Statistical Inference

STA 732 – Lecture Notes

Lasse Vuursteen

Duke
UNIVERSITY

Foreword

These lecture notes accompany the course STA 732 - Statistical Theory at Duke University. They provide an exposition of a mathematical theory of statistics. Sections and exercises with ♠ are more advanced and optional course material. This is a work in progress.

The text assumes students have a background in real analysis, measure theory, and linear algebra. Appendices A, B, and C provide condensed refreshers on the most important concepts. Some of the proofs (marked ♠) rely on techniques from functional analysis. To keep the notes self-contained, Appendix C introduces these techniques. For those familiar with the subject, it may be interesting to see how these tools are applied; for those who are not, it provides a helicopter overview but is not required for the course.

Literature and Acknowledgments

Several sources have been used, but some so extensively that they deserve upfront acknowledgment. I am grateful for the course material shared by Peter Hoff, Surya Tokdar, Yuansi Chen, and Li Ma. These notes are largely based on their material and the source material from which they draw:

- Keener, *Theoretical Statistics: Topics for a Core Course* (2010).
- E.L. Lehmann and G. Casella, *Theory of Point Estimation* (1998).
- G. Casella and R.L. Berger, *Statistical Inference* (2002).
- J.H. van Zanten, *Statistics for High- and Infinite-Dimensional Models* (unpublished).

Part two of the notes draws significantly from A.W. van der Vaart, *Asymptotic Statistics* (1998). Harry van Zanten's lecture notes have been a stylistic model, which, for me personally, set a standard for how to write a brief but rigorous exposition that I have tried to emulate.

Notation

- $\mathbb{N} = \{1, 2, \ldots\}$ denotes the set of natural numbers, and $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$.
- For a set A, its power set is denoted as $2^A = \{S : S \subseteq A\}$, the set of all subsets of A.
- The indicator function of a set A is denoted as $\mathbb{1}_A$ or $x \mapsto \mathbb{1}\{x \in A\}$, which takes value 1 if $x \in A$ and 0 otherwise.
- Given measurable spaces $(\mathcal{X}, \mathscr{X})$ and $(\mathcal{Y}, \mathscr{Y})$, a measurable map $f : \mathcal{X} \to \mathcal{Y}$ is to be understood as being measurable with respect to the σ-algebras $\mathscr{X}$ and $\mathscr{Y}$. If f is measurable and real valued, but no sigma-algebra is specified, the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ is what is meant.
- For probability measures, we will sometimes forego the set notation whenever no ambiguity arises: both $P(X \in A)$ and $P(x : X(x) \in A)$ are shorthand for $P(\{x : X(x) \in A\})$.
- We use $X \sim P$ to denote that the random variable X has distribution P.
- The n-fold product measure of a probability measure P is denoted by $P^{\otimes n}$.
- We use the notation $a_n \lesssim b_n$ to indicate that $a_n \leqslant C b_n$ for some constant $C > 0$ independent of n. We write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

Part I

Statistical Decision Theory

1 Models, Statistics and Decisions

Inference is the process of drawing conclusions from evidence. In deductive inference, the conclusions follow with certainty by reasoning from the premises. In inductive inference, the conclusions are uncertain; they are at best probable.

Statistical inference is inductive inference in which the evidence consists of data generated by some unknown data-generating process involving randomness. This randomness can arise from several sources: we may be randomly sampling a subset from a larger population, our measurements may contain error, or the phenomenon itself may be governed by inherently stochastic mechanisms.

To describe the data-generating process, we need to describe the randomness that underlies it. Probability theory provides a mathematical language for describing randomness. It allows us to formally reason about the question: given a data-generating process, what is the distribution of the observable data? In statistics, however, we wish to formally reason about probable cause based on observed effects. This concerns 'the inverse' of the previous question: what does the observed data tell us about certain unknown features of the data-generating process?

We pursue statistical inference about unknown features of a data-generating process because they govern the real-world consequences of our actions. Whether a treatment saves lives, whether an investment succeeds, or a policy achieves its intended effect – all depend on the true nature of that process. *Statistical decision theory* is a mathematical framework for reasoning about optimal actions when consequences depend on an uncertain process we can only observe indirectly.

1.1 Statistical Models

The central object of statistical decision theory is a *statistical model*, which is a collection of probability distributions. Each of these probability distributions is a possible description of the data-generating process.

Definition 1.1. A *statistical model* is a collection of probability measures $\mathcal{P}$ defined on a measurable space $(\mathcal{X}, \mathscr{X})$:

$$\text{for all } P \in \mathcal{P},\ P : \mathscr{X} \to [0,1] \text{ is a measure that satisfies } P(\mathcal{X}) = 1.$$

The objects accompanying the statistical model typically carry the special names and interpretations in statistical literature.

- The space $(\mathcal{X}, \mathscr{X})$ is the *sample space*. It represents the set of all possible data.
- The accompanying sigma-algebra $\mathscr{X}$ are the *events*. The collections of outcomes to which the model can assign probabilities.
- The collection $\mathcal{P}$ specifies the possible 'theories' that could have generated the data.
- The triple $(\mathcal{X}, \mathscr{X}, \mathcal{P})$ can be referred to as the *statistical experiment* (or simply the *experiment*), we revisit this terminology in Definition 1.6 below.
- The *outcome* of the experiment is represented by an element $x \in \mathcal{X}$. Equivalently (and more informatively), it is the list of all $A \in \mathscr{X}$ for which $x \in A$. In other words, the outcome tells us exactly which events have occurred and which have not.

Measure theory provides a rigorous framework ensuring the intuitive properties we expect from probabilities hold without exception. Some of these properties follow almost immediately from the definition of a probability measure and a sigma-algebra (see Definitions B.1 and B.2):

- If the event A implies the event B (i.e. $A \subseteq B$), then $P(A) \leqslant P(B)$ for every $P \in \mathcal{P}$.
- If we can assign probability to the event A, we can assign it to its complement A^c and we have $P(A^c) = 1 - P(A)$.

Less intuitive[1] but highly desirable mathematically, is the ability to assign probabilities to countable unions of events. Without it, paradoxes and inconsistencies can arise in uncountable sample spaces.

Besides its desirable properties in terms of formalizing probabilities, the sigma-algebra formalizes the information that can be extracted from the data. This allows us to compare models with the same underlying sample space but where different events are observable.

Example 1.2. Consider an experiment of rolling two six-sided dice and observing the eyes on top each die. Formally, we could model this as $(\mathcal{X}, \mathscr{X}, \mathcal{P})$ where the sample space is given by

$$\mathcal{X} = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\},$$

and its powerset $\mathscr{X} = 2^{\mathcal{X}}$ are the observable events, and $\mathcal{P}$ should consist of a subset of the probability measures on $(\mathcal{X}, \mathscr{X})$. The sum of the eyes on each die is observable: $S : \mathcal{X} \to \mathbb{R}$ given by $S(x, y) = x + y$ for $(x, y) \in \mathcal{X}$, and so is whether the first die is larger than the second die: $L(x, y) = \mathbb{1}_{\{x>y\}}$.

[1] and in the eyes of some, controversial Regazzini 2013.

Consider another statistical model which models the case where we only observe the sum of the eyes on each die:

$$(\mathcal{X}, \sigma(S), \mathcal{Q}),$$

where $\sigma(S)$ is the sigma-algebra generated by S (see Definition B.9) and the collection $\mathcal{Q}$ consists of the probability measures $P \in \mathcal{P}$ restricted to $\sigma(S)$. The sample space is the same as in the first experiment, but the sigma-algebra is strictly smaller.

In the first experiment, we can determine the value of the first die, the second die, whether they are equal, whether the first is larger, etc. In the second experiment, the observables are a subset of the observables in the first experiment. That is, certain events that we could assign probabilities to in the first experiment, we cannot assign probabilities to in the second experiment, such as the event that the first die is larger than the second die. ◊

The two experiments in Example 1.2 model different observational scenarios of the same underlying random phenomenon. The first experiment provides more information: knowing the individual outcomes of each die, we can reconstruct their sum. Conversely, knowing only the sum, we cannot recover the individual outcomes. Whether the first experiment is more suitable of the inference problem at hand depends on the question we are interested in. For certain inferential goals, knowing the sum of the individual dice is all we need. We will formalize this idea in Section 1.2, where we will discuss the concept of sufficient statistics. For now, let us note that the sigma-algebra captures precisely which features of the outcome are observable, making it possible to formalize such comparisons. The definition of the sample space allows for a lot of freedom: it need not match the minimal description of the data; in principle, it could be the whole universe, provided the sigma-algebra correctly captures the information available in the experiment.

To close this section, we will discuss the idea of a parameter space. It is common that we are only interested in particular characteristics of the data-generating process, such as its mean, certain quantiles, and so on. These are typically functionals on $\mathcal{P}$: they map each P to some value, for example its mean $\int x dP(x)$. We will call these characteristics *parameters*. Consider a set Θ of possible values of those parameters for our statistical model. We call this set the *parameter space*.

Definition 1.3. A *parameter space* for the model $\mathcal{P}$ is a set Θ together with a map $P \mapsto \theta(P)$ from $\mathcal{P}$ onto Θ.

Example 1.4. Let $(\mathcal{X}, \mathscr{X}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$ (see Definition B.10). Consider the collection of probability measures $\mathcal{P} = \{N(\theta, 1) : \theta \in \mathbb{R}\}$,

where $N(\theta, \sigma^2)$ denotes the normal distribution with mean θ and variance σ^2:

$$N(\theta, \sigma^2)(A) := \int_A \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} dx.$$

Further, consider the collection

$$\mathcal{Q} = \{\text{all probability measures on } (\mathbb{R}, \mathcal{B}(\mathbb{R})) \text{ with finite mean}\}.$$

Both $(\mathcal{X}, \mathscr{X}, \mathcal{P})$ and $(\mathcal{X}, \mathscr{X}, \mathcal{Q})$ are valid statistical models. The set $\Theta = \mathbb{R}$ is a valid parameter space for $\mathcal{P}$ and $\mathcal{Q}$ under the map $Q \mapsto \int x dQ(x)$. The set $\Theta = [-1, 1]$ is not a valid parameter space for either $\mathcal{P}$ or $\mathcal{Q}$ under the same map (why?). $\lozenge$

There is an important distinction between the models $\mathcal{P}$ and $\mathcal{Q}$ in Example 1.4. In the first model, the mean uniquely identifies its distribution: there is a one-to-one correspondence between the parameter space and the collection of probability measures. In the second model, the mean does not uniquely identify its distribution: many different probability measures have the same mean. The same parameter (here the mean) can identify the entire distribution in one model but fail to do so in another.

Definition 1.5. A statistical model $\mathcal{P}$ is *identifiable* by a parameter space Θ if for all $P, P' \in \mathcal{P}$, it holds that if $\theta(P) = \theta(P')$, then $P = P'$.

The identifiability condition $\theta(P) = \theta(P') \implies P = P'$ means that the parameter space Θ forms a 'coordinate system' for the collection of probability measures $\mathcal{P}$. Since the map $P \mapsto \theta(P)$ is surjective, it means that every probability measure in $\mathcal{P}$ is uniquely determined by its parameter value in Θ.

Every statistical model $\mathcal{P}$ admits a trivial identifiable parameterization: simply take $\Theta = \mathcal{P}$ and $\theta(P) = P$. Whilst always possible, this parametrization is not always the most useful. Typically, the introduced parameter space brings along useful extra structure on the model that the bare set $\mathcal{P}$ does not 'directly' possess. In the vast majority of examples in this course we choose Θ to be a open, convex subset of $\mathbb{R}^d$. This endows the model with the rich Euclidean structure: vector-space operations, a natural notion of a distance metric, an inner product, differentiability, and so on.

When a model is identifiable, we may (and usually do) identify the parameter value θ with the distribution P_θ itself, so that "knowing θ" is equivalent to "knowing which distribution generated the data". That is, we can index the distributions by the parameter value:

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}.$$

where the subscript θ uniquely labels the distribution P_θ.

This type of parametrization is what we will mostly be concerned with in this course, leading us to the definition of a statistical experiment.

Definition 1.6. A *statistical experiment* is a tuple $(\mathcal{X}, \mathscr{X}, \mathcal{P}, \Theta)$ where the parameter space Θ indexes the collection of probability measures $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ on the sample space $(\mathcal{X}, \mathscr{X})$.

There is often an intricate interplay between the parametrization of the model and the formulation of the sample space, leading to multiple ways to write down what is effectively the same model. Sufficiency is one concept that allows us to formalize this idea, which we will discuss next.

Remark 1.7 (But wait... isn't my data supposed to be a random variable?). In the current framework, there is no random variable explicitly representing 'the data' in the experiment. This may appear to differ from the typical introductory setting; where a statistical setting is defined by a random variable with a given distribution depending on some parameter: "let $X \sim N(\theta, 1)$ for $\theta \in [-1, 1]$". Alternatively, one may be used to the following formal setting from probability courses, in which one considers an (implicit) probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and defines a random variable (or rather a random element) $X : \Omega \to \mathcal{X}$ representing 'the data before it is observed'. The law of X is then defined as $P(A) = \mathbb{P}(\omega : X(\omega) \in A)$ for some unknown probability measure $P \in \mathcal{P}$ on $(\mathcal{X}, \mathscr{X})$.

In our framework, we work directly with $(\mathcal{X}, \mathscr{X}, \mathcal{P})$ without introducing any underlying probability space. We can always recover the setup in which 'the data is a random variable' (and it is often linguistically and pedagogically useful to do so). Simply take the target space $(\mathcal{X}, \mathscr{X})$ of the random variable that is supposed to represent the data and consider the identity map $X : \mathcal{X} \to \mathcal{X}$, $X(x) = x$ for all $x \in \mathcal{X}$. It is easy to see that this map is measurable with respect to $\mathscr{X}$. Further, the collection $\mathcal{P}$ describes the possible laws of this 'random element': $P(A) = P(x : X(x) \in A)$ for all $A \in \mathscr{X}$ in $P \in \mathcal{P}$.

With this understanding in place, we will frequently use the familiar language of random variables—for instance, "let $X \sim N(\theta, 1)$ for $\theta \in [-1, 1]$" should be understood as shorthand for the statistical model $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{N(\theta, 1) : \theta \in [-1, 1]\})$. In this context, there can be little to no ambiguity which sigma-algebra we are referring to (recall the definition of the normal distribution in Example 1.4). Similarly — recalling that independent random variables are distributed according to the product measure (see Definition B.30 in Appendix B) — $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$ with $\theta \in \Theta$ is to be understood as shorthand for the statistical model $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{N(\theta, 1)^{\otimes n} : \theta \in \Theta\})$.

1.2 Statistics, Sufficiency and Likelihoods

In Example 1.2, we noted that knowing the sum of two dice provides less information than knowing each die individually: from the sum alone, we cannot recover the individual outcomes. Yet for certain inferential goals, the sum may contain all the information we need. This section makes precise the notion of a statistic that captures "all the information about the parameter." This will allow us to compare different formulations of models and judge when they are effectively the same for all intents and purposes. We start by defining what a statistic is.

Definition 1.8. A *statistic* is a measurable map T from the sample space $(\mathcal{X}, \mathscr{X})$ to some measurable space $(\mathcal{T}, \mathscr{T})$.

Given a statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ defined on a measurable space $(\mathcal{X}, \mathscr{X})$, a statistic is a measurable map $T : \mathcal{X} \to \mathcal{T}$ into another measurable space $(\mathcal{T}, \mathscr{T})$. The statistic T induces a new statistical model on $(\mathcal{T}, \mathscr{T})$, which we denote $\mathcal{P}^T = \{P_\theta^T : \theta \in \Theta\}$, where each P_θ^T is the push-forward measure of P_θ under T:

$$P_\theta^T(B) = P_\theta(T^{-1}(B)) \quad \text{for all } B \in \mathscr{T}.$$

The map $T : \mathcal{X} \to \mathcal{T}$ sends each possible outcome of an experiment to a 'summary' of the data. Ideally, the summary $T(X)$ is more 'compressed' than the original data X, while retaining all relevant information about the parameter θ.

This brings us to the idea of sufficiency. The key idea is that a statistic $T(X)$ is sufficient if, once we know $T(X)$, the remaining randomness in X tells us nothing further about which P_θ generated the data. Formally, the conditional distribution of X given $T(X)$ should not depend on θ.

Definition 1.9 (Sufficiency)**.** Consider an identifiable model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ on a sample space $(\mathcal{X}, \mathscr{X})$. A statistic $T : \mathcal{X} \to \mathcal{T}$ is *sufficient* for $\mathcal{P}$ if for every $A \in \mathscr{X}$, the conditional probability $P_\theta(A \mid T)$ admits a version that does not depend on θ.

Recall that the conditional probability $P_\theta(A \mid T)$ is formally defined as a conditional expectation $\mathbb{E}_\theta[\mathbb{1}_A \mid T]$ (see Definition B.32 in Appendix B). This expectation is a random variable, measurable with respect to the σ-algebra generated by T, satisfying the condition

$$\int_B \mathbb{E}_\theta[\mathbb{1}_A \mid T](x)\, dP_\theta(x) = P_\theta(x : x \in A,\, x \in B)$$

for all $B \in \sigma(T)$.

Because conditional expectations are only unique up to P_θ-null sets, saying that "$P_\theta(A \mid T)$ admits a version that does not depend on θ" means: for each $A \in \mathscr{X}$ there

exists a measurable function $h_A : \mathcal{X} \to [0,1]$, independent of θ, such that

$$\mathbb{E}_\theta[\mathbb{1}_A \mid T](x) = h_A(x) \quad P_\theta\text{-a.s. for all } \theta \in \Theta.$$

For 'nice' σ-algebras (for example, the Borel σ-algebra), we can go a step further and find a measurable function $h_A : \mathcal{T} \to [0,1]$, independent of θ, such that for all $C \in \mathscr{T}$,

$$\int_C h_A(t)\, dP_\theta^T(t) \;=\; P_\theta\big(A \cap T^{-1}(C)\big) \;=\; P_\theta\big(x : x \in A,\ T(x) \in C\big), \qquad \forall\, \theta \in \Theta.$$

That is, $\mathbb{E}_\theta[\mathbb{1}_A \mid T]$ can be represented as a function of T (a so called 'regular' conditional probability), not depending on θ: $E_\theta[\mathbb{1}_A \mid T](x) = h_A(T(x))$. Regardless of the representation, the key point is that h_A does not depend on θ: *after conditioning on the 'information of T' (the $\sigma(T)$-algebra), whatever we know about the event A does not depend on θ* (up to sets of zero measure).

Sufficiency is a property of how the parameter enters the distribution of the data. The same statistic may be sufficient for one model but not another, and crucially depends on how the model is parametrized (we will see an illustration of this in Example 1.13).

Before studying interesting cases, we note that sufficiency is trivially achieved when no information is discarded. A statistic that is appropriately invertible is sufficient: by inverting the map, we can recover the data from the statistic.

Proposition 1.10. *Consider a statistic $T : \mathcal{X} \to \mathcal{T}$ that is bijective, and assume its inverse is also measurable. Then, T is sufficient.*

Proof. Consider the σ-algebra generated by T, $\sigma(T)$. This is the smallest σ-algebra containing all the preimages $T^{-1}(B)$ of the sets in $B \in \mathscr{T}$, so

$$\sigma(T) \subseteq \mathscr{X}.$$

As the inverse of T is measurable, $T(A) \in \mathscr{T}$ for all $A \in \mathscr{X}$. Hence, for any event $A \in \mathscr{X}$, measurability of T and its inverse implies that $A = T^{-1}(T(A)) \in \sigma(T)$, so $\mathbb{1}_A$ is $\sigma(T)$-measurable. Conclude that $\sigma(T) = \mathscr{X}$.

Hence, we have that for all $A \in \mathscr{X}$,

$$P_\theta(A \mid T) = \mathbb{E}_\theta[\mathbb{1}_A \mid T] = \mathbb{1}_A,$$

where the second equality holds because conditioning a $\sigma(T)$-measurable random variable on T returns itself. Since $\mathbb{1}_A$ does not depend on θ, T is sufficient. □

The interesting cases of sufficiency are when the statistic is not invertible: A non-

invertible statistics compresses the data in a strict sense, without losing information about the parameter.

If models have densities with respect to a common measure, we have a very useful characterization of sufficiency that allows us to check sufficiency by looking at the form of their *probability density functions*. We need these densities to be well-defined with respect to a common measure.

Definition 1.11. A statistical model $\mathcal{P}$ on a measurable space $(\mathcal{X}, \mathscr{X})$ is *dominated* by a σ-finite measure μ on $(\mathcal{X}, \mathscr{X})$ if every $P \in \mathcal{P}$ is absolutely continuous with respect to μ (denoted $P \ll \mu$). That is, for every $A \in \mathscr{X}$, if $\mu(A) = 0$, then $P(A) = 0$ for all $P \in \mathcal{P}$.

If a model is dominated by μ, the Radon–Nikodym theorem (see Theorem B.28 in Appendix B) guarantees the existence of a non-negative measurable function $p = dP/d\mu$, called the *Radon–Nikodym derivative* of P with respect to μ, such that for all $A \in \mathscr{X}$,

$$P(A) = \int_A p(x)\, d\mu(x).$$

When the model is parameterized as $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, we denote the density of P_θ by $p(\cdot \mid \theta)$ or $p_\theta(\cdot)$. The choice of dominating measure is not unique; if μ dominates $\mathcal{P}$, then any measure equivalent to μ also dominates $\mathcal{P}$.

For a fixed parameter value θ, the map $x \mapsto p(x \mid \theta)$ is the probability density function (with respect to μ). If we instead fix the observation x, the map $\theta \mapsto p(x \mid \theta)$ is called the *likelihood function*. Note that since the density is defined only up to a set of μ-measure zero, the likelihood function is also only defined up to a μ-null set of x's. For a fixed x, different versions of the density may yield different likelihood functions, but they will agree for μ-almost all x. In practice, we usually work with a specific, canonical version of the density (e.g., one that is continuous in x), which makes the likelihood unique.

The following theorem says that a statistic is sufficient if and only if the likelihood function can be factorized into a function of the statistic and a function of the data.

Theorem 1.12 (Fisher–Neyman Factorization)**.** *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be dominated by a σ-finite measure μ, with densities $p(x \mid \theta) = dP_\theta/d\mu$. A statistic T is sufficient for $\mathcal{P}$ if and only if*

$$p(x \mid \theta) = g(T(x), \theta)\, h(x) \quad \mu\text{-a.e.}, \quad \text{for all } \theta \in \Theta$$

for some non-negative measurable functions $g : \mathcal{T} \to [0, \infty)$ and $h : \mathcal{X} \to [0, \infty)$.

