(o)

}
a <- 0.5
dprop <- function(theta, theta0) dunif(theta, theta0 - a, theta0 + a)
rprop <- function(theta0) runif(1, theta0 - a, theta0 + a)
den <- integrate(lik, -pi, pi)$value
dpost <- function(theta) lik(theta) / den
x <- seq(-pi, pi, len=150); dpost.x <- dpost(x)
ylim <- c(0, 1.05 * max(dpost.x))
N <- 10000
B <- 5000
theta.mcmc <- mh(runif(1, -pi, pi), lik, dprop, rprop, N, B)
hist(theta.mcmc$x, freq=FALSE, col="gray", border="white", ylim=ylim, xlab=expression(theta), main="")
lines(x, dpost.x)
plot(theta.mcmc$x, type="l", col="gray", xlab="Iteration", ylab=expression(theta))
lines(1:N, cumsum(theta.mcmc$x) / (1:N))
print(quantile(theta.mcmc$x, c(0.05, 0.95)))
```

Figure 4.1: R codes for the Metropolis–Hastings algorithm in Example 4.3.



(a) Posterior distribution

(b) Trace plot

Figure 4.2: Panel (a): Histogram of the Monte Carlo sample from the posterior Π_x in Example 4.3, with the true posterior density overlaid; Panel (b): trace plot of the Monte Carlo sample, with running mean, suggesting that the Markov chain "mixed well."

Last, it is also interesting to consider the benefits of a Bayesian approach based on the classical criteria. Roughly speaking, if the prior is "reasonable," then a Bayesian procedure that uses this information will generally beat a non-Bayesian procedure that ignores it. This is essentially a decision theory problem, but below is one simple illustration of the main idea.

Example 4.5. Suppose $\omega = 0$, $\tau = 1$, and $\sigma = 1$. In this case, the posterior mean of θ is $\hat{\theta} = n\bar{X}/(n+1)$. Let's see how this compares to the usual estimate $\bar{X}$, the MLE. A criterion by which these can be compared is the mean square error $\mathsf{mse}(\theta; \hat{\theta}) := \mathsf{E}_\theta(\hat{\theta} - \theta)^2$ as a function of the true θ. It is easy to check that

$$\mathsf{mse}(\theta; \bar{X}) = \frac{1}{n} \quad \text{and} \quad \mathsf{mse}(\theta; \hat{\theta}) = \frac{\theta^2 + n}{(n+1)^2}.$$

It is easy to see that if the true θ is close to zero (the prior mean), then the Bayes estimate is better; however, if the prior is way off and the true θ is far from zero, the Bayes rule can be beaten badly by the MLE. The take away message is that, for suitably chosen priors, Bayes procedures generally outperform non-Bayes procedures. The catch is that the needed prior(s) depend on the true value of the unknown parameter.

4.4 Motivations for the Bayesian approach

4.4.1 Some miscellaneous motivations

- There are sets of rationality axioms and it has been shown that if one wants to be "rational" then one must be Bayesian. Some description of these ideas can be found in Ghosh et al. (2006). Utility theory is discussed in Keener (2010, Chap. 7.3).

- There is one specific kind of "rationality" axioms, called *coherence*, that's relatively easy to understand. It comes from a sort of gambling perspective. The idea is that a reasonable summary of uncertainty should be such that one is willing to make bets based on it. For the sake of space, I won't get into the details here, but the idea is that, under some basic assumptions,[6] unless the uncertainties you specify satisfy the rules of probability, i.e., Kolmogorov's axioms,[7] I have a betting strategy that will make you a sure loser. Of course, this indicates that your system of uncertainties is flawed in some way—or *incoherent*. The message here is that, in a statistical point of view, if you're not a Bayesian, and you summarize your uncertainties about θ by something other than a genuine probability, then there's something wrong with your uncertainty assessments. Thus, only the Bayesian approach is coherent. Chapter 1 in Kadane (2011) gives a lovely description of this idea.
- In a Bayesian approach, it is easy to incorporate any known information about the parameter into the analysis. For example, suppose it is known that the mean θ of a normal population satisfies $a \leq \theta \leq b$. Then the Bayesian approach can easily handle this by choosing a prior supported on $[a, b]$ reflecting this known information. The classical (frequentist) approach cannot handle such information so gracefully.
- There are theorems in decision theory (called *complete class theorems*) which state that, for a given inference problem, for any decision rule there is an (approximate) Bayes rule which is as good or better.[8] So, in other words, there is no real reason to look outside the class of (approximate) Bayes procedures since, for any non-Bayes procedure, there is a Bayes procedure that's just as good.

4.4.2 Exchangeability and deFinetti's theorem

In introductory courses we are used to seeing assumptions of "independent and indentically distributed" data. But we also know that there are other types of dependence structures that are often more realistic but not as easy to deal with. In this section we will discuss the notion of *exchangeable* random variables, which includes iid as a special case, and a remarkable consequence due to deFinetti and later built upon by others. This material comes from Schervish, Chapter 1.

Definition 4.1. A finite set of random variables $X_1, \ldots, X_n$ is *exchangeable* if all permutations of $(X_1, \ldots, X_n)$ have the same joint distribution. An infinite collection of random variables is exchangeable if every finite subset is exchangeable.

[6] The one key assumption here is that you are equally willing to buy tickets from me or to sell similar tickets to me. This may or may not be reasonable.

[7] One actually does not need Kolmogorov's countable additivity; coherence theorems are available for measures which are only finitely additive.

[8] For example, the sample mean $\bar{X}$ is known to be a good estimate of a normal mean, and this corresponds to the Bayesian posterior mean under an "improper" uniform prior on $(-\infty, \infty)$, which can be viewed as a limit of a sequence of "proper" priors.

For example, suppose $X_1, \ldots, X_{50}$ are exchangeable. Then all X_i's have the same marginal distribution. Moreover, (X_1, X_2) and (X_{33}, X_{44}) have the same joint distribution, as do (X_2, X_7, X_5) and (X_{47}, X_{21}, X_{15}). In fact, in Exercise 10 you are asked to prove that a set $X_1, \ldots, X_n$ are exchangeable if and only if all finite subsets of have the same joint distribution. It is also easy to see that iid random variables are exchangeable.[9]

It's important to understand, intuitively, what exchangeability means. Exchangeability implies nothing more than distributional symmetry. That is, if the order of the observations is irrelevant, then the data is exchangeable. This is obviously a very weak assumption. The assumption of iid data is quite strong and, moreover, there are some philosophical difficulties in assuming the existence of a fixed but unknown parameter is such a context.[10] It turns out that exchangeability is almost as simple as iid, and leads to a very nice motivation for Bayesian analysis. The next example gets us closer to the main theorem.

Example 4.6. Consider random variables $X_1, \ldots, X_n$ which we model as "conditionally iid." That is, there is a random variable Θ such that $X_1, \ldots, X_n$ are iid given the value θ of Θ. More formally,

$$\Theta \sim \Pi \quad \text{and} \quad (X_1, \ldots, X_n) \mid (\Theta = \theta) \overset{\text{iid}}{\sim} p_\theta. \tag{4.3}$$

Then $(X_1, \ldots, X_n)$ are exchangeable; see Exercise 11.

This "conditionally iid" structure is exactly like what we encountered in the examples in Section 4.3. That is, the Bayesian model implies an exchangeable model for the data (marginally). The surprising fact is that the relationship goes the other way too—*an infinite collection of exchangeable random variables are conditionally iid* (with respect to some prior Π and density p_θ). This is a version of deFinetti's theorem for binary observables.

Theorem 4.2. *A sequence X_n of binary random variables is exchangeable if and only if there exists a random variable Θ, taking values in $[0, 1]$, such that, given $\Theta = \theta$, the X_n's are iid $\mathsf{Ber}(\theta)$. Furthermore, if the sequence is exchangeable, then the distribution of Θ is unique and $n^{-1} \sum_{i=1}^n X_i$ converges almost surely to Θ.*

In other words, exchangeability implies that, for some probability measure Π on $[0, 1]$, the joint distribution of $(X_1, \ldots, X_n)$ can be written as

$$\mathsf{P}(X_1 = x_1, \ldots, X_n = x_n) = \int \theta^{t(x)} (1 - \theta)^{n - t(x)} \, d\Pi(\theta),$$

where $t(x) = \sum_{i=1}^n x_i$. One may interpret this Π is a prior in a Bayesian sense, along with the Bernoulli likelihood. However, Π is determined by the limit of the X sequence, which is exactly our intuition in a coin-flipping problem: the parameter θ represents the limiting proportion of heads in infinitely many flips. So the point is that a simple assumption of exchangeability is enough to imply that the Bayesian/hierarchical model is in play.

[9] Joint distributions are products and multiplication is commutative.

[10] For example, in a coin flipping experiment, the parameter θ representing the probability the coin lands on heads is usually thought of as the limiting proportion of heads in an infinite sequence of tosses. How can such a parameter exist before the sequence of flips has been performed?

There are more-general versions of the deFinetti theorem. While these are more mathematically complicated, the intuition is the same as in Theorem 4.2. Here is one such result.

Theorem 4.3 (Hewitt–Savage)**.** *A sequence of random variables X_n in $(\mathbb{X}, \mathcal{A})$ is exchangeable if and only if there exists a random probability measure $\mathscr{P}$ such that, given $\mathscr{P} = \mathsf{P}$, $X_1, X_2, \ldots$ are iid with distribution P. Moreover, if the model is exchangeable, then $\mathscr{P}$ is unique and determined by the limit $\mathscr{P}_n(A) := n^{-1} \sum_{i=1}^n I_A(X_i) \to \mathscr{P}(A)$ almost surely for each $A \in \mathcal{A}$.*

Another way to view this "random probability measure" business is through a mixture. Theorem 4.3 says that the sequence X_n is exchangeable if and only if there exists a probability measure Π on the set of all distributions on $(\mathbb{X}, \mathcal{A})$ and the marginal distribution of $(X_1, \ldots, X_n)$ is given by

$$\int \prod_{i=1}^n \mathsf{P}(X_i \in A_i) \, d\Pi(\mathsf{P}),$$

a mixture of iid models. This is more general, but makes a connection to the notion of conditionally iid random variables. This illustration also sheds light on the implications of this theorem for Bayesians. Indeed, the theorem states that a simple assumption of exchangeability implies that there exists a hierarchical model like (4.1) which can be interpreted as a prior and a likelihood to be updated via Bayes theorem. The one caveat, however, regarding the interpretation of deFinetti's theorem in a Bayesian context is that exchangeability does not say what the prior and likelihood should be, only that there is such a pair.

An alternative view of de Finetti's theorem as a motivation for a Bayesian approach, communicated to me by Stephen Walker, and focusing on *prediction*, is as follows. Let $X_1, X_2, \ldots$ be a sequence of independent observables, and suppose the goal is to predict the next one based on what has already been observed. The analyst starts the process with two things:

- A (guess of) the predictive distribution for X_1, and
- a rule for updating the guess based on an observation.

It is important to mention that, at least not yet, the updating rule need not be the Bayesian predictive updating discussed briefly in Section 4.2.4. When $X_1 = x_1$ is observed, the analyst updates the initial guess based on the aforementioned rule to get a predictive distribution for X_2. Now $X_2 = x_2$ is observed and a predictive distribution for X_3 is obtained. The process can go on indefinitely but we will stop the process here and think about the structure. In particular, does the predictive distribution of X_3 depend on the order of the observations (x_1, x_2)? In other words, would the predictive distribution of X_3 be the same had we observed (x_2, x_1) instead? This question can be answered only by knowing the analyst's rule for updating. However, if it happens that the order of the previous observations does not matter, which is quite intuitive, given that the data source is independent, then it follows from de Finetti's theorem that the analyst's updating rule must be the Bayesian rule discussed in Section 4.2.4.

4.5 Choice of priors

We have mentioned above several reasons to adopt a Bayesian approach. However, none of these justifications says what prior to choose for a given problem—at best, the results say simply that there is a "reasonable" prior. What can be done if one wants to be a Bayesian but does not know what prior to choose? Here's a few ideas.

4.5.1 Prior elicitation

Prior elicitation means having discussions with experts to encode their prior knowledge about the problem at hand into a probability distribution. This is a challenging endeavor for a number of reasons. First, this can be very time consuming. Second, even experts (who often will have little to no knowledge of probability and statistics) can have a hard time communicating their beliefs about the unknown parameter in a precise (and consistent) enough way that a statistician can turn this into a prior distribution. So, suffice it to say, this elicitation step is difficult to carry out and is rarely done to the fullest extent.

4.5.2 Convenient priors

As we saw in Section 4.3 there are some priors which are particularly convenient for the model in question. Conjugate priors are one set of convenient priors. To expand on the set of conjugate priors, one can consider mixtures of conjugate priors. The trouble is that it can be difficult to trust the results of a Bayesian analysis that's based on an unrealistic assumption to start with. For years, this was the only kind of Bayesian analysis that could be done since, otherwise, the computations were too difficult. Nowadays, with fast computers and advanced algorithms, there is really no need to limit oneself to a set of "convenient" priors. So conjugate priors, etc, are somewhat of a thing of the past.[11]

4.5.3 Many candidate priors and robust Bayes

An alternative to choosing a convenient prior is to consider a class of reasonable and relatively convenient priors, to look at each candidate posterior individually, and to decide if the results are sensitive to the choice of prior. Here is a nice example taken from Ghosh et al. (2006, Sec. 3.6).

Example 4.7. Suppose X follows a $\mathsf{Pois}(\theta)$ distribution. Suppose that it is believed that the prior for θ is continuous with 50th and 75th percentiles 2 and 4, respectively. If these are the only prior inputs, then the following are three candidates for Π:

1. Π_1: $\Theta \sim \mathsf{Exp}(a)$ with $a = \log(2)/2$.
2. Π_2: $\log \Theta \sim \mathsf{N}(\log(2), (\log(2)/0.67)^2)$.
3. Π_3: $\log \Theta \sim \mathsf{Cauchy}(\log 2, \log 2)$.

[11] They do occasionally appear in higher levels of hierarchical priors...

x	0	1	2	3	4	5	10	15	20	50
Π_1	0.75	1.49	2.23	2.97	3.71	4.46	8.17	11.88	15.60	37.87
Π_2	0.95	1.48	2.11	2.81	3.56	4.35	8.66	13.24	17.95	47.02
Π_3	0.76	1.56	2.09	2.63	3.25	3.98	8.87	14.07	19.18	49.40

Table 4.1: Posterior means $\mathsf{E}(\Theta \mid x)$ for various priors and x's.

Under these choices of prior, the posterior mean can be calculated. Table 4.1 lists these values for several different x's. Here we see that when x is relatively small (i.e., $x \leq 10$) the choice of prior doesn't matter much. However, when x is somewhat large, the posterior means seem to vary a lot.

There are other related approaches which define a large class Γ of priors which are somehow reasonable and attempt to derive upper and lower bounds on certain posterior quantities of interest. We shall look at one such result that bounds the posterior mean $\psi(\theta)$ over a class of symmetric unimodal priors. See Ghosh et al. (2006, Theorem 3.6) for details.

Theorem 4.4. *Suppose data X has a density p_θ, $\theta \in \mathbb{R}$, and consider a class Γ of symmetric unimodal priors about θ_0, i.e.,*

$$\Gamma = \{\pi : \pi \text{ is symmetric and unimodal about } \theta_0\}.$$

For a fixed real-valued function ψ, we have the bounds

$$\sup_{\pi \in \Gamma} \mathsf{E}_\pi(\psi(\Theta) \mid x) = \sup_{r>0} \frac{\int_{\theta_0 - r}^{\theta_0 + r} \psi(\theta) p_\theta(x)\, d\theta}{\int_{\theta_0 - r}^{\theta_0 + r} p_\theta(x)\, d\theta}$$

$$\inf_{\pi \in \Gamma} \mathsf{E}_\pi(\psi(\Theta) \mid x) = \inf_{r>0} \frac{\int_{\theta_0 - r}^{\theta_0 + r} \psi(\theta) p_\theta(x)\, d\theta}{\int_{\theta_0 - r}^{\theta_0 + r} p_\theta(x)\, d\theta}.$$

4.5.4 Objective or non-informative priors

The main idea behind objective priors is to choose a prior that has minimal impact on the posterior—in other words, objective priors allow the data to drive the analysis. There are basically three approaches to objective Bayes:

- Define a "uniform distribution" with respect to the geometry of the parameter space,
- Minimize a suitable measure of information in the prior, and
- Choose a prior so that the resulting posterior inferences (e.g., credible intervals) have some desirable frequentist properties.

Surprisingly, in single-parameter problems, one prior accomplishes all three.

Definition 4.2. The Jeffreys prior for θ has density $\pi(\theta) \propto (\det\{I_X(\theta)\})^{1/2}$, where $I_X(\cdot)$ denotes the Fisher information matrix.

It is interesting that the Jeffreys prior is a uniform distribution on Θ if, instead of the usual Euclidean geometry, one looks at the geometry induced by the Riemannian metric, which is determined by the Fisher information; for details, see Ghosh and Ramamoorthi (2003). When θ is a location parameter, the Fisher information is constant, and the geometry on induced by the Fisher information is exactly the usual geometry; hence, the Jeffreys' prior for θ is, in this case, a usual uniform distribution, though it's usually improper. Also, it can be shown that the Jeffreys prior minimizes the asymptotic Kullback–Leibler divergence between prior and posterior. There are also results on how Jeffreys prior produces posterior credible sets with approximately the nominal frequentist coverage. Ghosh et al. (2006) gives a clear description of all of these facts.

There are other notions of objective priors (e.g., invariant priors) but one thing these guys often have in common is that they do not integrate to one—that is, they're improper. For example, in a location problem, both the Jeffreys and invariant priors are Lebesgue measure on $(-\infty, \infty)$, which is not a finite measure. More generally, the (left and right) invariant Haar priors in group transformation problems are often improper. This raises a natural question: can probability theory in general, and Bayes theorem in particular be extended to the improper case? There are essentially two ways to deal with this: allow probabilities to be infinite, or remove the countable additivity assumption. In both cases, many of the known results on probability must be either scraped or re-proved. But, there are versions of Bayes' theorem which hold for improper of finitely additive probabilities. These are too technical for us though.

4.6 Bayesian large-sample theory

4.6.1 Setup

Large-sample theory in the classical setting is helpful for deriving statistical procedures in cases where exact sampling distributions are not available. Before computing power was so readily available, this kind of asymptotic theory was the only hope for handling non-trivial problems. In the Bayesian setting, there is a corresponding asymptotic theory. Besides the obvious asymptotic approximation of posterior probabilities, which can simplify computations in several ways, an important consequence of Theorem 4.6 is that, under minimal conditions, the choice of the prior is irrelevant when the sample size is large. Since the choice of prior is the only real obstacle to Bayesian analysis, this is a fundamentally important result. Other weaker Bayesian convergence results (e.g., posterior consistency, Exercise 19) are available, and we may discuss these things later on.

Another application of Bayesian asymptotic theory is as a tool for identifying "bad priors" that should not be used. That is, if a particular prior does not admit desirable behavior of the posterior as $n \to \infty$, then there is something wrong with that prior. This is particularly helpful in Bayesian nonparametric problems, where it is not so easy to cook up a "good

prior" based on intuition or experience. In fact, the primary motivation for the surge of work recently in Bayesian asymptotic theory is to help identify good priors to use in challenging nonparametric problems.

4.6.2 Laplace approximation

The *Laplace approximation* is a wonderfully simple, yet very powerful technique for approximating certain kinds of integrals; see Ghosh et al. (2006, Sec. 4.3). The approximation itself has nothing to do with Bayesian analysis, but the kinds of integration problems it's useful for are frequently encountered in Bayesian statistics. It is also the basis for the popular BIC (Bayesian Information Criterion) in model selection.

Consider an integral of the form

$$\text{integral} = \int q(\theta) e^{nh(\theta)}\, d\theta,$$

where both q and h are smooth functions of a p-dimensional quantity θ; that is, the integral above is taken over $\mathbb{R}^p$. Here it is assumed that n is large or increasing to ∞. Let $\hat\theta$ be the unique maximum of h. Then the Laplace approximation provides a way to calculate the integral without integration—only optimization!

Theorem 4.5. *Let h' and h'' denote derivatives of h and let* $\det(\cdot)$ *stand for matrix determinant. Then, as $n \to \infty$,*

$$\int q(\theta) e^{nh(\theta)}\, d\theta = q(\hat\theta) e^{nh(\hat\theta)} (2\pi)^{p/2} n^{-p/2} \det\{-h''(\hat\theta)\}^{-1/2} \{1 + O(n^{-1})\}.$$

Note that $h''(\hat\theta)$ is negative definite, by assumption, so $-h''(\hat\theta)$ is positive definite.