(♠)Proof. ($\Longleftarrow$) If $p(x \mid \theta) = g(T(x), \theta) \cdot h(x)$ for all $\theta \in \Theta$, then on the support of any $P_{\theta'}$ (i.e. where $h > 0$),

$$\frac{dP_\theta}{dP_{\theta'}} = \frac{g(T, \theta)}{g(T, \theta')}$$

is $\sigma(T)$-measurable. Hence, by Bayes formula (Theorem B.35 in Appendix B),

$$P_\theta(A \mid T) = E_\theta[\mathbb{1}_A \mid T] = \frac{E_{\theta'}[\mathbb{1}_A \frac{dP_\theta}{dP_{\theta'}} \mid T]}{E_{\theta'}[\frac{dP_\theta}{dP_{\theta'}} \mid T]} = E_{\theta'}[\mathbb{1}_A \mid T].$$

Since this is true all $\theta' \in \Theta$, we find that $P_\theta(A \mid T)$ admits a version that does is constant in θ. Hence, T is sufficient.

($\Longrightarrow$) Fix any $\theta_0 \in \Theta$. By the Halmos–Savage theorem, T being sufficient implies that dP_θ/dP_{θ_0} is $\sigma(T)$-measurable. Setting $g(T(x), \theta) := dP_\theta/dP_{\theta_0}(x)$ and $h := dP_{\theta_0}/d\mu$ gives the factorization:

$$\frac{dP_\theta}{d\mu}(x) = \frac{dP_\theta}{dP_{\theta_0}}(x) \cdot \frac{dP_{\theta_0}}{d\mu}(x) = g(T(x), \theta) \cdot h(x), \quad \mu - \text{a.e.}$$

□

The factorization says: the likelihood splits into a part g that depends on θ but only through $T(x)$, and a part h that depends on x directly but not on θ. All the θ-dependence is mediated by T.

Equipped with the Fisher–Neyman factorization theorem, we can revisit the example of two dice from Example 1.2.

Example 1.13. Consider again rolling two dice as in Example 1.2. We will discover that whether the sum $S(x, y) = x + y$ is sufficient depends critically on the assumed model.

Model 1 (Nonparametric): Suppose both dice are i.i.d. with unknown probability density[2] p on $\{1, \ldots, 6\}$, so the model is

$$\mathcal{P} = \{P_p : P_p(\{(x, y)\}) = p(x)p(y),\ p \text{ a probability density on } \{1, \ldots, 6\}\}.$$

The sum is *not* sufficient for the above model (no matter which parameterization is used). Consider the conditional probability of $(1, 6)$ given $S = 7$:

$$P_p\big((1, 6) \mid S = 7\big) = \frac{p(1)p(6)}{\sum_{k=1}^{6} p(k)p(7-k)}.$$

This depends on p. If p is uniform, this equals $1/6$. If the die is loaded toward extreme

[2] A Radon-Nikodym derivative with respect to the counting measure.

faces, it is larger. Knowing the sum is 7 does not pin down the conditional distribution of the outcome; the particular realization $(1, 6)$ versus $(3, 4)$ carries information about p.

Model 2 (A scalar family): Suppose each die follows a tilted distribution

$$p_\theta(x) = \frac{e^{\theta x}}{\sum_{k=1}^{6} e^{\theta k}}, \quad x \in \{1, \dots, 6\}, \quad \theta \in \mathbb{R}. \tag{1.1}$$

Here $\theta = 0$ gives fair dice, $\theta > 0$ biases toward higher faces, and $\theta < 0$ biases toward lower faces. The joint density is

$$p_\theta(x, y) = \frac{e^{\theta(x+y)}}{\left(\sum_{k=1}^{6} e^{\theta k}\right)^2} = \underbrace{\frac{e^{\theta(x+y)}}{\left(\sum_{k=1}^{6} e^{\theta k}\right)^2}}_{g(x+y,\,\theta)} \cdot \underbrace{1}_{h(x,y)},$$

which factors through the sum. By the Fisher–Neyman factorization theorem, S is sufficient for θ.

The contrast is instructive. Model 2 is indexed by a scalar parameter. Model 1 is nonparametric—a vastly larger model in which no reduction beyond the full data (the pair of eyes) is possible. Both models are identifiable (see Exercise 1.5), yet the sum is only sufficient in Model 2. Sufficiency is determined by the choice of the model. ♢

The example illustrates that sufficiency depends on the model $\mathcal{P}$: the same statistic can be sufficient for one family of distributions and insufficient for another.

Sometimes, two models may have sample spaces that look different, but they can be mapped to the same common sample space via sufficient statistics. This is called *observational equivalence*.

Definition 1.14. Two statistical models $(\mathcal{X}, \mathscr{X}, \{P_\theta : \theta \in \Theta\})$ and $(\mathcal{Y}, \mathscr{Y}, \{Q_\theta : \theta \in \Theta\})$ with a common parameter space Θ are *observationally equivalent* if there exist sufficient statistics $T : \mathcal{X} \to \mathcal{T}$ and $S : \mathcal{Y} \to \mathcal{T}$ such that $P_\theta^T = Q_\theta^S$ for all $\theta \in \Theta$.

Models being observationally equivalent means we can transform them to a common sample space, without losing information about the parameter. In particular, when one model can be mapped to the other, they are observationally equivalent.

Example 1.15. Let $\mathcal{P} = \{N(\mu, \sigma^2)^{\otimes n} : \mu \in \mathbb{R}\}$ on $\mathbb{R}^n$ and $\mathcal{Q} = \{N(\mu, \sigma^2/n) : \mu \in \mathbb{R}\}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for a fixed $\sigma^2 > 0$. Then $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ is sufficient for $\mathcal{P}$, the identity is sufficient for $\mathcal{Q}$, and both have distribution $N(\mu, \sigma^2/n)$. Thus, the two models are observationally equivalent. ♢

Intuitively, if two experiments are observationally equivalent, we should be able to simulate the outcome of one experiment using the outcome of the other (possibly

with some independent randomization), without knowing the true parameter value. However, under the notion of observational equivalence introduced in Definition 1.14, this is not always possible: it does not allow for randomization. Later, in Chapter 5, we will introduce a slightly more general notion of equivalence called *simulation equivalence* (sometimes called *Blackwell sufficiency*). This notion is more general than observational equivalence: it says that the two models are simulation equivalent if given the data of one model, we can simulate data as if it were generated by the other model. For most models (those defined on 'nice' sigma-algebras), observational equivalence implies simulation equivalence. For now, we will just illustrate this idea using the following example.

Example 1.16 (Simulation Equivalence of Normal Models)**.** Let $\mathcal{P} = \{N(\mu, \sigma^2)^{\otimes n} : \mu \in \mathbb{R}\}$ on $\mathbb{R}^n$ and $\mathcal{Q} = \{N(\mu, \sigma^2/n) : \mu \in \mathbb{R}\}$ on $\mathbb{R}$. These models are simulation equivalent:

- From $\mathcal{P}$ to $\mathcal{Q}$: Given $(X_1, \ldots, X_n) \sim P_\mu$, output $\bar{X} = n^{-1} \sum_{i=1}^n X_i$.
- From $\mathcal{Q}$ to $\mathcal{P}$: Given $Y \sim Q_\mu$, generate $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$ and output

$$X_i = Y + Z_i - \bar{Z}.$$

It can be shown that $P_\mu(\bar{X} \in A) = Q_\mu(A)$ for all $A \in \mathcal{B}(\mathbb{R})$ and $\mu \in \mathbb{R}$, and similarly, for $Y \sim Q_\mu$, $Y + Z_i - \bar{Z}$ can be shown to be distributed as $i = 1, \ldots, n$ i.i.d. $N(\mu, \sigma^2)$ random variables (see Exercise 1.6). ♢

A sufficient statistic always exists: the identity map $X(x) = x$ itself is trivially sufficient by Proposition 1.10. We typically seek a sufficient statistic that achieves maximal reduction of the data. This brings us to the notion of *minimal sufficiency*.

Definition 1.17. A sufficient statistic T is *minimal sufficient* for an experiment $(\mathcal{X}, \mathscr{X}, \{P_\theta : \theta \in \Theta\})$ if for any other sufficient statistic S, it satisfies $\sigma(T) \subseteq \sigma(S)$ modulo P_θ-null sets:

$$\sigma(T) \subseteq \sigma(\sigma(S) \cup \mathcal{N}), \qquad \mathcal{N} := \{N \in \mathscr{X} : P_\theta(N) = 0 \ \forall \theta \in \Theta\}.$$

Minimal sufficient statistics partition the sample space into the coarsest equivalence classes that preserve all information about θ. Any other sufficient statistic S generates a larger σ-algebra than T, except for sets which have probability zero under all P_θ.

Another way to think about minimal sufficiency is that a sufficient statistic T is *minimal sufficient* if it is a function of every other sufficient statistic. We can formalize this idea if T takes values in a measure space with a nice σ-algebra, like the Borel σ-algebra. For such nice σ-algebras, Definition 1.17 is equivalent to the following: for

any sufficient statistic S, there exists a measurable function f such that $T = f(S)$ almost surely under all P_θ, $\theta \in \Theta$ (Doob-Dynkin, Lemma B.36 in Appendix B).

Just as with sufficiency, minimal sufficiency can be difficult to verify. Luckily, we have the following useful tool to check minimal sufficiency.

Proposition 1.18. *Let $(\mathcal{X}, \mathscr{X}, \{P_\theta : \theta \in \Theta\})$ be a dominated model. A statistic T is minimal sufficient if and only if $T(x) = T(x')$ whenever the likelihood ratio $p(x \mid \theta)/p(x' \mid \theta)$ is constant in θ (except for a μ-null set).*

♠ *Proof.* ($\Longrightarrow$) Define an equivalence relation $\sim$ on $\mathcal{X}$ by $x \sim x'$ if and only if $p(x \mid \theta)/p(x' \mid \theta)$ is constant in θ, and let $S(x)$ denote the equivalence class of x. We claim S is sufficient (it is measurable with respect to the quotient σ-algebra $\mathscr{X}/\sim$, see Definition B.16 in Appendix B). For each equivalence class s, fix a representative x_s. For any x with $S(x) = s$, we have $p(x \mid \theta)/p(x_s \mid \theta) = h(x)$ for some function h not depending on θ. Thus,

$$p(x \mid \theta) = p(x_s \mid \theta)h(x) = g(S(x), \theta)h(x).$$

By the factorization theorem, S is sufficient. Since T is minimal sufficient, T is a function of S, so $S(x) = S(x')$ implies $T(x) = T(x')$. The equality $S(x) = S(x')$ is precisely the condition that the likelihood ratio is constant in θ.

($\Longleftarrow$) Suppose $T(x) = T(x')$ whenever the likelihood ratio is constant in θ. Let S be any sufficient statistic. By the factorization theorem, $p(x \mid \theta) = \tilde{g}(S(x), \theta)\tilde{h}(x)$. If $S(x) = S(x')$, then

$$\frac{p(x \mid \theta)}{p(x' \mid \theta)} = \frac{\tilde{h}(x)}{\tilde{h}(x')},$$

which is constant in θ. We have found that $S(x) = S(x')$ implies that $T(x) = T(x')$.

This aforementioned fact is sufficient to construct a function $f : S(\mathcal{X}) \to \mathcal{T}$ such that $f \circ S = T$. For each $s \in S(\mathcal{X})$, choose any $x_s \in S^{-1}(\{s\})$ and define $f(s) := T(x_s)$. This is well-defined: if $x' \in S^{-1}(\{s\})$ is another choice, then $S(x') = s = S(x_s)$, so by assumption $T(x') = T(x_s)$. For any $x \in \mathcal{X}$, we have $x \in S^{-1}(\{S(x)\})$, hence $f(S(x)) = T(x)$. Moreover, f is measurable: we have $S^{-1}(f^{-1}(A)) = T^{-1}(A) \in \mathscr{X}$.

Thus, T is a measurable function of S. Since S was arbitrary, T is minimal sufficient. □

To illustrate minimal sufficiency, we consider the following examples.

Example 1.19 (Minimal Sufficiency for Uniform)**.** Let $\text{Uniform}(0, \theta)$ be the uniform distribution on the interval $[0, \theta]$ with $\theta > 0$, that is, probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined through the Lebesgue density $p(x \mid \theta) = \frac{1}{\theta}\mathbf{1}\{0 \leqslant x \leqslant \theta\}$.

Consider the statistical model corresponding to $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$ with $\theta > 0$. The likelihood is

$$p(x \mid \theta) = \frac{1}{\theta^n} \mathbf{1}\{x_{(n)} \leqslant \theta\},$$

where $x_{(n)}$ is the n-th order statistic. By Theorem 1.12, the n-th order statistic $X_{(n)}$ is sufficient. The likelihood ratio is

$$\frac{p(x \mid \theta)}{p(y \mid \theta)} = \frac{\mathbf{1}\{x_{(n)} \leqslant \theta\}}{\mathbf{1}\{y_{(n)} \leqslant \theta\}}.$$

This is constant in θ if and only if $x_{(n)} = y_{(n)}$ (check this!). Thus, $X_{(n)}$ is minimal sufficient. ◊

Example 1.20 (Minimal Sufficiency via Likelihood Ratios)**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$.

Case 1: σ^2 known. The likelihood ratio is

$$\frac{p(x \mid \mu)}{p(y \mid \mu)} = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - \sum_i y_i^2 - 2\mu n(\bar{x} - \bar{y})\right)\right).$$

This is constant in μ if and only if $\bar{x} = \bar{y}$. Thus $\bar{X}$ is minimal sufficient.

Case 2: Both μ and σ^2 unknown. The likelihood ratio is

$$\frac{p(x \mid \mu, \sigma^2)}{p(y \mid \mu, \sigma^2)} = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - \sum_i y_i^2\right) + \frac{\mu}{\sigma^2} n(\bar{x} - \bar{y})\right).$$

This is constant in (μ, σ^2) if and only if $\bar{x} = \bar{y}$ and $\sum_i x_i^2 = \sum_i y_i^2$. Thus $(\bar{X}, \sum_i X_i^2)$ is minimal sufficient. ◊

Next, we introduce an additional type of sufficiency called *completeness*. A property that rules out redundancy in a statistic completely.

Definition 1.21. A statistic T is *complete* for $\mathcal{P}$ if for every $\sigma(T)$-measurable integrable random variable U,

$$\mathbb{E}_P[U] = 0 \text{ for all } P \in \mathcal{P} \implies U = 0 \quad P\text{-a.s. for all } P \in \mathcal{P}.$$

The definition of completeness is of a technical nature: if T is complete, there is no non-trivial function of T whose expectation is constant across all $P \in \mathcal{P}$. Given an identified model $\{P_\theta : \theta \in \Theta\}$, the idea is this: T contains no component that varies with the data but carries zero information about θ on average.

Completeness and sufficiency are logically independent properties: neither implies the other. Completeness is not in and of itself useful; the trivial statistic $T : x \mapsto c$ for

a constant c is complete for any model (check). However, when combined, they yield a powerful result: a complete sufficient statistic is automatically minimal sufficient.

Theorem 1.22 (Bahadur)**.** *If T is complete and sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, then T is minimal sufficient.*

Proof. Let S be any sufficient statistic. Fix $B \in \sigma(T)$. By sufficiency of S, there exists $H_B : \mathcal{X} \to \mathbb{R}$ such that

$$H_B := \mathbb{E}_\theta[\mathbb{1}_B \mid \sigma(S)] \quad P_\theta\text{-a.s. for all } \theta \in \Theta.$$

Fix such a version H_B, so H_B is $\sigma(S)$-measurable and note that $0 \leqslant H_B \leqslant 1$.

The random variable

$$U := \mathbb{E}_\theta[H_B \mid \sigma(T)] - \mathbb{1}_B$$

is $\sigma(T)$-measurable (for each θ), integrable, and satisfies $\mathbb{E}_\theta[U] = 0$ for all θ. By completeness of T (Definition 1.21), we conclude that $U = 0$ P_θ-a.s. for all θ, i.e.

$$\mathbb{E}_\theta[H_B \mid \sigma(T)] = \mathbb{1}_B \qquad P_\theta\text{-a.s. for all } \theta.$$

Since $0 \leqslant H_B \leqslant 1$, this identity forces $H_B = \mathbb{1}_B$ P_θ-a.s. (indeed, on B we have $\mathbb{E}_\theta[1 - H \mid \sigma(T)] = 0$, and on B^c we have $\mathbb{E}_\theta[H \mid \sigma(T)] = 0$). Since H_B is $\sigma(S)$-measurable, $H_B = \mathbb{1}_B$ P_θ-a.s. implies that $\mathbb{1}_B$ is $\sigma(S)$-measurable modulo P_θ-null sets. As $B \in \sigma(T)$ was arbitrary, $\sigma(T) \subseteq \sigma(S)$ modulo P_θ-null sets for every θ. □

The previous theorem shows that the notion of completeness is stronger than minimal sufficiency: every complete sufficient statistic is minimal sufficient. The converse is not true: the following example shows that minimal sufficiency does not imply completeness.

Example 1.23. Let $X_1, \ldots, X_n \sim \text{Uniform}(\theta, \theta + 1)$ for $\theta \in \mathbb{R}$. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \mathbf{1}_{\{\theta \leqslant x_i \leqslant \theta+1\}} = \mathbf{1}_{\{x_{(n)}-1 \leqslant \theta \leqslant x_{(1)}\}}.$$

The likelihood is non-zero if and only if the interval $[x_{(n)} - 1, x_{(1)}]$ is non-empty and contains θ. The pair $T = (X_{(1)}, X_{(n)})$ determines the likelihood function (as a function of θ) and is therefore minimal sufficient.

However, T is not complete.

Consider the statistic $R = X_{(n)} - X_{(1)}$. The expectation of R is

$$\mathbb{E}_\theta[R] = \mathbb{E}_\theta[X_{(n)}] - \mathbb{E}_\theta[X_{(1)}] = (\theta + \frac{n}{n+1}) - (\theta + \frac{1}{n+1}) = \frac{n-1}{n+1} \quad \text{(check)}.$$

Pick arbitrary $\theta_0 \in \mathbb{R}$. Since R is not a constant, $g(T) = R - \mathbb{E}_{\theta_0}[R]$ is a non-zero function of T. However, R has zero expectation for all θ. Thus, T is not complete. $\lozenge$

Next, we introduce the concept of *ancillarity*. This is a property of a statistic that is independent of the parameter.

Definition 1.24. Consider statistic V mapping $(\mathcal{X}, \mathscr{X})$ to $(\mathcal{V}, \mathscr{V})$. V is *ancillary* for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if $P_\theta(x : V(x) \in A)$ does not depend on θ for all $A \in \mathscr{V}$. That is, if the distribution of $V(X)$ does not depend on θ.

While a sufficient statistic carries all the information about θ, an ancillary statistic carries none—its distribution is the same regardless of which P_θ generated the data.

Example 1.25 (Ancillary in Scale Families)**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$ for unknown $\theta > 0$. The maximum $X_{(n)}$ is sufficient for θ (Example 1.19). The ratios

$$\left(\frac{X_1}{X_{(n)}}, \ldots, \frac{X_{n-1}}{X_{(n)}}\right)$$

are ancillary: their joint distribution does not depend on θ. To see this, write $X_i = \theta U_i$ where $U_i \overset{\text{iid}}{\sim} \text{Uniform}(0, 1)$. Then $X_i/X_{(n)} = U_i/U_{(n)}$, which involves only the U_i. $\lozenge$

A sufficient statistic captures all information about θ; an ancillary statistic carries none. One might hope these two types of statistics are "orthogonal" in some sense —- the sufficient part and the ancillary part of the data do not interact. This is not true in general: a minimal sufficient statistic can be dependent on an ancillary statistic. However, when the sufficient statistic is also complete, this independence is guaranteed. This is the content of Basu's theorem.

Theorem 1.26 (Basu)**.** *Consider a statistical model $(\mathcal{X}, \mathscr{X}, \mathcal{P})$ with $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Let T be complete and sufficient for $\mathcal{P}$, and let V be ancillary for $\mathcal{P}$. Then T and V are independent under every $P_\theta \in \mathcal{P}$.*

Proof. Fix $B \in \mathscr{V}$ and set $A := \{V \in B\}$. By sufficiency of T, the conditional expectation

$$H_B := \mathbb{E}_\theta[\mathbb{1}_A \mid \sigma(T)]$$

admits a version that is the same for all θ (i.e. H_B is $\sigma(T)$-measurable and does not depend on θ up to P_θ-a.s. equality). By ancillarity, $c_B := P_\theta(A)$ is constant in θ. Hence for every θ,

$$\mathbb{E}_\theta[H_B] = \mathbb{E}_\theta[\mathbb{1}_A] = c_B,$$

so with $G_B := H_B - c_B$ we have G_B $\sigma(T)$-measurable and $\mathbb{E}_\theta[G_B] = 0$ for all θ. By completeness of T (equivalently, of $\sigma(T)$), $G_B = 0$ P_θ-a.s. for all θ, i.e.

$$\mathbb{E}_\theta[\mathbb{1}_{\{V \in B\}} \mid \sigma(T)] = P_\theta(V \in B) \qquad P_\theta\text{-a.s.}$$

for all $B \in \mathscr{V}$. This is exactly the independence of V and $\sigma(T)$, hence of V and T. □

Basu's theorem is very useful for showing independence between statistics without deriving their joint distribution directly.

Example 1.27. Revisiting Example 1.25, where $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$, we identified that $X_{(n)}$ is sufficient and the ratios $V = (X_1/X_{(n)}, \dots, X_{n-1}/X_{(n)})$ are ancillary.

It turns out that $X_{(n)}$ is complete. Hence, by Basu's theorem, $X_{(n)}$ and V are – perhaps surprisingly – independent.

To check completeness of $T := X_{(n)}$, note that its density is $f_T(t) = nt^{n-1}/\theta^n$ for $0 < t < \theta$. Suppose $\mathbb{E}_\theta[g(T)] = 0$ for all $\theta > 0$. Then

$$\int_0^\theta g(t) t^{n-1}\, dt = 0 \quad \text{for all } \theta > 0.$$

Differentiating with respect to θ gives $g(\theta)\theta^{n-1} = 0$, which implies $g(\theta) = 0$ for almost all θ. Thus, $X_{(n)}$ is complete. ◊

There is a particular class of models for which it is easy to check completeness of sufficient statistics: the class of exponential families, which we will introduce in the next section.

Exponential Families

Many common statistical models share a structure that leads to elegant sufficiency and completeness results.

Definition 1.28. A family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ on $(\mathcal{X}, \mathscr{X})$ dominated by a σ-finite measure μ is an *exponential family* if the densities can be written as

$$p(x \mid \theta) = \exp\big\{\eta(\theta)^\top T(x) - B(\theta)\big\} h(x), \tag{1.2}$$

where $T : \mathcal{X} \to \mathbb{R}^k$ and $h : \mathcal{X} \to [0, \infty)$ are measurable functions, and $\eta : \Theta \to \mathbb{R}^k$.

The map $T : \mathcal{X} \to \mathbb{R}^k$ is the *natural sufficient statistic*: by the Fisher–Neyman factorization theorem, T is sufficient for θ in any exponential family – the density (1.2) factors as $g(T(x), \theta) \cdot h(x)$. For i.i.d. observations $X_1, \dots, X_n$ from a distribution in an exponential family, the sufficient statistic is the sum $\sum_{i=1}^n T(X_i)$.

Proposition 1.29. *If $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is an exponential family with natural sufficient statistic T, then the model $\{P_\theta^{\otimes n} : \theta \in \Theta\}$ is also an exponential family, with natural sufficient statistic $(x_1, \dots, x_n) \mapsto \sum_{i=1}^n T(x_i)$.*

Proof. The joint density is

$$\prod_{i=1}^{n} p(x_i \mid \theta) = \exp\Big\{\eta(\theta)^\top \sum_{i=1}^{n} T(x_i) - nB(\theta)\Big\} \prod_{i=1}^{n} h(x_i),$$

which is of the form (1.2) with $\tilde{T}(x_1, \ldots, x_n) = \sum_{i=1}^{n} T(x_i)$, $\tilde{B}(\theta) = nB(\theta)$, and $\tilde{h}(x_1, \ldots, x_n) = \prod_{i=1}^{n} h(x_i)$. □

The exponential family is in *natural* (or *canonical*) *parameterization* if $\Theta \subseteq \mathbb{R}^k$ and $\eta(\theta) = \theta$ is the identity map. The *natural parameter space* is

$$\mathcal{H} = \Big\{\eta \in \mathbb{R}^k : \int_{\mathcal{X}} \exp\big\{\eta^\top T(x)\big\} h(x)\, d\mu(x) < \infty\Big\}.$$

The set $\mathcal{H}$ is convex (verify this). A naturally parameterized exponential family with $\mathcal{H}$ is called *full-rank* if $\mathcal{H}$ contains an open subset of $\mathbb{R}^k$. For full-rank exponential families, there is a convenient route to minimal sufficiency: we show that T is complete using the proposition below, after which minimality is implied by Theorem 1.22.

Proposition 1.30. *Let $\mathcal{P} = \{P_\eta : \eta \in \mathcal{H}\}$ be an exponential family in natural parameterization with natural sufficient statistic T. If $\mathcal{H}$ contains an open subset of $\mathbb{R}^k$, then T is complete.*

♠*Proof.* The density of T (note T takes values in a regular Borel space) with respect to some base measure ν is

$$p_T(t \mid \eta) = \exp(\eta^\top t - B(\eta)).$$

Suppose $E_\eta[g(T)] = 0$ for all $\eta \in \mathcal{H}$. Then

$$\int g(t) \exp(\eta^\top t - B(\eta))\, d\nu(t) = 0$$

for all $\eta \in \mathcal{H}$. Since $e^{-B(\eta)} \neq 0$, this implies

$$\int g(t) \exp(\eta^\top t)\, d\nu(t) = 0$$

for all η in an open subset of $\mathcal{H}$. The left side is the Laplace transform of the (signed) measure $g\, d\nu$. Since Laplace transforms are analytic and this one vanishes on an open set, it vanishes on its entire domain. By uniqueness of the Laplace transform, $g\, d\nu = 0$, so $g(T) = 0$ ν-a.s., hence P_η-a.s. for all $\eta \in \mathcal{H}$. □

What if $\{P_\theta : \theta \in \Theta\}$ is not a family in natural parameterization? If η is injective, we may re-index the family as follows. Set $\Xi := \eta(\Theta)$ and define a re-parameterized

family

$$Q_\xi := P_{\eta^{-1}(\xi)}, \qquad \xi \in \Xi.$$

Then, by injectivity of η, we have

$$\{Q_\xi : \xi \in \Xi\} = \{P_\theta : \theta \in \Theta\}$$

as sets of probability measures. Given this re-indexing, if $\eta(\Theta)$ has nonempty interior in $\mathbb{R}^k$ (equivalently; contains an open set), Proposition 1.30 implies that the natural sufficient statistic T is complete for this family, hence also complete under the original parameterization. Indeed, completeness is a property of the family of distributions, not of the parameterization: if two parameterizations define the same collection of probability measures and T is complete for one, it is complete for the other as well.

The exponential family encompasses a wide range of models. Many of the models we will see in this course belong to this class.

Example 1.31 (Common exponential families)**.** Most identifiably-parameterized commonly-used exponential families are full rank.

- **Poisson model:** Consider the Poisson(θ) distribution for $\theta > 0$, defined by its density with respect to the counting measure:

  $$p(x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!} = \frac{1}{x!}\exp\big(x \log\theta - \theta\big), \quad x \in \{0, 1, \dots\}.$$

  This constitutes an exponential family with sufficient statistic $T(x) = x$ and natural parameter $\eta = \log\theta$. Since $\theta > 0$, the natural parameter η ranges over all of $\mathbb{R}$, which is an open set, so Proposition 1.30 yields that T is complete.

  Similarly, for a sample $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$, the sum $T = \sum_{i=1}^n X_i$ is a complete sufficient statistic.

- **Binomial model:** The Bernoulli(p) distribution for $p \in (0, 1)$ has counting measure density

  $$p(x \mid p) = p^x(1-p)^{1-x}, \quad x \in \{0, 1\}.$$

  This is an exponential family with $T(x) = x$ and natural parameter $\eta(p) = \log(p/(1-p))$. Since $\eta(p)$ ranges over all of $\mathbb{R}$ for $p \in (0, 1)$, the family is full-rank. For the n i.i.d. draws model – $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ – the sum $\sum_{i=1}^n X_i$ is complete and sufficient. Since the map $p \mapsto \eta(p)$ is a bijection between the parameter space $(0, 1)$ and the natural parameter space $\mathbb{R}$, the statistic is also complete and sufficient for the original parameterization.

- **Multinomial model:** Consider a categorical distribution on k categories with

probabilities $p = (p_1, \dots, p_k)$ satisfying $p_j > 0$ and $\sum_{j=1}^k p_j = 1$. A single observation is a basis vector $x = e_j$ indicating category j, equivalently represented as $x \in \{0,1\}^k$ with $\sum_{j=1}^k x_j = 1$. The density with respect to counting measure is

$$p(x \mid p) = \prod_{j=1}^{k} p_j^{x_j}.$$

Using the constraint $p_k = 1 - \sum_{j=1}^{k-1} p_j$ and $x_k = 1 - \sum_{j=1}^{k-1} x_j$:

$$\log p(x \mid p) = \sum_{j=1}^{k-1} x_j \log p_j + \Big(1 - \sum_{j=1}^{k-1} x_j\Big) \log p_k = \sum_{j=1}^{k-1} x_j \log \frac{p_j}{p_k} + \log p_k.$$

This is an exponential family with sufficient statistic $T(x) = (x_1, \dots, x_{k-1}) \in \mathbb{R}^{k-1}$ and natural parameter $\eta_j = \log(p_j/p_k)$ for $j = 1, \dots, k-1$. Since $p_j > 0$ for all j, the ratios p_j/p_k can take any positive value, so $\eta \in \mathbb{R}^{k-1}$. The natural parameter space is all of $\mathbb{R}^{k-1}$, which is open, so the family is full-rank.

For $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Categorical}(p)$, the sufficient statistic is $\sum_{i=1}^n T(X_i) = (N_1, \dots, N_{k-1})$, where $N_j = \sum_{i=1}^n X_{ij}$ counts observations in category j. This statistic is complete.

- **Gamma model:** Consider the Gamma(a, b) distribution with shape parameter $a > 0$ and rate parameter $b > 0$: its density (with respect to Lebesgue measure on $(0, \infty)$) is

$$p(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} = \exp\big((a-1)\log x - bx + a \log b - \log \Gamma(a)\big).$$

This constitutes an exponential family with sufficient statistic $T(x) = (\log x, x)$ and natural parameter $\eta = (a-1, -b)$. The natural parameter space is $(-1, \infty) \times (-\infty, 0)$, which is an open subset of $\mathbb{R}^2$.

For a sample $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(a, b)$, the statistic $T = (\sum_{i=1}^n \log X_i, \sum_{i=1}^n X_i)$ is sufficient and complete.

- **Multivariate normal model:** Consider the $N_d(\mu, \Sigma)$ distribution for $\mu \in \mathbb{R}^d$ and Σ positive definite. The density with respect to Lebesgue measure on $\mathbb{R}^d$ is

$$p(x \mid \mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\Big(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)\Big).$$

Expanding the quadratic form gives

$$p(x \mid \mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\Big((\Sigma^{-1}\mu)^\top x - \frac{1}{2} x^\top \Sigma^{-1} x - \frac{1}{2} \mu^\top \Sigma^{-1} \mu\Big).$$

This is an exponential family with sufficient statistic $T(x) = (x, xx^\top)$ and natural parameters $\eta_1 = \Sigma^{-1}\mu \in \mathbb{R}^d$ and $\eta_2 = -\frac{1}{2}\Sigma^{-1}$, a negative definite $d \times d$ matrix. The natural parameter space is $\mathbb{R}^d \times \{M \in \mathbb{R}^{d\times d}_{\text{sym}} : M < 0\}$, which is open in $\mathbb{R}^{d+d(d+1)/2}$.

For i.i.d. observations $X_1, \ldots, X_n$, the sufficient statistic is $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i X_i^\top)$. Since $\sum_{i=1}^n X_i X_i^\top = S + n\bar{X}\bar{X}^\top$ where $S = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top$, the pair $(\bar{X}, S)$ is an equivalent (and complete) sufficient statistic.

The map $(\mu, \Sigma) \mapsto (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1})$ is a bijection between the original parameter space $\mathbb{R}^d \times \{M \in \mathbb{R}^{d\times d}_{\text{sym}} : M > 0\}$ and the natural parameter space, so $(\bar{X}, S)$ is also complete for the original parameterization $\mu \in \mathbb{R}^d$ and $\Sigma > 0$.

A non-example (a curved exponential family): An exponential family model can fail to have a complete sufficient statistic if the parameter space has a different dimension than the minimal sufficient statistic. For example, suppose $X_1, \ldots, X_n \sim$ i.i.d. $N(\theta, \theta^2)$. Then the density can be expressed as

$$p(x_1, \ldots, x_n|\theta) = c(\theta)h(x_1, \ldots, x_n) \exp\{-\sum x_i^2/[2\theta^2] + \sum x_i/\theta\}.$$

In this case, $T(x) = (\sum x_i, \sum x_i^2)$ is a minimal sufficient statistic (why?), but $\eta(\Theta)$ equals $\{(-1/[2\theta^2], 1/\theta) : \theta \in \Theta\}$, which is a curve in $\mathbb{R}^2$ – it does not contain an open set.

$\Diamond$

We have so far discussed sufficiency and completeness as properties of the statistical model, allowing us to identify when different mathematical formulations of models are the same 'for all intents and purposes'. This development, however, has been independent of any specific statistical task. To proceed, we must define what we aim to achieve with the data—whether to estimate a parameter, test a hypothesis, or predict a future value—and how to evaluate our success. In the next section, we introduce the framework of decision theory, which allows us to formalize these goals and rigorously compare statistical procedures.

1.3 Decision Problems

Given all the possiblities in terms of writing down models, which one is the 'correct' one? One might be tempted to adopt a very large model on the grounds that it is "most likely to contain the true data-generating process". We will see that this reasoning is flawed. Larger models typically come with costs: more parameters to estimate, higher variance, etc. To compare models meaningfully, we must first specify *what we intend*

to do with the data: what decision or action we will take, and how we quantify the consequences of making that decision under different possible states of nature. Only once the decision problem (and an associated utility or loss function) has been fixed can we meaningfully compare statistical procedures, models, or parameterizations according to their expected performance.

Given observed data x and a statistical model $\mathcal{P}$, we want to determine which decision to take. For example, we may want to infer which probability distribution in the collection $\mathcal{P}$ is 'most likely' to have generated the data. Perhaps we are interested in testing whether a particular $P_0 \in \mathcal{P}$ gave rise to the observed data versus it is more likely that the data was generated by distribution $P_1 \in \mathcal{P} \backslash \{P_0\}$. Perhaps we are interested in predicting a future observed data from the data-generating process, and we want to make sure that our prediction is 'good' in the sense that if given that the true data-generating process is $P_0 \in \mathcal{P}$, then the prediction is 'close' to future data drawn from P_0. We will develop the formal framework that encapsulates these different inferential goals in a unified way.

To formalize this, we first need to specify the ingredients of a decision problem. Besides a statistical experiment $(\mathcal{X}, \mathscr{X}, \mathcal{P})$, a decision problem consists of set of possible actions (decisions), and a way to quantify the consequences of each action under each possible data-generating process. The set of possible actions forms the *decision space* $\mathcal{D}$, equipped with a σ-algebra $\mathscr{D}$ to ensure measurability when we later define expectations and integrals over decisions.

Definition 1.32. A *decision space* is a measurable space $(\mathcal{D}, \mathscr{D})$.

A *(deterministic) decision rule* is a measurable function that assigns an action to each possible data outcome.

Definition 1.33. A *(deterministic) decision rule* is a measurable function δ mapping the sample space $(\mathcal{X}, \mathscr{X})$ into $(\mathcal{D}, \mathscr{D})$.

Later on, we will also consider *randomized decision rules*, which allow for probabilistic mixing of different deterministic decision rules.

To evaluate the quality of decisions, we need a *loss function* that measures the penalty for choosing action $d \in \mathcal{D}$ when the true data-generating process is P_θ.

Definition 1.34. A *loss function* is a function $L : \Theta \times \mathcal{D} \to [0, \infty)$ such that $d \mapsto L(\theta, d)$ is measurable for each $\theta \in \Theta$.

The loss function $L(\theta, d)$ quantifies the penalty incurred by taking decision d when the 'true state of nature' is θ. The choice of decision space $\mathcal{D}$ and loss function L depends on the inferential goal and the consequences of errors in the application at hand.

Two of the most common types of problems which we will study extensively with corresponding loss functions are estimation and hypothesis testing.

Example 1.35 (Estimation)**.** Suppose we wish to estimate an unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^k$. A natural choice is to take the decision space $\mathcal{D} = \Theta$ and to equip it with the Borel sigma-algebra $\mathscr{D} = \mathcal{B}(\Theta)$.

Examples of loss functions for estimation:

- Euclidean distance; $L(\theta, d) = \|\theta - d\|$.
- Squared Euclidean distance; $L(\theta, d) = \|\theta - d\|^2$.
- Sup-norm loss; $L(\theta, d) = \|\theta - d\|_\infty = \max_i |\theta_i - d_i|$.
- Zero-one loss; $L(\theta, d) = \mathbb{1}_{\{\theta \neq d\}}$.

$\Diamond$

Example 1.36 (Hypothesis Testing)**.** Suppose we wish to test between two hypotheses: $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, where $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \varnothing$. The decision space is $\mathcal{D} = \{0, 1\}$, where $d = 0$ means "accept H_0" and $d = 1$ means "accept H_1". A simple loss function is:

$$L(\theta, d) = \begin{cases} 0 & \text{if } d = 0 \text{ and } \theta \in \Theta_0, \\ 0 & \text{if } d = 1 \text{ and } \theta \in \Theta_1, \\ a & \text{if } d = 0 \text{ and } \theta \in \Theta_1, \\ b & \text{if } d = 1 \text{ and } \theta \in \Theta_0, \end{cases}$$

where $a, b > 0$ are the costs of Type II and Type I errors, respectively. When $a = b = 1$, this is the zero-one loss. When $a \neq b$, we reflect asymmetric consequences—for instance, in medical testing, falsely declaring a patient healthy (Type II error) might be far more costly than falsely declaring them sick (Type I error). $\Diamond$

Combining all the ingredients, we can now define a decision problem.

Definition 1.37. A (statistical) *decision problem* is a tuple $(\mathcal{X}, \mathscr{X}, \mathcal{P}, \Theta, (\mathcal{D}, \mathscr{D}), L)$ where $(\mathcal{X}, \mathscr{X}, \mathcal{P}, \Theta)$ is a statistical experiment and $(\mathcal{D}, \mathscr{D})$ is a decision space and L is a loss function.

For decision problems with identifiable models, the expected loss (under P_θ) of the decision rule δ given the 'true state of nature θ' is called its *risk*.

Definition 1.38 (Risk Function)**.** Given an identifiable statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ on a measurable space $(\mathcal{X}, \mathscr{X})$, a decision space $(\mathcal{D}, \mathscr{D})$ and a loss function

$L : \Theta \times \mathcal{D} \to \mathbb{R}$, the *risk* of a decision rule $\delta : \mathcal{X} \to \mathcal{D}$ is defined as

$$\mathcal{R}(\theta, \delta) := \int_{\mathcal{X}} L(\theta, \delta(x)) dP_\theta(x), \tag{1.3}$$

whenever this integral exists in $[-\infty, \infty]$.

Instead of the integral in Equation (1.3), we will frequently write

$$\mathcal{R}(\theta, \delta) = \mathbb{E}_{P_\theta}[L(\theta, \delta(X))] \equiv \mathbb{E}_\theta[L(\theta, \delta(X))],$$

where the expectation is to be understood as the expectation under the probability distribution P_θ and the random element X is simply the identity map on $\mathcal{X}$ (see Remark 1.7).

Throughout this course, we evaluate decision rules based on their risk, the expected loss. One might object that focusing on the first moment of the loss is a limiting choice made for mathematical convenience. However, there is considerable flexibility in the choice of loss function itself. For instance, if our goal is for an estimator to be ϵ-close to the true parameter with high probability, the loss function $L(\theta, d) = \mathbb{1}_{\{\|\theta - d\| > \epsilon\}}$ captures exactly this objective.

Beyond flexibility in L, there are deeper reasons to focus on expected loss. One appeals to frequency under (hypothetical) repetitions and betting interpretations: if we use a decision rule δ repeatedly under the same conditions, the average loss converges to the risk $\mathcal{R}(\theta, \delta)$ by the law of large numbers. A more philosophical justification, grounded in rational preferences over random outcomes, is given in Section 1.3.3.

So far, our decision rules $\delta : \mathcal{X} \to \mathcal{D}$ have been deterministic: given data x, the action $\delta(x)$ is fully determined. Just as mixed strategies in game theory allow players to randomize over actions, we can allow decision rules to incorporate auxiliary randomness independent of the data.

Definition 1.39. A *randomized decision rule* is a measurable function $\delta : \mathcal{X} \times [0, 1] \to \mathcal{D}$. Given data $x \in \mathcal{X}$ and an independent random variable $U \sim \text{Uniform}(0, 1)$, the decision is $\delta(x, U)$.

The risk of a randomized rule averages over both the data and the auxiliary randomness:

$$\mathcal{R}(\theta, \delta) = \int_{\mathcal{X}} \int_0^1 L(\theta, \delta(X, u)) du dP_\theta(x) = \int_0^1 \int_{\mathcal{X}} L(\theta, \delta(X, u)) dP_\theta(x) du. \tag{1.4}$$

Often, we will simply write this as $\mathbb{E}_\theta\big[L(\theta, \delta(X, U))\big]$, but it is important to remember that U is ancillary to the model.

Why allow randomization? In most estimation problems, deterministic rules suffice—randomization cannot improve expected performance when the loss is convex (as we will see in Section 1.3.1 below). However, randomization becomes important in two settings we will study later:

- *Hypothesis testing* (Chapter 3): To achieve exactly a prescribed significance level α, we may need to randomize when the test statistic falls on the boundary of the rejection region (see Exercise 1.14).
- *Bayesian decision theory* (Chapter 4).

Furthermore, for general (non-convex) loss functions, randomization can strictly reduce the *worst-case risk*, as shown in the following example.

Example 1.40 (Matching pennies)**.** Let $X \in \{0, 1\}$ be a single observation from a model with $\Theta = \{0, 1\}$ and

$$P_\theta(X = 1) = \begin{cases} 0.3 & \text{if } \theta = 0, \\ 0.6 & \text{if } \theta = 1. \end{cases}$$

The decision space is $\mathcal{D} = \{0, 1\}$ with 0-1 loss $L(\theta, d) = \mathbb{1}\{\theta \neq d\}$. Consider the deterministic rule $\delta(x) = x$. Its risk is

$$R(0, \delta) = P_0(X = 1) = 0.3, \qquad R(1, \delta) = P_1(X = 0) = 0.4.$$

The worst-case risk is $\max_\theta R(\theta, \delta) = 0.4$.

Now consider a randomized rule that follows $\delta(x) = x$ except when $X = 0$, where it randomizes:

$$\delta(x, u) = \begin{cases} \mathbb{1}\{u \leqslant \gamma\} & \text{if } x = 0, \\ 1 & \text{if } x = 1. \end{cases}$$

The risks are $R(0, \delta) = 0.3 + 0.7\gamma$ and $R(1, \delta) = 0.4(1 - \gamma)$. Setting these equal gives $\gamma = 1/11$, yielding

$$\max_\theta R(\theta, \delta) = \frac{4}{11} \approx 0.364 < 0.4.$$

Randomization strictly improves the risk for the worst-case θ. ♢

The example above shows that randomization allows us to 'hedge' against the worst-case scenario. We will return to such worst-case analyses in later chapters.

For convex loss functions, however, randomized decision rules do not outperform deterministic decision rules. The theorem below makes this precise: for any randomized decision rule, there exists a deterministic decision rule with at most the same risk if the loss is convex.