Proof. Here is a sketch for the case of $p = 1$. The first observation is that, if h has a unique maximum at $\hat\theta$ and n is very large, then the primary contribution to the integral is in a small interval around $\hat\theta$, say $\hat\theta \pm a$. Second, since this interval is small, and $q(\theta)$ is smooth, it is reasonable to approximate $q(\theta)$ by the constant function $q(\hat\theta)$ for $\theta \in (\hat\theta - a, \hat\theta + a)$. Now the idea is to use a Taylor approximation of $h(\theta)$ up to order two around $\theta = \hat\theta$:

$$h(\theta) = h(\hat\theta) + h'(\hat\theta)(\theta - \hat\theta) + (1/2) h''(\hat\theta)(\theta - \hat\theta)^2 + \text{error}.$$

Since $h'(\hat\theta) = 0$ by definition of $\hat\theta$, plugging this into the exponential term in the integral (and ignoring the error terms) gives

$$\begin{aligned}
\text{integral} &\approx \int_{\hat\theta - a}^{\hat\theta + a} q(\theta) \exp\{n[h(\hat\theta) + (1/2) h''(\hat\theta)(\theta - \hat\theta)^2]\}\, d\theta \\
&\approx \int_{\hat\theta - a}^{\hat\theta + a} q(\hat\theta) \exp\{nh(\hat\theta) - (\theta - \hat\theta)^2 / 2\sigma^2\}\, d\theta \\
&= q(\hat\theta) e^{nh(\hat\theta)} \int_{\hat\theta - a}^{\hat\theta + a} e^{-(\theta - \hat\theta)^2 / 2\sigma^2}\, d\theta,
\end{aligned}$$

where $\sigma^2 = [-nh''(\hat{\theta})]^{-1}$, small. The last integrand looks almost like a normal density function, except that it's missing $(2\pi\sigma^2)^{-1/2}$. Multiply and divide by this quantity to get

$$\text{integral} \approx q(\hat{\theta}) e^{nh(\hat{\theta})} (2\pi)^{1/2} n^{-1/2} [-h''(\hat{\theta})]^{-1/2},$$

which is what we were looking for. □

One simple yet interesting application of the Laplace approximation is Stirling's approximation of $n!$ (Exercise 16). The connection to the normal distribution in the above proof sketch is the key to the main theorem on posterior normality, discussed next.

4.6.3 Bernstein–von Mises theorem

Let $\hat{\theta}_n$ be a consistent sequence of solutions to the likelihood equation, and let $I(\theta)$ denote the Fisher information. The Bernstein–von Mises theorem states that the posterior distribution of $n^{1/2}(\Theta - \hat{\theta}_n)$ is approximately normal with mean zero and variance $I(\theta^\star)^{-1}$ in $\mathsf{P}_{\theta^\star}$-probability as $n \to \infty$. The conditions required for such a result are essentially the same as those used to show posterior normality of the MLE. In particular, in addition to conditions C1–C4 from Chapter 3, assume the following:

C5. For any $\delta > 0$, with $\mathsf{P}_{\theta^\star}$-probability 1, there exists $\varepsilon > 0$ such that

$$\sup_{\theta : |\theta - \theta^\star| > \delta} n^{-1} \{\ell_n(\theta) - \ell_n(\theta^\star)\} \leq -\varepsilon$$

for all sufficiently large n, where $\ell_n = \log L_n$; and

C6. The prior density $\pi(\theta)$ is continuous and positive at $\theta^\star$.

Condition C6 is easy to verify and holds for any reasonable prior. Condition C5, on the other hand, is a bit more challenging, though it does hold for most examples; see Exercise 17. We shall focus here on the one-dimensional Θ case, though a similar result holds for d-dimensional Θ with obvious modifications.

Theorem 4.6. *Assume Conditions C1–C6, and let $\hat{\theta}_n$ be a consistent sequence of solutions to the likelihood equation. Write $Z_n = n^{1/2}(\Theta - \hat{\theta}_n)$ and let $\tilde{\pi}_n(z)$ be the posterior density of Z_n, given $X_1, \ldots, X_n$. Then*

$$\int |\tilde{\pi}_n(z) - \mathsf{N}(z \mid 0, I(\theta^\star)^{-1})| \, dz \to 0 \quad \textit{with } \mathsf{P}_{\theta^\star}\textit{-probability 1.}$$

The message here is that, under suitable conditions, if the sample size is large, then the posterior distribution for Θ will look approximately normal with mean $\hat{\theta}_n$ and variance $[nI(\theta^\star)]^{-1}$. Under the same conditions, $I(\theta^\star)$ can be replaced by n^{-1} times the observed Fisher information; see the Laplace approximation argument below. The kind of convergence being discussed here is L_1-convergence of the densities, which is stronger than the usual "in distribution" convergence. That is, by Theorem 4.6, the expectation of any function of Θ

can be approximated by that same expectation under the limiting normal. Besides as a tool for approximate posterior inference, such results can be useful for developing computational methods for simulating from the exact posterior.

Details of the proof are given in Ghosh et al. (2006, Sec. 4.1.2). Here I will give a sketch of the proof based on the Laplace approximation. The idea is to approximate the log-likelihood by a quadratic function via Taylor approximation; recall that we did a similar thing in the proof of asymptotic normality of the MLE. Let $Z = n^{1/2}(\Theta - \hat{\theta})$ be the rescaled parameter value. Then

$$\Pi_n(-a < Z < a) = \Pi_n(\hat{\theta} - an^{-1/2} < \Theta < \hat{\theta} + an^{-1/2}) = \frac{\text{num}}{\text{den}}.$$

Letting

$$q(\theta) = \pi(\theta) \quad \text{and} \quad h(\theta) = \frac{1}{n} \log L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i),$$

then the denominator above can be approximated (via Laplace) by

$$\text{den} = \int L_n(\theta)\pi(\theta)\, d\theta = \int \pi(\theta) e^{nh(\theta)}\, d\theta \approx L_n(\hat{\theta})\pi(\hat{\theta})(2\pi/nv)^{1/2},$$

where $v = -h''(\hat{\theta}) = -n^{-1} \sum_{i=1}^{n} (\partial^2/\partial\theta^2) \log p_\theta(X_i)\big|_{\theta=\hat{\theta}}$ is the observed Fisher information. The numerator can be similarly approximated:

$$\text{num} = \int_{\hat{\theta}-an^{-1/2}}^{\hat{\theta}+an^{-1/2}} L_n(\theta)\pi(\theta)\, d\theta \approx L_n(\hat{\theta})\pi(\hat{\theta})n^{-1/2} \int_{-a}^{a} e^{-vu^2/2}\, du.$$

Taking the ratio of num to den gives

$$\Pi_n(-a < Z < a) = \frac{\text{num}}{\text{den}} \approx \frac{L_n(\hat{\theta})\pi(\hat{\theta})n^{-1/2} \int_{-a}^{a} e^{-vu^2/2}\, du}{L_n(\hat{\theta})\pi(\hat{\theta})(2\pi/nv)^{1/2}} = \int_{-a}^{a} \frac{\sqrt{v}}{\sqrt{2\pi}} e^{-vu^2/2}\, du,$$

and this latter expression is the probability that a normal random variable with mean zero and variance v^{-1} is between $-a$ and a, which was what we set out to show.

It is intuitively clear from the Bernstein–von Mises theorem that the posterior mean ought to behave just like the consistent sequence of solutions $\hat{\theta}_n$. The following result makes this intuition rigorous.

Theorem 4.7. *In addition to Conditions C1–C6 above, assume that the prior has a finite mean. Then the posterior mean $\tilde{\theta}_n = \mathsf{E}(\Theta \mid X)$ satisfies*

$$n^{1/2}(\tilde{\theta}_n - \hat{\theta}_n) \to 0 \quad \textit{and} \quad n^{1/2}(\tilde{\theta}_n - \theta^\star) \to \mathsf{N}(0, I(\theta^\star)^{-1}).$$

Proof. See Ghosh and Ramamoorthi (2003, p. 39). Exercise 18 outlines the proof of a weaker result using the Laplace approximation. □

4.7 Concluding remarks

4.7.1 Lots more details on Bayesian inference

Here, we do not attempt to get too deep into Bayesian methodology and philosophy. A formal Bayesian inference course would take these points more seriously. The book by Berger (1985) provides a nice sort of "philosophical" justification of Bayesian analysis, along with many other things. The same can be said for Ghosh et al. (2006). Bayesian modeling and methodology is challenging. A good place to start to learn about these things in the book by Gelman et al. (2004), which is a fairly modern look at Bayesian analysis, from an applied point of view. There are methodological challenges beyond the philosophical justification of the Bayesian approach. One important example is in the case of hypothesis testing when the null is a singleton (or some other set with Lebesgue measure zero). In this case, the simple Bayesian hypothesis testing formulation breaks down and different ideas are needed. *Bayes factors* are the typical choice in such cases, see Kass and Raftery (1995) and Ghosh et al. (2006), but, interestingly, since these are not functions of the posterior distribution, they are not really Bayesian. Anyway, the point is just that, although the ideas presented here are rather simple, it is an oversimplification in general. Finally, I should mention that computation is crucial to Bayesians because really nothing admits closed-form expressions. Various forms of Monte Carlo, which is a method for numerical integration, are used, and the best resource to learn about these things is the book Robert and Casella (2004). In these notes, the goal is just to introduce some of the key ideas in the Bayesian framework. Some of these points will be useful in our coverage of statistical decision theory that's coming up. In any case, all statisticians should be familiar with the basics of all the main ideas in statistics; limiting oneself to just one perspective is just that, a limitation.

4.7.2 On Bayes and the likelihood principle

Recall the brief comments on the *likelihood principle* at the end of Chapter 3. The likelihood principle is a philosophical position that the final inference should depend on data/model only through the observed likelihood function. While this is a rather extreme position, Birnbaum (1962) showed that the likelihood principle was implied by two quite reasonable principles. Although this theorem has recently been refuted (e.g., Evans 2013; Martin and Liu 2014; Mayo 2014), the fact remains that Birnbaum's result helped to convince many statisticians to seriously consider the Bayesian approach. The quick conclusion of Birnbaum's theorem is that the likelihood principle is desirable (because it is equivalent to other desirable properties, according to the theorem), and since the only the classical Bayesian approach (with a "subjective prior," see below) satisfies the likelihood principle, the only logical approach is a Bayesian one. This attention given to Bayesian methods following Birnbaum's discovery might be considered as a catalyst for the theoretical, methodological, and computational developments in the last 20–30 years.

Here I want to make two remarks about the likelihood principle and Bayesian analysis. First, the claim that "the Bayesian approach obeys the likelihood principle" is incomplete.

The claim is true if the prior is based off of some subjective considerations. However, the standard approach now is to use default "non-informative" priors, such as Jeffreys prior, which depends on the model itself. The use of Bayesian inference with such a default prior *does not* satisfy the likelihood principle.[12] Second, in some cases, that the Bayesian approach satisfies the likelihood principle might be considered a disadvantage, or at least not intuitively appealing. One example is in the case of finite-population sampling and inference. Good books on Bayesian and non-Bayesian approaches to this problem are Ghosh and Meeden (1997) and Hedayat and Sinha (1991), respectively. For such problems, considerable effort is taken to design a good sampling scheme, so that the obtained sample is "representative" in some sense. However, the fact that the Bayesian posterior depends only on the observed likelihood function means that *the sampling design is irrelevant to inference.* This is somewhat counterintuitive and maybe even controversial.[13] So, in this case, some might say that obeying the likelihood principle is a disadvantage to the Bayesian approach, though I don't think it's that simple.

4.7.3 On the "Bayesian" label

Statisticians often throw around the labels "Bayesian" and "non-Bayesian." I personally do not like this because I think (a) it is somewhat divisive, and (b) it can give a false indication that some problems can be solved with Bayesian analysis while others cannot. We are all working on the same problems, and we should be happy to consider different perspectives. For that reason, though I consider myself knowledgeable about the Bayesian perspective, I would not classify myself as a Bayesian, per se. In fact, I could be considered very much non-Bayesian because I have questions concerning the meaningfulness of Bayesian posterior probabilities, etc, in general, and I have done work to develop some new framework, different from (but not orthogonal to) existing Bayesian ideas.

4.7.4 On "objectivity"

Remember, at the very beginning of the course, I mentioned that the statistics problem is ill-posed, and there is no "right answer." Every approach makes assumptions to be able to proceed with a theory. In the Bayesian setting, there is an assumption about existence of a prior distribution. In the frequentist setting, there is an assumption that the observed data is

[12] A simple example is in the case of binary data. The likelihood functions for binomial and negative binomial models are proportional, so inference should not depend on the choice between these two models; a frequentist inference, e.g., with a p-value, does depend on the model, hence does not satisfy the likelihood principle. In a classical Bayesian approach with a subjective prior, the Bayesian posterior depends on data/model only through the observed likelihood function and, therefore, does satisfy the likelihood principle. However, the Jeffreys prior for the two models is different, so the corresponding Bayesian analysis *would not* satisfy the likelihood principle.

[13] I can see both sides of the argument. On one hand, if inference doesn't depend on the design, then it suggests that a good design and bad design are equivalent, which is not reasonable. On the other hand, if we have a "good" sample, then inference shouldn't care how exactly that sample was obtained. To me, the issue boils down to one of defining what it means for a sample to be "representative"...

one of those typical outcomes. So, no matter what, we cannot escape without making some kind of assumptions. Be skeptical about arguments, statistical or otherwise, that claim to be "objective." The next subsection briefly describes some concerns about the use of objective priors and, more generally, the use of probability for inference.

4.7.5 On the role of probability in statistical inference

Recall that the basic motivation for the Bayesian approach is that probability is the correct way to summarize uncertainty. Is there any justification for probability being the correct tool? For me, the justification is clear when a meaningful prior distribution for θ is available. Then the statistical inference problem boils down to a probability calculation. However, the typical scientific problem is one where little/nothing is known about θ, which means there is no meaningful prior distribution available, or one is reluctant to use what little information is available for fear of influencing the results.

In such cases, one can introduce a default non-informative prior for θ and carry out a Bayesian analysis. This kind of Bayesian approach has been applied successfully in many problems, but that doesn't mean we can't question its use. The main concern is as follows: the prior distribution effectively establishes the scale on which the posterior probabilities are interpreted. This point is clear if you think of the posterior as an updated prior based on the observed data; another extreme way to understand this is that the prior and the posterior have the same null sets. So, at least in finite samples, the prior plays a role in scaling the numerical values of the posterior. When the prior itself has no meaningful scale, e.g., it's improper, then how can one assign any meaning to the corresponding posterior probabilities? One can only assign meaning to the posterior probabilities (of certain subsets of Θ) asymptotically, which is basically what the Bernstein–von Mises theorem says. So, when the prior is non-informative, the posterior probabilities lack any meaningful interpretation, at least for finite samples.

Another concern with probability is the complementation rule, i.e., $\mathsf{P}(A^c) = 1 - \mathsf{P}(A)$. Such a property makes perfect sense when the goal is to predict the outcome of some experiment to be performed—X will be in exactly one of A or A^c. This is the context in which probability was first developed, that is, when the goal is to predict the outcome of some experiment when all the information about the experiment is available. However, I would argue that this is very different from the statistical inference problem. The quantity we are after is not a realization of some experiment, but rather a fixed quantity about which we have limited information, in the form of a model and observed data. Why then should probability be the right tool for summarizing our uncertainty? For example, I don't think the complementation rule is logical in the inference problem. In reality, θ is in exactly one of A and A^c, but with the limited information in the data, it may not be reasonable to make such a sharp conclusion that A is strongly supported and A^c is weakly supported, or vice versa. In fact, it seems quite reasonable that data cannot strongly support either A or A^c, in which case, the "probability" of these events should add to a number less than 1. No probability can satisfy this property, so maybe probability isn't the correct tool. You might ask if there is something else that can accommodate this, and the answer is YES, a *belief*

function. This is part of the motivation behind the *inferential model* (IM) framework; see Martin and Liu (2013, 2015a,b,c) and Liu and Martin (2015).

4.8 Exercises

1. (a) Consider some generic σ-finite measures, μ, ν, and λ, all defined on the measurable space $(\mathbb{X}, \mathcal{A})$. Suppose that $\mu \ll \nu$ and $\nu \ll \lambda$. Show that $\mu \ll \lambda$ and that the Radon–Nikodym derivative of μ with respect to λ satisfies a chain rule, i.e.,

$$\frac{d\mu}{d\lambda} = \frac{d\mu}{d\nu} \cdot \frac{d\nu}{d\lambda} \quad (\lambda\text{-almost everywhere}).$$

   (b) Use part (a) to prove Corollary 4.1.

2. Given $\theta \in (0,1)$, suppose $X \sim \mathsf{Bin}(n, \theta)$.

   (a) Show that the $\mathsf{Beta}(a,b)$ family is conjugate.

   (b) Under a $\mathsf{Beta}(a,b)$ prior, find the posterior mean and variance. Give an interpretation of what's happening when $n \to \infty$.

   (c) Consider the prior $\pi(\theta) = [\theta(1-\theta)]^{-1}$ for $\theta \in (0,1)$. Show that the prior is improper but, as long as $0 < x < n$, the posterior turns out to be proper.

3. Suppose $X = (X_1, \ldots, X_n)$ are iid $\mathsf{Unif}(0, \theta)$ and θ has a $\mathsf{Unif}(0,1)$ prior.

   (a) Find the posterior distribution of θ.

   (b) Find the posterior mean.

   (c) Find the posterior median.

   (d) Find the posterior mode.

4. Consider the setting in Example 4.2. Design a simulation study to assess the coverage probability of the 95% credible interval for θ for various choices of θ, sample size n, and gamma hyperparameters (a, b).

5. Find the 95% marginal credible interval for μ based on the calculations in Example 4.4. Prove that the frequentist coverage probability of the 95% credible interval is 0.95.

6. Problem 7.11 in Keener (2010, p. 126).

7. Problem 7.12 in Keener (2010, p. 126).

8. Consider the Poisson setup in Example 4.2.

   (a) Describe an approach to simulate from the predictive distribution. How do you use this posterior sample to produce a 95% prediction interval for the next observation X_{n+1}.

(b) Propose a non-Bayesian 95% prediction interval for X_{n+1}.

(c) Design a simulation study to assess the performance (coverage and length) of your Bayesian and non-Bayesian prediction intervals. Try several values of θ, sample size n, and prior hyperparameters (a, b).

9. Consider a model with density $p_\theta(x) = h(x)\exp\{\theta x - A(\theta)\}$, i.e., a simple one-parameter exponential family. Let $X_1, \ldots, X_n$ be iid from p_θ.

   (a) Consider a prior for θ, with density $\pi(\theta) = g(\theta)e^{\eta\theta - B(\eta)}$. Show that this is conjugate, and write down the corresponding posterior density.

   (b) Consider the special case where $\pi(\theta) \propto e^{\eta\theta - mA(\theta)}$ for fixed (η, m), and assume that $\pi(\theta)$ vanishes on the boundary of the parameter space. (i) Show that the prior mean of $A'(\Theta)$ is η/m; (ii) Show that the posterior mean for $A'(\Theta)$ is a weighted average of the prior mean and the sample mean.

10. Prove that $X_1, \ldots, X_n$ are exchangeable if and only if, for all $k \leq n$, all k-tuples $(X_{i_1}, \ldots, X_{i_k})$ have the same joint distribution.

11. Prove that conditionally iid random variables, satisfying (4.3), are exchangeable. You may assume existence of densities if that makes it easier.

12. Problem 7.14(b) in Keener (2010, p. 127). [Hint: Use "iterated covariance."]

13. Suppose $X|\theta \sim \mathsf{N}(\theta, 1)$ and the goal is to test $H_0 : \theta \leq \theta_0$. Consider the class Γ of symmetric and unimodal priors about θ_0. Use Theorem 4.4 to get upper and lower bounds on $\Pi_x(H_0)$, the posterior probability of H_0. [Hint: P-value will be involved.]

14. Let h be a one-to-one differentiable mapping and consider the reparametrization $\xi = h(\theta)$. Let π_θ and π_ξ are the Jeffreys priors for θ and ξ, respectively. Prove that

$$\pi_\theta(u) = \pi_\xi(h(u))|h'(u)|.$$

    This property says that Jeffreys' prior is invariant to smooth reparametrizations.

15. Monte Carlo integration is an important part of modern Bayesian analysis. The key idea is to replace difficult integration with simulation and simple averaging. In other words, if the goal is to evaluate $\mathsf{E}[h(X)] = \int h(x)\, d\mathsf{P}(x)$ for a probability measure P, then a Monte Carlo strategy is to simulate $\{X_t : t = 1, \ldots, T\}$ independent[14] from P and approximate $\mathsf{E}[h(X)]$ by $T^{-1}\sum_{t=1}^T h(X_t)$.

    Suppose $X \sim \mathsf{Unif}(0, 1)$. The goal is to use Monte Carlo integration to approximate the moment-generating function $M_X(u)$ of X on the interval $u \in (-2, 2)$.

[14] independence is not really necessary

(a) Describe your algorithm and use Hoeffding's inequality and the Borel–Cantelli lemma (Chapter 1) to prove consistency of your Monte Carlo estimator. That is, if $\hat{M}_X(u)$ is your Monte Carlo estimator, then prove that $\hat{M}_X(u) \to M_X(u)$ with probability 1 for each fixed u.

(b) Implement your method, draw a plot of your approximation with the true moment generating function overlaid, and remark on the quality of the approximation. [Hints: (i) 1000 Monte Carlo samples should be enough; (ii) you can use the same Monte Carlo samples for each point u on the grid.]

16. Use the Laplace approximation to derive the *Stirling's formula*:

$$n! \approx n^{n+(1/2)} e^{-n} \sqrt{2\pi}, \quad \text{for large } n.$$

[Hint: Use the gamma function: $n! = \Gamma(n+1) = \int_0^\infty e^{-u} u^n \, du$.]

17. Check that $\mathsf{N}(\theta, 1)$ satisfies Condition C5 in the Bernstein–von Mises setup.

18. (a) Write $\pi_n(\theta) \propto L_n(\theta)\pi(\theta)$, the posterior density of a real-valued parameter. For a function $g(\theta)$, the posterior mean is defined as

$$\mathsf{E}\{g(\Theta) \mid X\} = \frac{\int_{-\infty}^{\infty} g(\theta) L_n(\theta) \pi(\theta) \, d\theta}{\int_{-\infty}^{\infty} L_n(\theta) \pi(\theta) \, d\theta}.$$

If $g(\theta)$ is sufficiently smooth, use Laplace approximation in both the numerator and denominator to get a formula for $\mathsf{E}\{g(\Theta) \mid X\}$ in terms of the MLE $\hat{\theta}_n$.