Theorem 1.41. *If $d \mapsto L(\theta, d)$ is convex for all $\theta \in \Theta$, then for any randomized decision rule δ, there exists a deterministic decision rule δ^* with $\mathcal{R}(\theta, \delta^*) \leqslant \mathcal{R}(\theta, \delta)$ for all $\theta \in \Theta$.*

Proof. Let $\delta : \mathcal{X} \times [0,1] \to \mathcal{D}$ be a randomized decision rule. Define the deterministic decision rule $\delta^*(x) = \mathbb{E}^U[\delta(x, U)]$. By convexity of $d \mapsto L(\theta, d)$ and Jensen's inequality,

$$L(\theta, \delta^*(x)) = L\big(\theta, \mathbb{E}^U[\delta(x, U)]\big) \leqslant \mathbb{E}^U[L(\theta, \delta(x, U))].$$

Taking expectations over $X \sim P_\theta$ yields

$$\mathcal{R}(\theta, \delta^*) = \mathbb{E}_\theta[L(\theta, \delta^*(X))] \leqslant \mathbb{E}_\theta\big[\mathbb{E}^U[L(\theta, \delta(X, U))]\big] = \mathcal{R}(\theta, \delta). \qquad \square$$

Remark 1.42 (Why a uniform random variable?). The choice of $U \sim \text{Uniform}(0,1)$ as the source of randomness may seem restrictive. Does a single uniform provide enough randomness? For decision spaces with standard measurability properties (e.g., $\mathcal{D} \subseteq \mathbb{R}^k$ with the Borel σ-algebra), the answer is yes: any conditional distribution on $\mathcal{D}$ can be generated from a uniform random variable. We revisit this in Section 4.2.

1.3.1 Sufficiency and loss

Intuitively, a sufficient statistic T contains all the information about the parameter θ that is relevant to making decisions. So, if we are able to attain a certain level of risk in one model, we should be able to attain the same level of risk in the forward model induced by sufficient statistic T.

The Rao-Blackwell theorem formalizes this intuition. It states that for any convex loss function, we can improve (or at least match) the performance of any decision rule by conditioning on a sufficient statistic.

Theorem 1.43 (Rao-Blackwell, convex loss)**.** *Let T be a $(\mathcal{T}, \mathscr{T})$-valued sufficient statistic for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, and let $L : \Theta \times \mathcal{D} \to [0, \infty)$ be a loss function.*

If $d \mapsto L(\theta, d)$ is convex for each θ and $\mathcal{D} \subseteq \mathbb{R}^m$ is convex, closed and equipped with the Borel σ-algebra, then for any decision rule $\delta : \mathcal{X} \to \mathcal{D}$ with $\mathbb{E}_\theta[\|\delta(X)\|] < \infty$, there exists a decision rule δ^* satisfying $\delta^*(X) = \mathbb{E}_\theta[\delta(X) \mid T]$ P_θ-a.s. for all $\theta \in \Theta$ and

$$\mathcal{R}(\theta, \delta^*) \leqslant \mathcal{R}(\theta, \delta) \quad \textit{for all } \theta \in \Theta.$$

If L is strictly convex, the inequality is strict unless δ is already a function of T.

Proof. Fix any $\theta \in \Theta$. The random vector $\delta^*(X) = \mathbb{E}_\theta[\delta(X) \mid T]$ is $\sigma(T)$-measurable (and hence $\mathscr{X}$-measurable) and admits a version not depending on θ; meaning that

there exists a version of the conditional expectation that does not depend on θ (through similar arguments as in Exercise 1.13) and is $\mathcal{D}$-valued. Hence, $\delta^*(X)$ is a valid decision rule.

By Jensen's inequality (see Lemma B.41 in Appendix B),

$$L(\theta, \delta^*(X)) = L(\theta, \mathbb{E}_\theta[\delta(X) \mid T]) \leqslant \mathbb{E}_\theta[L(\theta, \delta(X)) \mid T]$$

as $d \mapsto L(\theta, d)$ is convex. Taking expectations gives $\mathcal{R}(\theta, \delta^*) \leqslant \mathcal{R}(\theta, \delta)$. If $L(\theta, \cdot)$ is strictly convex and δ is not $\sigma(T)$-measurable, then $\delta(X)$ is non-constant conditional on T with positive probability, and Jensen's inequality is strict on that event, giving $\mathcal{R}(\theta, \delta^*) < \mathcal{R}(\theta, \delta)$. □

This theorem is powerful because it gives us a constructive way to improve estimators. If you have an estimator δ and a sufficient statistic T, you should consider $\delta^* = E[\delta \mid T]$. For example, if T is minimal sufficient, this often leads to the "best" possible reduction. If T is complete and sufficient, and we restrict ourselves to unbiased estimators, the Lehman-Scheffé theorem (which we will cover later) tells us δ^* is the unique best unbiased estimator.

For general loss functions, we have a randomized version of the Rao-Blackwell theorem. The theorem says: *we lose nothing by restricting to decision rules based on T alone.*

Theorem 1.44 (Rao-Blackwell, general loss)**.** *Let T be a $(\mathcal{T}, \mathscr{T})$-valued sufficient statistic for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, and let $L : \Theta \times \mathcal{D} \to [0, \infty)$ be a loss function. Consider $(\mathcal{D}, \mathscr{D})$ a standard Borel measurable space. For general loss functions, let $\delta : \mathcal{X} \to \mathcal{D}$ be any decision rule. There exists a randomized decision rule δ^* of the form $\delta^*(X, U) = f(T(X), U)$ for some measurable function $f : \mathcal{T} \times [0, 1] \to \mathcal{D}$ such that*

$$\mathcal{R}(\theta, \delta^*) = \mathcal{R}(\theta, \delta) \quad \textit{for all } \theta \in \Theta.$$

Proof. By sufficiency, the conditional distribution of X given $T(X)$ admits a version not depending on θ: we denote its conditional expectation by $\mathbb{E}[h(X) \mid T]$ for functions $h : \mathcal{X} \to \mathbb{R}$. Since $(\mathcal{D}, \mathscr{D})$ is standard Borel, the conditional distribution of $\delta(X)$ given $T(X) = t$ can be represented via a measurable function (see Theorem B.38 in Appendix B). By sufficiency, this function does not depend on θ: there exists $f : \mathcal{T} \times [0, 1] \to \mathcal{D}$ such that for each $t \in \mathcal{T}$, the random variable $f(t, U)$ with $U \sim \text{Uniform}(0, 1)$ has the same distribution as $\delta(X)$ given $T(X) = t$. Define $\delta^*(x, u) = f(T(x), u)$.

For each $t \in \mathcal{T}$, by construction,

$$\int_0^1 L(\theta, f(t, u))du = \mathbb{E}[L(\theta, \delta(X)) \mid T = t].$$

Therefore,

$$\begin{aligned}\mathcal{R}(\theta, \delta^*) &= \mathbb{E}_\theta\Big[\int_0^1 L(\theta, f(T(X), u))du\Big] \\ &= \mathbb{E}_\theta\Big[\mathbb{E}[L(\theta, \delta(X)) \mid T]\Big] \\ &= \mathbb{E}_\theta[L(\theta, \delta(X))] = \mathcal{R}(\theta, \delta).\end{aligned}$$ □

Together, the Rao-Blackwell theorems can be summarized as follows. For convex losses, deterministic conditioning on a sufficient statistic improves performance. For general losses, a sufficient statistic combined with randomization based on the conditional distribution attains equally good performance. Sufficiency means all decision-relevant information is contained in T.

1.3.2 Comparing decision problems

Now that the key infrastructure of statistical decision theory is in place, we are able to formalize several fundamental questions.

1. **Comparing decision rules:** Given a model $\mathcal{P}$ and loss function L, which decision rule δ has the 'best' risk function $\mathcal{R}(\theta, \delta)$?
2. **Comparing loss functions:** For a given model $\mathcal{P}$, how does the choice of loss function L affect which decision rules are optimal?
3. **Comparing models:** Given models $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and $\mathcal{Q} = \{Q_\theta : \theta \in \Theta\}$, a decision space $(\mathcal{D}, \mathscr{D})$ and loss function $L : \Theta \times \mathcal{D} \to \mathbb{R}$, how does the choice of model $\mathcal{P}$ (or $\mathcal{Q}$) affect inference?

We will do an in-depth study of each of these questions in this course in the order of the list above. To give a flavor, we now preview each of these questions in turn.

Comparing decision rules

Given a statistical model $\mathcal{P}$ and a loss function L, we seek to identify decision rules with small risk. Ideally, we would find a rule δ^* with *uniformly minimum risk*:

$$\mathcal{R}(\theta, \delta^*) \leqslant \mathcal{R}(\theta, \delta) \quad \text{for all } \theta \in \Theta \text{ and all decision rules } \delta.$$

It turns out that uniformly optimal rules rarely exist (see Exercise 1.15). Given that a uniformly optimal rules does not exist, the question "which rule has the best risk function?" is ill-posed. We must refine our criterion for comparing decision rules. In these notes, we will consider two main strategies:

1. **Global risk comparisons.** Compare rules based on summaries of their entire risk function:

   - *Admissibility*: Eliminate rules that are uniformly dominated by another rule.
   - *Bayes risk*: Average the risk over $\theta \in \Theta$ with respect to a prior distribution ('weighing' the risk accross the parameter space).
   - *Minimax risk*: Find the best rule with respect to the worst-case risk; i.e. δ^* such that
   $$\sup_{\theta \in \Theta} \mathcal{R}(\theta, \delta^*) = \inf_{\delta \in \mathcal{C}} \sup_{\theta \in \Theta} \mathcal{R}(\theta, \delta). \tag{1.5}$$
   where $\mathcal{C}$ is the class of all decision rules.

2. **Restricted decision rules.** Impose additional structure or constraints:

   - *Unbiasedness*: Require for example that $\mathbb{E}_{P_\theta}[\delta(X)] = \theta$ for estimation problems.
   - *Invariance*: Require decision rules to respect symmetries in the problem (e.g. translation invariance for location parameters, or invariant to scale: unit of measurement does not matter).
   - *Level constraints*: For testing, consider tests with Type I error rate at most α and find the most powerful test within this class.

These approaches are not mutually exclusive. In some problems, the best unbiased estimator coincides with a Bayes rule, or the minimax rule can be found within the class of invariant procedures. In other cases, we will see that we have to choose between being e.g. minimax or unbiased. There is a deep connection between admissibility and Bayes' risk which we will explore in Chapter 4. In many problems, the best unbiased estimator coincides with a Bayes rule, or the minimax rule can be found within the class of invariant procedures. Understanding these connections is a central goal of this course.

Comparing loss functions

Once we have a deepened understanding of optimal decision rules given statistical model, a decision space and a loss function, we can start to compare different loss functions and see how they affect the optimal decision rules. Different loss functions encode different priorities: squared error loss heavily penalizes large errors and leads to estimators sensitive to outliers; absolute error loss treats all errors more equally and yields more robust estimators; zero-one loss distinguishes only between correct and incorrect decisions, ignoring the magnitude of errors entirely.

We will see that the quality of an inference depends fundamentally on the chosen loss function; what constitutes an optimal decision rule changes as the loss function changes. A decision rule that is optimal for one loss function may not be optimal for another. Understanding the interplay between loss function and model allows us to reflect on the consequences of errors in specific applications. A medical diagnostic test, a financial trading algorithm, and a scientific hypothesis test may all involve the same statistical model but call for different loss functions.

In some cases, it makes sense to study a model for a collection of loss functions. These considerations will not be the main focus of this course, but we will touch upon them here and there.

Comparing models

After we have a deepened understanding of performance 'within' the context of a given statistical model, a decision space and loss function(s), we can revisit the question posed at the end of Section 1.1 (see Example 1.4): which model is the right one? The concept of sufficiency introduced in Section 1.2 allows us to say when models are effectively the *same for all intents and purposes*. However, in many cases of practical interest, we are interested in comparing models that are fastly different, and knowing which is better for a particular task at hand.

Given a parameter space Θ, a decision space $\mathcal{D}$ and a loss function $L : \Theta \times \mathcal{D} \to \mathbb{R}$, consider two models

$$\begin{aligned} \mathcal{P} = \{P_\theta : \theta \in \Theta\} \quad &\text{with each} \quad P_\theta \ \text{ defined on sample space } (\mathcal{X}, \mathscr{X}) \\ \text{and } \mathcal{Q} = \{Q_\theta : \theta \in \Theta\} \quad &\text{with each} \quad Q_\theta \ \text{ defined on sample space } (\mathcal{Y}, \mathscr{Y}). \end{aligned}$$

The parameter space Θ, which represents the phenomenon of interest, is the same for both models. The distributions P_θ and Q_θ could be different. Their sample spaces $(\mathcal{X}, \mathscr{X})$ and $(\mathcal{Y}, \mathscr{Y})$ could be vastly different. When the two models are observationally equivalent, we expect there not to be any difference in terms of inference. But when they are not, how do we decide which model is "better"? Or more generally, how do we quantify how much information is lost by choosing one model over the other?

Le Cam and Yang 1986 gives the following example:

Example 1.45 (Estimating the half-life of Carbon 14)**.** A physicist wants to estimate the half-life of Carbon 14, assuming the lifetime of a C^{14} atom follows an exponential distribution with rate parameter $\theta > 0$. To do so, the physicist considers two possible experimental designs.

In the first setup, the physicist takes a sample of n atoms and observes the number of disintegrations $x \in \mathbb{N}_0$ over a fixed time period of 2 hours. Under this model,

$P_\theta = \text{Poisson}(2n(1 - e^{-2\theta}))$: the distribution of the count in fixed time. This defines the statistical experiment $\mathcal{P} = \{P_\theta : \theta \in (0, \infty)\}$, where P_θ is defined on the sample space of non-negative integers.

In the second setup, the physicist observes the waiting time $y \geqslant 0$ until a fixed number of disintegrations, say $m = 10^6$, occurs. Here, $Q_\theta = \text{Gamma}(m, \theta)$, defining another experiment $\mathcal{Q} = \{Q_\theta : \theta \in (0, \infty)\}$, with Q_θ on the positive real line. Which setup is more informative? Explore this further in Exercise 1.9. ♢

Given a loss function $L : \Theta \times \mathcal{D} \to \mathbb{R}$, we could compare best possible performance of the two models. We could find for example that one model has a strictly better performance in terms of minimax risk:

$$\inf_\delta \sup_{\theta \in \Theta} \mathbb{E}_{P_\theta}[L(\theta, \delta(X))] = \inf_\delta \sup_{\theta \in \Theta} \mathbb{E}_{Q_\theta}[L(\theta, \delta(Y))] + \epsilon$$

for some $\epsilon > 0$, which is means that the model $\mathcal{P}$ is 'ϵ-deficient' for the loss function L compared to the model $\mathcal{Q}$ in terms of its best worst-case performance. We could even go a step further and compare the best worst-case performance of the two models with respect to a large collection of loss functions, to see if the one model is deficient across loss functions compared to the other. In other cases, we might find that two models with different sample spaces and distributions lead to exactly the same best possible performance. Sometimes, we might find that this to be true for all loss functions.

The above notion of deficiency allows us to think about situations where models are *not* observationally equivalent: given that a statistic is not sufficient for a model, how much information is lost? Note that being 'ϵ-deficient' might not mean that the model $\mathcal{P}$ is 'bad'; it might be the better model to work with for practical purposes. Finding its deficiency with respect to another model is a way to quantify how much information is lost when approximating one model by something that is perhaps more tractable, or more affordable in terms of experimental design.

This line of thinking extends to perhaps the most powerful theoretical tool developed in Part II: The ability to compare models asymptotically. Under certain regularity conditions, complicated models can be shown to 'tend asymptotically' —in various precise senses—to much simpler experiments whose performance is well understood. This allows us to reason about performance in complicated models by reasoning about performance in simpler models, enabling meaningful analysis of performance that would otherwise be intractable.

Lastly, another reason to compare models is *misspecification*. We might want to know how robust decision procedure is if in reality, the model $\mathcal{Q}$ is the correct one, but we are using the model $\mathcal{P}$ to make decisions.

1.3.3 ♠ Why (Expected) Loss?

One might reasonably ask: why focus on expected loss rather than, say, the median loss, or some quantile of the loss distribution, or the maximum loss? And why consider loss functions at all?

Suppose that if θ were known, you could provide a preference ordering over possible decisions in $\mathcal{D}$. We write $d_1 \preceq d_2$ if we prefer decision d_1 to decision d_2 (or are indifferent between them) when the true parameter is θ. For instance, in hypothesis testing with $\theta \in \Theta_0$, we would prefer deciding H_0 over deciding H_1. In estimation, we typically prefer decisions closer to the true value of the estimand $g(\theta)$.

These preferences naturally extend to randomized decisions. Consider now a comparison between two randomized decision rules: We write $\delta_1 \preceq \delta_2$ if we prefer (or are indifferent to) the randomized decision rule δ_1 over δ_2.

A1 (Transitivity): If $\delta_1 \preceq \delta_2$ and $\delta_2 \preceq \delta_3$, then $\delta_1 \preceq \delta_3$.

A2 (Independence): If $\delta_1 \preceq \delta_2$, then

$$\lambda\delta_1 + (1-\lambda)\delta_3 \preceq \lambda\delta_2 + (1-\lambda)\delta_3 \quad \text{for all } \lambda \in (0,1],\ \delta_3.$$

A3 (Continuity): If $\delta_1 \prec \delta_2 \prec \delta_3$, then there exist $\lambda_a, \lambda_b \in (0,1)$ such that

$$\lambda_a\delta_1 + (1-\lambda_a)\delta_3 \preceq \delta_2 \preceq \lambda_b\delta_1 + (1-\lambda_b)\delta_3.$$

The first axiom says that if we prefer δ_2 over δ_1 and δ_3 over δ_2, then we should also prefer δ_3 over δ_1. This makes the comparison $\preceq$ a partial order on the space of decision rules. The second axiom says that mixing between decision rules in an irrelevant alternative should not reverse preferences. The third axiom says there is no decision rule that is infinitely preferable to another; every decision rule can be made comparable through appropriate randomization. We will skip the philosophical discussion of why these axioms could be considered 'rational'.

These axioms are enough to guarantee that our preferences over the decision space can be represented by risk in the sense of Definition 1.38: there exists a loss function such that the corresponding risk function captures our preferences.

Theorem 1.46 (Representation Theorem)**.** *If the space of decision rules equipped with the comparison $\preceq$ satisfies axioms A1, A2 and A3, then there exists a measurable function $L : \Theta \times \mathcal{D} \to [-\infty, \infty]$ such that*

$$\delta_1 \preceq \delta_2 \iff \mathcal{R}(\theta, \delta_1) \leqslant \mathcal{R}(\theta, \delta_2).$$

See Ferguson 1967 for a proof. In words, if our preferences over the decision space are rational in this sense, then they can be represented by minimizing expected loss for some loss function L.

Exercises

Exercise 1.1. Consider an experiment in which we observe $Y = \mu + \epsilon$, where $\mu \in \mathbb{R}^d$ is an unknown vector and $\epsilon \sim N_d(0, \sigma^2 I_d)$ is independent noise with unknown $\sigma > 0$.

1. Write down a corresponding statistical model and verify that the set $\Theta := \mathbb{R}^d \times (0, \infty)$ is identifiable under an appropriate parameterization.

2. Suppose we instead believe μ lies on a ray through the origin: $\mu = \alpha v$ for some unknown $\alpha \in \mathbb{R}$ and direction $v \in S^{d-1}$ in the unit sphere. Consider the statistical model $\{N_d(\alpha v, \sigma^2 I_d) : \alpha \in \mathbb{R}, v \in S^{d-1}, \sigma > 0\}$ and the set $\Theta_{\text{ray}} := \mathbb{R} \times S^{d-1} \times (0, \infty)$.

   Is there a parameterization $\vartheta : \mathcal{P}_{\text{ray}} \to \Theta_{\text{ray}}$ such that for every $P \in \mathcal{P}_{\text{ray}}$ with $\vartheta(P) = (\alpha, v, \sigma)$ we have $P = N_d(\alpha v, \sigma^2 I_d)$? If not, give a subset $\Theta'_{\text{ray}} \subseteq \Theta_{\text{ray}}$ for which such a parameterization exists and is identifiable.

Exercise 1.2. In Example 1.2, show that the observable L (whether the first die is larger than the second) is not $\sigma(S)$-measurable. What is the smallest sigma-algebra containing $\sigma(S)$ that makes L measurable?

Exercise 1.3. Consider $\mathcal{X} = \mathbb{R}$, $\mathscr{X} = \mathcal{B}(\mathbb{R})$, and the model

$$\mathcal{P} = \left\{ \tfrac{1}{2}N(\theta_1, 1) + \tfrac{1}{2}N(\theta_2, 1) : (\theta_1, \theta_2) \in \mathbb{R}^2 \right\}.$$

(a) Can we take the inverse of the indexing map $(\theta_1, \theta_2) \mapsto P_{\theta_1,\theta_2}$ (as a well defined map) onto $\mathbb{R}^2$ to obtain a valid parameter space for $\mathcal{P}$?

(b) Show that $\Theta_{\leqslant} := \{(\theta_1, \theta_2) \in \mathbb{R}^2 : \theta_1 \leqslant \theta_2\}$ *is* a valid parameter space under the map $\frac{1}{2}N(\theta_1, 1) + \frac{1}{2}N(\theta_2, 1) \mapsto (\theta_1, \theta_2)$ and that the induced parameterization is identifiable in the sense of Definition 1.5.

Exercise 1.4. Let $x, z \in \mathbb{R}^n$ and consider statistical model $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathcal{P})$ where $\mathcal{P} = \{P_{\alpha,\beta} : \alpha, \beta \in \mathbb{R}\}$ and $P_{\alpha,\beta}$ is the multivariate normal distribution with mean $\alpha x + \beta z$ and variance I_n. Find a sufficient condition on x and z and an explicit parameterization (a map from $\mathcal{P}$ to $\mathbb{R}^2$) that makes the model $\mathcal{P}$ identifiable with $\Theta = \mathbb{R}^2$.

Exercise 1.5. Revisiting Example 1.13, prove that the sum S is not sufficient in Model 1 but is sufficient in Model 2. Specifically:

(a) In Model 1 (nonparametric), show that the conditional distribution of the outcome given $S = 7$ depends on the unknown distribution p.

(b) In Model 2, use the definition of sufficiency (or the Factorization Theorem) to prove that S is sufficient for p_θ.

Exercise 1.6. Verify the claims in Example 1.16.

(a) Show that if $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \sigma^2/n)$.

(b) Show that if $Y \sim N(\mu, \sigma^2/n)$ and $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$ independent of Y, then the variables $X_i = Y + Z_i - \bar{Z}$ are i.i.d. $N(\mu, \sigma^2)$.

Exercise 1.7. Verify the claims in Example 1.23. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(\theta, \theta + 1)$ for $\theta \in \mathbb{R}$.

1. Show that $T = (X_{(1)}, X_{(n)})$ is minimal sufficient.
2. Show that T is not complete.

Exercise 1.8. Let $\{P_\eta : \eta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^k$ be an exponential family with density $p(x \mid \eta) = \exp\{\eta^\top T(x) - A(\eta)\}h(x)$ with respect to a σ-finite measure μ.

1. Show that $P_{\eta_1} = P_{\eta_2}$ if and only if $(\eta_1 - \eta_2)^\top T(x)$ is constant μ-a.e.
2. Conclude that $P_\eta = P_{\eta'} \iff \eta = \eta'$ if and only if there do not exist distinct $\eta_1, \eta_2 \in \Theta$ with $(\eta_1 - \eta_2)^\top T(x)$ constant μ-a.e.

Exercise 1.9. Revisiting Example 1.45, suppose we are interested in estimating the mean lifetime $\tau = 1/\theta$.

(a) In the first setup, let X be the number of disintegrations in time $t = 2$. Show that the maximum likelihood estimator for τ is $\hat{\tau}_1 = \frac{-2}{\log(1 - X/(2n))}$.

(b) In the second setup, let Y be the time until m disintegrations. Show that the maximum likelihood estimator for τ is $\hat{\tau}_2 = Y/m$.

(c) Compare the variances of these two estimators (you may use a heuristic argument, considering what happens for m, n and τ). Which experiment seems more informative if we want to estimate τ, particularly for large τ (long lifetimes)?

Exercise 1.10. The *empirical distribution* of a sample $X_1, \dots, X_n$ is given by $\hat{P}$ satisfying

$$\hat{P}(A) = \sum_i \mathbf{1}(X_i \in A)/n$$

for all measurable sets $A \subseteq \mathcal{X}$. Suppose $\mathcal{X} = \mathbb{R}$.

(a) Show that observing the empirical distribution $\hat{P}$ is observationally equivalent to observing the sample cumulative distribution function

$$\hat{F}(x) = \sum_i \mathbf{1}(X_i \leqslant x)/n.$$

(b) Show that observing the empirical distribution is observationally equivalent to observing the order statistics $(X_{(1)}, \ldots, X_{(n)})$.

Exercise 1.11. Consider the following **definition**: A model $\mathcal{P}_n$ on a product space $(\mathcal{X}^n, \mathcal{A}^n)$ is *exchangeable* if for all $P_n \in \mathcal{P}_n$, sets $A_1 \in \mathcal{A}, \ldots, A_n \in \mathcal{A}$, and permutation π of $\{1, \ldots, n\}$,

$$P_n(A_1 \times \cdots \times A_n) = P_n(A_{\pi_1} \times \cdots \times A_{\pi_n}).$$

Let $\mathcal{P}_n$ be an exchangeable model on $(\mathcal{X}^n, \mathcal{A}^n)$, where $\mathcal{X}$ is a finite set. Prove that the empirical distribution $\hat{P}_n$, defined by $\hat{P}_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_A(X_i)$, is a sufficient statistic.

Exercise 1.12. Consider the statistical model $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathcal{P})$ where $\mathcal{P} = \{P_f : f \in \Theta\}$ and $\Theta \subseteq L_2[0,1]$ (see Definition B.24) is such that $Y = (Y_1, \ldots, Y_n) \sim P_f$ satisfies

$$Y_i = f(i/n) + \epsilon_i, \quad i = 1, \ldots, n,$$

for $f \in \Theta$ and $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} N(0,1)$.

(a) Is the map $\vartheta : \Theta \to \mathcal{P}$, $f \mapsto P_f$ injective?

(b) Consider instead Θ equal to the space $L_2([0,1], \mathcal{B}[0,1], \mathbb{P}_n)$ where the measure $\mathbb{P}_n : \mathcal{B}[0,1] \to [0,1]$ is to be understood as

$$\mathbb{P}_n(A) = \frac{|\{i \in \{1, \ldots, n\} : i/n \in A\}|}{n}.$$

Show that this makes the previous map injective and provide a map from $\mathcal{P}$ to Θ that makes the parameterization identifiable.

Exercise 1.13. Let T be a sufficient statistic for the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Show that if $h(X)$ is any bounded measurable function that does not depend on θ, then the conditional expectation $\mathbb{E}_\theta[h(X) \mid T]$ admits a version that does not depend on θ.

Hint: Use the definition of conditional expectation and the fact that T is sufficient and use the standard machine of measure theory (Appendix B Section B.2.1).

Exercise 1.14 (Deterministic and randomized tests). Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$ with $\theta > 0$ unknown, and consider testing $H_0 : \theta \leqslant \theta_0$ versus $H_1 : \theta > \theta_0$. The decision space is $\mathcal{D} = \{0, 1\}$, where $d = 1$ means "reject H_0". A deterministic test is

$$\delta(x) = \mathbb{1}\{\textstyle\sum_{i=1}^n x_i > c\}$$

for some threshold $c \in \mathbb{N}$.

1. Show that there may be no c such that $P_{\theta_0}(\sum_{i=1}^n X_i > c) = \alpha$ for a given significance level α.
2. Consider the randomized test

$$\delta(x, u) = \begin{cases} 1 & \text{if } \sum_{i=1}^n x_i > c, \\ \mathbb{1}\{u \leqslant \gamma\} & \text{if } \sum_{i=1}^n x_i = c, \\ 0 & \text{if } \sum_{i=1}^n x_i < c. \end{cases}$$

   Show that $c \in \mathbb{N}$ and $\gamma \in [0, 1]$ can be chosen such that $E_{\theta_0}[\delta(X, U)] = \alpha$.

Exercise 1.15 (Nonexistence of uniformly optimal rules). Consider the statistical model corresponding to $X \sim P_\theta$ where $P_{\theta_0} = N(0, 1)$ and $P_{\theta_1} = N(1, 1)$ and let Δ denote the set of all (possibly randomized) decision rules. Define the *risk set*

$$\mathcal{R} = \big\{(R(\theta_0, \delta), R(\theta_1, \delta)) : \delta \in \Delta\big\} \subseteq \mathbb{R}^2.$$

1. Consider estimating θ under squared error loss. Compute the risk pair for:
   (a) the estimator $\delta_0(X) = 0$,
   (b) the estimator $\delta_1(X) = 1$,
   (c) the estimator $\delta_{1/2}(X) = 1/2$.
   (d) the estimator $\delta(X) = X$.

   Which of these decision rules do you prefer? Can you think of a rule that is better than all of them?
2. A decision rule δ^* is *uniformly optimal* if $(R(\theta_0, \delta^*), R(\theta_1, \delta^*))$ is componentwise smaller than or equal to $(R(\theta_0, \delta), R(\theta_1, \delta))$ for all $\delta \in \Delta$. Using the risk pairs (specifically, δ_0 and δ_1), argue that no uniformly optimal rule exists.

Exercise 1.16 (♠). Consider an experiment of flipping a coin infinitely many times.

(a) What should the sample space $\mathcal{X}$ be? What is its cardinality?

(b) What sigma-algebra $\mathscr{X}$ would naturally represent the observable events (e.g., "the n-th flip is heads")?

(c) Explain why it is impossible to define a countably additive probability measure on $(\mathcal{X}, 2^{\mathcal{X}})$ that consistently assigns probabilities to cylinder sets $C_n = \{x \in \mathcal{X} : x_n = 1\}$ as if the coin flips were independent.

Hint: Consider what happens if all singleton sets have probability zero versus positive probability.

Exercise 1.17 (♠ Empirical distribution equivalence). Let $(\mathcal{X}, \mathscr{X}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and consider the i.i.d. model

$$\mathcal{P} = \{P^{\otimes n} : P \in \Theta\}, \qquad \Theta := \mathcal{M}_1(\mathbb{R}),$$

where $\mathcal{M}_1(\mathbb{R})$ denotes the set of all probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Equip $\mathcal{M}_1(\mathbb{R})$ with the Borel σ-algebra generated by the total variation distance (see Definitions 2.2 and B.10 in the appendix).

For $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ define the *empirical measure* $\hat{P}_x \in \mathcal{M}_1(\mathbb{R})$ by

$$\hat{P}_x(A) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \in A\}, \qquad A \in \mathcal{B}(\mathbb{R}).$$

Let $\mathcal{C}$ denote the set of all cumulative distribution functions (c.d.f.'s), and equip $\mathcal{C}$ with the Borel σ-algebra generated by the sup-norm on $\mathcal{C}$, and define the *empirical c.d.f.* $\hat{F}_x \in \mathcal{C}$ by

$$\hat{F}_x(t) := \hat{P}_x((-\infty, t]) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \leqslant t\}, \qquad t \in \mathbb{R}.$$

Lastly, consider the order statistics $T(X) = (X_{(1)}, \dots, X_{(n)})$ as a $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ valued statistic.

(a) Show that $\sigma(\hat{P}) = \sigma(\hat{F}) = \sigma(T)$.

(b) Show that the order statistics are complete.

(c) Conclude that $\hat{P}$ and $\hat{F}$ are complete sufficient statistics for $\mathcal{P}$.

2 Point Estimation

In this chapter, we study the problem of *estimation*: constructing a decision rule that approximates an unknown parameter or functional of the parameter based on observed data. Recall from Chapter 1 that a statistical model is a family of probability distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ indexed by a parameter θ in a parameter space Θ. Often, we wish to estimate either the parameter θ itself or a functional of it, $\theta \mapsto \phi(\theta)$. An *estimator* is a decision rule $\delta : \mathcal{X} \to \phi(\Theta)$ that maps the observed data X to an estimate $\delta(X)$ of the target quantity. This means our decision space is the same as our parameter space (or a function thereof, $\phi(\Theta)$).

The central questions of this chapter are: what makes an estimator good, and how do we construct estimators with desirable properties? And how do we compare estimators? Intuitively, good performance means that when we observe data $X \sim P_\theta$, the estimator $\delta(X)$ is "close" to the true value $\phi(\theta)$. To formalize this, we could equip the target space $\phi(\Theta)$ with a metric distance d and define the loss as a function in terms of this distance, for example $L(\delta(X), \theta) = \mathsf{d}(\delta(X), \phi(\theta))^2$. In this sense, specifying the functional ϕ is part of specifying the loss and the decision space: it fixes what quantity the estimator is judged against, and hence what counts as estimation error. Taking the Borel σ-algebra on $\phi(\Theta)$ induced by d ensures that loss functions and estimators can be defined as measurable functions. In the common setting where $\phi(\Theta) \subseteq \mathbb{R}^k$, the Euclidean metric is a natural choice.

Example 2.1 (Estimation (and prediction) in a linear model)**.** Suppose we observe a random vector Y in $\mathbb{R}^n$ generated from the linear model

$$\mathcal{P} = \{N_d(X\beta, \sigma^2 I_n) : (\beta, \sigma^2) \in \Theta\},$$

with fixed design $X \in \mathbb{R}^{n \times p}$ (with $X^\top X$ invertible) and parameter space $\Theta = \mathbb{R}^p \times (0, \infty)$. If we wish to estimate $\phi(\theta) = \beta$, an estimator is any decision rule $\delta : \mathbb{R}^n \to \mathbb{R}^p$, e.g.

$$\delta(Y) := (X^\top X)^{-1} X^\top Y.$$

A sensible loss function is the squared Euclidean distance:

$$L((\beta, \sigma^2), \delta) = \|\delta - \beta\|^2.$$

If instead we wish to predict the mean response at a known new covariate $x_{\text{new}} \in \mathbb{R}^p$, the target is the functional $\phi(\theta) = x_{\text{new}}^\top \beta \in \mathbb{R}$. That means that our decision space is

$\mathbb{R}$, and as a loss function, we could consider

$$L((\beta, \sigma^2), \delta) = |\delta - x_{\text{new}}^\top \beta|.$$

$\Diamond$

However, estimation problems sometimes involve more abstract target spaces. For example, the target of estimation could be itself the probability distribution generating the data — i.e. when $\phi(\theta) = P_\theta$. In this case, the loss function could be a metric on the space of probability measures. A possible choice for such a metric is the total variation distance.

Definition 2.2. The *total variation distance* between two probability measures P and Q on a measurable space $(\mathcal{X}, \mathscr{X})$ is defined as

$$\mathsf{d}_{TV}(P, Q) = \sup_{A \in \mathscr{X}} |P(A) - Q(A)|.$$

The total variation metric allows us to study the parameter space where Θ is (a subset of) the space of probability measures, equipped with the Borel sigma-algebra of the total variation metric. Given a statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, we could take ϕ to be the map $\theta \mapsto P_\theta$ and the loss function to be the total variation distance:

$$L(\delta(X), \theta) = \mathsf{d}_{TV}(\delta(X), P_\theta).$$

For some models, various losses can be related to each other in a simple way. In other cases, they do not. For example, when $P_\theta \mapsto \phi(\theta)$ does not identify the model, we cannot generally hope that estimating the functional $\phi(\theta)$ allows for an estimate of P_θ. The example below illustrates that in estimation problems, we are often not trying to estimate the entire distribution P_θ, but rather a lower-dimensional summary such as the mean, variance, or quantile, depending on what we are interested in. Only in special cases, these lower-dimensional summaries translate back to the data generating process.

Example 2.3 (Estimating a functional vs. the distribution)**.** We revisit Example 1.4 from Chapter 1. Consider two statistical models on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$:

(i) $\mathcal{P} = \{P_\theta := N(\theta, 1) : \theta \in \mathbb{R}\}$.

(ii) $\mathcal{Q} = \{\text{all probability measures on } (\mathbb{R}, \mathcal{B}(\mathbb{R})) \text{ with variance at most } 1\}$.

For the model $\mathcal{P}$, it can be shown (see Exercise 2.18) that the total variation

distance is bounded by the distance between the means for $\theta, \theta' \in \mathbb{R}$:

$$\mathsf{d}_{TV}(P_\theta, P_{\theta'}) \leqslant \frac{1}{2}|\theta - \theta'|.$$

This implies that estimating the parameter θ well (in Euclidean distance) automatically ensures that we estimate the distribution P_θ well (in total variation distance).

For the model $\mathcal{Q}$, the mean parameter $\phi : Q \mapsto \int x dQ(x)$ does not identify the distribution in the sense of Definition 1.5. Estimating the mean $\phi(Q) = \int x dQ(x)$ is still almost equally 'doable' as in the case of normals (see Exercise **??**). However, estimating the distribution $Q \in \mathcal{Q}$ itself turns out to be much more difficult: it is impossible to find a 'good' estimate of the distribution in total variation distance uniformly over $\mathcal{Q}$, even under repeated sampling (see Exercise 2.17). That is, for the model $\mathcal{Q}$, estimating the distribution $Q \in \mathcal{Q}$ itself is a very different estimation problem compared to estimating $\int x dQ(x)$. ◊

This example also motivates a useful (informal) taxonomy of estimation problems. The labels *parametric*, *semiparametric*, and *nonparametric* are best thought of as describing the *complexity of the model in relation to the estimand*: the same model can lead to different types of problems depending on whether we aim to estimate a low-dimensional functional (like a mean) or a high/infinite-dimensional object (like an entire distribution).

- **Parametric estimation:** the model is indexed by a *finite-dimensional* parameter, typically $\Theta \subseteq \mathbb{R}^d$ with fixed d, and the data-generating distribution is fully determined (up to θ). In Example 2.3(i), $\mathcal{P} = \{N(\theta, 1) : \theta \in \mathbb{R}\}$ is a one-dimensional parametric model. In such settings, estimating θ is often closely related to estimating P_θ itself because the parameter identifies the distribution (and here even controls it in total variation).

- **Semiparametric estimation:** the model is *infinite-dimensional*, but the target $\phi(\theta)$ is *finite-dimensional* (typically in $\mathbb{R}^k$ for fixed k). The remaining aspects of the distribution act as an infinite-dimensional *nuisance*. In Example 2.3(ii), if the goal is only to estimate the mean functional $g(Q) = \int x\, dQ(x) \in \mathbb{R}$, then we are in this regime: many different $Q \in \mathcal{Q}$ share the same mean, yet the target itself is one-dimensional.

- **Nonparametric estimation:** the model is *infinite-dimensional* (e.g. a large class of distributions, densities, regression functions, etc.), and the target is typically itself an *infinite-dimensional object* such as the distribution P, its CDF, or its density. In Example 2.3(ii), if the goal is to estimate $Q \in \mathcal{Q}$ (as a distribution), then this is a nonparametric estimation problem.

The example above illustrates that the labels *parametric*, *semiparametric*, and *nonparametric* are best understood as describing an *estimation problem*—in particular, the target $g(\theta)$ and the loss function—and not only the "size" or complexity of the model class. Moreover, these labels are only *loose distinctions*: different statistics books (and different subfields) use them in slightly different and sometimes inconsistent ways.

There is also a fourth, even less sharply defined regime that we will encounter throughout the chapter: *high-dimensional* estimation problems, where the parameter is technically finite-dimensional (e.g. $\Theta \subseteq \mathbb{R}^d$), but the dimension d is large relative to the other relevant aspects of the problem (such as the sample size n) in a way that drastically changes which decision rules are reasonable. In this sense, high-dimensional problems often behave more like nonparametric problems than classical parametric ones, despite having a finite-dimensional parameter space. We return to this theme in Section 2.3.

Returning to the central question of this chapter: given a statistical model $\mathcal{P}$, a target quantity $\phi(\theta)$ and a loss function L, what is a good estimator for $\phi(\theta)$? For typical loss functions (e.g. if L is related to some distance metric) and a fixed θ, we might observe that if we knew P_θ, the "best estimator" would simply be the constant function $\delta(x) = \phi(\theta)$ for all x. This estimator is measurable and achieves zero loss if θ is the parameter underlying the data-generating process. However, since inference of θ is the whole point, this is not a sensible estimator. The challenge of formulating what is 'a good estimator' is to construct a data-dependent rule that 'performs well across the parameter space'. There are multiple criteria for measuring what 'performance across the parameter space' means, which we explore throughout this chapter.

2.1 Unbiasedness

Throughout this section, we consider a statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and a target quantity $\phi : \Theta \to \mathbb{R}^k$. We will study estimators of $\phi(\theta)$ satisfying the following property.

Definition 2.4 (Unbiasedness)**.** An estimator $\delta : \mathcal{X} \to \mathbb{R}^k$ is an *unbiased estimator* of $\phi(\theta)$ if for all $\theta \in \Theta$,

$$\mathbb{E}_\theta[\delta(X)] = \phi(\theta).$$

Unbiasedness is an appealing property: on average, the estimator produces the correct value. If we, or others, were to repeat the experiment many times, the average of the estimates would converge to the true parameter value.

Example 2.5 (Averaging i.i.d. estimators)**.** Suppose we have m independent replications of a study, yielding estimators $\delta_1, \ldots, \delta_m$ of a parameter $\theta \in \mathbb{R}$. Assume these

are independent and identically distributed with unit variance (i.e., $\mathrm{Var}(\delta_j) = 1$). A natural estimator is the average:

$$\delta(X) = \frac{1}{m}\sum_{j=1}^{m} \delta_j.$$

If the individual estimators are unbiased, then $\delta(X)$ converges to θ over repeat replications. However, if there is systematic bias ($\mathbb{E}[\delta_j] \neq \theta$), the convergence fails. See Exercise 2.5. ♢

However, it does not guarantee that any single estimate is close to the true parameter. A large variance implies that the estimator fluctuates significantly, making individual estimates unreliable. By minimizing variance, we maximize the probability that the estimator is close to the target $\phi(\theta)$.

Definition 2.6. Let X and Y be random vectors in $\mathbb{R}^k$ and $\mathbb{R}^m$, respectively, with means $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. The *covariance matrix* of X and Y is the $k \times m$ matrix defined by

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^\top].$$

The *variance matrix* (or simply variance) of a random vector $X \in \mathbb{R}^k$ is the $k \times k$ covariance matrix of X with itself:

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \mathbb{E}[(X - \mu_X)(X - \mu_X)^\top].$$

For the Euclidean metric, it turns out that minimizing the variance across all unbiased estimators is equivalent to minimizing the expected squared error of the estimator.

Lemma 2.7 (Bias-Variance Decomposition)**.** *Let $\delta(X)$ be an estimator of $\phi(\theta)$ with finite second moments. Then,*

$$\mathbb{E}_\theta\|\delta(X) - \phi(\theta)\|^2 = \|\mathbb{E}_\theta[\delta(X)] - \phi(\theta)\|^2 + \mathrm{Trace}(\mathrm{Var}_\theta(\delta(X))). \tag{2.1}$$

Proof. See Exercise 2.4. □

The first term in (2.1) is called the (squared) bias of the estimator, and the second term is the 'variance term'. The lemma states that the expected squared error of any estimator is the sum of the squared bias and the variance. For unbiased estimators, (the trace of) the variance effectively measures the average squared Euclidean distance from the true parameter value. This is one of the reasons to compare unbiased estimators based on their variance. This leads us to the concept of a UMVUE.

Definition 2.8. An estimator δ is a *uniformly minimum variance unbiased estimator (UMVUE)* of $\phi(\theta)$ if it is unbiased, i.e., $\mathbb{E}_\theta[\delta(X)] = \phi(\theta)$ for all $\theta \in \Theta$, and if for any other unbiased estimator δ',

$$\operatorname{Var}_\theta(\delta(X)) \leqslant \operatorname{Var}_\theta(\delta'(X)) \quad \text{for all } \theta \in \Theta.$$

For a formal definition of the matrix ordering $\leqslant$ (the Loewner order), see Definition C.10 in Appendix C. The notation $\operatorname{Var}_\theta$ denotes the variance operator with respect to the expectation operator $\mathbb{E}_\theta$.

Finding a UMVUE is a challenging problem in general. However, for certain models, those where complete sufficient statistics exist, the UMVUE can be found using the following theorem.