(b) Let $\tilde{\theta}_n = \int \theta \pi_n(\theta) \, d\theta$ be the posterior mean based on iid data $X_1, \ldots, X_n$. Use the Laplace approximation to prove that $\tilde{\theta}_n$ is a consistent estimator of θ. [Hint: Show that $(\tilde{\theta}_n - \theta^\star)^2 \to 0$ in $\mathsf{P}_{\theta^\star}$-probability. First use Jensen's inequality, then take $g(\theta) = (\theta - \theta^\star)^2$ for the Laplace approximation.]

19. Write Π_n for the posterior distribution based on an iid sample of size n; for simplicity, suppose that the parameter is a scalar, but the ideas are much more general. The posterior is said to be *consistent* at $\theta^\star$ if

$$\Pi_n(\{\theta : |\theta - \theta^\star| > \varepsilon\}) \to 0 \quad \text{in } \mathsf{P}_{\theta^\star}\text{-probability, for all } \varepsilon > 0, \text{ as } n \to \infty.$$

As an application, let $X_1, \ldots, X_n$ be iid $\mathsf{N}(\theta, 1)$, and consider a prior $\pi(\theta) \propto 1$, a constant prior. Use Markov's inequality and basic properties of the normal samples to prove that the corresponding posterior is consistent at all $\theta^\star \in \mathbb{R}$.

20. For posterior consistency, etc, the prior Π must put sufficient mass near the true parameter value, say, $\theta^\star$. One way to ensure this is to assume that

$$\Pi(\{\theta : K(p_{\theta^\star}, p_\theta) < \varepsilon\}) > 0, \quad \forall\, \varepsilon > 0,$$

where K denotes the Kullback–Leibler divergence introduced in Notes I. The condition above reads as "Π satisfies the Kullback–Leibler property at $\theta^\star$."

(a) Let p_θ be an exponential family. Find $K(p_{\theta^\star}, p_\theta)$.

(b) Assuming regularity of p_θ, argue that a prior Π satisfies the Kullback–Leibler property at $\theta^\star$ if it has a positive density π in a neighborhood of $\theta^\star$.

Chapter 5

Statistical Decision Theory

5.1 Introduction

An important part of a statistical analysis is making decisions under uncertainty. In many cases, there is a cost to making incorrect decisions and so it may be a good strategy to incorporate these costs into the statistical analysis and to seek the decision which will minimize (in some sense) the expected cost. This is the focus of *statistical decision theory*.

Decision theory itself is part of the more general game theory setup that originated with von Neumann and Morgenstern in the 1950s in an economics context. In the game theory context, there are two (or more) players competing against one another and the viewpoint is typically that a win by one player is a loss to the other. Since neither player generally knows the strategy to be taken by the other, the goal for each player to pick a strategy that guarantees he/she cannot to "too bad," in some sense. The movie *A Beautiful Mind*, inspired by the life of mathematician John F. Nash, highlights the development of his "Nash equilibrium," a result in game theory, now taught to undergraduates in economics. In the statistical decision theory context, the players are the statistician and "Nature"—a hypothetical character who knows the true value of the parameter.

Statistical decision theory setup starts with the familiar ingredients: there is a sample space $(\mathbb{X}, \mathcal{A})$, a parameter space Θ, and a family of probability measures $\{\mathsf{P}_\theta : \theta \in \Theta\}$ defined on $(\mathbb{X}, \mathcal{A})$ indexed by Θ. In some cases, there may also be a prior Π on $(\Theta, \mathcal{B})$, where $\mathcal{B}$ is a σ-algebra on Θ, but this is not always necessary. The two "new" ingredients are as follows:

- An action space $\mathbb{A}$. When we think of the results of the statistical analysis as determining an action to be taken or a decision to be made by the decision-maker, then there must be a set of all such actions.
- A (non-negative) loss function $L(\theta, a)$ defined on $\Theta \times \mathbb{A}$. This function is meant to encode the "costs" of making wrong decisions. In particular, $L(\theta, a)$ represents the cost of taking action a when the parameter is θ.

Example 5.1 (Hypothesis testing)**.** In a hypothesis testing problem, we may view the parameter space as $\Theta = \{0, 1\}$, where "0" means H_0 is true and "1" means H_1 is true. Then

the action space is also $\mathbb{A} = \{0, 1\}$ with 0 and 1 corresponding to "accept H_0" and "reject H_0," respectively. Here a typical loss function is what's called 0–1 loss, i.e.,

$$L(0,0) = L(1,1) = 0 \quad \text{and} \quad L(1,0) = L(0,1) = 1.$$

That is, correct decisions cost nothing but Type I and Type II errors both cost 1 unit. But it is not always the case that Type I and Type II errors are equally costly—it is easy to extend the loss function above to such cases.

Example 5.2 (Point estimation)**.** Suppose the goal is to estimate $\psi(\theta)$, where θ is unknown but ψ is a known real-valued function. Then $\mathbb{A} = \psi(\Theta)$ is the image of Θ under ψ. The typical loss function is squared-error loss, i.e.,

$$L(\theta, a) = \big(a - \psi(\theta)\big)^2.$$

However, other loss functions like $L(\theta, a) = |\psi(\theta) - a|$ can also be considered.

Now suppose that data $X \sim \mathsf{P}_\theta$ is observed. We would like to use the information $X = x$ to help choose an action in $\mathbb{A}$ to take. A choice of action in $\mathbb{A}$ based on data x is called a *decision rule*.

Definition 5.1. A *non-randomized* decision rule δ is a function mapping $\mathbb{X}$ to $\mathbb{A}$, and the loss incurred by using $\delta(x)$ is simply given by $L(\theta, \delta(x))$. A *randomized* decision rule δ is a mapping from $\mathbb{X}$ to the set of probability measures defined on $\mathbb{A}$. In this case, the loss incurred by using $\delta(x)$ is given by the expectation

$$L(\theta, \delta(x)) = \int_{\mathbb{A}} L(\theta, a)\, \delta(x)(da).$$

A non-randomized decision rule δ is a special case of a randomized one, where $\delta(x)$ is viewed as a point mass at the number $\delta(x) \in \mathbb{A}$.

We are familiar with the concept of randomized decision rules in the context of hypothesis testing. In that setting, recall that in some (usually discrete) problems, it is not possible to achieve a specified Type I error with a non-randomized test. So the idea is to flip a coin with probability $\delta(x) \in [0, 1]$ to decide between accept and reject. For obvious reasons, non-randomized decision rules are preferred over randomized ones. We shall show below (Theorem 5.2) that for "nice" loss functions, we can safely ignore randomized decision rules.

With a given $\mathbb{A}$, L and x, the goal of decision theory is to choose a decision rule δ to minimize $L(\theta, \delta(x))$. But this typically cannot be done without the knowledge of θ. To make life simpler, we shall seek decision rules which have good properties on average, as $X \sim \mathsf{P}_\theta$. In this direction, one defines the *risk function*

$$R(\theta, \delta) = \int_{\mathbb{X}} L(\theta, \delta(x))\, \mathsf{P}_\theta(dx), \tag{5.1}$$

which is just the expected loss incurred by using decision rule δ. The goal of classical decision theory is to find a decision rule δ that minimizes $R(\theta, \delta)$ in some sense. The trouble is that there is generally no single δ that minimizes $R(\theta, \delta)$ for all θ. In such cases, one looks for other ways to minimize risk which are weaker than uniformly. These include various ways of removing θ from the risk making it just a function of δ (see Section 5.3) or by introducing constraints on δ (see Section 5.4). In these cases, one can speak of a risk-minimizing decision rule. As a first step, it helps to reduce the search to decision rules which are *admissible*, which we discuss next in Section 5.2.

5.2 Admissibility

When searching for an "optimal" decision rule, it can be helpful to rule out some procedures which are known to be suboptimal, thus reducing the size of the search space.

Definition 5.2. A decision rule δ is *inadmissible* if there exists another decision rule δ' such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all θ with strict inequality for some θ. We say that δ' *dominates* δ. If there is no such δ', then δ is *admissible*.

Roughly speaking, only those admissible decision rules need to be considered. However, not all admissible decision rules are reasonable. For example, if estimating θ under square-error loss is the goal, the decision rule $\delta(x) \equiv \theta_0$ is admissible since its the only decision rule if zero risk at $\theta = \theta_0$. But, clearly, the rule $\delta(x) \equiv \theta_0$ which focuses only on one possible value of θ will pay a huge price, in terms of risk, for $\theta \neq \theta_0$.

It turns out that admissibility of a decision rule is closely related to properties of the loss function. In particular, when the loss function $L(\theta, a)$ is *convex* in a, there are some nice properties. Here is an important result.

Theorem 5.1 (Rao–Blackwell)**.** *Let $X \sim \mathsf{P}_\theta$ and T be a sufficient statistic. Let δ_0 be a non-randomized decision rule, taking values in a convex $\mathbb{A} \subseteq \mathbb{R}^d$, with $\mathsf{E}_\theta \|\delta_0(X)\| < \infty$ for all θ. If $\mathbb{A}$ is convex and $L(\theta, \cdot)$ is a convex function for each θ, then*

$$\delta_1(x) = \delta_1(t) = \mathsf{E}\{\delta_0(X) \mid T = t\}$$

satisfies $R(\theta, \delta_1) \leq R(\theta, \delta_0)$ for all θ.

Proof. From Jensen's inequality,

$$L(\theta, \delta_1(t)) \leq \mathsf{E}\{L(\theta, \delta_0(X)) \mid T = t\} \quad \forall\, \theta.$$

Now taking expectation of both sides (with respect to the distribution of T under $X \sim \mathsf{P}_\theta$) shows that $R(\theta, \delta_1) \leq R(\theta, \delta_0)$. □

This theorem shows that, for a convex loss function, only decision rules that are functions of sufficient statistics can be admissible.[1] Furthermore, it shows how to improve on a given decision rule—just "Rao–Blackwellize" it by taking a conditional expectation given a sufficient statistic.

[1] But the rule δ_1 in the theorem need not be admissible—it requires its own proof.

Example 5.3. Suppose $X_1, \ldots, X_n$ are independent $\mathsf{N}(\theta, 1)$ random variables. The goal is to estimate $\Phi(c - \theta)$, the probability $X_1 \leq c$, for some fixed constant c, under square-error loss. That is, $L(\theta, a) = \{a - \Phi(c - \theta)\}^2$. A straightforward estimator is $\delta_0(x) = n^{-1} \sum_{i=1}^n I_{(-\infty, c]}(x_i)$. But this is not a function of the sufficient statistic $T = \bar{X}$. Since $L(\theta, \cdot)$ is convex (prove it!), the Rao–Blackwell theorem says we can improve δ_0 by taking its conditional expectation given $T = t$. That is,

$$\begin{aligned} \mathsf{E}\{\delta_0(X) \mid T = t\} &= \frac{1}{n} \sum_{i=1}^n \mathsf{E}\{I_{(-\infty, c]}(X_i) \mid T = t\} \\ &= \mathsf{P}(X_1 \leq c \mid T = t) = \Phi\Big(\frac{c - t}{\sqrt{(n-1)/n}}\Big), \end{aligned} \tag{5.2}$$

where the last equality follows from the fact that $X_1 \mid (T = t) \sim \mathsf{N}(t, \frac{n-1}{n})$; see Exercise 5. Therefore, δ_0 is inadmissible and the right-hand side of (5.2) is one estimator that beats it.

It follows (almost) immediately from the Rao–Blackwell theorem that, in the case of a convex loss function, one need not consider randomized decision rules.

Theorem 5.2. *Suppose the loss function is convex. Then for any randomized decision rule δ_0, there exists a non-randomized rule δ_1 which is no worse than δ_0 in terms of risk.*

Proof. A randomized decision rule δ_0 defines a distribution on $\mathbb{A}$ in the sense that, given $X = x$, an action is taken by sampling $A \sim \delta_0(x)$. One can express this randomized rule as a non-randomized one, call it $\delta_0'(X, U)$, that's a function of X as well as an independent random variable U whose distribution Q is free of θ. The idea is to write $A = f_x(U)$, for some suitable data-dependent function f_x. That is, U plays the role of the randomization mechanism. But we know that $T(X, U) = X$ is sufficient, so Rao–Blackwell tells us how to improve δ_0:

$$\delta_1(x) = \mathsf{E}\{\delta_0'(X, U) \mid X = x\} = \int f_x(u) \, \mathsf{Q}(du) = \int_{\mathbb{A}} a \, \delta_0(x)(da).$$

That is, the dominating non-randomized rule $\delta_1(x)$ is the expectation under $\delta_0(x)$. □

This result explains why we almost never consider randomized *estimators*—in estimation problems, the loss function (e.g., square-error) is almost always convex, so Theorem 5.2 says there's no need to consider randomized estimators.

Admissibility is an interesting property in the sense that one should only use admissible rules, but there are many admissible rules which are lousy. So admissibility alone is not enough to justify the use of a decision rule—compare this to Type I error probability control in a hypothesis testing problem, where one must also consider Type II error probabilities. There are some other general admissibility properties (e.g., all proper Bayes rules are admissible) which we will discuss below. Interestingly, there are some surprising results which state that some "standard" procedures are, in some cases, inadmissible. The most famous example of this is *Stein's paradox*: if $X \sim \mathsf{N}_d(\theta, I)$ and $d \geq 3$, then the maximum likelihood/least-squares estimator $\hat{\theta} = X$ is in fact *inadmissible*.

5.3 Minimizing a "global" measure of risk

Finding δ to minimize $R(\theta, \delta)$ uniformly over θ is an impossible task. The trouble is that there are ridiculous decision rules which do very well for certain values of θ but very bad for others. If we introduce a global measure of risk—one that does not depend on a single value of θ—then it's a bit easier to find optimal decision rules. The two common ways to eliminate θ from the risk are to integrate it out (average risk) or to maximize over θ (maximum risk), and each of these has its own rich body of theory.

5.3.1 Minimizing average risk

The first way one might consider removing θ from the risk function $R(\theta, \delta)$ is to average/integrate it out with respect to some probability distribution[2] Π on Θ. This leads to the notion of *Bayes rules*.

Definition 5.3. For a decision problem as above, suppose there is also a probability measure Π on Θ. Then for a decision rule δ, the *Bayes risk* is

$$r(\Pi, \delta) = \int R(\theta, \delta)\, \Pi(d\theta), \tag{5.3}$$

the average risk of δ with respect to Π. If there exists a $\delta = \delta_\Pi$ that minimizes $r(\Pi, \delta)$, then δ_Π is called the *Bayes rule* (with respect to Π).

In this context, the probability measure Π is not necessarily meant to describe one's prior knowledge about θ. Instead, a suitably chosen Π can help in the construction of (non-Bayes) procedures with good properties. This is the big idea behind the modern work on shrinkage estimation via penalties based on prior distributions.

To find Bayes rules, it is important to see that

$$r(\Pi, \delta) = \int_{\mathbb{X}} \int_{\Theta} L(\theta, \delta(x))\, \Pi_x(d\theta)\, \mathsf{P}_\Pi(dx),$$

where the inner integral, $\int_\Theta L(\theta, \delta(x))\, \Pi_x(d\theta)$, is known as the posterior risk, and P_Π is the marginal distribution of X. This is a consequence of Fubini's theorem. It can be shown that (in most cases) a minimizer of the posterior risk will also be the Bayes rule. This is important because minimizing the posterior risk is often easy.

Example 5.4. Consider estimation of a real parameter θ under squared error loss. Then the posterior risk is

$$\int_\Theta (\theta - \delta(x))^2\, \Pi_x(d\theta) = \mathsf{E}(\theta^2 \mid x) - 2\delta(x)\mathsf{E}(\theta \mid x) + \delta(x)^2.$$

[2]When we talk about a prior probability distribution, there is nothing special about the total mass being 1. If the mass is some other finite number, then everything goes through just fine; if the mass is infinite, however, then there's problems to be concerned about.

From this it is clear that (if all the expectations exist) the posterior risk is minimized by taking $\delta(x) = \mathsf{E}(\theta \mid x)$. Similar arguments can be used to show that, if the loss is absolute error, then the Bayes rule is a posterior median.

Example 5.5. Suppose $X_1, \ldots, X_n$ are independent $\mathsf{Ber}(\theta)$ observations. The goal is to test $H_0 : \theta \leq 0.5$ versus $H_1 : \theta > 0.5$ under 0–1 loss. For simplicity, assume n is even and that $T = \sum_{i=1}^n X_i$. If the prior is $\mathsf{Unif}(0,1)$, then the posterior is $\mathsf{Beta}(t+1, n-t+1)$. The posterior risk, call it $r_t(H_0)$, for choosing H_0 is the posterior probability of H_1, i.e., $r_t(H_0) = \Pi_t(H_1)$. Likewise, the posterior risk of choosing H_1 is the posterior probability of H_0, i.e., $r_t(H_1) = \Pi_t(H_0) = 1 - r_t(H_0)$. The Bayes rule chooses H_0 or H_1 depending on which of $r_t(H_0)$ and $r_t(H_1)$ is the smaller. That is, the Bayes rule rejects H_0 if and only if $r_t(H_0) < 0.5$. However, if $t = n/2$, then $r_t(H_0) = r_t(H_1) = 0.5$ and so it's not clear which to choose. Therefore, the Bayes rule is a randomized one, given by

$$\delta(x) = \begin{cases} \text{choose } H_0 & \text{if } t(x) < n/2 \\ \text{choose } H_1 & \text{if } t(x) > n/2 \\ \text{flip a fair coin to decide} & \text{if } t(x) = n/2. \end{cases}$$

The main results we will consider here are that, under certain conditions, Bayes rules are admissible. Therefore, one cannot rule out being a Bayesian based on admissibility constraints alone. What is even more interesting is that there are theorems which state that essentially all admissible decision rules are Bayes; see Section 5.5.

Theorem 5.3. *Suppose Θ is a subset of $\mathbb{R}^d$ such that every neighborhood of every point in Θ intersects the interior of Θ. Let Π be a measure on Θ such that $\lambda \ll \Pi$, where λ is Lebesgue measure. Suppose that $R(\theta, \delta)$ is continuous in θ for each δ for which the risk is finite. If the Bayes rule δ_Π has finite risk, then it is admissible.*

Proof. Suppose δ_Π is not admissible. Then there is a δ_1 such that $R(\theta, \delta_1) \leq R(\theta, \delta_\Pi)$ for all θ with strict inequality for some, say for θ_0. By continuity of the risk function, there is an open neighborhood N of θ_0, which intersects the interior of Θ, such that $R(\theta, \delta_1) < R(\theta, \delta_\Pi)$ for all $\theta \in N$. Since Lebesgue measure is is dominated by Π and N is open, we must have $\Pi(N) > 0$. This implies that $r(\Pi, \delta_1) < r(\Pi, \delta_\Pi)$, which is a contradition. Therefore, δ_Π is admissible. □

Example 5.6. Consider an exponential family distribution with natural parameter space Θ containing an open set. For estimating $g(\theta)$ for some continuous function g, consider square-error loss $L(\theta, a) = (a - g(\theta))^2$. Since Θ is convex (Chapter 2), the neighborhood condition is satisfied. Also, the risk function for any δ with finite variance will be continuous in θ. Finally, if Π has a positive density π on Θ with respect to Lebesgue measure, then Lebesgue $\ll \Pi$ also holds. Therefore, the Bayes rule δ_Π is admissible.

Theorem 5.4. *Suppose $\mathbb{A}$ is convex and all P_θ are absolutely continuous with respect to each other. If $L(\theta, \cdot)$ is strictly convex for each θ, then for any probability measure Π on Θ, the Bayes rule δ_Π is admissible.*

Proof. Suppose δ_Π is not admissible. Then there exists δ_0 such that $R(\theta, \delta_0) \leq R(\theta, \delta_\Pi)$ with strict inequality for some θ. Define a new decision rule $\delta_1(x) = \frac{1}{2}\{\delta_\Pi(x) + \delta_0(x)\}$, which is valid since $\mathbb{A}$ is convex. Then for all θ we have

$$\begin{aligned} R(\theta, \delta_1) &= \int_{\mathbb{X}} L(\theta, \tfrac{1}{2}\{\delta_\Pi(x) + \delta_0(x)\})\, \mathsf{P}_\theta(dx) \\ &\leq \int_{\mathbb{X}} \tfrac{1}{2}\{L(\theta, \delta_\Pi(x)) + L(\theta, \delta_0(x))\}\, \mathsf{P}_\theta(dx) \\ &= \tfrac{1}{2}\{R(\theta, \delta_\Pi) + R(\theta, \delta_0)\} \\ &\leq R(\theta, \delta_\Pi). \end{aligned}$$

The first inequality will be strict unless $\delta_\Pi(X) = \delta_0(X)$ with P_θ-probability 1. However, since the P_θ's are absolutely continuous with respect to each other, it follows that the first inequality is strict unless $\delta_\Pi(X) = \delta_0(X)$ with P_θ-probability 1 for all θ. Hence the first inequality above is strict unless $\delta_\Pi(X)$ and $\delta_0(X)$ have exactly the same distribution. Since this would violate the supposition that δ_1 dominates δ_Π, it must be that the first inequality above is strict for all θ. That is, $R(\theta, \delta_1) < R(\theta, \delta_\Pi)$ for all θ. Averaging both sides over Π leads to the conclusion that $r(\Pi, \delta_1) < r(\Pi, \delta_\Pi)$, which violates the assumption that δ_Π is the Bayes rule. Therefore, δ_Π must be admissible. □

One can extend the definition of Bayes rules to cases where Π is a measure, not necessarily a probability measure. In my opinion, the use of improper priors is more reasonable in this case (compared to the purely Bayes case), because the goal here is simply to construct decision rules with good risk properties.[3]

Definition 5.4. Let $(d\mathsf{P}_\theta/d\mu)(x) = p_\theta(x)$. Let Π be a measure on Θ and suppose that, for every x, there exists $\delta(x)$ such that

$$\int_\Theta L(\theta, \delta(x)) p_\theta(x)\, \Pi(d\theta) = \min_{a \in \mathbb{A}} \int_\Theta L(\theta, a) p_\theta(x)\, \Pi(d\theta).$$

Then $\delta = \delta_\Pi$ is called a *generalized Bayes rule* with respect to Π.