Theorem 2.9 (Lehmann-Scheffé)**.** *Let T be a complete sufficient statistic for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. If δ_0 is any unbiased estimator of $\phi(\theta)$ with finite variance, then $\delta(X) = \mathbb{E}[\delta_0(X) \mid T(X)]$ is the a.s. unique UMVUE of $\phi(\theta)$.*

Proof. By sufficiency, $\delta(X) = \mathbb{E}[\delta_0(X) \mid T]$ admits a version not depending on θ, so it is a valid estimator. By the tower property, $\mathbb{E}_\theta[\delta(X)] = \mathbb{E}_\theta[\delta_0(X)] = \phi(\theta)$, so δ is unbiased. For any $v \in \mathbb{R}^k$, the loss function $L_v(\delta, \theta) = (v^\top(\delta - \phi(\theta)))^2$ is convex in δ. By Rao-Blackwell (Theorem 1.43),

$$\mathbb{E}_\theta[(v^\top(\delta(X) - \phi(\theta)))^2] \leqslant \mathbb{E}_\theta[(v^\top(\delta_0(X) - \phi(\theta)))^2].$$

Since both estimators are unbiased, this gives

$$v^\top \operatorname{Var}_\theta(\delta)\, v \leqslant v^\top \operatorname{Var}_\theta(\delta_0)\, v \quad \text{for all } v \in \mathbb{R}^k,$$

i.e., $\operatorname{Var}_\theta(\delta) \leqslant \operatorname{Var}_\theta(\delta_0)$ in the positive semidefinite ordering.

Now suppose δ' is any other unbiased estimator of $\phi(\theta)$. Define $\psi(T) = \mathbb{E}[\delta'(X) \mid T] - \delta(X)$. Then

$$\mathbb{E}_\theta[\psi(T)] = \phi(\theta) - \phi(\theta) = 0 \quad \text{for all } \theta \in \Theta.$$

By completeness, $\psi(T) = 0$ almost surely, so $\mathbb{E}[\delta'(X) \mid T] = \delta(X)$ almost surely. By Rao-Blackwell, $\operatorname{Var}_\theta(\delta) \leqslant \operatorname{Var}_\theta(\delta')$. Since δ' was arbitrary, δ is UMVUE. □

The Lehmann-Scheffé theorem provides a strategy for finding a unique, best unbiased estimator:

1. Start with an arbitrary unbiased estimator δ_0;
2. Find a complete sufficient statistic T;

3. Apply the Rao-Blackwell theorem to obtain the UMVUE $\delta(X) = \mathbb{E}[\delta_0(X) \mid T]$. This is sometimes called "Rao-Blackwellization".

We illustrate the use of the Lehmann-Scheffé theorem with two examples below: estimating the CDF of a distribution at a fixed point in a parametric setting and in a semiparametric setting.

Example 2.10 (Normal mean with known variance)**.** Consider the model $\mathcal{P} = \{N(\theta, \sigma^2)^{\otimes n} : \theta \in \mathbb{R}\}$, corresponding to observing $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\theta, \sigma^2) =: P_\theta$, where $\sigma^2 > 0$ is known. The sample mean $\bar{X}$ is a complete sufficient statistic for θ (see Example 1.16 combined with Proposition 1.29). Suppose we want to estimate the CDF at a point $t \in \mathbb{R}$, i.e., $\phi(\theta) = P_\theta((-\infty, t])$. Consider the unbiased estimator $\delta_0(X) = \mathbb{1}_{\{X_1 \leqslant c\}}$. Since $\bar{X}$ is complete sufficient, the UMVUE is given by $\delta(X) = \mathbb{E}[\delta_0(X) \mid \bar{X}]$ by the Lehmann-Scheffé theorem.

We can compute the UMVUE explicitly (Exercise 2.2):

$$\delta(X) = \Phi\left(\sqrt{\frac{n}{n-1}}\frac{t - \bar{X}}{\sigma}\right).$$

$\lozenge$

Example 2.11 (Estimating the cumulative distribution function)**.** Consider observing $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$, where P is any probability distribution on $\mathbb{R}$. The corresponding statistical experiment is $(\mathbb{R}^n, \mathcal{B}(\mathbb{R})^n, \mathcal{P}, \mathcal{P})$ where

$$\mathcal{P} = \{P : P \text{ is a probability distribution on } \mathbb{R}\}.$$

Given $P \in \mathcal{P}$, let F_P be the cumulative distribution function of P: $F_P(t) = P((-\infty, t])$. To estimate $\phi(P) = F_P(t)$ at a fixed point $t \in \mathbb{R}$, a natural estimator is the empirical distribution function:

$$\delta(X) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leqslant t\}}.$$

For a fixed t, the random variable $Y_i = \mathbb{1}_{\{X_i \leqslant t\}}$ is Bernoulli distributed with parameter $p = P(X_i \leqslant t) = F_P(t)$. Thus,

$$\mathbb{E}_P[\delta(X)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_P[Y_i] = \frac{1}{n} \cdot nF_P(t) = F_P(t),$$

showing that $\delta(X)$ is an unbiased estimator for $F_P(t)$. Its variance is given by

$$\operatorname{Var}_P(\delta(X)) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}_P(Y_i) = \frac{1}{n}F_P(t)(1 - F_P(t)).$$

By Exercise 1.17, the order statistics $T(X) = (X_{(1)}, \ldots, X_{(n)})$ are a complete sufficient statistic for this model. Furthermore,

$$\mathbb{E}_P\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i \leqslant t\}} \mid T(X)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i \leqslant t\}}.$$

Hence, $\delta(X)$ is the UMVUE for $\phi(P) = F_P(t)$ by the Lehmann-Scheffé theorem. $\lozenge$

Returning briefly to our earlier discussion concerning parametric vs semiparametric estimation problems, a further investigation the above examples reveal an important phenomenon: in both problems, it can be shown that the accuracy of the estimator is of the order $1/\sqrt{n}$: the rate as a function of the sample size at which we can expect the estimator to be accurate is the same. However, the (in both cases optimal!) variances differ between the two models (see Exercise 2.2). This is expected: the parametric model is more informative and allows us to estimate the parameter with more precision. In the semiparametric model, we are paying a price for the flexibility of the model; its infinite dimensional nature.

2.1.1 The Cramér-Rao lower bound

In some statistical models, a differentiable relationship between Θ and $\mathcal{P}$ allows us to derive fundamental limits on estimation accuracy for smooth functionals of the parameter. The key idea is that if the distribution P_θ changes smoothly with θ, we can quantify how much information the data carries about the parameter.

Consider a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ dominated by a measure μ with densities $p_\theta = dP_\theta/d\mu$, where Θ is an open subset of $\mathbb{R}^d$. If the map $\theta \mapsto p_\theta(x)$ is differentiable for each x, we can define the *score function*

$$S_\theta(x) = \nabla_\theta \log p_\theta(x).$$

Under regularity conditions that permit interchanging differentiation and integration, the expected score is zero: $\mathbb{E}_\theta[S_\theta(X)] = 0$. The *Fisher information matrix* is then defined as the covariance of the score:

$$I(\theta) = \mathbb{E}_\theta[S_\theta(X)S_\theta(X)^\top],$$

which (under regularity conditions) can equivalently be computed as

$$I(\theta) = -\mathbb{E}_\theta[\nabla_\theta^2 \log p_\theta(X)].$$

The Cramér-Rao lower bound states that the variance of any unbiased estimator of $g(\theta)$

is at least $\nabla g(\theta)^\top I(\theta)^{-1} \nabla g(\theta)$. This result is fundamental: it shows that estimation precision is governed by the Fisher information, which quantifies how sensitively the distribution responds to changes in θ.

The classical approach requires verifying regularity conditions for each model—conditions that ensure differentiation under the integral sign is valid. It turns out that a weaker notion of differentiability "on average" suffices and leads to a cleaner and much more general theory. This is the concept of *differentiability in quadratic mean*.

Definition 2.12 (Differentiability in Quadratic Mean)**.** A statistical model $\{P_\theta : \theta \in \Theta\}$ with densities p_θ is *differentiable in quadratic mean (DQM)* at θ if there exists a measurable function $S_\theta : \mathcal{X} \to \mathbb{R}^d$ such that

$$\int \left(\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2} h^\top S_\theta(x) \sqrt{p_\theta(x)} \right)^2 d\mu(x) = o(\|h\|^2) \quad \text{as } h \to 0.$$

DQM implies that the Fisher information matrix $I(\theta) = \mathbb{E}_\theta[S_\theta(X) S_\theta(X)^\top]$ exists and roughly speaking allows for exchange of differentiation and integration. It effectively replaces the "standard" regularity conditions that we might be familiar with from e.g. undergraduate textbooks on statistics.

Lemma 2.13. *Let the model $\{P_\theta : \theta \in \Theta\}$ be differentiable in quadratic mean at θ with score S_θ. Then:*

(i) The Fisher information $I(\theta) = E_\theta[S_\theta S_\theta^\top]$ is well-defined with all entries finite.

(ii) If $T : \mathcal{X} \to \mathbb{R}$ is a measurable function with T^2 uniformly integrable under $\mathbb{E}_{\theta'}$ for all θ' in a neighborhood of θ, then $\psi(\theta') = E_{\theta'}[T]$ is differentiable at θ with

$$\nabla \psi(\theta) = E_\theta[T \cdot S_\theta].$$

Proof. Exercise 2.8. □

The DQM condition allows us to differentiate expectations of statistics, which is key to establishing a fundamental limit on the variance of any unbiased estimator. This limit is determined by the Fisher information, quantifying the intuition that estimation is harder when the distribution P_θ changes slowly with θ (low information). This leads us to the famous Cramér-Rao Lower Bound.

Theorem 2.14 (Cramér-Rao Lower Bound – Biased Case)**.** *Consider a model $\{P_\theta : \theta \in \Theta\}$ that is DQM at $\theta \in \Theta$ with positive definite Fisher information matrix $I(\theta)$. Let $\delta(X)$ be an $\mathbb{R}^k$-valued estimator that is uniformly square-integrable under $P_{\theta'}$ for all θ' in a neighborhood of θ and write $\psi(\theta) = \mathbb{E}_\theta[\delta(X)]$ for the expectation of the estimator.*

It holds that

$$\mathrm{Var}_\theta(\delta(X)) \geqslant \nabla\psi(\theta)^\top I(\theta)^{-1}\nabla\psi(\theta).$$

Proof. By definition, we have $\psi(\theta) = \mathbb{E}_\theta[\delta(X)] = \int \delta(x)p_\theta(x)\,d\mu(x)$. Using Lemma 2.13, we have

$$\nabla\psi(\theta) = \mathbb{E}_\theta[\delta(X)S_\theta(X)],$$

and

$$\mathbb{E}_\theta[S_\theta(X)] = \mathbb{E}_\theta[1 \cdot S_\theta(X)] = \nabla 1 = 0.$$

Combining these, we obtain

$$\nabla\psi(\theta) = \mathrm{Cov}_\theta(\delta(X), S_\theta(X)).$$

Now, for any constant vector $a \in \mathbb{R}^d$, consider the scalar random variable $Z = a^\top S_\theta(X)$. The covariance between $\delta(X)$ and Z is

$$\mathrm{Cov}_\theta(\delta(X), Z) = \mathrm{Cov}_\theta(\delta(X), a^\top S_\theta(X)) = a^\top \mathrm{Cov}_\theta(\delta(X), S_\theta(X)) = a^\top\nabla\psi(\theta).$$

Applying the Cauchy-Schwarz inequality to the covariance squared, we have

$$(\mathrm{Cov}_\theta(\delta(X), Z))^2 \leqslant \mathrm{Var}_\theta(\delta(X))\,\mathrm{Var}_\theta(Z).$$

Substituting the expressions for covariance and variance, noting that $\mathrm{Var}_\theta(Z) = \mathrm{Var}_\theta(a^\top S_\theta(X)) = a^\top I(\theta)a$, we get

$$(a^\top\nabla\psi(\theta))^2 \leqslant \mathrm{Var}_\theta(\delta(X))(a^\top I(\theta)a).$$

This inequality holds for any vector a. To obtain the tightest bound, we choose $a = I(\theta)^{-1}\nabla\psi(\theta)$. With this choice:

$$a^\top\nabla\psi(\theta) = \nabla\psi(\theta)^\top I(\theta)^{-1}\nabla\psi(\theta)$$

and

$$a^\top I(\theta)a = \nabla\psi(\theta)^\top I(\theta)^{-1}I(\theta)I(\theta)^{-1}\nabla\psi(\theta) = \nabla\psi(\theta)^\top I(\theta)^{-1}\nabla\psi(\theta).$$

The inequality becomes

$$(\nabla\psi(\theta)^\top I(\theta)^{-1}\nabla\psi(\theta))^2 \leqslant \mathrm{Var}_\theta(\delta(X))(\nabla\psi(\theta)^\top I(\theta)^{-1}\nabla\psi(\theta)).$$

Assuming $\nabla\psi(\theta)^\top I(\theta)^{-1}\nabla\psi(\theta) > 0$ (otherwise the bound is trivial), we can multiply

by the inverse on both sides to obtain

$$\mathrm{Var}_\theta(\delta(X)) \succcurlyeq \nabla\psi(\theta)^\top I(\theta)^{-1}\nabla\psi(\theta).$$

□

The bound in Theorem 2.14 applies to any estimator, regardless of whether it is biased or unbiased. For unbiased estimators, the bound simplifies and takes a particularly interpretable form.

Corollary 2.15. *Assume the setting of Theorem 2.14. If $\delta(X)$ is an unbiased estimator of $\phi(\theta)$ in a neighborhood of θ and $\phi : \Theta \to \mathbb{R}^d$ is differentiable, then*

$$\mathrm{Var}_\theta(\delta(X)) \succcurlyeq \nabla\phi(\theta)^\top I(\theta)^{-1}\nabla\phi(\theta).$$

If ϕ is the identity function, then this reduces to the familiar inequality:

$$\mathrm{Var}_\theta(\delta(X)) \succcurlyeq I(\theta)^{-1}. \tag{2.2}$$

Proof. Note that unbiasedness implies $\psi(\theta) = \phi(\theta)$. If ϕ is the identity function on $\mathbb{R}^d$, it follows that $\nabla\phi(\theta) = I_d$. □

The following example shows that the requirement of unbiasedness in a neighborhood of θ_1 is critical for the result of Corollary 2.15 to hold.

Example 2.16. Consider $X \sim N(\theta, I_d)$, $\theta \in \mathbb{R}^d$ and the estimator

$$\delta(X) = \omega\theta_1 + (1-\omega)X$$

for some $\omega \in [0,1]$ and $\theta_1 \in \mathbb{R}^d$. Estimator is unbiased at θ_1. However, its variance is

$$\mathrm{Var}_{\theta_1}(\delta(X)) = (1-\omega)^2 I_d.$$

For $\omega > 0$, this is strictly smaller than the right-hand side of (2.2), which evaluates to $I(\theta_1)^{-1} = I_d$ (check). Indeed, for $\omega > 0$, the estimator is not unbiased over any neighborhood of θ_1. Setting $\omega = 0$ gives the UMVUE. ♢

If $\delta(X)$ is unbiased, the inequalities of Corollary 2.15 hold for all $\theta \in \Theta$. For unbiased estimators, the Cramér-Rao lower bound provides a target in terms of what is the best possible variance to achieve. Clearly, if $\delta(X)$ is unbiased and attains the Cramer-Rao lower bound, it is a UMVUE. The converse is not true in general: in certain problems, the UMVUE might not attain the Cramér-Rao lower bound.

However, in certain problems, attaining the Cramér-Rao lower bound is not only a sufficient condition for the estimator to be UMVUE, but also a necessary condition. The implication goes really far: it also tells us the form that our decision rule should have, given that it is unbiased and attains the Cramér-Rao lower bound. This form is affine function of the score. This insight will prove to be useful later when we study asymptotic properties of maximum likelihood estimators in Part II of the course.

Proposition 2.17 (Attainment of the Cramér-Rao bound)**.** *Under the conditions of the Cramér-Rao theorem, equality holds if and only if*

$$\delta(X) = \psi(\theta) + \nabla\psi(\theta)^\top I(\theta)^{-1} S_\theta(X) \quad P_\theta\text{-a.s.}$$

In particular, the bound is attained if and only if $\delta(X)$ is an affine function of the score.

Proof. Fix θ and $v \in \mathbb{R}^k$. Consider the scalar estimator $\delta_v(X) = v^\top \delta(X)$ with mean $\psi_v(\theta) = v^\top \psi(\theta)$. Applying the (scalar) Cramér–Rao inequality to δ_v gives

$$\mathrm{Var}_\theta(\delta_v(X)) \geqslant \nabla\psi_v(\theta)^\top I(\theta)^{-1} \nabla\psi_v(\theta) = v^\top \Big(\nabla\psi(\theta)^\top I(\theta)^{-1} \nabla\psi(\theta) \Big) v.$$

Since this holds for all v, it is equivalent to the stated matrix inequality.

Moreover, in the scalar proof the inequality comes from Cauchy–Schwarz applied to $\mathrm{Cov}_\theta(\delta_v(X), a^\top S_\theta(X))$ with the choice $a = I(\theta)^{-1}\nabla\psi_v(\theta)$. Equality in Cauchy–Schwarz holds if and only if

$$\delta_v(X) - \psi_v(\theta) = a^\top S_\theta(X) \qquad P_\theta\text{-a.s.}$$

With $a = I(\theta)^{-1}\nabla\psi(\theta)\, v$, this becomes

$$v^\top(\delta(X) - \psi(\theta)) = v^\top \nabla\psi(\theta)^\top I(\theta)^{-1} S_\theta(X) \qquad P_\theta\text{-a.s.}$$

for every $v \in \mathbb{R}^k$, which implies the vector identity in the statement. The converse direction is immediate by substitution. □

Let us consider what this result implies. For an unbiased estimator with $\psi(\theta) = \phi(\theta)$, Proposition 2.17 tells us that attaining the bound requires

$$\delta(X) = \phi(\theta) + \nabla\phi(\theta)^\top I(\theta)^{-1} S_\theta(X) \quad P_\theta\text{-a.s.}$$

At first glance, this seems problematic: the right-hand side depends on the unknown parameter θ through $\phi(\theta)$, $\nabla\phi(\theta)$, $I(\theta)$, and $S_\theta(X)$. For the estimator to be a valid statistic—a function of the data alone—these θ-dependent terms must combine in a way that eliminates the dependence on θ. This places strong constraints on the model:

only in special cases does such cancellation occur. Exponential families provide the canonical example.

Example 2.18 (Some exponential families attain the bound naturally)**.** Consider a natural exponential family with density

$$p_\theta(x) = h(x) \exp\left(\theta^\top T(x) - A(\theta)\right),$$

where $\theta \in \Theta \subseteq \mathbb{R}^d$ is the natural parameter and $A(\theta)$ is twice differentiable. The score is

$$S_\theta(X) = \nabla_\theta \log p_\theta(X) = T(X) - \nabla A(\theta).$$

Since $E_\theta[S_\theta(X)] = 0$, we have $E_\theta[T(X)] = \nabla A(\theta)$. The Fisher information is

$$I(\theta) = \mathrm{Cov}_\theta(S_\theta(X)) = \mathrm{Cov}_\theta(T(X)) = \nabla^2 A(\theta).$$

Now consider estimating $\phi(\theta) = \nabla A(\theta) = E_\theta[T(X)]$ by the estimator $\delta(X) = T(X)$. This estimator is unbiased, and satisfies the attainment condition:

$$\delta(X) - \psi(\theta) = T(X) - \nabla A(\theta) = S_\theta(X) = I(\theta)^{-1} \nabla\psi(\theta)^\top S_\theta(X),$$

where the last equality uses $\nabla\psi(\theta) = \nabla^2 A(\theta) = I(\theta)$. Thus, the sufficient statistic $T(X)$ achieves the Cramér-Rao lower bound for estimating its own expectation. $\lozenge$

2.2 Invariance

In Section 2.1.1, we saw that models with a differentiable structure allow us to derive fundamental limits on estimation accuracy through the Fisher information. Another type of structure that proves useful is that of *symmetry*: if transforming the data in a certain way corresponds to a transformation of the parameter that leaves the model's form unchanged, we say the model is invariant under that transformation.

When the loss function respects the same symmetry, it is often natural to restrict attention to decision rules that also respect it. The word "natural" here has both technical and conceptual interpretations. On the technical side, invariance can simplify the analysis: as we will see, equivariant estimators in invariant problems have constant risk, reducing the comparison of decision rules to a single number. On the conceptual side, invariance captures the intuition that our estimates should not depend on arbitrary choices such as the coordinate system (rotation invariance) or unit of measurement (scale invariance).

This section formalizes these ideas and illustrates them in classical location, location-

scale, and covariance models. There is much more to say on this topic than fits into this section. The interested reader is referred to Chapter 3 of Lehmann and Romano 2005, Chapter 6 of Lehmann and Casella 2006 and Berger 2013.

2.2.1 Invariant models

The idea of invariance, or that of symmetries generally, is closely related to the concept of a group.

Definition 2.19. A *group* is a set G with an operation $\cdot : G \times G \to G$ such that

1. $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in G$ (associativity);
2. there exists $e \in G$ with $e \cdot a = a \cdot e = a$ for all $a \in G$ (identity);
3. for each $a \in G$ there exists $a^{-1} \in G$ with $a \cdot a^{-1} = a^{-1} \cdot a = e$ (inverse).

Groups are common objects in mathematics, and many of the most important groups in statistics are related to groups in mathematics that we are already familiar with.

Example 2.20.

- $(\mathbb{Z}, + : (a, b) \mapsto a + b)$ is a group with identity 0 and inverse $-a$.
- $(\mathbb{R}_{>0}, \times : (a, b) \mapsto ab)$ is a group with identity 1 and inverse $1/a$.
- $\mathrm{GL}(p)$ (invertible $p \times p$ matrices) is a group under matrix multiplication.
- The permutation group $\mathfrak{S}_n$ acts on $\mathcal{X}^n$ by permuting coordinates.

$\lozenge$

Definition 2.21 (Group action)**.** A group G *acts on* a set A if there is a map $G \times A \to A$, written $(g, x) \mapsto gx$, such that $ex = x$ and $g(hx) = (gh)x$ for all $g, h \in G$ and $x \in A$. The action is *transitive* if for all $x, y \in A$, there exists $g \in G$ such that $gx = y$.

In statistical models, we sometimes have actions on the sample space $\mathcal{X}$ and the parameter space Θ. Sometimes, actions on the parameter space have an 'inverse' action on the sample space that 'respects' the data generating process: whether we act on the data, or perform the same corresponding action on the parameter, the data generating process remains the same.

Definition 2.22 (Equivariance of a model)**.** Consider statistical experiment with $\mathcal{P} = \{P_\theta : \theta \in \mathbb{R}\}$ defined on a sample space $(\mathcal{X}, \mathscr{X})$. Let G act on $\mathcal{X}$ and Θ, such that its action is measurable. The model $\mathcal{P}$ is *equivariant* under G if

$$P_{g\theta}(A) = P_\theta(g^{-1}A) \qquad \text{for all } g \in G,\ \theta \in \Theta,\ A \in \mathscr{X}.$$

Equivalently: if $X \sim P_\theta$, then $gX \sim P_{g\theta}$.

One of the most important examples of an equivariant model is the *location family*.

Example 2.23 (Location family)**.** Consider a statistical experiment $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mathcal{P}, \Theta)$ with parameter space $\Theta = \mathbb{R}^d$ where $\mathcal{P}$ is dominated with respect to the Lebesgue measure, with Lebesgue density $p(x|\theta)$ a.e. equal to $f(x-\theta)$ for some measurable function $f : \mathbb{R}^d \to [0, \infty)$.

The model $\mathcal{P}$ is equivariant under the group $G = (\mathbb{R}^d, +)$ acting on $\mathbb{R}^d$ by translation, i.e. $g_c x = x + c$ for all $c \in \mathbb{R}^d$ and $x \in \mathbb{R}^d$.

Indeed, if $X \sim P_\theta$, then $X + c \sim P_{\theta+c}$, since

$$P_{\theta+c}(A) = \int_A f(x - (\theta + c))dx = \int_{x-c\in A} f(x-\theta)dx = P_\theta(A - c).$$

where we used the change of variables $y = x + c$ and translation invariance of Lebesgue measure. ◊

Many common distributions are location families, such as the normal distribution, Laplace distribution, Cauchy distribution, etc. Another important example of an equivariant model is the *scale family*.

Example 2.24 (Scale family)**.** Consider a statistical experiment $(\mathbb{R}_{>0}, \mathcal{B}(\mathbb{R}_{>0}), \mathcal{P}, \Theta)$ with parameter space $\Theta = \mathbb{R}_{>0}$, where $\mathcal{P}$ is dominated with respect to Lebesgue measure on $\mathbb{R}_{>0}$, with density

$$p(x \mid \sigma) = \frac{1}{\sigma} f\Big(\frac{x}{\sigma}\Big)$$

for some measurable function $f : \mathbb{R}_{>0} \to [0, \infty)$ integrating to one.

The model $\mathcal{P}$ is equivariant under the multiplicative group $G = (\mathbb{R}_{>0}, \times)$ acting on $\mathbb{R}_{>0}$ by scaling: $g_c x = cx$ for $c > 0$. Indeed, if $X \sim P_\sigma$, then $cX \sim P_{c\sigma}$, since

$$P_{c\sigma}(A) = \int_A \frac{1}{c\sigma} f\Big(\frac{x}{c\sigma}\Big)\, dx = \int_{c^{-1}A} \frac{1}{\sigma} f\Big(\frac{y}{\sigma}\Big)\, dy = P_\sigma(c^{-1}A).$$

Common examples include the exponential family $\mathrm{Exp}(1/\sigma)$ and the chi-squared distribution scaled by σ^2. ◊

Another canonical example of an equivariant model is the multivariate standard normal distribution; which is spherically symmetric, on top of being a location and scale family.

Example 2.25 (Spherically symmetric normal)**.** Consider observing $X \sim N_d(\theta, \sigma^2 I_d)$ with $\theta \in \mathbb{R}^d$ and $\sigma^2 > 0$ known. The group G of $d \times d$ orthonormal matrices under matrix multiplication acts on both $\mathbb{R}^d$ and $\Theta = \mathbb{R}^d$ by matrix-vector multiplication: $g_Q x = Qx$.

The model is equivariant under this action. If $X \sim N_d(\theta, \sigma^2 I_d)$, then

$$QX \sim N_d(Q\theta, \sigma^2 Q I_d Q^\top) = N_d(Q\theta, \sigma^2 I_d),$$

using $QQ^\top = I_d$. Thus $QX \sim P_{Q\theta}$, as required.

$\Diamond$

2.2.2 Invariance and estimation

For models that are equivariant under a group action, it is natural to consider decision rules that respect the same symmetry. The intuition is straightforward: if the statistical problem is unchanged by a transformation, the solution should transform accordingly. Put differently, if two scientists analyze the same data but use different coordinate systems—one rotated relative to the other, or one using meters while the other uses feet—their estimates should be related by the same transformation. An estimator that violates this principle would give answers that depend on arbitrary choices having nothing to do with the data.

Suppose we are interested in estimating $\theta \in \Theta$, taking $\mathcal{D} = \Theta$ (with some suitable σ-algebra). The action of G on Θ *induces* an action on $\mathcal{D}$ by $g_{\mathcal{D}}(d) = gd$ for all $g \in G$ and $d \in \mathcal{D}$. In some applications it would be unnatural for the loss function to depend on the orientation in the parameter space for which the estimation error occurs. For example, for a GPS system, the loss of predicting a certain location should not depend on one's initial orientation relative to the true location. This brings us to the concept of invariant loss.

That is, if we decide d based on data X and the true state turns out to θ (for which we incur loss $L(\theta, d)$), this loss should be the same as someone who decides $g_{\mathcal{D}}(d)$ based on data gX and the true state turning out to be $g\theta$ (for which they incur loss $L(g\theta, g_{\mathcal{D}}(d))$). This motivates the following definition.

Definition 2.26 (Invariant loss)**.** Consider a decision problem $(\mathcal{X}, \mathscr{X}, \mathcal{P}, \Theta, (\mathcal{D}, \mathscr{D}), L)$. Suppose G acts on $\mathcal{X}$, Θ, and $\mathcal{D}$, and these actions are measurable. Write $g_{\mathcal{D}}$ for the induced action on $\mathcal{D}$.

A loss function $L(\theta, d)$ is *invariant* under G if

$$L(g\theta, g_{\mathcal{D}}\, d) = L(\theta, d) \qquad \text{for all } g \in G,\ \theta \in \Theta,\ d \in \mathcal{D}.$$

A decision problem is *invariant* under G if the model and loss function are invariant under G.

Example 2.27 (Revisiting the normal location model)**.** Recall the equivariant normal location model from Example 2.25: $X \sim N_d(\theta, \sigma^2 I_d)$ with $\theta \in \mathbb{R}^d$ and $\sigma^2 > 0$ known,

with the group G of $d \times d$ orthonormal matrices under matrix multiplication acts on both $\mathbb{R}^d$ and $\Theta = \mathbb{R}^d$ by matrix-vector multiplication: $g_Q x = Qx$.

If we are interested in estimating θ, we can consider the loss function $L(\theta, d) = \|\theta - d\|^2$, defined on $\mathbb{R}^d \times \mathbb{R}^d$. This loss is invariant under G since $\|Q\theta - Qd\|^2 = \|\theta - d\|^2$, turning the corresponding decision problem into an invariant one. ♢

For an invariant decision problem, it can be natural to restrict attention to estimators that respect the same symmetry. If the data X lead us to the estimate $\delta(X)$, then the transformed data gX should lead us to the correspondingly transformed estimate $g_{\mathcal{D}}\delta(X)$. This motivates the following definition.

Definition 2.28 (Equivariant decision rule)**.** Consider a decision problem with decision space $(\mathcal{D}, \mathscr{D})$. Suppose G acts on $\mathcal{X}$, Θ, and $\mathcal{D}$, and these actions are measurable. Write $g_{\mathcal{D}}$ for the induced action on $\mathcal{D}$. A decision rule $\delta : \mathcal{X} \to \mathcal{D}$ is *equivariant* if

$$\delta(gx) = g_{\mathcal{D}}\, \delta(x) \qquad \text{for all } g \in G,\ x \in \mathcal{X}.$$

Equivariant estimators in invariant problems have constant risk, which greatly simplifies the task of finding optimal procedures.

Theorem 2.29. *Assume G acts transitively on Θ (i.e. for all $\theta, \theta' \in \Theta$ there exists $g \in G$ with $\theta' = g\theta$). If the model $\mathcal{P}$ is equivariant, the loss is invariant, and δ is equivariant, then the risk $\mathcal{R}(\theta, \delta)$ is constant in θ.*

Proof. Fix $\theta_0 \in \Theta$. For any $\theta = g\theta_0$, using model equivariance and then equivariance and invariance,

$$\begin{aligned}
\mathcal{R}(\theta, \delta) &= \mathbb{E}_{g\theta_0}[L(g\theta_0, \delta(X))] \\
&= \mathbb{E}_{\theta_0}[L(g\theta_0, \delta(gX))] \\
&= \mathbb{E}_{\theta_0}[L(g\theta_0, \tilde{g}\, \delta(X))] \\
&= \mathbb{E}_{\theta_0}[L(\theta_0, \delta(X))] = \mathcal{R}(\theta_0, \delta).
\end{aligned}$$

□

If we are convinced that equivariant decision rules are the natural ones to consider in an invariant problem, then the goal becomes finding the best among them. This is formalized by the uniformly minimum risk equivariant estimator (UMREE), which plays a role analogous to the UMVUE in the class of unbiased estimators. Theorem 2.29 shows that every equivariant estimator has constant risk, so comparing two equivariant estimators reduces to comparing a single number rather than two functions on Θ.

Definition 2.30 (Uniformly Minimum Risk Equivariant Estimator)**.** Consider an invariant decision problem: a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ that is equivariant under G, and

a loss function L that is invariant under G. An estimator δ^* is a *uniformly minimum risk equivariant estimator (UMREE)* if:

(i) δ^* is equivariant: $\delta^*(gx) = g_{\mathcal{D}}\,\delta^*(x)$ for all $g \in G$ and $x \in \mathcal{X}$, and

(ii) for any other equivariant estimator δ,

$$\mathcal{R}(\theta, \delta^*) \leqslant \mathcal{R}(\theta, \delta) \qquad \text{for all } \theta \in \Theta.$$

We now apply the general theory to one of the most important invariant problems: estimation in a location family under squared error loss. This setting illustrates how the UMREE can be characterized explicitly as a Bayesian posterior mean under an improper prior.

Consider observing $X = (X_1, \dots, X_n)$ with joint density $\prod_{i=1}^n f(x_i - \theta)$ with respect to Lebesgue measure, where $\theta \in \mathbb{R}^d$ and $x_i \in \mathbb{R}^d$. The translation group $G = (\mathbb{R}^d, +)$ acts on the sample space $\mathbb{R}^{nd}$ by $g_c x = (x_1 + c, \dots, x_n + c)$ and on the parameter space by $g_c\theta = \theta + c$. Since Lebesgue measure is translation-invariant, the model is equivariant: if $X \sim P_\theta$, then $X + c\mathbf{1} \sim P_{\theta+c}$.

For squared error loss $L(\theta, d) = \|d - \theta\|^2$, the induced action on decisions is $g_c d = d + c$, and the loss is invariant since $\|(d+c) - (\theta+c)\|^2 = \|d - \theta\|^2$. An estimator δ is equivariant if and only if $\delta(x + c\mathbf{1}) = \delta(x) + c$ for all $c \in \mathbb{R}^d$. This is a substantial restriction: for instance, the constant estimator $\delta(x) = 0$ is not equivariant, while the sample mean and geometric median are.

Since G acts transitively on $\Theta = \mathbb{R}^d$, Theorem 2.29 implies that every equivariant estimator has constant risk. The UMREE is therefore the equivariant estimator minimizing $\mathcal{R}(\theta_0, \delta)$ for any fixed θ_0; taking $\theta_0 = 0$ is conventional. The following theorem identifies this optimal estimator.

Theorem 2.31 (Pitman estimator in $\mathbb{R}^d$)**.** *Let $X = (X_1, \dots, X_n)$ with $X_i \in \mathbb{R}^d$ have joint density $\prod_{i=1}^n f(x_i - \theta)$ with respect to Lebesgue measure on $\mathbb{R}^{nd}$, where $\theta \in \mathbb{R}^d$. Under squared error loss $L(\theta, \delta) = \|\delta - \theta\|^2$, the ($P_\theta$-a.s. unique) UMREE is*

$$\delta^*(x) = \frac{\displaystyle\int_{\mathbb{R}^d} \theta \prod_{i=1}^n f(x_i - \theta)\, d\theta}{\displaystyle\int_{\mathbb{R}^d} \prod_{i=1}^n f(x_i - \theta)\, d\theta},$$

provided the integrals are finite.

Proof. The translation group $G = (\mathbb{R}^d, +)$ acts on $\mathbb{R}^{nd}$ by $g_c x = (x_1 + c, \dots, x_n + c)$ and on $\Theta = \mathbb{R}^d$ by $g_c\theta = \theta + c$. Since Lebesgue measure on $\mathbb{R}^d$ is translation-invariant, the model is equivariant. The squared error loss is invariant since $\|\delta + c - (\theta + c)\| = \|\delta - \theta\|$.

By Theorem 2.29, every translation-equivariant estimator has constant risk, so it suffices to minimize $\mathcal{R}(0,\delta) = \mathbb{E}_0[\|\delta(X)\|^2]$ over equivariant δ.

First, δ^* is equivariant: substituting $\eta = \theta - c$,

$$\delta^*(x_1 + c, \dots, x_n + c) = \frac{\int_{\mathbb{R}^d} (\eta + c) \prod_i f(x_i - \eta)\, d\eta}{\int_{\mathbb{R}^d} \prod_i f(x_i - \eta)\, d\eta} = \delta^*(x) + c.$$

Let δ be any other equivariant estimator and write $h = \delta - \delta^*$. Then h is translation-invariant. We claim $\mathbb{E}_0[\langle \delta^*(X), h(X) \rangle] = 0$.

By definition, $\delta^*(x)$ minimizes $\int_{\mathbb{R}^d} \|\theta - d\|^2 \prod_i f(x_i - \theta)\, d\theta$ over $d \in \mathbb{R}^d$. The first-order condition gives

$$\int_{\mathbb{R}^d} (\delta^*(x) - \theta) \prod_i f(x_i - \theta)\, d\theta = 0.$$

Taking the inner product with $h(x)$, integrating over x under P_0, and applying Fubini's theorem with translation invariance of h yields $\mathbb{E}_0[\langle \delta^*(X), h(X) \rangle] = 0$.

Finally,

$$\begin{aligned} \mathbb{E}_0[\|\delta(X)\|^2] &= \mathbb{E}_0[\|\delta^*(X) + h(X)\|^2] \\ &= \mathbb{E}_0[\|\delta^*(X)\|^2] + 2\mathbb{E}_0[\langle \delta^*(X), h(X) \rangle] + \mathbb{E}_0[\|h(X)\|^2] \\ &\geqslant \mathbb{E}_0[\|\delta^*(X)\|^2], \end{aligned}$$

with equality if and only if $h = 0$ a.s., implying δ^* is P_θ-a.s. unique. □

The Pitman estimator admits an elegant Bayesian interpretation: it is the 'posterior mean' under the 'uniform prior' $\pi(\theta) \propto 1$ on $\mathbb{R}^d$. Although this prior is improper (it does not integrate to a finite value), the posterior is proper whenever the likelihood is integrable, and the resulting estimator is well-defined. The uniform prior is the *right Haar measure* for the translation group—the unique (up to scale) measure on $\mathbb{R}^d$ that is invariant under the group action – explored in more generality in Section 2.2.3.

We now illustrate the Pitman estimator in two classical location families.

Example 2.32 (Normal location)**.** For $X_i \overset{\text{iid}}{\sim} N_d(\theta, \sigma^2 I_d)$ with σ^2 known, the joint density is proportional to $\exp(-\frac{1}{2\sigma^2} \sum_i \|x_i - \theta\|^2)$. Completing the square in θ, the Pitman estimator evaluates to $\delta^*(X) = \bar{X}$, which coincides with both the MLE and the UMVUE. ◊

Example 2.33 (Uniform location)**.** For $X_i \overset{\text{iid}}{\sim} \text{Uniform}(\theta, \theta + 1)$, the joint density is $\prod_i \mathbb{1}_{\{\theta \leqslant x_i \leqslant \theta + 1\}} = \mathbb{1}_{\{X_{(n)} - 1 \leqslant \theta \leqslant X_{(1)}\}}$. This is constant (equal to 1) on the interval $[X_{(n)} - 1, X_{(1)}]$ and zero elsewhere. The Pitman estimator is therefore the midpoint of this interval:

$$\delta^*(X) = \frac{X_{(1)} + X_{(n)} - 1}{2}.$$

This differs from the MLE, which is any point in $[X_{(n)} - 1, X_{(1)}]$ (conventionally taken as $\hat{\theta} = X_{(n)} - 1$). The Pitman estimator uses information from both extremes, while the MLE uses only one. ♢

These examples highlight that the Pitman estimator may or may not coincide with other familiar estimators, depending on the model. For the Cauchy location family, the Pitman estimator takes a more complex form; see Exercise 2.11.

Remark 2.34. The concept of UMREE differs from the concept of UMVUE in that the latter is defined in the context of unbiased estimators and their variance, whilst the UMREE is defined in the context of equivariant estimators and *a specific loss function*. For different loss functions, we obtain different UMREE's.

2.2.3 ♠ Haar measures and the general UMREE construction

The Pitman estimator for location families relied on the fact that Lebesgue measure is translation-invariant. This observation generalizes: for any locally compact group, there exists a canonical "invariant measure" called the Haar measure, which allows us to construct best equivariant estimators via the same 'Bayesian recipe'.

Definition 2.35 (Haar measure)**.** Let G be a locally compact topological group. A *left Haar measure* on G is a nonzero Borel measure ν_L satisfying

$$\nu_L(gA) = \nu_L(A) \quad \text{for all } g \in G \text{ and all Borel sets } A \subseteq G.$$

A *right Haar measure* ν_R satisfies $\nu_R(Ag) = \nu_R(A)$ for all $g \in G$ and Borel $A \subseteq G$.

Equivalently, in terms of integrals: ν_L is left-invariant if

$$\int_G f(gh)\, d\nu_L(h) = \int_G f(h)\, d\nu_L(h) \qquad \text{for all } g \in G \text{ and integrable } f.$$

and similarly for right invariance.

Theorem 2.36 (Haar, 1933)**.** *Let G be a locally compact topological group. Then:*

(i) A right Haar measure exists.

(ii) Any two right Haar measures differ by a positive multiplicative constant.

The analogous statements hold for left Haar measures.

The proof of existence is nontrivial and relies on techniques from functional analysis; see Folland 2016 for a complete treatment. For our purposes, the key point is that Haar measures exist and are essentially unique, so they provide a canonical choice of "uniform" measure on any group.

Example 2.37 (Common Haar measures)**.** (i) **Translation group** $G = (\mathbb{R}^d, +)$: Lebesgue measure $d\theta$ is Haar (both left and right as the group is abelian).

(ii) **Multiplicative group** $G = (\mathbb{R}_{>0}, \times)$: The measure $d\nu(\sigma) = d\sigma/\sigma$ is both left and right Haar.

(iii) **Location-scale group** $G = \{(a, b) : a > 0, b \in \mathbb{R}\}$ with operation $(a_1, b_1) \cdot (a_2, b_2) = (a_1 a_2, a_1 b_2 + b_1)$: The left Haar measure is $da\, db/a^2$, and the right Haar measure is $da\, db/a$.

(iv) **Orthogonal group** $G = \mathrm{O}(d)$: Since $\mathrm{O}(d)$ is compact, the Haar measure is finite and can be normalized to a probability measure (the "uniform distribution" on $\mathrm{O}(d)$).

(v) **General linear group** $G = \mathrm{GL}(p)$: The left and right Haar measures are $d\nu(A) = |\det A|^{-p}\, dA$, where dA denotes Lebesgue measure on $\mathbb{R}^{p\times p}$.

$\Diamond$

A group G is called *unimodular* if its left and right Haar measures coincide. Compact groups and abelian groups are all unimodular. The location-scale group in Example 2.37(iii) is a standard example of a non-unimodular group.

When a group G acts transitively on a parameter space Θ, a Haar measure on G induces a natural "uniform" measure on Θ by *pushing it forward* through the orbit map. Concretely, fix a reference point $\theta_0 \in \Theta$ and define $\tau : G \to \Theta$ by $\tau(g) = g\theta_0$; then the induced measure on Θ is the push-forward $\tau_{\#}\nu$ given by $\tau_{\#}\nu(A) = \nu(\tau^{-1}(A))$ for measurable $A \subseteq \Theta$ (and we often denote $\tau_{\#}\nu$ simply by ν). This measure is invariant under the group action, and different choices of θ_0 change it only by a multiplicative constant.

We now present the general recipe for constructing best equivariant estimators using Haar measures.

Theorem 2.38 (UMREE via Haar measure)**.** *Consider an invariant decision problem with model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ equivariant under a locally compact group G that acts transitively on Θ. Let $L(\theta, d)$ be an invariant loss function, and let ν denote the right Haar measure on G, and consider the induced measure (also denoted ν) on Θ via the group action.*

Define the generalized Bayes estimator

$$\delta^*(x) = \mathrm{argmin}_{d\in\mathcal{D}} \int_\Theta L(\theta, d)\, p(x|\theta)\, d\nu(\theta),$$

provided the integral is finite. Then δ^* is equivariant, and if it exists, it is the UMREE.

Proof. See Berger 2013, Chapter 6. □

Given an invariant decision problem:

1. **Identify the group** G under which the model is equivariant and the loss is invariant.
2. **Verify transitivity:** Check that G acts transitively on Θ.
3. **Compute the right Haar measure** ν on G (or equivalently, on Θ via the action).
4. **Form the 'generalized Bayes' estimator** using ν as an improper prior:

$$\delta^*(x) = \operatorname{argmin}_{d\in\mathcal{D}} \int_\Theta L(\theta, d)\, p(x|\theta)\, d\nu(\theta).$$

If the latter integral is finite, δ^* is the UMREE.

2.3 Admissibility

The idea behind admissibility is simple: we should not use a decision rule if another rule is strictly better. The idea behind admissibility is simple: we wish to only consider decision rules that are not strictly dominated by some other decision rule.

Definition 2.39. A decision rule δ is *admissible* if there exists no other estimator δ' such that

1. $\mathcal{R}(\theta, \delta') \leqslant \mathcal{R}(\theta, \delta)$ for all $\theta \in \Theta$, and
2. $\mathcal{R}(\theta, \delta') < \mathcal{R}(\theta, \delta)$ for at least one $\theta \in \Theta$.

If such a δ' exists, we say that δ is *inadmissible* and that δ' *dominates* δ.

Admissibility captures a minimal requirement: we should reject any decision rule that is strictly dominated by another. Considering admissibility as a criterion leads to some surprising facts and insights. Perhaps one of the most impactful insights is the so-called *Stein's shrinkage phenomenon*, which has shaped the way we think about estimation in high-dimensional models.