The difference between Definition 5.4 and Definition 5.3 is that the former does not require Π to be a probability/finite measure. Things look a little different in this case because, if the prior is improper, then the posterior might also be improper, which can make defining the Bayes rule as a minimizer of posterior risk problematic.

For example, if $X_1, \ldots, X_n$ are iid $\mathsf{N}(\theta, 1)$ and Π is Lebesgue measure on $(-\infty, \infty)$, then $\delta_\Pi(x) = \bar{x}$ is the corresponding generalized Bayes rule. But note that the admissibility result in Theorem 5.3 does not generally hold for generalized Bayes rules. The additional condition is that the risk function is Π-integrable.

Theorem 5.5. *Suppose Θ is as in Theorem 5.3. Assume $R(\theta, \delta)$ is continuous in θ for all δ. Let Π be a measure that dominates Lebesgue measure, and δ_Π the corresponding generalized Bayes rule. If $(x, \theta) \mapsto L(\theta, \delta_\Pi(x)) p_\theta(x)$ is $\mu \times \Pi$-integrable, then δ_Π is admissible.*

[3] If the goal is to get a rule with good properties, who really cares *how* the rule is constructed...?

Proof. Use Fubini's theorem and Definition 5.4. □

Back to the normal example, consider again the generalized Bayes rule $\delta_\Pi(x) = \bar{x}$ (with respect to prior Π equal to Lebesgue measure) for estimating θ under square-error loss. The risk function $R(\theta, \delta_\Pi)$ for $\delta_\Pi(x) = \bar{x}$ is constant (equal to $1/n$) and so is not integrable with respect to Π = Lebesgue. Therefore, Theorem 5.5 is not enough to prove admissibility of the maximum likelihood estimator.

An alternative approach is to consider a sequence $\{\Pi_s : s \geq 1\}$ of proper or improper priors, and the corresponding sequence of Bayes rules $\{\delta_{\Pi_s} : s \geq 1\}$. The next theorem is a powerful and general tool for proving admissibility. A similar result is given in Schervish (1995, p. 158–159).

Theorem 5.6. *Assume that each decision rule δ has a continuous risk function $R(\theta, \delta)$. Suppose there are finite measures $\{\Pi_s : s \geq 1\}$ on Θ such that a generalized Bayes rule $\delta_s = \delta_{\Pi_s}$ exists for all s, and $\liminf_s \Pi_s(B) > 0$ for every open ball $B \subset \Theta$. If δ is a decision rule such that*

$$\lim_{s\to\infty} \big\{ r(\Pi_s, \delta) - r(\Pi_s, \delta_s) \big\} = 0,$$

then δ is admissible.

Proof. See Keener (2010, p. 215). □

To illustrate the use of this theorem, we will prove that, for $X \sim \mathsf{N}(\theta, 1)$, the maximum likelihood estimator $\hat{\theta} = x$ is admissible under square-error loss. The more general $n > 1$ case follows from this by making a change of scale.

Example 5.7. Consider a sequence of measures $\Pi_s = \sqrt{s}\mathsf{N}(0, s)$, $s \geq 1$; these are finite but not probability measures. Let $\delta(x) = x$ be the maximum likelihood estimator. The generalized Bayes rules are given by $\delta_s(x) = sx/(s+1)$ and the Bayes risks are

$$r(\Pi_s, \delta) = \sqrt{s} \quad \text{and} \quad r(\Pi_s, \delta_s) = s^{3/2}/(s+1).$$

The difference $r(\Pi_s, \delta) - r(\Pi_s, \delta_s) = s^{1/2}/(s+1)$ goes to zero as $s \to \infty$. The only thing left to do is check that $\Pi_s(B)$ is bounded away from zero for all open intervals $x \pm m$. For this, we have

$$\Pi_s(x \pm m) = \frac{1}{\sqrt{2\pi}} \int_{x-m}^{x+m} e^{-u^2/2s}\, du,$$

and since the integrand is bounded by 1 for all s and converges to 1 as $s \to \infty$, the dominated convergence theorem says that $\Pi_s(x \pm m) \to 2m(2\pi)^{-1/2} > 0$. It follows from Theorem 5.6 that $\delta(x) = x$ is admissible.

Example 5.8. Let $X \sim \mathsf{Bin}(n, \theta)$, so that the MLE of θ is the sample mean $\delta(X) = X/n$. The goal is to show, using Theorem 5.6, that δ is admissible under squared-error loss. For this, we need a sequence of proper priors $\{\Pi_s : s \geq 1\}$ for θ. Since beta priors are conjugate

for the binomial model, a reasonable starting point is to consider $\theta \sim \mathsf{Beta}(s^{-1}, s^{-1})$. For such a model, the Bayes rule is

$$\mathsf{E}(\theta \mid X) = \frac{X + s^{-1}}{n + 2s^{-1}}.$$

It is clear that, as $s \to \infty$, the Bayes rule $\delta_{\Pi_s}(X)$ converges to the sample mean $\delta(X) = X/n$. But the limit of the beta priors is not a proper prior for θ; in fact, the limit of the priors has a density that is proportional to $\{\theta(1-\theta)\}^{-1}$, which is improper.

The beta priors themselves do not satisfy the conditions of Theorem 5.6 (Exercise 7). Fortunately, there is a simple modification of the beta prior that does work. Take Π_s to have density π_s which is just the $\mathsf{Beta}(s^{-1}, s^{-1})$ without the normalizing constant:

$$\pi_s(\theta) = \{\theta(1-\theta)\}^{1/s-1}$$

or, in other words,

$$\pi_s(\theta) = \lambda(s)\mathsf{Beta}(\theta \mid s^{-1}, s^{-1}), \quad \text{where} \quad \lambda(s) = \frac{\Gamma(s^{-1})^2}{\Gamma(2s^{-1})}.$$

Since Π_s is just a rescaling of the beta prior from before, it is clear that the Bayes rule $\delta_{\Pi_s}(X)$ for the new prior is the same as for the beta prior above. Then the Bayes risk calculations are relatively simple (Exercise 8). With the sequence of priors Π_s, which are proper but not probability measures, it follows from Theorem 5.6 that the sample mean $\delta(X) = X/n$ is an admissible estimator of θ.

There are also some general theorems about the admissibility of the "standard estimators" in exponential families that cover the result proved in Example 5.7. The conditions of such theorems are rather technical so we won't cover them here. See Schervish (1995), pages 160–161, for a detailed statement and proof of one such theorem.

Another important use of Bayes rules will be seen in the next section, where Bayes rules with respect to certain "worst-case scenario" priors will produce minimax rules.

5.3.2 Minimizing maximum risk

In the previous section we measured the global performance of a decision rule by averaging its risk function with respect to a probability measure Π on Θ. But this is not the only way one can summarize a decision rule's performance. Another criterion is to consider the maximum of the risk function $R(\theta, \delta)$ as θ ranges over Θ. This maximum risk represents the worst the decision rule δ can do. Then the idea is to choose δ so that this worst-case performance is as small as possible.

Definition 5.5. For a decision problem with risk function $R(\theta, \delta)$, a minimax decision rule δ_0 satisfies

$$\sup_\theta R(\theta, \delta_0) \le \sup_\theta R(\theta, \delta)$$

for all decision rules δ. The minimax rule protects against the worst-case scenario in the sense that it MINImizes the MAXimum risk.

The origin of this approach is in the game theory scenario where one is playing against an opponent. The opponent's goal is to maximize your own loss, so one approach to game play would be to choose a strategy that minimizes the maximum loss, i.e., the minimax strategy. But in a statistical decision problem, one might consider this strategy too conservative, or pessimistic, since it does not consider how likely the value of θ at which the maximum occurs. Nevertheless, minimax decision rules have a rich history.

Theorem 5.7. *Let Π be a probability measure and δ_Π the corresponding Bayes rule. If $r(\Pi, \delta_\Pi) = \sup_\theta R(\theta, \delta_\Pi)$, then δ_Π is minimax.*

Proof. Let δ be some other procedure. Then $\sup_\theta R(\theta, \delta) \geq r(\Pi, \delta) \geq r(\Pi, \delta_\Pi)$. Since $r(\Pi, \delta_\Pi) = \sup_\theta R(\theta, \delta_\Pi)$ by assumption, minimaxity follows. □

Corollary 5.1. *A Bayes rule δ_Π with constant risk is minimax.*

Proof. If the risk is constant, then the conditions of Theorem 5.7 hold trivially. □

As one might guess, a prior Π for which the average risk equals the maximum risk is a particular weird prior. That is, it must put all its mass on θ values where the risk of the Bayes rule δ_Π is large. These are the "worst-case scenario" priors mentioned at the end of the previous section. Such a prior is called a *least favorable prior* and it satisfies

$$r(\Pi, \delta_\Pi) \geq r(\Pi', \delta_{\Pi'}) \quad \text{for all priors } \Pi'.$$

To see this, let Π be a prior satisfying $r(\Pi, \delta_\Pi) = \sup_\theta R(\theta, \delta_\Pi)$. For another prior Π' we have

$$r(\Pi', \delta_{\Pi'}) \leq r(\Pi', \delta_\Pi) \leq \sup_\theta R(\theta, \delta_\Pi) = r(\Pi, \delta_\Pi),$$

hence Π is least favorable. Practically, least favorable priors are not particularly useful. However, this connection between least favorable priors and minimax estimators provides a powerful technique for finding minimax estimators.

Example 5.9. Let $X \sim \mathsf{Bin}(n, \theta)$. The goal is to find a minimax estimator of θ under square-error loss. Consider a conjugate $\mathsf{Beta}(\alpha, \beta)$ prior. Then the posterior mean is

$$\delta(X) = \mathsf{E}(\theta \mid X) = aX + b = \frac{1}{\alpha + \beta + n} X + \frac{\alpha}{\alpha + \beta + n}.$$

The risk function for δ is

$$R(\theta, \delta) = \mathsf{V}_\theta\{aX + b - \theta\} + \mathsf{E}_\theta^2\{aX + b - \theta\} = A\theta^2 + B\theta + C,$$

where A, B, and C depend on (α, β, n) and you're invited to find the exact expressions in Exercise 9. The risk function is constant iff $A = B = 0$, which holds iff $\alpha = \beta = \frac{1}{2}\sqrt{n}$. Therefore, the Bayes rule with constant risk is

$$\delta(x) = \frac{x + \frac{1}{2}\sqrt{n}}{n + \sqrt{n}},$$

and so this guy is minimax according to Corollary 5.1.

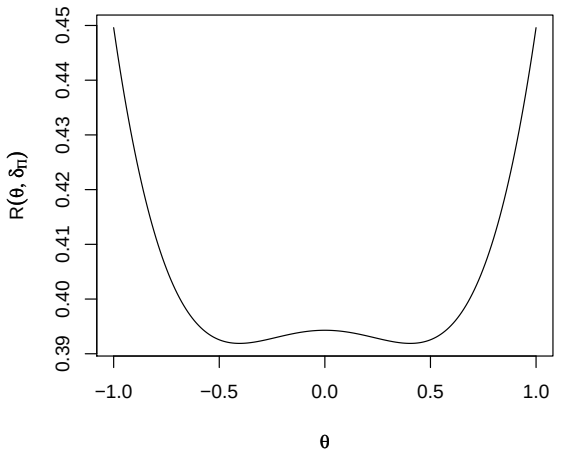


Figure 5.1: Plot of risk function in Example 5.10.

The next example sheds some light on the nature of a least favorable prior in the case of a constrained parameter space.

Example 5.10. Let $X \sim \mathsf{N}(\theta, 1)$ where it is known that $|\theta| \leq 1$. It can be shown (Exercise 12) that the MLE $\hat{\theta} = X$ is inadmissible in this case. Here we shall find a minimax estimator δ of θ under square-error loss. Consider a probability measure Π that puts probability 0.5 on the endpoints of the interval $[-1, 1]$. That is, $\Pi(\{-1\}) = \Pi(\{1\}) = 0.5$. In this case, the posterior distribution is determined by

$$\Pi_x(\{1\}) = \frac{\varphi(x-1)/2}{\varphi(x-1)/2 + \varphi(x+1)/2} = \frac{\varphi(x-1)}{\varphi(x-1) + \varphi(x+1)},$$

where φ is the standard normal density function. Then the posterior mean is

$$\delta_\Pi(x) = \frac{\varphi(x-1) - \varphi(x+1)}{\varphi(x-1) + \varphi(x+1)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \tanh(x).$$

It can be shown that the risk function $R(\theta, \delta_\Pi)$ is symmetric and maximized at $\theta = \pm 1$; see Figure 5.1. In this case the maximum risk equals the average of $R(\pm 1, \delta_\Pi)$ so, by Theorem 5.7, δ_Π is minimax.

In Exercise 10(b) you're invited to show using direct arguments that, in a normal mean problem, the sample mean is a minimax estimator under squared error loss. The multivariate case with more general loss function is considered in Section 5.6. This argument rests on an interesting result in convex analysis called Anderson's lemma.

Minimax procedures are pessimistic in nature and, in the relatively simple problems considered so far, it is not so difficult to find better procedures, e.g., maximum likelihood estimators. However, if we move away from these "relatively simple" problems, maximum likelihood may not work and we need some different criteria for constructing estimators, etc.

In these problems, typically the parameter has dimension increasing at least as fast as the sample size, and *asymptotically minimax* procedures provide a benchmark for comparison. This notion of asymptotic minimaxity is covered in more detail in Chapter ??.

5.4 Minimizing risk under constraints

Earlier we saw that finding a δ that minimizes the risk $R(\theta, \delta)$ uniformly over θ is impossible. In Section 5.3 we saw two common strategies for introducing a global risk and finding optimal decision rules. An alternative approach is to introduce a "reasonable" constraint on the set of decision rules one is willing to consider. In this case, it can be possible to find a (constrained) δ for which $R(\theta, \delta)$ is minimized uniformly over θ.

5.4.1 Unbiasedness constraints

We are familiar with unbiasedness in the estimation context from Stat 411. However, unbiasedness is a general condition for decision rules. That is, for a loss function $L(\theta, a)$, a decision rule δ is *unbiased* if

$$\mathsf{E}_\theta\{L(\theta', \delta(X))\} \geq \mathsf{E}_\theta\{L(\theta, \delta(X))\}, \quad \forall\, \theta'. \tag{5.4}$$

In Exercise 13 you are invited to show that, if the goal is to estimate $g(\theta)$ under squared-error loss, then the unbiasedness condition (5.4) is equivalent to the familiar definition, i.e., $\mathsf{E}_\theta\{\delta(X)\} = g(\theta)$ for all θ. Though unbiasedness is more general (see Section 5.4.3), we shall focus here on the estimation problem.

First is an interesting result that says we need not look at Bayes estimators in this context, because (except in weird cases) they cannot be unbiased.

Theorem 5.8. *No unbiased estimator $\delta(X)$ can be a Bayes estimator unless the prior Π satisfies $\Pi\{\theta : R(\theta, \delta) = 0\} = 1$.*

Proof. Suppose δ is a Bayes rule (under square-error loss with respect to Π) and is unbiased. Then we know

$$\delta(X) = \mathsf{E}\{g(U) \mid X\} \quad \text{and} \quad g(U) = \mathsf{E}\{\delta(X) \mid U\}.$$

Then depending on the order in which we condition, we get

$$\mathsf{E}[g(U)\delta(X)] = \begin{cases} \mathsf{E}[g(U)\mathsf{E}\{\delta(X) \mid U\}] = \mathsf{E}[g(U)^2] & \text{conditioning on } U \\ \mathsf{E}[\delta(X)\mathsf{E}\{g(U) \mid X\}] = \mathsf{E}[\delta(X)^2] & \text{conditioning on } X. \end{cases}$$

Therefore, $\mathsf{E}[g(U)^2] = \mathsf{E}[\delta(X)^2]$ and, hence,

$$r(\Pi, \delta) = \mathsf{E}[\delta(X) - g(U)]^2 = \mathsf{E}[\delta(X)^2] - 2\mathsf{E}[g(U)\delta(X)] + \mathsf{E}[g(U)^2] = 0.$$

But the Bayes risk also satisfies $r(\Pi, \delta) = \int R(\theta, \delta)\, d\Pi(\theta)$. Since $R(\theta, \delta) \geq 0$ for all θ, the only way the Π-integral can be zero is if Π assigns probability 1 to the set of θ where $R(\theta, \delta)$ vanishes. This proves the claim. □

So restricting to unbiased estimators necessarily rules out (reasonable) Bayes estimators. But this doesn't help to find the best estimator, nor does it even suggest that there is a "best" estimator. Fortunately, there is a very powerful result—the Lehmann–Scheffe theorem—which states that there is indeed a rule that uniformly minimizes risk and, moreover, gives easily verifiable sufficient conditions to identify this best estimator.

Theorem 5.9 (Lehmann–Scheffe)**.** *Let $X \sim \mathsf{P}_\theta$ and suppose that T is a complete sufficient statistic. Suppose the goal is to estimate $g(\theta)$ under convex loss, and that an unbiased estimator exists. Then there exists an essentially unique unbiased estimator that is a function of T and uniformly minimizes the risk.*

Proof. See Keener (2010, p. 62). □

There's a few things worth mentioning about this important theorem. First, notice that square-error loss is not important; all that really matters is that the loss function is convex. Second, it does not guarantee that an unbiased estimator exists; there are examples where no unbiased estimator exists (e.g., estimating $1/\theta$ in a $\mathsf{Bin}(n, \theta)$). If there is an unbiased estimator, then the Rao–Blackwell theorem shows how to improve it by conditioning on T. It's the fact that T is also complete which leads to the uniqueness. Indeed, if there are two unbiased estimators which are functions of T, then their difference $f(T) = \delta_1(T) - \delta_2(T)$ satisfies $\mathsf{E}_\theta f(T) = 0$ for all θ. Completeness of T implies that $f = 0$ a.e., which implies δ_1 and δ_2 are actually (a.e.) the same.

5.4.2 Equivariance constraints

For a sample space $\mathbb{X}$, let $\{\mathsf{P}_\theta : \theta \in \Theta\}$ be a group transformation model with respect to a group $\mathscr{G}$ of transformations $g : \mathbb{X} \to \mathbb{X}$. That is, if $X \sim \mathsf{P}_\theta$, then $gX \sim \mathsf{P}_{\theta'}$ for some θ' in Θ. This particular θ' is determined by θ and the transformation g. In other words, the transformations g also act on Θ, but possibly in a different way than they act on $\mathbb{X}$. We shall refer to this transformation on Θ determined by g as $\bar{g}$, and $\bar{\mathscr{G}}$ the collection of all such $\bar{g}$'s. It can be shown that $\bar{\mathscr{G}}$ is also a group. To summarize, we have groups $\mathscr{G}$ and $\bar{\mathscr{G}}$ acting on $\mathbb{X}$ and Θ, respectively, which are related to the distribution P_θ in the following way:

$$X \sim \mathsf{P}_\theta \iff gX \sim \mathsf{P}_{\bar{g}\theta},$$

where $\bar{g} \in \bar{\mathscr{G}}$ is determined by $g \in \mathscr{G}$.

This is a special structure imposed on the distribution P_θ. We are familiar with the special location and scale structures, where $\mathscr{G}$ and $\bar{\mathscr{G}}$ are the same, but there are others; for example, you saw the Weibull distribution as a group transformation model, and you wrote down explicitly the $\bar{g}$ for a given g.

In this section we will investigate the effect of such a structure in the statistical decision problem. First we need to impose this structure on some of the other ingredients, such as action space, loss function, and decision rules. We shall do this quickly here; see Schervish (1995, Sec. 6.2.1) or Lehmann and Casella (1998, Sec. ??) for more details.

- The loss function is called invariant (with respect to $\mathscr{G}$ or $\bar{\mathscr{G}}$...) if, for each $g \in \mathscr{G}$ (or each $\bar{g} \in \bar{\mathscr{G}}$) and each $a \in \mathbb{A}$, there exists a (unique) $a' \in \mathbb{A}$ such that $L(\bar{g}\theta, a') = L(\theta, a)$ for all θ. In this case, the group also (in)directly acts on the action space $\mathbb{A}$, i.e., a' is determined by a and g. Write $a' = \tilde{g}a$. Then it can be shown that the collection $\tilde{\mathscr{G}}$ of transformations $\tilde{g} : \mathbb{A} \to \mathbb{A}$ also forms a group.
- A function h defined on $\mathbb{X}$ (or some other space like Θ or $\mathbb{A}$ equipped with a group of transformations) is called *invariant* if $h(gx) = h(x)$ for all $x \in \mathbb{X}$ and all $g \in \mathscr{G}$. Alternatively, a function $f : \mathbb{X} \to \mathbb{A}$ is *equivariant* if $f(gx) = \tilde{g}f(x)$.
- We shall focus on decision rules δ which are equivariant. The intuition is that our decision rules should be consistent with the assumed structure. For example, in a location parameter problem, a shift of the data by a constant should cause a shift of our estimator of the location by the same amount.

A decision problem whose ingredients satisfy all these properties will generically be called an *invariant decision problem.*

We shall view the insistence that the decision rule be equivariant as a constraint on the possible decision rules, just like unbiasedness is a constraint. Then the question is if there is an *equivariant* rule that uniformly minimizes the risk. The first result is a step in this direction.

Theorem 5.10. *In an invariant decision problem, the risk function $R(\theta, \delta)$ of an equivariant decision rule δ is an invariant function on Θ, i.e., constant on orbits of $\bar{\mathscr{G}}$.*

The orbits referred to in Theorem 5.10 are the sets $O_\theta = \{\theta' \in \Theta : \theta' = \bar{g}\theta, \bar{g} \in \bar{\mathscr{G}}\}$. This O_θ consists of all possibly images of θ under transformations $\bar{g} \in \bar{\mathscr{G}}$. Then an equivalent definition of an invariant function is one that's constant on orbits. An invariant function is called *maximal* if the constant values are different on different orbits. Maximal invariants are important, but we won't discuss these further here.

An interesting special case is when the group $\bar{\mathscr{G}}$ has only one orbit, in which case, the risk function in Theorem 5.10 is everywhere a constant. Groups that have just a single orbit are called transitive. One-dimensional location transformations correspond to transitive groups; same for scale transformations. In this case, it is easy to compare risk functions of equivariant decision rules.

The question is if there is an equivariant rule to minimize risk. There is a general result along these lines. I will not give a precise statement.

Theorem 5.11. *Consider an invariant decision problem. Under some assumptions, if the formal Bayes rule with respect to the right invariant Haar prior on $\bar{\mathscr{G}}$ exists, then it is the minimum risk equivariant rule.*

For a precise statement and proof of this theorem, see Schervish (1995), Theorem 6.59. The major challenge in understanding this theorem is the definition of Haar measure.[4] This is perhaps too far beyond our scope, but we can look at a simple but important example, namely, equivariant estimation of a location parameter.

[4]Technically, the Haar measure is obtained by equipping $\bar{\mathscr{G}}$, a locally compact topological group, with

Example 5.11. Consider a location parameter problem, where the density of $X_1, \ldots, X_n$ under P_θ has the form $p_\theta(x_1, \ldots, x_n) = p_0(x_1 - \theta, \ldots, x_n - \theta)$, $\theta \in \mathbb{R}$. That is, $X_i = \theta + Z_i$, where $Z_1, \ldots, Z_n$ have distribution P_0. In this case, all the groups $\mathscr{G}$, $\bar{\mathscr{G}}$, and $\tilde{\mathscr{G}}$ are (isomorphic to) the group of real numbers under addition. For the real numbers under addition, the (left and right) invariant measure is Lebesgue measure λ (why?). An invariant loss function is of the form $L(\theta, a) = L(a - \theta)$. Then the theorem says that the minimum risk equivariant estimator δ_λ is the formal Bayes rule based on a formal Lebesgue measure prior. That is, $\delta_\lambda(x)$ is the $\delta(x)$ that minimizes

$$\frac{\int_\Theta L(\delta(x) - \theta) p_0(x - \theta 1)\, d\theta}{\int_\Theta p_0(x - \theta 1)\, d\theta}.$$

In the case where $L(a-\theta) = (a-\theta)^2$, squared-error loss, we know that δ_λ is just the posterior mean under the (formal) Lebesgue measure prior λ, i.e.,

$$\delta_\lambda(x) = \frac{\int_\Theta \theta p_0(x - \theta 1)\, d\theta}{\int_\Theta p_0(x - \theta 1)\, d\theta}.$$

This estimator—*Pitman's estimator*, $\hat{\theta}_{\text{pit}}$—is the minimum risk equivariant estimator. Note that in the case where $\mathsf{P}_0 = \mathsf{N}(0, 1)$, Pitman's estimator is $\hat{\theta}_{\text{pit}}(x) = \bar{x}$.

5.4.3 Type I error constraints

In a testing problem with 0–1 loss, it can be shown (Exercise 3) that the risk function is the sum of the Type I and Type II error probabilities. It should be clear from our previous knowledge of hypothesis tests that when we make Type I error probability small then the Type II error probability increases, and vice versa, so it's not clear how to strictly minimize this risk. Therefore, the usual strategy is to fix the Type I error probability at some $\alpha \in (0, 1)$ and try to find a test, satisfying this constraint, that minimizes Type II error probability (or maximizes power). This is the idea behind *most powerful* tests.

Here we will focus on the simplest situation. Suppose X is a realization from one of two models P_0 and P_1, both having densities p_0 and p_1 on $\mathbb{X}$ with respect to μ. Then the goal is to test

$$H_0 : X \sim \mathsf{P}_0 \quad \text{versus} \quad H_1 : X \sim \mathsf{P}_1.$$

This is called the simple-versus-simple testing problem. In this case, a decision rule is a function δ mapping $\mathbb{X}$ to $[0, 1]$. In a non-randomized problem, δ maps $\mathbb{X}$ to $\{0, 1\}$. The important theorem along these lines is as follows.

---

a σ-algebra of measurable subsets. Then a measure can be defined as usual. There is also a very elegant theory of integration in the context; see Eaton (1989). The left Haar measure λ is one that is invariant under actions on Θ, $\lambda(\bar{g}B) = \lambda(B)$ for all measurable $B \subset \bar{\mathscr{G}}$; the right Haar measure ρ is defined similarly. These two measures are the same iff $\bar{\mathscr{G}}$ is an Abelian group. For example, in location groups and scale groups, left and right Haar measures are equal; in a location-scale group, they are not. Haar measures are generally not finite, which explains why we use the adjective "formal" in the statement of Theorem 5.11.

Theorem 5.12 (Neyman–Pearson)**.** *For fixed $\alpha \in (0,1)$, the most powerful α-level test is is given by*

$$\delta(x) = \begin{cases} 0 & \text{if } p_1(x) < k_\alpha p_0(x) \\ \gamma & \text{if } p_1(x) = k_\alpha p_0(x) \\ 1 & \text{if } p_1(x) > k_\alpha p_0(x), \end{cases} \tag{5.5}$$

where γ and $k(\alpha)$ are uniquely determined by the constraint

$$\alpha = \mathsf{P}_0\Big\{\frac{p_1(X)}{p_0(X)} > k_\alpha\Big\} + \gamma \mathsf{P}_1\Big\{\frac{p_1(X)}{p_0(X)} = k_\alpha\Big\}.$$

Proof. See my Stat 411 notes on hypothesis testing. □

Note that the γ part of the theorem allows for randomized decision rules. That is, if the particular $X = x$ observed satisfies $p_1(x) = k_\alpha p_0(x)$, then the rule says to flip a coin with success probability γ and reject H_0 iff the coin lands on heads. This randomization mechanism is typically not needed in continuous data problems since the event that the likelihood ratio exactly equals k_α has probability 0. But we cannot rule out randomized test at the outset because the 0–1 loss is not convex.

Here's an alternative interpretation of the familiar Neyman–Pearson lemma. We could index the tests δ in (5.5) by the particular α; write them as δ_α. Now take any other test δ' for this particular problem. It has some Type I error probability α'. Then the theorem shows that $\delta_{\alpha'}$ dominates δ' in terms of risk. Therefore, δ' is inadmissible.

In the discussion above, we focused on the case of simple-versus-simple hypothesis testing. Next are a few remarks related to some more general problems.

- If the alternative is one-sided (e.g., $H_1 : \theta > \theta_0$), then it is often the case that the simple-versus-simple test coming from the Neyman–Pearson lemma is still the best one. The point is that the Neyman–Pearson test will not actually depend on the value θ_1 in the simple alternative.
- When the alternative is two-sided, there is the well-known fact that there is generally no uniformly most powerful test. To account for this, one can further focus on *unbiased* tests that satisfy (5.4). In particular, in many cases, there is a uniformly most powerful unbiased test; Lehmann and Romano (2005) for a careful treatment of uniformly most powerful tests and the unbiasedness condition.

5.5 Complete class theorems

A class of decision rules is called a complete class, denoted by $\mathcal{C}$, if for any $\delta_1 \notin \mathcal{C}$, there exists a rule $\delta_0 \in \mathcal{C}$ such that $R(\theta, \delta_0) \le R(\theta, \delta_1)$ for all θ with strict inequality for some θ. In other words, no δ outside of $\mathcal{C}$ is admissible. Here's a few interesting facts:

- If the loss function is convex, then the set of all decision rules which are functions of a sufficient statistic forms a complete class.

- If the loss function is convex, then the set of all non-randomized decision rules forms a complete class.
- The set of tests of the form (5.5) (indexed by α) forms a complete class.

Although a complete class $\mathcal{C}$ contains all admissible decision rules, there may be many rules in $\mathcal{C}$ which are inadmissible. Therefore, it would be interesting to identify the smallest complete class. A complete class $\mathcal{C}$ is called *minimal* if there is no proper subset of $\mathcal{C}$ that is complete. It can be shown (see Exercise 20) that a minimal complete class is exactly the set of admissible decision rules.

The result we will focus on here is one which says that (limits of) Bayes rules form a complete class or, in other words, for any decision rule δ, there is an "approximately Bayes" rule $\delta^\star$ such that the risk of $\delta^\star$ is not everywhere greater than the risk of δ. One can give this result a topological flavor—roughly, the proper prior Bayes rules form a dense subset of all admissible rules.

Theorem 5.13. *Estimators that satisfy the conditions of Theorem 5.6 form a complete class.*

As a special case of this theorem, if the model is part of an exponential family and if δ is a limit of Bayes rules, then there exists a subsequence $\{\Pi_{s'}\}$ such that $\Pi_{s'} \to \Pi$ and δ is the Bayes rule δ_Π corresponding to this limit. That is, the class of all generalized Bayes rules forms a complete class in the exponential family case.

5.6 On minimax estimation of a normal mean

Here we are interested in minimax estimation of a normal mean vector θ, under loss function more general than squared error, based on a normal sample $X \sim \mathsf{N}_d(\theta, \Sigma)$, where the covariance matrix Σ is known. The kind of loss function we shall consider are those of the form $L(\theta, a) = W(a - \theta)$, where W is a "bowl-shaped" function.

Definition 5.6. A function $W : \mathbb{R}^d \to [0, \infty]$ is bowl-shaped if $\{x : W(x) \leq \alpha\}$ is convex and symmetric about the origin for all $\alpha \geq 0$.

In the case $d = 1$, the function $W(x) = x^2$ is bowl-shaped; so, the results to be developed below will specialize to the case of estimating a scalar normal mean under regular squared error loss. The d-dimensional analogue of squared error loss is $L(\theta, a) = \|a - \theta\|^2$, where $\|\cdot\|$ is the usual Euclidean norm on $\mathbb{R}^d$. In Exercise 21 you're invited to show that the corresponding $W(x) = \|x\|^2$ is bowl-shaped. An important result related to bowl-shaped functions is the following result, known as *Anderson's lemma*.

Lemma 5.1. *Let f be a Lebesgue density of $\mathbb{R}^d$, with $\{x : f(x) \geq \alpha\}$ convex and symmetric about the origin for all $\alpha \geq 0$. If W is a bowl-shaped function, then*

$$\int W(x - c) f(x)\, dx \geq \int W(x) f(x)\, dx \quad \forall\, c \in \mathbb{R}^d.$$

Proof. The proof of Anderson's lemma uses a rather specialized result, called the Brunn–Minkowski inequality. Keener (2010, Sec. 16.4) gives a proof of all this stuff. □

The key point is that the function $\int W(x-c)f(x)\,dx$ is minimized at $c=0$. This fact will be useful in our derivation of a minimax estimator for θ below. Before this, I'd like to mention one application of Anderson's lemma.

Example 5.12. Let $X \sim \mathsf{N}_d(0,\Sigma)$, and let A be a convex set symmetric about the origin. Then the density f of X and $W(x) = 1 - I_A(x)$ satisfy the conditions of Lemma 5.1 (see Exercise 22). One example of a set A is a ball centered at the origin. Then it follows that

$$\mathsf{P}(X \in A) \geq \mathsf{P}(X + c \in A) \quad \forall\, c \in \mathbb{R}^d. \tag{5.6}$$

In other words, the normal distribution with mean zero assigns the largest probability to the convex symmetric set A. This is perhaps intuitively obvious, but the proof isn't easy. Results such as this have been used recently in applications of Bayesian methods in high-dimensional normal mean problems (e.g., Bhattacharya et al. 2014; Castillo and van der Vaart 2012).

Below is the main result of this section, i.e., that $\delta(X) = X$ is minimax for estimating θ under any loss function $L(\theta, a) = W(a-\theta)$ with W bowl-shaped.

Theorem 5.14. *Let $X \sim \mathsf{N}_d(\theta,\Sigma)$ where Σ is known. Then X is a minimax estimator of θ under loss $L(\theta,a) = W(a-\theta)$ for bowl-shaped W.*

Proof. Consider a Bayes setup and take a prior $\Theta \sim \Pi_\psi \equiv \mathsf{N}_d(0, \psi\Sigma)$ for a generic scale $\psi > 0$. Then the posterior distribution of Θ, given X, is

$$\Theta \mid X \sim \mathsf{N}_d\Big(\frac{\psi}{\psi+1}X, \frac{\psi}{\psi+1}\Sigma\Big).$$

Write $f(z)$ for the $\mathsf{N}_d(0, \{\psi/(\psi+1)\}\Sigma)$ density. For any estimator $\delta(X)$, the posterior risk is

$$\mathsf{E}\{W(\Theta - \delta(X)) \mid X = x\} = \int W\Big(z + \frac{\psi}{\psi+1}x - \delta(x)\Big) f(z)\,dz.$$

Since W is bowl-shaped and f satisfies the convexity requirements, Anderson's lemma says that the posterior risk is minimized at $\delta_\psi(x) = \psi x/(\psi+1)$; therefore, this $\delta(x)$ is the Bayes rule. Under the Bayes model, the distribution of X is the same as that of $\Theta + Z$, where $Z \sim \mathsf{N}_d(0,\Sigma)$ and Z independent of Θ. Then

$$\Theta - \delta_\psi(X) = \Theta - \delta_\psi(\Theta + Z) = \frac{\Theta - \psi Z}{\psi+1} \quad \text{(in distribution)},$$

and the distribution of the right-hand side is the same as that of $\{\psi/(\psi+1)\}^{1/2}Z$. Then the corresponding Bayes risk is

$$r(\Pi_\psi, \delta_\psi) = \mathsf{E}W(\Theta - \delta_\psi(X)) = \mathsf{E}W(\{\psi/(\psi+1)\}^{1/2}Z).$$

For any estimator δ, we have $\sup_\theta R(\theta, \delta) \geq r(\Pi_\psi, \delta) \geq r(\Pi_\psi, \delta_\psi)$. This holds for all ψ, so it also holds in the limit $\psi \to \infty$, which implies

$$\sup_\theta R(\theta, \delta) \geq \lim_{\psi \to \infty} \mathsf{E} W\Big(\Big\{ \frac{\psi}{\psi + 1} \Big\}^{1/2} Z \Big).$$

Since $\psi/(\psi + 1) \to 1$ as $\psi \to \infty$, it follows from the monotone convergence theorem that the lower bound above is $\mathsf{E}W(Z)$, which is exactly $\sup_\theta R(\theta, \hat{\theta})$, where $\hat{\theta} = X$. Since $\sup_\theta R(\theta, \delta) \geq \sup_\theta R(\theta, \hat{\theta})$ for all δ, it follows that $\hat{\theta} = X$ is minimax. □

5.7 Exercises

1. Suppose $X_1, \ldots, X_n$ are independent $\mathsf{Ber}(\theta)$ random variables. The goal is to estimate θ under square-error loss.

   (a) Calculate the risk for the maximum likelihood estimator $\hat{\theta}_{\text{mle}} = \bar{X}$.

   (b) Find the posterior mean $\hat{\theta}_{\text{Bayes}} = \mathsf{E}(\Theta \mid X)$ under a $\mathsf{Unif}(0, 1)$ prior for θ and calculate its risk function. [Hint: You've already found the formula for the posterior mean in Homework 04—just use the fact that $\mathsf{Unif}(0, 1)$ is a special case of $\mathsf{Beta}(a, b)$.]

   (c) Compare the two risk functions.

2. Suppose $X_1, \ldots, X_n$ are independent $\mathsf{N}(\theta, 1)$ random variables.

   (a) Find the risk function of the MLE $\bar{X}$ (under square-error loss).

   (b) Find the risk function for the Bayesian posterior mean under a $\mathsf{N}(0, 1)$ prior.

   (c) Compare the two risk functions, e.g., where do they intersect?

3. Let $X \sim \mathsf{P}_\theta$ and consider testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$. Find the risk function for a non-randomized test δ based on the 0–1 loss. [Hint: It will involve Type I and Type II error probabilities.]

4. Let X be a random variable with mean θ and variance σ^2. For estimating θ under square-error loss, consider the class $\delta_{a,b}(x) = ax + b$. Show that if

   $$a > 1 \quad \text{or} \quad a < 0 \quad \text{or} \quad a = 1 \text{ and } b \neq 0,$$

   then $\delta_{a,b}$ is inadmissible.

5. Suppose $X_1, \ldots, X_n$ are iid $\mathsf{N}(\theta, 1)$. If T is the sample mean, show that the conditional distribution of X_1 given $T = t$ is $\mathsf{N}(t, \frac{n-1}{n})$. [Hint: You can do this using Bayes theorem or using properties of the multivariate normal distribution.]

6. Reconsider the problem in Exercise 2 and assume a $\mathsf{N}(0, 1)$ prior Π.

(a) Find the Bayes risk of the MLE $\bar{X}$.

(b) Find the Bayes risk of the Bayes rule.

(c) Which estimator has smaller Bayes risk?

7. Consider $\theta \sim \Pi_a = \mathsf{Beta}(a, a)$, and let $D \subset [0, 1]$ be an open interval that does not contain 0 or 1. Show that $\Pi_a(D) \to 0$ as $a \to 0$. Hint: Use the fact that $\Gamma(x) = \Gamma(x + 1)/x$.

8. Consider admissibility of the sample mean as discussed in Example 5.8.

   (a) Show that the Bayes risk of the sample mean $\delta(X) = X/n$ with respect to Π_s is

   $$r(\Pi_s, \delta) = \frac{1}{4n}\Big(3 - \frac{1}{2s^{-1} + 1}\Big).$$

   (b) Show that the Bayes risk of the posterior mean $\delta_{\Pi_s}(X) = \mathsf{E}(\theta \mid X)$ with respect to the prior Π_s is

   $$r(\Pi_s, \delta_{\Pi_s}) = \Big(\frac{1}{2s^{-1} + 1}\Big)^2\Big[\frac{3n}{4} - \frac{n}{4(2s^{-1} + 1} + \frac{s^{-2}}{2s^{-1} + 1}\Big].$$

   (c) Show that $\{r(\Pi_s, \delta) - r(\Pi_s, \delta_{\Pi_s}\} \to 0$ as $s \to \infty$. Hint: You'll probably need the property of the gamma function from Exercise 7.

9. Consider the binomial problem in Example 5.9.

   (a) Find expressions for A, B and C (involving α, β, and n).

   (b) Show that $A = B = 0$ iff $\alpha = \beta = \frac{1}{2}\sqrt{n}$.

   (c) Plot the risk function of the minimax rule and that of the maximum likelihood estimate $\delta(x) = x/n$ for $n \in \{10, 25, 50, 100\}$. Compare the performance of the two estimators in each case.

10. (a) Show that if a decision rule is admissible and has constant risk, then it's minimax.

    (b) Use part (a) and Example 5.7 to argue that, if $X_1, \ldots, X_n$ are iid $\mathsf{N}(\theta, 1)$, then the sample mean $\bar{X}$ is a minimax estimator of θ under square-error loss.

    (c) Suppose $X \sim \mathsf{Bin}(n, \theta)$. Show that $\delta(x) = x/n$ is minimax for estimating θ under the loss function

    $$L(\theta, a) = \frac{(a - \theta)^2}{\theta(1 - \theta)}.$$

    [Hint: Find a proper prior Π so that $\delta(x)$ is a Bayes rule, hence it is admissible. For minimaxity, use part (a).]

11. Minimax estimates are not unique. Indeed, show that if $X \sim \mathsf{Pois}(\theta)$, then every estimator of θ is minimax under squared error loss. [Hint: To show that every estimator δ has unbounded risk function $R(\theta, \delta)$, demonstrate that there are priors Π and corresponding Bayes rules δ_Π with Bayes risk $r(\Pi, \delta_\Pi)$ arbitrarily large.]

12. In Example 5.10, show that the the MLE $\delta(x) = x$ is inadmissible. [Hint: Find another rule $\delta'(x)$ with risk everywhere no larger than that of $\delta(x) = x$; the trick is to incorporate the constraint—think about truncating $\delta(x)$.]

13. Consider the estimation problem with loss function $L(\theta, a) = (a - \theta)^2$, square-error loss. Show that, in this case, the unbiasedness condition (5.4) on an estimator $\delta(X)$ of θ reduces to the familiar definition, i.e., $\mathsf{E}_\theta\{\delta(X)\} = \theta$ for all θ.

14. Problem 4.6 in Keener (2010, p. 78). [Hint: $\delta + cU$ is an unbiased estimator for all c.]

15. Let $X_1, \ldots, X_n$ be iid $\mathsf{Pois}(\theta)$. Find the UMVU estimator of $\mathsf{P}_\theta(X_1 \text{ is even})$.

16. Prove that Pitman's estimator $\hat{\theta}_{\text{pit}}$ is location equivariant.

17. For each location problem below, find Pitman's estimator of θ.

    (a) $X_1, \ldots, X_n$ iid $\mathsf{Unif}(\theta - 1, \theta + 1)$.

    (b) $X_1, \ldots, X_n$ iid with density $\frac{1}{2}e^{-|x-\theta|}$ for $x \in \mathbb{R}$. There is no closed-form expression for $\hat{\theta}_{\text{pit}}$, but it can be found numerically. Write a computer program to do it, and apply it to data (6.59, 4.56, 4.88, 6.73, 5.67, 4.26, 5.80).

18. Problem 10.2(a) and 10.3 in Keener (2010, p. 201).

19. Problem 10.8 in Keener (2010, p. 202).

20. Show that if $\mathcal{C}$ is a minimal complete class, then it is exactly the class of all admissible decision rules.

21. Define $W(x) = \|x\|^2$ for $x \in \mathbb{R}^n$. Show that W is bowl-shaped.

22. Verify the claims in Example 5.12 leading to (5.6).

Chapter 6

More Asymptotic Theory

This chapter is still a work-in-progress. The part on M- and Z-estimators is in pretty good shape but lots more work is needed on the asymptotic normality and optimality results; the section on Bayesian posterior consistency and rates is in pretty good shape, but examples and details about the Bernstein–von Mises theorem needs to be added.

6.1 Introduction

In Chapter 3 we discussed certain asymptotic distribution properties for the maximum likelihood estimator (MLE) and the likelihood ratio test. Also, in Chapter 4, we discussed a bit about the asymptotic properties of Bayesian posterior distributions. In this chapter we want to expand our horizon a bit to see some generalizations of the results so far. There are two (related) general classes of estimators, called M- and Z-estimators, which contain the MLE as a special case. In Section 6.2, we will discuss some general asymptotic properties, such as consistency, convergence rates, and asymptotic normality, of M- and Z-estimators which, in particular, will reveal some sufficient conditions for consistency of the MLE, a point which was glossed over in Chapter 3. Then, in Section 6.3, we want to take a closer look at the regularity conditions used in Chapter 3 (and again in Chapter 4). It turns out that these can be replaced by some more mild conditions, at the expense of some additional conceptual and technical difficulties. More importantly, we will see that those desirable properties that the MLE possesses actually have nothing to do with the MLE. Specifically, these properties are consequences of the "niceness" of the sampling model, and can be described outside the context of a particular estimator. This makes the discussion of (asymptotically) optimal estimators, etc possible. We will also revisit the Bayesian problem in Section 6.4 and discuss what these new asymptotic considerations have to say concerning the behavior of the posterior distribution. Finally, in Section 6.5, I will make some cautionary remarks on how much emphasis should be given to these asymptotic results, based on their scientific/practical importance. Much of the material in this chapter is taken from, or at least inspired by, the relevant chapters in van der Vaart (1998). The work in this chapter will also highlight the importance of *empirical process* results in the study of asymptotics.

6.2 M- and Z-estimators

6.2.1 Definition and examples

Start with some notation, which is a bit different then that used previously. Let P be a probability measure defined on the measurable space $(\mathbb{X}, \mathcal{A})$. If f is a measurable function on $\mathbb{X}$, write Pf for $\int f\, dP$; this is called *deFinetti notation*. Next, let $X_1, \ldots, X_n$ be an iid sample from some distribution P on $\mathbb{X}$. Write $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ for the empirical distribution. For a measurable function f on $\mathbb{X}$, we write $\mathbb{P}_n f$ for the empirical average $\int f\, d\mathbb{P}_n = n^{-1} \sum_{i=1}^n f(X_i)$. As we will see, interesting questions about asymptotic performance of certain statistical procedures can be formulated in terms of questions about uniform convergence of $\mathbb{P}_n f$ to Pf over sets of functions f.

Two popular ways to define estimators $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ is by:

- Maximizing $M_n(\theta) = \mathbb{P}_n m_\theta$ for some known functions $m_\theta : \mathbb{X} \to \mathbb{R}$;
- Solving $Z_n(\theta) = 0$, where $Z_n(\theta) = \mathbb{P}_n z_\theta$ for some known functions $z_\theta : \mathbb{X} \to \mathbb{R}$.

The former produces an estimator $\hat{\theta}_n$ called an *M-estimator*, and the latter gives an estimator called a *Z-estimator*. Obviously, an M-estimator could be based on minimizing M_n instead of maximizing. Also, if $\theta \mapsto m_\theta$ are sufficiently smooth, then a M-estimator can be re-expressed as a Z-estimator via differentiation. Here are a few examples.

- *Maximum likelihood.* A familiar example of M/Z-estimators is maximum likelihood. In this case, we require that the underlying model P depend on a parameter θ in a specified way, and we write $P = P_\theta$ to represent this dependence. Then we can take $m_\theta = \log p_\theta$, where p_θ is the density of P_θ, and the corresponding M-estimator is the MLE. Also, if $\theta \mapsto p_\theta$ is smooth, then one could write $z_\theta = (\partial/\partial\theta) \log p_\theta$ so that the corresponding Z-estimator is the MLE.
- *Least squares.* Another familiar example of M/Z-estimation is least-squares. While this is usually presented first for normal linear models (e.g., regression), where it actually agrees with maximum likelihood, this procedure can be used in lots of applications outside the class of normal linear models. Here we consider a non-linear least squares. Write the mean of Y, given $X = x$, as $\mathsf{E}(Y \mid x) = f_\theta(x)$, where θ is some finite-dimensional parameter and f_θ is a fixed (possibly non-linear) function. An estimate of θ can be obtained via least squares, which is an M-estimator, with $m_\theta(x, y) = -\{y - f_\theta(x)\}^2$. This M-estimation approach is standard in certain machine learning applications. If $\theta \mapsto f_\theta$ is sufficiently smooth, then the non-linear least squares can be formulated as a Z-estimation method by taking derivatives.
- *Quantiles.* For simplicity, let's look at the median. A median θ is a solution to the equation $Z(\theta) := P(X > \theta) - P(X < \theta) = 0$. Therefore, the sample median $\hat{\theta}_n$ can be viewed as a Z-estimator of θ, with $z_\theta(x) = 1_{x>\theta} - 1_{x<\theta}$. Other quantiles besides the median can be defined similarly as solutions to certain equations involving the distribution function under P, and this representation is fundamental to developments in *quantile regression*; see Koenker (2005).

- *Robust estimation of a location parameter.* Both sample mean and sample median are Z-estimators of a location parameter, based on $z_\theta(x) = x - \theta$ and $z_\theta(x) = \text{sign}(x - \theta)$, respectively. More general versions of Z-estimators in this case have $z_\theta(x) = g(x - \theta)$ for some function g. These kinds of estimators are popular in a "robust" estimation context; one example is Huber's estimator corresponding to

$$g(u) = g_k(u) = \begin{cases} -k & \text{if } u \le -k \\ u & \text{if } |u| \le k \\ k & \text{if } u \ge k; \end{cases}$$

  the motivation for this kind of function is that it controls the influence any extreme observations, which leads to the robustness properties; see Huber (1981) for details.

One of the main benefits of M- and Z-estimation methods is that one does not need a full model or likelihood to produce a good, or at least reasonable, estimator. This is particularly important because, in some cases, a full model might not be available. Moreover, even if a model is available, it may not be so easy to marginalize over nuisance parameters to get at the actual parameter of interest. M- and Z-estimation techniques are frequently used by folks in *machine learning*, primarily because they can avoid modeling, model assumptions, and the bias that can be incurred when the posited model is wrong.

6.2.2 Consistency

Let's focus on M-estimators for the moment. One should ask the question: *why is maximizing M_n a good idea for producing an estimator?* Often, a law of large numbers will be applicable, and we'll have

$$M_n(\theta) \to M(\theta) := Pm_\theta \quad \text{as } n \to \infty, \quad \text{in } P\text{-probability}, \quad \textit{pointwise in } \theta.$$

So, we can think of maximizing M_n as approximately maximizing M. In fact, unless we have a specific model in mind, i.e., $P = P_\theta$, the interpretation of θ is as a maximizer of some function M. More importantly, M-estimators have nice properties under general (and relatively mild) conditions.

One point to be made is that the pointwise convergence $M_n(\theta) \to M(\theta)$ is not enough to guarantee that the maximizer $\hat{\theta}_n$ of M_n converges to the maximizer $\theta^\star$ of M. See Exercise 1. One needs to strengthen pointwise convergence to *uniform convergence*. That is, the sample size n required for $M_n(\theta)$ to be within a specified distance of its target $M(\theta)$, in P-probability, does not depend on θ. We will have more to say about uniform convergence following the statement and proof of the main result of this subsection.

Theorem 6.1. *Equip the parameter space Θ with a metric d. Let M_n be random functions and M a fixed function such that*

$$M_n \to M \quad \textit{uniformly in } P\textit{-probability, as } n \to \infty, \tag{6.1}$$

and assume that the maximizer $\theta^\star$ of M is well-separated in the sense that

$$\sup_{\theta: d(\theta,\theta^\star)\geq\varepsilon} M(\theta) < M(\theta^\star), \quad \forall\, \varepsilon > 0. \tag{6.2}$$

If $\hat{\theta}_n$ is a sequence such that $M_n(\hat{\theta}_n) \geq M_n(\theta^\star) - o_P(1)$, then $\hat{\theta}_n \to \theta^\star$ in P-probability.

Proof. First, we have that $M_n(\hat{\theta}_n) \geq M(\theta^\star)$ in P-probability. To see this, by the condition on $\hat{\theta}_n$, we know that $M_n(\hat{\theta}_n) \geq M_n(\theta^\star) - o_P(1)$. By the usual law of large numbers, we know that $M_n(\theta^\star) \geq M(\theta^\star) - o_P(1)$ so the claim follows since $o_p(1) + o_P(1) = o_P(1)$. From this, we immediately get

$$\begin{aligned} M(\theta^\star) - M(\hat{\theta}_n) &\leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_P(1) \\ &\leq \sup_\theta |M_n(\theta) - M(\theta)| + o_P(1). \end{aligned}$$

By uniform convergence, the supremum term is $o_P(1)$, so we can conclude that $M(\hat{\theta}_n) \geq M(\theta^\star)$ in P-probability. Now is the key step of the proof were we write the event $d(\hat{\theta}_n, \theta^\star) > \varepsilon$ in terms of M. By the separation condition, we have that, for any $\varepsilon > 0$, there exists $\delta > 0$, which does not depend on data, such that

$$d(\hat{\theta}_n, \theta^\star) > \varepsilon \implies M(\hat{\theta}_n) < M(\theta^\star) - \delta.$$

To complete the proof, note that by Step 2, the P-probability of the right-most event in the previous display vanishes as $n \to \infty$. Therefore,

$$P\{d(\hat{\theta}_n, \theta^\star) > \varepsilon\} \to 0, \quad \text{as } n \to \infty \text{ for any } \varepsilon > 0,$$

so the M-estimator is consistent, i.e., $\hat{\theta}_n \to \theta^\star$ in P-probability. □

Here are some remarks about the conditions of the theorem.

- van der Vaart (1998, p. 46) states that the uniform convergence condition (6.1) holds whenever $\{m_\theta : \theta \in \Theta\}$ is a *Glivenko–Cantelli* class. This is an important property in the empirical process literature, but I do not want to discuss the details here. Fortunately, there are some general and relatively simple sufficient conditions for this. In particular, if Θ is compact, $\theta \mapsto m_\theta(x)$ is continuous for each x, and $\theta \mapsto |m_\theta(x)|$ is bounded by a function (of x only) that is P-integrable. These are conditions under which we have a "uniform law of large numbers," which is just a law of large numbers for iid random functions (with respect to a metric of uniform convergence). See Keener (2010), Sec. 9.1, for some details.
- The separation condition (6.2) requires that the M function have a maximum that is clearly identified in the sense that there are no θ values such that $M(\theta)$ is too close to $M(\theta^\star)$. Since the M-estimator only cares about maximizing, if there are multiple θ's with nearly the same (large) M values then there is no hope for consistency.

- We do not require that $\hat{\theta}_n$ is a strict maximizer of M_n, it suffices that it is only a "near maximizer." That is, we only need $M_n(\hat{\theta}_n) \geq M_n(\theta^\star) - o_P(1)$, which would hold for some local maximizers. Also, is clearly satisfied if $\hat{\theta}_n$ is the global maximizer.

Example 6.1. Consider a parametric model P_θ for θ a scalar. Let $\theta^\star$ denote the true parameter value and write $P = P_{\theta^\star}$. Suppose that P_θ admits a density p_θ with respect to, say, Lebesgue measure, and define $m_\theta(x) = \log\{p_\theta(x)/p_{\theta^\star}(x)\}$. The expectation Pm_θ is the negative Kullback–Leibler divergence between $P = P_{\theta^\star}$ and P_θ, which is maximized uniquely at $\theta = \theta^\star$ provided that the model is identifiable, i.e., if $\theta \mapsto P_\theta$ is one-to-one or, in other words, if $p_\theta(x) = p_{\theta^\star}(x)$ for Lebesgue-almost all x implies $\theta = \theta^\star$.

To simplify things, consider an exponential family model with density

$$p_\theta(x) = h(x)e^{\theta T(x) - A(\theta)}.$$

In this case, m_θ simplifies to

$$m_\theta(x) = (\theta - \theta^\star)T(x) - \{A(\theta) - A(\theta^\star)\}.$$

If η is differentiable, then the expectation $M(\theta)$ also simplifies:

$$M(\theta) = (\theta - \theta^\star)\dot{A}(\theta^\star) - \{A(\theta) - A(\theta^\star)\}.$$

Then, clearly $\dot{M}(\theta^\star) = 0$ so $\theta^\star$ is a critical point; moreover, by standard results on exponential families and convex functions discussed previously, $M(\theta)$ is concave (see Exercise 3), so the separation condition holds. Let $X_1, \ldots, X_n$ be an iid sample, and $\mathbb{P}_n$ the corresponding empirical measure. Then the empirical mean $M_n(\theta) = \mathbb{P}_n m_\theta$ is given by

$$M_n(\theta) = (\theta - \theta^\star)\bar{T}_n - \{A(\theta) - A(\theta^\star)\},$$

where $\bar{T}_n = n^{-1}\sum_{i=1}^n T(X_i)$. Then we have

$$|M_n(\theta) - M(\theta)| = |\theta - \theta^\star|\,|\bar{T}_n - \dot{A}(\theta^\star)|.$$

The uniform convergence clearly holds, by the ordinary LLN, if θ is restricted to a bounded interval. So, consistency of the MLE follows from Theorem 6.1 if either θ is naturally restricted to a compact set or if it can be shown that the MLE resides in a compact set with probability approaching 1 as $n \to \infty$.

Example 6.2. Consider the non-linear least squares example above. That is, let P be the joint distribution for (X, Y) and write $Y = f_\theta(X) + \varepsilon$, where $\mathsf{E}(\varepsilon \mid X) = 0$ and $\mathsf{E}(\varepsilon^2) < \infty$. Using the $m_\theta(x, y) = -\{y - f_\theta(x)\}^2$ function and law of iterated expectation, we have

$$Pm_\theta = \cdots = -P(f_\theta - f_{\theta^\star})^2 + \mathsf{E}(\varepsilon^2), \tag{6.3}$$

where $\theta^\star$ is the "true" parameter. Clearly $Pm_\theta \leq \mathsf{E}(\varepsilon^2)$ and equality holds if and only if $P(f_\theta - f_{\theta^\star})^2 = 0$ if and only if $f_\theta - f_{\theta^\star} = 0$ P-almost surely. Therefore, if θ is identifiable, then $M(\theta) \equiv Pm_\theta$ is uniquely maximized at $\theta = \theta^\star$. Consistency of the M-estimator follows from the theorem if θ lives in a compact set and $\theta \mapsto f_\theta$ is smooth.

6.2.3 Rates of convergence

A subtle point in the consistency result above is that the parameter θ can be *anything*, i.e., it does not have to be a scalar or a vector, even an infinite-dimensional object, like a function, is covered by the theorem. In the next section we discuss the finest kind of convergence result, namely, limit distributions, which precisely describe the rate at which the convergence holds. A shortcoming of the analysis presented there is that it is focused on scalar or vector parameters only. There must be a middle-ground where we can get some idea about the speed of convergence in the consistency theorem in its most general form. Here we will present some results on the rate of convergence in the consistency theorem.

Of course, to get a stronger form of convergence we need stronger conditions. Here I will take an approach which is just slightly more sophisticated than one that assumes the convergence rate result to prove the same convergence rate result. More precisely, I will show that the convergence rate result is an easy consequence of another result, namely, one that provides uniform control on the objective function away from the maximizer/root. Taking this approach will highlight the importance of being able to derive the needed uniform bounds, which will motivate serious students to delve into the empirical process theory.

Consider the M-estimation problem, i.e., where we seek to estimate the maximizer of $M(\theta) = Pm_\theta$ by maximizing an empirical version $M_n(\theta) = \mathbb{P}_n m_\theta$.

Theorem 6.2. *Let $d(\cdot,\cdot)$ be a metric defined on Θ, and let $\theta^\star$ be the maximizer of $M(\theta)$. Suppose that for some positive sequence $\varepsilon_n \to 0$ and some constant $K > 0$,*

$$P\Big(\sup_{\theta: d(\theta,\theta^\star) > \varepsilon_n} \{M_n(\theta) - M_n(\theta^\star)\} \geq -K\varepsilon_n^2\Big) \to 0, \quad n \to \infty. \tag{6.4}$$

If $\hat\theta_n$ is any "approximate maximizer" of $M_n(\theta)$, i.e., if $M_n(\hat\theta_n) \geq M_n(\theta^\star) - \eta_n$ for any $\eta_n \leq K\varepsilon_n^2$, then $d(\hat\theta_n, \theta^\star) = o_P(\varepsilon_n)$ as $n \to \infty$, i.e.,

$$P\{d(\hat\theta_n, \theta^\star) > \varepsilon_n\} \to 0, \quad n \to \infty.$$

Proof. By definition of $\hat\theta_n$, we have that $M_n(\hat\theta_n) - M_n(\theta^\star) \geq -\eta_n$. So, the event $\{d(\hat\theta_n, \theta^\star) > \varepsilon_n\}$ implies

$$\sup_{\theta: d(\theta,\theta^\star) > \varepsilon_n} \{M_n(\theta) - M_n(\theta^\star)\} \geq -\eta_n.$$

Since $\eta_n \leq K\varepsilon_n^2$, the implied event has vanishing probability by assumption and, therefore, so does the original event, proving the claim. □

The conclusion of the theorem is that $\hat\theta_n \to \theta^\star$ at rate ε_n as $n \to \infty$, in the sense that if $d(\hat\theta_n, \theta^\star)$ is divided by ε_n, then the ratio still vanishes. There is nothing particularly special about "ε_n^2" in (6.4), some other non-negative function $f(\varepsilon_n)$ would work, provided that f is continuous and equals zero at 0. The proof above is similar to that of Theorem 2 in Wong and Shen (1995), a classic paper, but most of their hard work goes into establishing the sufficient condition (6.4). A slightly different result is proved in Theorem 5.52 of van der Vaart (1998), and there he gives sufficient conditions in terms of the behavior of M outside

a neighborhood of $\theta^\star$ and (indirectly) in terms of the complexity of m_θ and the space Θ. The latter part is where the empirical process tools are needed. The reader is encouraged to look at Theorem 5.52 in van der Vaart (1998), the proof, and the relevant discussion.

6.2.4 Asymptotic normality

Beyond consistency and convergence rates, a finer result is a kind of limit distribution, as this gives a precise characterization of the rate of convergence. A familiar result of this form is asymptotic normality of the MLE, which says that $n^{1/2}(\hat{\theta}_n - \theta^\star)$ is approximately normal for n large. Besides giving a $O_p(n^{-1/2})$ rate of convergence for $\hat{\theta}_n$, it allows one to construct asymptotically approximate confidence intervals, etc. There are analogous asymptotic normality results for general M- and Z-estimators and, perhaps surprisingly, the conditions given here are considerably weaker than those given for asymptotic normality of the MLE in Chapter 3. We'll have more to say about asymptotic normality of MLEs in Section 6.3.

To understand what's going on with the asymptotic normality result, let's start with some informal calculations. Consider a Z-estimator $\hat{\theta}_n$ that (approximately) solves the equation $Z_n(\theta) = 0$, where $Z_n(\theta) = \mathbb{P}_n z_\theta$ for suitable functions z_θ; also, let $Z(\theta) = P z_\theta$, so that $\theta^\star$ satisfies $Z(\theta^\star) = 0$. To keep things simple, assume θ is a scalar. If $\hat{\theta}_n \to \theta^\star$, then we may consider a Taylor approximation of $Z_n(\theta)$ in a neighborhood of $\theta^\star$:

$$0 = Z_n(\hat{\theta}_n) = Z_n(\theta^\star) + \dot{Z}_n(\theta^\star)(\hat{\theta}_n - \theta^\star) + \tfrac{1}{2}\ddot{Z}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta^\star)^2,$$

where $\tilde{\theta}_n$ is a point between $\hat{\theta}_n$ and $\theta^\star$. Rewrite this equation as

$$n^{1/2}(\hat{\theta}_n - \theta^\star) = \frac{-n^{1/2} Z_n(\theta^\star)}{\dot{Z}_n(\theta^\star) + \frac{1}{2}\ddot{Z}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta^\star)}.$$

Under suitable conditions, it is fairly easy to show that

- $n^{1/2} Z_n(\theta^\star) \to \mathsf{N}(0, P z_{\theta^\star}^2)$ in distribution;
- $\dot{Z}_n(\theta^\star) \to P\dot{z}_{\theta^\star}$ in probability; and
- $\ddot{Z}_n(\tilde{\theta}_n) \to 0$ in probability.

The first two properties follow easily by CLT and LLN arguments, under suitable moment conditions. Only the third property is tricky, as it involves a random function evaluated at a random argument, but we dealt with this in Chapter 3. Given these three properties, by Slutsky's theorem, we get that

$$n^{1/2}(\hat{\theta}_n - \theta^\star) \to \mathsf{N}\Big(0, \frac{P z_{\theta^\star}^2}{(P\dot{z}_{\theta^\star})^2}\Big), \quad \text{in distribution.}$$

The more general p-dimensional parameter case is exactly the same, but the notation is more complicated; see Equation (5.20) in van der Vaart (1998). The above argument implicitly assumed that $\theta \mapsto z_\theta(x)$ has two continuous derivatives for each x; such an assumption often holds, but there are cases where they don't. One can prove asymptotic normality under much weaker conditions—the following theorem assumes less than one derivative of $\theta \mapsto z_\theta(x)$! Again, for simplicity, I assume θ is a scalar here.

Theorem 6.3. *Let $\theta \mapsto z_\theta(x)$ be a measurable function satisfying the Lipschitz condition*

$$|z_{\theta_1}(x) - z_{\theta_2}(x)| \le \dot{z}(x)|\theta_1 - \theta_2|, \quad \textit{for all } \theta_1,\, \theta_2,$$

where $\dot{z}$ satisfies $P\dot{z}^2 < \infty$. Suppose $Pz_{\theta^\star}^2 < \infty$ and $Z(\theta) = Pz_\theta$ is differentiable at a zero $\theta^\star$, with nonsingular derivative $V_{\theta^\star} = \dot{Z}(\theta^\star)$. If $\hat{\theta}_n \to \theta^\star$ in P-probability and $Z_n(\hat{\theta}_n) = o_p(n^{-1/2})$, then $n^{1/2}(\hat{\theta}_n - \theta^\star) \to \mathsf{N}(0, V_{\theta^\star}^{-2} P z_{\theta^\star}^2)$ in distribution.

The asymptotic normality development in terms of Z-estimators is convenient because of the parallels with the more familiar case of maximum likelihood estimators that solve a likelihood equation. However, a result similar to that above also holds for M-estimators.

Theorem 6.4. *For each θ in an open subset of Euclidean space, let $x \mapsto m_\theta(x)$ be measurable and $\theta \mapsto m_\theta(x)$ be differentiable at $\theta^\star$ for P-almost all x, with derivative $\dot{m}_{\theta^\star}(x)$. Suppose that there exists a measurable function $\dot{m}$ with $P\dot{m}^2 < \infty$ such that, for each θ_1 and θ_2,*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \le \dot{m}(x)\|\theta_1 - \theta_2\|.$$

Furthermore, assume that $M(\theta) = Pm_\theta$ admits a second-order Taylor approximation at the point $\theta^\star$ of maximum, with non-singular second derivative matrix $V_{\theta^\star}$. If $\hat{\theta}_n$ "approximately" maximizes $M_n(\theta) = \mathbb{P}_n m_\theta$ and $\hat{\theta}_n \to \theta^\star$ in P-probability, then

$$n^{1/2}(\hat{\theta}_n - \theta^\star) \to \mathsf{N}_p(0, V_{\theta^\star}^{-1} P\dot{m}_{\theta^\star}\dot{m}_{\theta^\star}^\top V_{\theta^\star}^{-1}), \quad \textit{in distribution.}$$

Example 6.3. Consider the non-linear least squares regression problem from before. That is $Y = f_\theta(X) + \varepsilon$ and f_θ is a known function depending on a finite-dimensional parameter θ. Using the notation from before, we have

$$m_\theta(x, y) = -\{y - f_\theta(x)\}^2 \quad \text{and} \quad M(\theta) = -P(f_\theta - f_{\theta^\star})^2 - \sigma^2,$$

where $\sigma^2 = \mathsf{E}(\varepsilon^2)$. For consistency, we could assume that Θ is a bounded subset and that f_θ is sufficiently smooth, so that $\{m_\theta : \theta \in \Theta\}$ is G–C. In fact, the Lipschitz condition in the theorem above is sufficient for Glivenko–Cantelli; see Example 19.7 in van der Vaart (1998). The Lipschitz condition of the theorem holds, e.g., if $\dot{f}_\theta(x)$ is uniformly bounded, in a neighborhood of $\theta^\star$, by a square P-integrable function, say, $\bar{f}(x)$. Next we check that M admits a second-order Taylor expansion. If $\theta \mapsto f_\theta$ is sufficiently smooth near $\theta^\star$, then $(f_\theta - f_{\theta^\star})^2$ can be expanded as

$$(f_\theta - f_{\theta^\star})^2 = \tfrac{1}{2}(\theta - \theta^\star)^\top\{2\dot{f}_{\theta^\star}\dot{f}_{\theta^\star}^\top\}(\theta - \theta^\star) + o(\|\theta - \theta^\star\|^2).$$

The little-oh term above is a function of x, and if its expectation is still little-oh, e.g., if the above little-oh is uniform in x, then

$$M(\theta) = -\sigma^2 - \tfrac{1}{2}(\theta - \theta^\star)^\top\{2P\dot{f}_{\theta^\star}\dot{f}_{\theta^\star}^\top\}(\theta - \theta^\star).$$

Therefore, if $V_{\theta^\star} := 2P\dot{f}_{\theta^\star}\dot{f}_{\theta^\star}^\top$ is non-singular, then the theorem holds.

Let's make some connection to the more-or-less standard conditions assumed for asymptotic normality of the MLE seen in Chapter 3. It is typical that to assume that $z_\theta(x) = (\partial/\partial\theta) \log p_\theta(x)$ has two continuous derivatives for all x to prove asymptotic normality, but here we see that, almost everywhere differentiability and a Lipschitz condition is enough. However, even the Lipschitz condition is stronger than necessary; see the discussion following the proof of Theorem 5.21 in van der Vaart (1998), and also Example 5.24.[1]

Like in the case of consistency, the classes of functions $\{z_\theta : \theta \in \Theta\}$ or $\{m_\theta : \theta \in \Theta\}$ for which the asymptotic normality result holds have a name—they are called *Donsker* classes. Roughly, Donsker classes are collections of that admit a "uniform CLT." A pointwise CLT holds with a second moment condition, but a uniform CLT requires something stronger.

An interesting but mathematically difficult observation is based on the idea of a uniform CLT. It will help to make an analogy to the LLN setup where things are simpler. If we had a uniform LLN then we can conclude consistency of the corresponding M- or Z-estimator. For asymptotic normality, however, we didn't think this way. The question is: *can asymptotic normality of M- or Z-estimators be deduced by a kind of uniform CLT?* The answer is YES, but the challenge is in defining what is meant by "uniform CLT." The details are rather complicated, and you can find some of the basics in Chapters 18–19 in van der Vaart (1998). Let me summarize the main idea. Consider a class of square P-integrable functions $\{m_\theta : \theta \in \Theta\}$, and define the *empirical process*

$$\mathbb{G}_n m_\theta = n^{1/2}(\mathbb{P}_n m_\theta - P m_\theta), \quad \theta \in \Theta.$$

If Θ is finite, then we can conclude that

$$\{\mathbb{G}_n m_\theta : \theta \in \Theta\} \to \mathsf{N}_{|\Theta|}(0, \Sigma), \quad \text{in distribution},$$

for some covariance matrix Σ; this is just the usual multivariate CLT. If Θ is not finite, then we can still reach the above conclusion for any finite subset. The asymptotic normality of these "finite-dimensional distributions" turns out to completely describe the limit distribution, which is called a *Gaussian process*. So, in a certain sense, there is a uniform CLT that says $\mathbb{G}_n m_\theta$ converges "uniformly" to a Gaussian limit. How does this translate to asymptotic normality of the corresponding M-estimator $\hat{\theta}_n$? Recall the delta theorem, which says that continuous functions of an asymptotically normal quantity are also asymptotically normal. It turns out that there is a more general "functional" delta theorem, which says that if a process is asymptotically Gaussian, then any continuous functional of that process is asymptotically normal. Then it remains to show that the "argmax" functional is continuous to reach an asymptotic normality result for the M-estimator. This is what the "argmax theorem" in Section 5.9 of van der Vaart (1998) is about.

[1]For problems with a likelihood, there is a really slick condition for very strong asymptotic normality results called *quadratic mean differentiability*, which we'll discuss later.

6.3 More on asymptotic normality and optimality

6.3.1 Introduction

Chapters 6–9 in van der Vaart (1998), as well as Chapter 16 in Keener (2010), provide some details about a deeper way of thinking about asymptotic theory. This deeper way of thinking is due to Lucian Le Cam and were developed in the 1960s and 1970s, and were later summarized in Le Cam (1986) and Le Cam and Yang (2000); van der Vaart (2002) gives a nice review of Le Cam's work. As far as I can understand, the key observation is that asymptotic properties are really functions of the models in question and not really about the choice of a particular sequence of tests, estimators, etc. That is, we should not have to pick a particular test or estimator and then study its properties—its properties should somehow be clear from some more general properties about the model itself. Le Cam termed the models as *statistical experiments*, and proposed a framework by which one can study the corresponding limit experiment which characterizes the asymptotic properties of any convergence sequence of tests or estimators. This leads to some very clear description of asymptotically efficiency and optimality.

A first step along the way is a relaxation of the usual regularity conditions or, more precisely, an efficient statement of what actually is needed to get some interesting limit results. This is called *differentiability in quadratic mean*, and is shown in Section 6.3.5 to be sufficient for a kind of normal limit experiment, from which many of our familiar asymptotic normality results can be derived.

6.3.2 Hodges's provocative example

To do...

6.3.3 Differentiability in quadratic mean

Consider an iid sample $X_1, \ldots, X_n$ of $\mathbb{X}$-values random variables with statistical model $\{P_\theta : \theta \in \Theta\}$. Assume that there exists a dominating σ-finite measure μ on $\mathbb{X}$ and let p_θ be the density (Radon–Nikodym derivative) of P_θ with respect to μ. This is the standard setup where μ is either counting or Lebesgue measure. For simplicity of notation, I'll generally assume here that θ a scalar.

It's clear that if p_θ is sufficiently smooth, then an asymptotic normality result should be possible. For example, assuming three continuous derivatives of $\theta \mapsto p_\theta$, pointwise, and a little more gets us what we need. The question is: do we need less? To start, let's look at a particular technical condition that will be central to much of the development in this section of material. It's a notion of functional differentiability. Typically we think of Taylor approximations coming after a notion of differentiability; however, in more complicated spaces, differentiation is usually defined through a type of Taylor approximation.

Definition 6.1. The model P_θ is *differentiable in quadratic mean* (DQM) at θ if $\theta \mapsto \sqrt{p_\theta}$ is $L_2(\mu)$-differentiable at θ, i.e., there exists a function $\dot{\ell}_\theta$ such that

$$\int_{\mathbb{X}} \bigl[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \tfrac{1}{2}\dot{\ell}_\theta \sqrt{p_\theta}\, h\bigr]^2 \, d\mu = o(h^2), \quad h \to 0.$$

To motivate the subsequent development, note that DQM implies a sort of "local asymptotic normality" of the statistical model/statistical experiment. To understand better what this $\dot{\ell}_\theta$ function is, ignore the distinction between pointwise and L_2 differentiability. Then $\frac{1}{2}\dot{\ell}_\theta\sqrt{p_\theta}$ should be the (regular) derivative of $\sqrt{p_\theta}$. Using ordinary calculus, we get that

$$\frac{\partial}{\partial\theta}\sqrt{p_\theta} = \frac{1}{2}\frac{1}{\sqrt{p_\theta}}\dot{p}_\theta = \frac{1}{2}\frac{\dot{p}_\theta}{p_\theta}\sqrt{p_\theta}.$$

So, we must have $\dot{\ell}_\theta = \dot{p}_\theta/p_\theta$, "in an L_2 sense." But this latter quantity is a familiar one, i.e., $\dot{p}_\theta/p_\theta = (\partial/\partial\theta)\log p_\theta$ is just the score function. In other words, $\dot{\ell}_\theta$ is just an "L_2-score function," and it's not surprising that this quantity (and its variance) will be crucial to the asymptotic normality developments to be discussed below.

An immediate question is if there are any convenient sufficient conditions for DQM. From the stated context, we are introducing DQM as a way to side-step those tedious regularity conditions. So, naturally, those regularity conditions are sufficient for DQM. But we don't need quite that much. The following result is stated in Lemma 7.6 in van der Vaart (1998):

> *Let $\theta \mapsto \sqrt{p_\theta(x)}$ be continuously differentiable for each x, and suppose that $\theta \mapsto I_\theta := \int (\dot{p}_\theta/p_\theta)^2 p_\theta \, d\mu$ is well-defined and continuous. Then the model is DQM at θ and $\dot{\ell}_\theta = \dot{p}_\theta/p_\theta$.*

Then the familiar exponential families are DQM under some mild integrability conditions, with $\dot{\ell}_\theta$ the usual score function; see Example 7.7 in van der Vaart (1998).

But there are some families that fail the usual regularity conditions but are still DQM. Standard examples of trouble-maker distributions are those whose support depends on the parameter. It turns out that DQM does not require common support for all θ, but something more relaxed. In the definition of DQM, split the integral to the disjoint subsets $\{p_\theta = 0\}$ and $\{p_\theta > 0\}$. Then the sum of the two integrals must be $o(h^2)$, hence, each must be $o(h^2)$. Focus only on the integral over $\{p_\theta = 0\}$. On this set, the integrand simplifies considerably, leaving the condition

$$\int_{\{p_\theta = 0\}} p_{\theta+h} \, d\mu = o(h^2).$$

This is a necessary condition for DQM. The simplest example of parameter-dependent support is $\mathsf{Unif}(0,\theta)$, and it's easy to check that it doesn't satisfy the necessary condition. Indeed, if $h > 0$, then

$$\int_{\{p_\theta=0\}} p_{\theta+h}\, d\mu = \int_\theta^{\theta+h} \frac{1}{\theta+h}\, dx = \frac{h}{\theta+h} \neq o(h^2) \implies \mathsf{Unif}(0,\theta) \text{ not DQM!}$$

Next, recall that the score function having mean zero was an important idea in much of the developments of the asymptotic theory in Chapter 3. Recall that this was arranged by requiring that certain integration and differentiation could be interchanged, which might be a strong assumption. It turns out, however, that the score function having mean zero is a consequence of DQM. To prove this, we require a basic fact about the function space $L_2(\mu)$. In particular, fix $g \in L_2(\mu)$ and take a sequence $\{f_t : t \geq 1\} \subset L_2(\mu)$ such that $f_t \to f$ in $L_2(\mu)$ as $t \to \infty$; that is, $\int (f_t - f)^2 \, d\mu \to 0$ as $t \to \infty$. Then the claim is that $\int fg \, d\mu = \lim_{t\to\infty} \int f_n g \, d\mu$. The proof is an easy consequence of the Cauchy–Schwartz inequality:

$$\begin{aligned} \Big| \int f_t g \, d\mu - \int fg \, d\mu \Big| = \Big| \int (f_t - f) g \, d\mu \Big| &\leq \int |f_t - f||g| \, d\mu \\ &\leq \Big(\int (f_t - f)^2 \, d\mu \Big)^{1/2} \Big(\int g^2 \, d\mu \Big)^{1/2} \to 0. \end{aligned}$$

This can immediately be generalized to show that, if $f_t \to f$ and $g_t \to g$, both in $L_2(\mu)$, then $\int f_t g_t \, d\mu \to \int fg \, d\mu$ as $t \to \infty$. This result will be used to prove that $P_\theta \dot{\ell}_\theta = 0$ under DQM. Next we need some notation: write $p = p_\theta$, $p_t = p_{\theta + ht^{-1/2}}$ for a fixed $h \neq 0$, and $\psi = \dot{\ell}_\theta$ for the score function defined by DQM. From DQM we can conclude the following:

- $\sqrt{t}(\sqrt{p_t} - \sqrt{p}) \to f := \frac{1}{2}\psi\sqrt{p}h$ in $L_2(\mu)$, as $t \to \infty$, and
- $\sqrt{p_t} \to \sqrt{p}$ in $L_2(\mu)$ as $t \to \infty$.

Then, using the interchange of limit with integral discussed above, we get

$$\begin{aligned} (P_\theta \dot{\ell}_\theta)h &= \int \dot{\ell}_\theta p_\theta \, h \, d\mu = 2 \int f \sqrt{p} \, d\mu \\ &= \int f \sqrt{p} \, d\mu + \int f \sqrt{p} \, d\mu \\ &= \lim_{t\to\infty} \Big[\int \sqrt{t}(\sqrt{p_t} - \sqrt{p}) \sqrt{p_t} \, d\mu + \int \sqrt{t}(\sqrt{p_t} - \sqrt{p}) \sqrt{p} \, d\mu \Big] \\ &= \lim_{t\to\infty} \int \sqrt{t}(\sqrt{p_t} - \sqrt{p})(\sqrt{p_t} + \sqrt{p}) \, d\mu \\ &= \lim_{t\to\infty} \Big[\sqrt{t} \Big\{ \int p_t \, d\mu - \int p \, d\mu \Big\} \Big] \\ &= 0. \end{aligned}$$

This holds for all $h \neq 0$, which implies $P_\theta \dot{\ell}_\theta = 0$, i.e., score function has mean zero.

It turns out that DQM also implies that the variance of the score is finite, i.e., $I_\theta := P_\theta \dot{\ell}_\theta^2 < \infty$; see the proof of Theorem 7.2 in van der Vaart (1998). This is nothing but the familiar Fisher information we saw in Chapter 2. Usually to properly define the Fisher information we have to make assumptions about interchange of derivative and expectation,

but here we can define it with only DQM.[2] Before getting to the asymptotic normality stuff, I want to discuss a version of the Cramer–Rao inequality under DQM.

Theorem 6.5. [3]*Suppose that the model is DQM at θ with score function $\dot{\ell}_\theta$ and non-zero Fisher information I_θ. Let T be a real-valued statistic whose variance V_θ exists and is bounded in a neighborhood of θ. Then $\gamma_\theta := P_\theta(T)$ has derivative $\dot{\gamma}_\theta = P_\theta(T\dot{\ell}_\theta)$ at θ, and $V_\theta \geq \dot{\gamma}_\theta^2 I_\theta^{-1}$.*

The proof of the second claim follows from the first claim and the standard result that the squared covariance is upper bounded by product of variances; we did this in Chapter 2. So, it's only the first that needs some attention. The result is true in general, but to simplify things I'll assume that the statistic T is bounded, i.e., $|T| \leq M$ for some $M > 0$.

The key to the proof is the Cauchy–Schwartz inequality and a kind of chain rule for L_2 derivatives. We need to show that

$$|\gamma_{\theta+h} - \gamma_\theta - P_\theta(T\dot{\ell}_\theta)h| = o(h), \quad h \to 0.$$

The quantity in the absolute values on the left-hand side is just a series of integrals involving T, i.e.,

$$\Big| \int T p_{\theta+h}\, d\mu - \int T p_\theta \, d\mu - \int T \dot{\ell}_\theta p_\theta \, d\mu \, h \Big|.$$

Collect all the integrals into one and move the absolute value to the inside. Then the quantity above is no bigger than

$$\int |T| \, |p_{\theta+h} - p_\theta - \dot{\ell}_\theta p_\theta \, h| \, d\mu.$$

Using the "difference of two squares" algebra trick, we can write this as

$$\int |T| \, |(\sqrt{p_{\theta+h}} - \sqrt{p_\theta})(\sqrt{p_{\theta+h}} + \sqrt{p_\theta}) - \tfrac{1}{2}\dot{\ell}_\theta \sqrt{p_\theta} h (2\sqrt{p_\theta})| \, d\mu.$$

For small h, we have $\sqrt{p_{\theta+h}} + \sqrt{p_\theta} \approx 2\sqrt{p_\theta}$, so the previous display is approximately

$$\int |T| \, |2\sqrt{p_\theta}| \, |\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \tfrac{1}{2}\dot{\ell}_\theta \sqrt{p_\theta} h| \, d\mu.$$

To simplify the argument, assume that T is bounded, i.e., $|T| \leq M$ (μ-almost everywhere); this is not essential to the result, just makes our lives easier here. Use this bound and the Cauchy–Schwartz inequality to bound the previous display above by

$$M \Big\{ \int 4 p_\theta \, d\mu \Big\}^{1/2} \Big\{ \int (\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \tfrac{1}{2}\dot{\ell}_\theta \sqrt{p_\theta} h)^2 \, d\mu \Big\}^{1/2}.$$

The first integral equals 4, but it doesn't matter: the second integral is $o(h^2)$ by DQM, so its square root must be $o(h)$. This is what we wanted to show, so γ_θ is differentiable and its derivative is $\dot{\gamma}_\theta = P_\theta(T\dot{\ell}_\theta)$. From here, the Cramer–Rao inequality follows as usual.

[2]We cannot expect an alternative definition of Fisher information in terms of second derivatives because DQM says nothing about even existence the of these second derivatives.

[3]Taken from Chapter 6.3 in David Pollard's *Asymptopia* draft, available online.

We can already see that some of those familiar results from earlier chapters follow from DQM, which is weaker than those standard regularity conditions. The question about asymptotic normality comes next. Specifically, we want to study asymptotic normality in a more general setting. For example, the classical approach is to deal with asymptotic normality for individual estimators, but is it possible to get these results all from a more precise study of properties of the model? The answer is YES and, in the iid case, DQM is a sufficient condition for this kind of "asymptotic normality of the sequence of models."

6.3.4 Contiguity

To do...

6.3.5 Local asymptotic normality

To do...

6.3.6 On asymptotic optimality

To do...

6.4 More Bayesian asymptotics

6.4.1 Consistency

Consider a Bayes model where $X_1, \ldots, X_n$ are iid P_θ, given $\theta \in \Theta$, and the prior distribute for θ is Π, a proper prior. If $L_n(\theta)$ is the likelihood function for θ based on data $X_1, \ldots, X_n$, then the posterior distribution is a probability measure Π_n that satisfies

$$\Pi_n(A) \propto \int_A L_n(\theta)\, \Pi(d\theta), \quad A \subseteq \Theta.$$

Our notation Π_n hides the dependence on the data $X_1, \ldots, X_n$, but it should be implicitly understood that the measure is random. We discussed some Bayesian large-sample theory in Section 4.6, but the focus there was on distributional approximations, in particular, a Bernstein–von Mises theorem. While this is a desirable result, maybe the best possible, there are some cases where no posterior normality can be possible. An example is in the case of infinite-dimensional parameters. Here we will give some more basic posterior asymptotic results that can applied across finite- and infinite-dimensional parameters alike. Our first consideration is posterior consistency.

Definition 6.2. The posterior distribution Π_n is consistent at $\theta^\star$ if $\Pi_n(U^c) \to 0$ with $P_{\theta^\star}$-probability 1 for any neighborhood U of $\theta^\star$.

The intuition is that, if $\theta^\star$ is the true value of the parameter, then a consistent posterior distribution will concentrate around $\theta^\star$ in the sense that, for large n, all the posterior mass will reside on an arbitrarily small neighborhood of $\theta^\star$.

Towards sufficient conditions for posterior consistency, we first want to rewrite the formula for the posterior distribution in a slightly different way:

$$\Pi_n(A) = \frac{N_n(A)}{D_n} = \frac{\int_A R_n(\theta)\,\Pi(d\theta)}{\int_\Theta R_n(\theta)\,\Pi(d\theta)},$$

where $R_n(\theta) = L_n(\theta)/L_n(\theta^\star)$ is the likelihood ratio. Dividing by $L_n(\theta^\star)$ inside the integrand has no effect because this term does not depend on θ. Then the idea is to show that, for $A = U^c$, the complement of a neighborhood of $\theta^\star$, $N_n(A)$ is not too large and D_n is not too small, so that the ratio is small. As a first step, we get a lower bound on the denominator D_n. For this we need an important condition on the prior Π.

Definition 6.3. The prior Π satisfies the *KL condition* at $\theta^\star$ if

$$\Pi(\{\theta : K(\theta^\star, \theta) < \varepsilon\}) > 0 \quad \forall\, \varepsilon > 0,$$

where $K(\theta^\star, \theta)$ is the Kullback–Leibler divergence of P_θ from $P_{\theta^\star}$.

Lemma 6.1. *If Π satisfies the KL-condition at $\theta^\star$, then, or sufficiently large n, $D_n \geq e^{-nc}$ for any $c > 0$ with $P_{\theta^\star}$-probability 1.*

Proof. The strategy is to show that $\liminf_{n\to\infty} e^{nc} D_n = \infty$ with $P_{\theta^\star}$-probability 1. By definition of D_n,

$$e^{nc} D_n = \int_\Theta e^{nc + \log R_n(\theta)}\,\Pi(d\theta).$$

Let's rewrite the log term as

$$\log R_n(\theta) = -nK_n(\theta^\star, \theta) = -n\mathbb{P}_n \log\{p_{\theta^\star}/p_\theta\}.$$

We know that $K_n(\theta^\star, \theta) \to K(\theta^\star, \theta)$, the Kullback–Leibler divergence, with $P_{\theta^\star}$-probability 1 as $n \to \infty$ by the law of large numbers. For the given c, let $B = \{\theta : K(\theta^\star, \theta) < c/2\}$. Then

$$e^{nc} D_n > \int_B e^{nc - nK_n(\theta^\star, \theta)}\,\Pi(d\theta).$$

For $\theta \in B$, we have that $K_n(\theta^\star, \theta)$ has a limit less than $c/2$. Take lim inf of both sides and move the lim inf inside the integral (via Fatou's lemma) to get

$$\liminf_{n\to\infty} e^{nc} D_n > \int_B \liminf_{n\to\infty} e^{nc - nK_n(\theta^\star, \theta)}\,\Pi(d\theta).$$

Of course, the integrand converges to ∞ and, since $\Pi(B) > 0$ by the KL-condition, we get the desired result. □

The next step is to bound the numerator N_n. For this, there are a variety of conditions that will suffice, but here I want to emphasize simplicity. So, I choose to keep a connection with the M-estimation material in previous sections. As before, let d be a metric on Θ, so that neighborhoods of $\theta^\star$ can be taken as balls $\{\theta : d(\theta, \theta^\star) < r\}$. Specifically, we will assume that, for the true $\theta^\star$, i.e., minimizer of $\theta \mapsto K(\theta^\star, \theta)$, the following. First, $\theta^\star$ is well-separated in the sense that

$$\inf_{\theta: d(\theta,\theta^\star) \geq \varepsilon} K(\theta^\star, \theta) > 0, \quad \forall\, \varepsilon > 0. \tag{6.5}$$

Second, assume that the log-likelihood ratios satisfies a uniform law of large numbers, i.e.,

$$Z_n(\theta^\star) := \sup_\theta |K_n(\theta^\star, \theta) - K(\theta^\star, \theta)| \to 0 \quad \text{with } P_{\theta^\star}\text{-probability 1.} \tag{6.6}$$

Together, these two are sufficient conditions for consistency of the MLE, which can be viewed as a minimizer of $\theta \mapsto K_n(\theta^\star, \theta)$, but not a necessary.

Theorem 6.6. *If the prior satisfies the KL-condition at $\theta^\star$ and the conditions* (6.5) *and* (6.6) *hold, then the posterior is consistent at $\theta^\star$.*

Proof. Given any $\varepsilon > 0$, let $A = \{\theta : d(\theta, \theta^\star) \leq \varepsilon\}^c$. Rewrite the numerator as

$$N_n(A) = \int_A e^{-nK_n(\theta^\star,\theta)}\, \Pi(d\theta) = \int_A e^{-n\{K_n(\theta^\star,\theta) - K(\theta^\star,\theta) + K(\theta^\star,\theta)\}}\, \Pi(d\theta).$$

The integrand is clearly bounded by

$$e^{nZ_n(\theta^\star)} e^{-nK(\theta^\star,\theta)}.$$

By (6.5), $K(\theta^\star, \theta)$ is bounded below by zero uniformly in A, i.e., there exists $\delta = \delta(\varepsilon, \theta^\star) > 0$ such that $K(\theta^\star, \theta) \geq \delta$ for all $\theta \in A$. Therefore, the integrand can be further bounded by

$$e^{nZ_n(\theta^\star)} e^{-n\delta}.$$

By (6.6), $Z_n(\theta^\star) < \delta/2$ for all large n with $P_{\theta^\star}$-probability 1. Combine this with the lower bound on D_n we get

$$\Pi_n(A) \leq e^{nc - n\delta/2} \quad \text{with } P_{\theta^\star}\text{-probability 1 for all large } n.$$

Since c is arbitrary, take $c < \delta/2$ and let $n \to \infty$ to complete the proof. □

Recall that posterior consistency means that, in the limit, the posterior puts all of its mass in an arbitrarily small neighborhood of $\theta^\star$. In other words, the posterior measure is converging to a point-mass at $\theta^\star$. Therefore, barring any mathematical irregularities, one should expect that a reasonable summary of the posterior distribution would yield a consistent point estimator. For example, the posterior mean

$$\tilde{\theta}_n := \int \theta\, \Pi_n(d\theta),$$

is a consistent estimator in the sense that $\tilde{\theta}_n \to \theta^\star$ as $n \to \infty$ with $P_{\theta^\star}$-probability 1 under under the conditions of Theorem 6.6 plus a little more.

Corollary 6.1. *Under the conditions of Theorem 6.6, if the prior mean exists, then the posterior mean satisfies $\tilde{\theta}_n \to \theta^\star$ as $n \to \infty$ with $P_{\theta^\star}$-probability 1.*

Proof. See Exercise 4. □

An important point is that this makes minimal requirements on the prior Π, i.e., it only needs to assign a sufficient amount of mass near $\theta^\star$. The challenge is that $\theta^\star$ is unknown, so one needs to put a sufficient amount of mass "everywhere." In cases where $\theta^\star$ is finite-dimensional, this can be accomplished by taking the prior to have a density (with respect to Lebesgue measure, say) which is strictly positive and continuous. In infinite-dimensional problems, the KL-condition is more difficult, but can be checked; see, e.g., Wu and Ghosal (2008). The main point here is that one can get Bayesian posterior consistency under basically the same conditions needed for MLE consistency.[4]

An interesting side point is that one could replace K and K_n in the above results with something else that has the same properties as these. In other words, there is nothing special about the use of likelihood in constructing a consistent posterior distribution. For example, Let $K(\theta) = Pk_\theta$ be a function minimized at $\theta = \theta^\star$, and define the empirical version $K_n(\theta) = \mathbb{P}_n k_\theta$. Provided that the properties (6.5) and (6.6) hold, then the pseudo-posterior $\tilde{\Pi}_n$ defined as

$$\tilde{\Pi}_n(A) \propto \int_A e^{-nK_n(\theta)}\, \Pi(d\theta)$$

would have the same consistency properties as the genuine Bayes posterior. Such a construction would be particularly useful in problems where a Bayesian analysis is desired but a full model/likelihood is not available. See Bissiri et al. (2016) for some general discussion on this, and Syring and Martin (2015) for an application in medical statistics.

6.4.2 Convergence rates

After seeing the results in the previous subsection, it should be no surprise that we can strengthen the posterior convergence results if we assume some stronger control on the log-likelihood ratios. That is, if we assume what was needed for a M-estimator or, in this case, MLE rate of convergence, then we can get the same rate for the posterior. Here we need to first explain what we mean by posterior rate of convergence. Recall that d is a metric on Θ.

Definition 6.4. Let ε_n be a positive vanishing sequence. The posterior distribution Π_n converges to $\theta^\star$ at rate ε_n if, for some constant $M > 0$, $\Pi_n(\{\theta : d(\theta, \theta^\star) > M\varepsilon_n\}) \to 0$ in $P_{\theta^\star}$-probability as $n \to \infty$.

Besides the more refined result on how quickly the posterior concentrates around $\theta^\star$, a small difference between this definition and that of consistency is that the mode of convergence here is "in probability" rather than "with probability 1." The stronger mode of convergence can be obtained, at the expense of stronger assumptions.

[4]This is not precisely true, i.e., there are examples where one is consistent but the other is not. The point is that we've made strong enough assumptions here that consistency of both would hold; in those examples where one fails to be consistent, the assumptions here are violated.

We saw in Chapter 4 that, in many examples, the posterior has an asymptotic normality property, which implies that the convergence rate is $\varepsilon_n = n^{-1/2}$. The conditions for this are (roughly) the same as for asymptotic normality (see Section 6.4.3) where the rate is also $n^{-1/2}$, so there is some common ground here. The results to come are applicable in problems where a posterior normality result is not possible, such as infinite-dimensional problems.

To start here, we modify the lower bound result for the denominator D_n. For this, we need a slightly strengthen the Kullback–Leibler property. Let p_θ be the density function for P_θ, and define

$$V(\theta^\star, \theta) = \mathsf{V}_{\theta^\star}\{\log p_\theta(X) - \log p_{\theta^\star}(X)\};$$

this is like a "Kullback–Leibler variance," since the Kullback–Leibler divergence $K(\theta^\star, \theta)$ equals $\mathsf{E}_{\theta^\star}\{\log p_\theta(X) - \log p_{\theta^\star}(X)\}$. Following Shen and Wasserman (2001), for $t > 0$, define

$$S(t) = \{\theta : \max[K(\theta^\star, \theta), V(\theta^\star, \theta)] \leq t\} \subset \Theta.$$

If the prior assigns a sufficient amount of mass on $S(t)$ for sufficiently small t—both "sufficiently"s depend on n—then we get a useful lower bound on the denominator D_n.

Lemma 6.2. *For a positive sequence t_n, let $S_n = S(t_n)$. Then*

$$P_{\theta^\star}\{D_n \leq \tfrac{1}{2}\Pi(S_n)e^{-2nt_n}\} \leq \frac{2}{nt_n}.$$

Therefore, if $nt_n \to \infty$, then

$$D_n \geq \tfrac{1}{2}\Pi(S_n)e^{-2nt_n} \quad \textit{with } P_{\theta^\star}\textit{-probability converging to 1.}$$

Proof. The proof in Shen and Wasserman (2001) is based on some basic tools, e.g., standardization and Chebyshev's inequality, and is very readable. A similar result is proved in Lemma 8.1 in Ghosal et al. (2000). □

As before, we opt to give a result which has a simple proof and is consistent with the assumptions made for M-estimator (e.g., maximum likelihood estimator) rate of convergence. The assumption here is basically the same as (6.4) in the M-estimator problem. In particular, we assume that, for a positive sequence ε_n and a constant $K > 0$,

Theorem 6.7. *Let $r_n \to 0$, and suppose that, for a constant $K > 0$ and all $s_n \geq r_n$,*

$$P_{\theta^\star}\Big(\sup_{\theta : d(\theta, \theta^\star) > s_n} \{K_n(\theta^\star) - K_n(\theta)\} \geq -Ks_n^2\Big) \to 0, \quad n \to \infty. \tag{6.7}$$

Suppose t_n is such that $\Pi\{S(t_n)\} \gtrsim e^{-2nt_n}$ and define

$$\varepsilon = \max(t_n^{1/2}, r_n).$$

If $n\varepsilon_n^2 \to \infty$, then for all sufficiently large $J > 0$,

$$\Pi_n(\{\theta : d(\theta, \theta^\star) > J\varepsilon_n\}) \lesssim e^{-(J^2K/2)n\varepsilon_n^2}.$$

Proof. To do... □

Examples...

6.4.3 Bernstein–von Mises theorem, revisited

Based on the LAN stuff...

6.5 Concluding remarks

To do...

6.6 Exercises

1. Two calculus/analysis problems.[5]

   (a) Give an example of a sequence of functions $f_n(x)$ that converge pointwise to a function $f(x)$ but not uniformly.

   (b) Show that if $f_n(x) \to f(x)$ uniformly in a compact set of x, if f_n have unique maximizers x_n, and if the unique maximizer $x^\star$ of f is "well-separated" [i.e., no point x away from $x^\star$ has $f(x)$ close to $f(x^\star)$], then $x_n \to x^\star$.

2. Fill in the missing details in (6.3).

3. Show that the function M in Example 6.1 is concave.

4. Prove Corollary 6.1. Hints?

[5] See, also, Problem 5.7 in van der Vaart (1998).

Bibliography

Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood 1912–1922. *Statist. Sci.*, 12(3):162–176.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, second edition.

Berger, J. O. (2014). Conditioning is the issue. In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present, and Future of Statistical Science*, chapter 23. Chapman & Hall/CRC Press.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2014). Dirichlet–Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.*, to appear, `arXiv:1212.6088`.

Birnbaum, A. (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.*, 57:269–326.

Bissiri, P., Holmes, C., and Walker, S. G. (2016). A general framework for updating belief distributions. *J. Roy. Statist. Soc. Ser. B*, to appear; `arxiv:1306.6430`.

Bølviken, E. and Skovlund, E. (1996). Confidence intervals from Monte Carlo tests. *J. Amer. Statist. Assoc.*, 91(435):1071–1078.

Boos, D. D. and Stefanski, L. A. (2013). *Essential Statistical Inference*. Springer Texts in Statistics. Springer, New York. Theory and methods.

Brazzale, A. R., Davison, A. C., and Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge University Press, Cambridge.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 9. Institute of Mathematical Statistics, Hayward, CA.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40(4):2069–2101.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer, New York.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum-likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38.

Dempster, A. P. (2008). The Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.*, 48(2):365–377.

DiCiccio, T. J. and Romano, J. P. (1995). On bootstrap procedures for second-order accurate confidence limits in parametric models. *Statist. Sinica*, 5(1):141–160.

Drton, M., Sturmfels, B., and Sullivant, S. (2009). *Lectures on algebraic statistics*, volume 39 of *Oberwolfach Seminars*. Birkhäuser Verlag, Basel.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669.

Eaton, M. L. (1989). *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics, Hayward, CA.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26.

Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*, 65(3):457–487. With comments by Ole Barndorff-Nielsen, A. T. James, G. K. Robinson and D. A. Sprott and a reply by the authors.

Evans, M. (2013). What does the proof of Birnbaum's theorem prove? *Electron. J. Stat.*, 7:2645–2655.

Fisher, R. A. (1973). *Statistical Methods for Research Workers*. Hafner Publishing Co., New York. Fourteenth edition—revised and enlarged.

Fraser, D. A. S. (1968). *The Structure of Inference*. John Wiley & Sons Inc., New York.

Fraser, D. A. S. (2004). Ancillaries and conditional inference. *Statist. Sci.*, 19(2):333–369. With comments and a rejoinder by the author.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, second edition.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis*. Springer, New York.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer-Verlag, New York.

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*, volume 79 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Ghosh, M., Reid, N., and Fraser, D. A. S. (2010). Ancillary statistics: A review. *Statist. Sinica*, 20:1309–1332.

Hannig, J. (2009). On generalized fiducial inference. *Statist. Sinica*, 19(2):491–544.

Hedayat, A. S. and Sinha, B. K. (1991). *Design and Inference in Finite Population Sampling*. John Wiley & Sons Inc., New York.

Hogg, R. V., McKean, J., and Craig, A. T. (2012). *Introduction to Mathematical Statistics*. Pearson, 7th edition.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Jaynes, E. T. (2003). *Probability Theory*. Cambridge University Press, Cambridge.

Kadane, J. B. (2011). *Principles of Uncertainty*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL. `http://uncertainty.stat.cmu.edu`.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, 90(430):773–795.

Keener, R. W. (2010). *Theoretical Statistics*. Springer Texts in Statistics. Springer, New York.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, Cambridge.

Kullback, S. (1997). *Information Theory and Statistics*. Dover Publications, Inc., Mineola, NY. Reprint of the second (1968) edition.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer-Verlag, New York.

Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics*. Springer Series in Statistics. Springer-Verlag, New York, second edition. Some basic concepts.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition.

Liu, C. and Martin, R. (2015). Frameworks for prior-free posterior probabilistic inference. *WIREs Comput. Stat.*, 7(1):77–85.

Martin, R. (2015). Plausibility functions and exact frequentist inference. *J. Amer. Statist. Assoc.*, 110:1552–1561.

Martin, R. (2016). On an inferential model construction using generalized associations. Unpublished manuscript, `arXiv:1511.06733`.

Martin, R. and Liu, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.*, 108(501):301–313.

Martin, R. and Liu, C. (2014). Discussion: Foundations of statistical inference, revisited. *Statist. Sci.*, 29:247–251.

Martin, R. and Liu, C. (2015a). Conditional inferential models: Combining information for prior-free probabilistic inference. *J. R. Stat. Soc. Ser. B*, 77(1):195–217.

Martin, R. and Liu, C. (2015b). *Inferential Models: Reasoning with Uncertainty*. Monographs in Statistics and Applied Probability Series. Chapman & Hall/CRC Press.

Martin, R. and Liu, C. (2015c). Marginal inferential models: Prior-free probabilistic inference on interest parameters. *J. Amer. Statist. Assoc.*, 110:1621–1631.

Martin, R. and Liu, C. (2016). Validity and the foundations of statistical inference. Unpublished manuscript, `arXiv:1607.05051`.

Mayo, D. (2014). On the Birnbaum argument for the strong likelihood principle. *Statist. Sci.*, 29(2):227–239.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32.

Reid, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.*, 31(6):1695–1731.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *J. Educational Statist.*, 6(4):377–401.

Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag, New York.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J.

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714.

Stigler, S. M. (2007). The epic story of maximum likelihood. *Statist. Sci.*, 22(4):598–620.

Syring, N. and Martin, R. (2015). Likelihood-free Bayesian inference on the minimum clinically important difference. unpublished manuscript, `arXiv:1501.01840`.

van der Vaart, A. (2002). The statistical work of Lucien Le Cam. *Ann. Statist.*, 30(3):631–682. Dedicated to the memory of Lucien Le Cam.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.*, 23(2):339–362.

Wu, Y. and Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat.*, 2:298–331.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge University Press, Cambridge.

Zabell, S. L. (1992). R. A. Fisher and the fiducial argument. *Statist. Sci.*, 7(3):369–387.