2.3.1 Stein's shrinkage phenomenon

Consider the normal means model where we observe $X \sim N_d(\theta, \sigma^2 I_d)$, for some $d \in \mathbb{N}$ and $\sigma > 0$, and the aim is to estimate the mean θ. In the case $d = 1$ it seems rather clear that if we do not know anything about the parameter θ, we can not do much

better than estimating it by the observation X. Proving this rigorously is actually not completely trivial, see Exercise 2.12.

For larger d it is in fact also not immediately clear whether if we assume no further structure on θ, we can do better than simply using the maximum likelihood estimator $\delta_{\mathrm{MLE}}(X) = X$. Clearly, X is a sufficient statistic and moreover a complete one (Example 1.31). It is unbiased over $\mathbb{R}^d$, and hence it is the UMVUE. Furthermore, it has invariance properties both in terms of location-shifts and rotations; it is the UMREE for the normal location model with Euclidian loss. It turns out, however, that it is possible to perform strictly better, in the sense of expected quadratic error.

To get a first indication of this fact, note that for any estimator δ with a finite covariance we have the bias-variance decomposition (Lemma 2.7)

$$\mathbb{E}_\theta \|\delta(X) - \theta\|^2 = \|\mathbb{E}_\theta \delta(X) - \theta\|^2 + \operatorname{Tr} \operatorname{Cov}_\theta \delta(X).$$

If we apply this to $\delta_c(X) = cX$ we find that $E_\theta \|\delta_c(X) - \theta\|^2 = (c-1)^2 \|\theta\|^2 + c^2 \sigma^2 d$, which, for given θ, is minimal for c equal to

$$c_\theta = \frac{\|\theta\|^2}{\|\theta\|^2 + \sigma^2 d},$$

and the minimal value is

$$\mathbb{E}_\theta \|\delta_{c_\theta}(X) - \theta\|^2 = \frac{\sigma^2 d \|\theta\|^2}{\|\theta\|^2 + \sigma^2 d} = \frac{\|\theta\|^2}{\|\theta\|^2 + \sigma^2 d} \mathbb{E}_\theta \|\delta_{\mathrm{MLE}}(X) - \theta\|^2.$$

Since $c_\theta < 1$, this indicates that it might be advantageous to shrink the estimator X towards 0, that is, to multiply it by a factor strictly smaller than 1. Since c_θ depends on the unknown parameter θ, one might argue that this is not a sensible estimator. However, it turns out that for $d \geqslant 3$, we can shrink by an appropriate data-dependent constant that leads to an estimator with an expected squared error that is *strictly smaller than that of the MLE for all* $\theta \in \mathbb{R}^d$.

Theorem 2.40 (James-Stein)**.** *Define*

$$\delta_{JS}(X) = \left(1 - \frac{\sigma^2 (d-2)}{\|X\|^2}\right) X.$$

For $d \geqslant 3$*, we have* $\mathbb{E}_\theta \|\delta_{JS}(X) - \theta\|^2 < \mathbb{E}_\theta \|\delta_{MLE}(X) - \theta\|^2$ *for all* $\theta \in \mathbb{R}^d$*.*

Proof. For the bias and variance of the ith component of the JS estimator we have

$$\mathbb{E}_\theta \delta_{\mathrm{JS},i}(X) - \theta_i = \sigma^2 (d-2) \mathbb{E}_\theta \frac{X_i}{\|X\|^2}$$

and

$$\mathrm{Var}_\theta \delta_{\mathrm{JS},i}(X) = \sigma^2 + \sigma^4(d-2)^2 \mathrm{Var}_\theta \frac{X_i}{\|X\|^2} - 2\sigma^2(d-2)\left(\mathbb{E}_\theta \frac{X_i^2}{\|X\|^2} - \mathbb{E}_\theta \frac{\theta_i X_i}{\|X\|^2}\right),$$

respectively. (Note that since $\mathbb{E}_\theta 1/\|X\|^p$ is finite if and only if $d > p$, all expectations here are finite for $d \geqslant 3$. See Exercise 2.13.) It follows that the mean squared error of the estimator is given by

$$\sigma^2 d + \sigma^4(d-2)^2 \mathbb{E}_\theta \frac{1}{\|X\|^2} - 2\sigma^2(d-2)\left(\sum_i \mathbb{E}_\theta \frac{X_i(X_i-\theta_i)}{\|X\|^2}\right)$$

(check!). By Lemma 2.41 below,

$$\mathbb{E}_\theta \frac{X_i(X_i-\theta_i)}{\|X\|^2} = \mathbb{E}_\theta \frac{\sigma^2}{\|X\|^2} - 2\mathbb{E}_\theta \frac{\sigma^2 X_i^2}{\|X\|^4}.$$

Hence, the mean squared error (MSE) $\mathbb{E}_\theta \|\delta_{\mathrm{JS}}(X) - \theta\|^2$ equals

$$\sigma^2 d - \sigma^4(d-2)^2 \mathbb{E}_\theta \frac{1}{\|X\|^2}.$$

Since the MSE of the MLE $\delta_{\mathrm{MLE}}(X) = X$ equals $d\sigma^2$, this completes the proof. □

The key tool used in the proof above is Stein's lemma, which provides a useful identity for expectations involving Gaussian random variables.

Lemma 2.41. *Let $X \sim N_d(\theta, I_d)$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be an absolutely continuous (in each coordinate a.e.) function such that $\mathbb{E}_\theta |(\partial f/\partial x_i)(X)| < \infty$ for $i = 1, \ldots, d$. Then for $i = 1, \ldots, d$,*

$$\mathbb{E}_\theta (X_i - \theta_i) f(X) = \mathbb{E}_\theta \frac{\partial f}{\partial x_i}(X).$$

Proof. Integration by parts, see Exercise 2.14. □

The James-Stein theorem gives a number of very interesting insights in statistics for 'high-dimensional' models. It shows that by shrinking the MLE towards zero, thereby reducing the variance at the cost of increasing the bias, we obtain an estimator with a strictly better risk $\mathbb{E}_\theta \|\delta(X) - \theta\|^2$. Moreover, although the observed X_i are independent by assumption, the shrinkage factor depends on all the observations. Hence, to estimate the ith component θ_i, we do not only use the information in X_i, but we also *borrow strength* from the other observations, even though they are independent coordinate wise.

One argument that Stein (1956) used to intuitively justify the concept of shrinkage is the observation that if $X \sim N_d(\theta, I_d)$, then by the law of large numbers it holds for

large d that $\|X\|^2 \approx \|\theta\|^2 + d$. So the norm of the MLE X is typically substantially larger than the norm of the parameter θ it is supposed to estimate. Therefore, it may be beneficial to shrink the vector X so that the norm is reduced.

Alternatively, we may argue that shrinking reduces the contributions of outliers, i.e. relatively large observations X_i, on the squared estimation error. This possibly comes at the cost of increasing the error made in the other coordinates, but the net effect is that shrinking improves the total squared error $\|\delta_{\mathrm{JS}}(X) - \theta\|^2$ of the estimator on average. Observe that this reasoning indicates that it is essential that we assess the quality of the estimator using a norm that simultaneously takes all coordinates of θ into account. This allows us to trade off gains in one coordinate with losses in others.

The James-Stein theorem can be generalized in many directions, for instance away from the normal distribution with unit variance, using other norms, other statistical models, et cetera. The precise form of the shrinking is not crucial either. Shrinking towards a fixed point $v \in \mathbb{R}^d$ other than 0 works just as well for instance (see Exercise 2.15). The general message is always that in high-dimensional settings it is typically advantageous to somehow reduce the variance by shrinking, or otherwise regularizing. We explore this further in the next section.

Theorem 2.40 shows that for $d \geqslant 3$, the MLE $\delta_{\mathrm{MLE}}(X) = X$ is inadmissible in the model $X \sim N_d(\theta, I_d)$, with respect to the squared Euclidean risk. By definition, this means that there exists another estimator δ such that $\mathbb{E}_\theta\|\delta(X) - \theta\|^2 \leqslant \mathbb{E}_\theta\|\delta_{\mathrm{MLE}}(X) - \theta\|^2$ for all $\theta \in \mathbb{R}^d$, with strict inequality for at least one $\theta \in \mathbb{R}^d$. The theorem asserts that the James-Stein estimator is such an estimator. It can be shown however that the James-Stein estimator itself is inadmissible as well. For example the positive part Stein estimator

$$\delta_{\mathrm{JS}^+}(X) = \left(1 - \frac{d-2}{\|X\|^2}\right)_+ X$$

is an estimator with strictly smaller risk for all $\theta \in \mathbb{R}^d$. See for instance Section 3.4 of Tsybakov (2009). Unfortunately, δ_{JS^+} is not admissible either. It turns out that finding an admissible estimator is easy if we take a Bayesian approach – both in terms of its construction and in terms of verifying its admissibility – we will discuss this in Chapter 4.

2.3.2 Bias-variance trade-off

The Stein-shrinkage phenomenon demonstrates that in high-dimensional settings, trading bias for variance can yield strict improvements over the best unbiased estimator. This raises a natural question: how far can we push this trade-off? Can we achieve arbitrarily good performance at a particular parameter value by accepting bias elsewhere?

In Example 2.16, we saw an estimator that achieves variance strictly below the

Cramér-Rao bound at a specific point θ_1 by being unbiased only at that point rather than in a neighborhood. Taken to the extreme, the 'guesstimator' $\delta(X) = \theta_1$ is admissible—it achieves a risk at θ_1 which no other estimator can beat. Of course, this estimator performs terribly elsewhere in the parameter space. Intuitively, there is a 'no-free-lunch' principle at play: exceptional performance at one parameter value must come at the cost of degraded performance elsewhere.

The following example demonstrates this phenomenon concretely: a pretest estimator that achieves dramatically reduced risk at $\theta = 0$ suffers substantially inflated risk at nearby parameter values.

Example 2.42 (Test first, then estimate)**.** Let $X \sim N(\theta, \sigma^2)$ and consider squared error loss. The MLE $\delta_{\mathrm{MLE}}(X) = X$ has constant risk $R(\theta, \delta_{\mathrm{MLE}}) = \sigma^2$.

Fix $t > 0$ and define the pretest (hard-threshold) estimator

$$\delta_t(X) = \begin{cases} 0, & |X| \leqslant t, \\ X, & |X| > t. \end{cases}$$

At $\theta = 0$, writing $Z \sim N(0,1)$ and taking $\sigma = 1$ for simplicity,

$$R(0, \delta_t) = \mathbb{E}[Z^2 \mathbb{1}\{|Z| > t\}] = 2\big(t\varphi(t) + \Phi(-t)\big),$$

so for instance $t = 3$ gives $R(0, \delta_3) \approx 0.029$. We know from the fact that the MLE is admissible in this setting (Exercise 2.12) that there must be some θ such that $R(\theta, \delta_3) > R(\theta, X)$. This is indeed the case: $R(2, \delta_3) \approx 3.766 > 1$. For $|\theta| \to \infty$, the $R(\theta, \delta_3)$ approaches that of $R(\theta, X)$. The cost for performance at $\theta = 0$ is paid for by a worse performance 'nearby' $\theta = 0$. ♢

Example 2.42 suggests that dramatic gains at one parameter value force losses nearby. Can we quantify this trade-off? The Cramér-Rao bound provides one such tool, but it requires differentiability of the model and is most informative for unbiased estimators. For biased estimators, or for models lacking smooth structure, we need a more general approach.

Understanding this cost is not merely of theoretical interest. Later in this chapter, we will encounter models where trading bias for variance is not optional but *necessary*—unbiased estimators may not exist, or may perform poorly. To navigate such settings, we need tools to characterize the fundamental limitations on estimation.

The *constraint risk inequality* offers exactly this. The idea is simple: if two distributions P_f and P_g are 'similar', yet the parameters f and g are far apart, then no estimator can perform well at both. An estimator that gets close to f under P_f will tend to be far from g under P_g, and vice versa.

To make this precise, we need to quantify two notions of distance: distance between parameters and similarity between distributions. For parameters, we use a (semi-)metric d on Θ. For distributions, we use the *Bhattacharyya coefficient*

$$\rho(P_f, P_g) = \int \sqrt{p_f\, p_g}\, d\mu,$$

which measures the overlap between two densities. Geometrically, $\rho(P_f, P_g)$ is the cosine of the angle between $\sqrt{p_f}$ and $\sqrt{p_g}$ viewed as unit vectors in $L^2(\mu)$. This geometric viewpoint on the space of densities has (implicitly) already appeared in our discussion of differentiability in quadratic mean.

Lemma 2.43 (Constraint Risk Inequality)**.** *Let (Θ, d) be a (semi-)metric space and let P_f, P_g be probability measures on $(\mathcal{X}, \mathscr{X})$ dominated by a common measure μ, with densities p_f and p_g. For any estimator $\delta : \mathcal{X} \to \Theta$ and any $f, g \in \Theta$,*

$$\sqrt{\mathbb{E}_f \mathsf{d}(\delta, f)^2} + \sqrt{\mathbb{E}_g \mathsf{d}(\delta, g)^2} \geqslant \mathsf{d}(f, g) \cdot \rho(P_f, P_g).$$

Proof. By the triangle inequality, for all $x \in \mathcal{X}$,

$$\mathsf{d}(f, \delta(x)) + \mathsf{d}(\delta(x), g) \geqslant \mathsf{d}(f, g).$$

Multiplying both sides by $\sqrt{p_f(x) p_g(x)}$ and integrating with respect to μ gives

$$\int \mathsf{d}(f, \delta) \sqrt{p_f\, p_g}\, d\mu + \int \mathsf{d}(\delta, g) \sqrt{p_f\, p_g}\, d\mu \geqslant \mathsf{d}(f, g) \cdot \rho(P_f, P_g).$$

For the first term, the Cauchy–Schwarz inequality yields

$$\int \mathsf{d}(f, \delta) \sqrt{p_f} \cdot \sqrt{p_g}\, d\mu \leqslant \sqrt{\int \mathsf{d}(f, \delta)^2\, p_f\, d\mu} \cdot \sqrt{\int p_g\, d\mu} = \sqrt{\mathbb{E}_f \mathsf{d}(f, \delta)^2}.$$

The same argument applied to the second term completes the proof. □

The constraint risk inequality reveals a fundamental tension in estimation. The right-hand side, $\mathsf{d}(f, g) \cdot \rho(P_f, P_g)$, captures the difficulty of the estimation problem between f and g: it is large when the parameters are far apart (large $\mathsf{d}(f, g)$) yet the distributions are similar (large ρ). When this product is large, the sum of the root-MSEs at f and g must also be large—no estimator can perform well at both.

The bound is most informative when $\rho(P_f, P_g)$ is not too small. If P_f and P_g are nearly orthogonal ($\rho \approx 0$), the bound becomes vacuous; but this is unsurprising, since very different distributions are easy to distinguish. The interesting regime is when *statistical similarity* coexists with *parameter separation*.

We apply the constraint risk inequality in more complicated settings in Section 2.4.1, but for now let us illustrate it in a model where the Cramér-Rao bound does not apply.

Example 2.44. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0, \theta)$ for $\theta > 0$. We use Lemma 2.43 to show that the θ^2/n MSE achieved by the unbiased estimator $\frac{n+1}{n} X_{(n)}$ cannot be improved by allowing an estimator to be biased.

For $\theta_1 < \theta_2$, the densities $p_{\theta_1} = \theta_1^{-1} \mathbb{1}_{[0,\theta_1]}$ and $p_{\theta_2} = \theta_2^{-1} \mathbb{1}_{[0,\theta_2]}$ overlap only on $[0, \theta_1]$, so

$$\rho(P_{\theta_1}, P_{\theta_2}) = \int_0^{\theta_1} \sqrt{\frac{1}{\theta_1 \theta_2}}\, dx = \sqrt{\frac{\theta_1}{\theta_2}}.$$

For the product measure of n observations, $\rho(P_{\theta_1}^{\otimes n}, P_{\theta_2}^{\otimes n}) = (\theta_1/\theta_2)^{n/2}$. Lemma 2.43 then gives, for any estimator δ,

$$\sqrt{\mathbb{E}_{\theta_1} |\delta - \theta_1|^2} + \sqrt{\mathbb{E}_{\theta_2} |\delta - \theta_2|^2} \geqslant (\theta_2 - \theta_1) \left(\frac{\theta_1}{\theta_2} \right)^{n/2}.$$

Writing $\theta_2 = \theta_1 + \epsilon$ for small $\epsilon > 0$, the right-hand side is approximately $\epsilon \cdot e^{-n\epsilon/(2\theta_1)}$. For any $\epsilon \in [\theta_1/n, 3\theta_1/n]$, this quantity is at least of order θ_1/n. In particular, for any estimator δ and any $\theta_2 \in [\theta_1 + \theta_1/n, \theta_1 + 3\theta_1/n]$, either $\mathbb{E}_{\theta_1} |\delta - \theta_1|^2$ or $\mathbb{E}_{\theta_2} |\delta - \theta_2|^2$ must be at least of order θ_1^2/n^2.

Since the above recipe works for arbitrary θ_1, this rules out estimators that attain MSE's of a smaller order than θ^2/n across the parameter space, no matter how large or small n and θ are.

$\lozenge$

2.4 Minimax paradigms

Admissibility is a minimal requirement: it rules out estimators that are uniformly dominated, but little else. The guesstimator $\delta(X) = \theta_1$ is admissible—no estimator can beat it at θ_1—yet it is clearly unsatisfactory. Admissibility tells us which estimators to avoid, but does not prescribe how to choose among the many that remain.

The constraint risk inequality shows that trade-offs across the parameter space are unavoidable: exceptional performance at one parameter value must be paid for elsewhere. But how should we navigate these trade-offs?

The *minimax paradigm* takes the pessimist's view: assume the worst and optimize accordingly. Rather than asking "is there any θ where this estimator is dominated?" (admissibility), we ask "what is the largest risk this estimator can incur?" and seek to minimize this worst-case risk. Where admissibility is permissive—accepting any estimator that is not uniformly beaten—minimaxity is demanding: it insists on the best possible guarantee against the least favorable parameter value.

Definition 2.45 (Minimax risk and minimax estimator)**.** Consider a decision problem $(\mathcal{X}, \mathscr{X}, \mathcal{P}, \Theta, (\mathcal{D}, \mathscr{D}), L)$ and let $\mathcal{C}$ denote the class of all (possibly randomized) decision rules $\delta : \mathcal{X} \to \mathcal{D}$.

The *minimax risk* is defined as

$$R^* := \inf_{\delta \in \mathcal{C}} \sup_{\theta \in \Theta} R(\theta, \delta).$$

A decision rule δ^* is called *minimax* if it achieves the minimax risk:

$$\sup_{\theta \in \Theta} R(\theta, \delta^*) = R^*.$$

The quantity $\sup_{\theta \in \Theta} R(\theta, \delta)$ is called the *maximum risk* (or *worst-case risk*) of δ.

The minimax criterion can be interpreted as a two-player zero-sum game. In an estimation problem, the statistician chooses an estimator δ, and then "nature" (or an adversary) chooses the parameter θ to maximize the risk. The minimax estimator is the statistician's optimal strategy in this game, guaranteeing the best possible worst-case performance.

Remark 2.46 (Pessimism or robustness?). The minimax approach is sometimes criticized as overly pessimistic (or overly conservative): why should we optimize for the worst case when it may rarely occur? However, this perspective has several compelling justifications:

(i) *Robustness:* The minimax estimator provides a strong *statistical guarantee*—its risk never exceeds R^*, regardless of the true θ.

(ii) *Ruling out super-efficiency:* As we saw in Example 2.42, achieving exceptionally low risk at some θ values necessarily inflates risk elsewhere (cf. the constraint risk inequality). Minimax estimation explicitly penalizes such greedy trade-offs, forcing estimators that do not sacrifice worst-case performance for gains at favorable parameter values.

(iii) *Unknown or adversarial settings:* In some applications, θ may be chosen by an adversary (e.g., in game theory or robust statistics) or may represent a "hard" instance. The minimax estimator is natural in such settings.

(iv) *Submodel flexibility:* Nothing prevents us from considering minimax risk over a subset $\Theta' \subset \Theta$:

$$\sup_{\theta \in \Theta'} R(\theta, \delta) \leqslant \sup_{\theta \in \Theta} R(\theta, \delta) \tag{2.3}$$

This allows us to calibrate our pessimism to the problem at hand. By considering various subsets Θ', we can study how the difficulty of estimation depends on the

region of the parameter space. Comparing minimax risks across nested subsets reveals which parts of the parameter space drive the difficulty of the problem. For certain models, the minimax risk is only non-trivial for such restriction – for example the uniform distribution studied in Example 2.44. Indeed, in example shows $\inf_\delta \sup_{\theta\in\Theta} R(\theta,\delta) = \infty$ an iid sample of size n from Uniform$[0,\theta]$, $\theta \in \Theta = (0,\infty)$.

(v) *Restriction to estimator classes:* Rather than optimizing over all decision rules $\mathcal{C}$, we may restrict to a subclass $\mathcal{C}' \subset \mathcal{C}$—for instance, unbiased estimators, equivariant estimators, or linear estimators. This yields

$$\inf_{\delta\in\mathcal{C}} \sup_{\theta\in\Theta} R(\theta,\delta) \leqslant \inf_{\delta\in\mathcal{C}'} \sup_{\theta\in\Theta} R(\theta,\delta).$$

The UMVUE and UMREE can be viewed through this lens: they are minimax within their respective estimator classes.

Finding minimax estimators and determining the minimax risk is generally difficult: the definition involves an infimum over all estimators and a supremum over the parameter space, neither of which admits a direct computation in most problems. We now present several tools that simplify this task in structured settings.

Our first tool connects back to the theory of equivariant estimation developed in Section 2.2. Recall that in invariant decision problems—where both the model and loss respect a group symmetry—equivariant estimators have constant risk (Theorem 2.29). This dramatically simplifies the minimax problem: among estimators with constant risk, the one with the smallest risk is automatically minimax.

Theorem 2.47 (Hunt-Stein for compact groups)**.** *Consider a decision problem where a locally compact abelian group G acts on $\mathcal{X}$, Θ, and $\mathcal{D} = \Theta$. Assume:*

(i) the action of G on Θ is transitive,

(ii) the model is equivariant under G,

(iii) the loss L is invariant under G and $d \mapsto L(\theta,d)$ is convex for all θ.

Then the UMREE δ^* is minimax.

(♠). Let δ^* be best equivariant with constant risk r^*. Let ν be the Haar measure on G, and let $G_1 \subset G_2 \subset \cdots$ be an increasing sequence of compact sets with $0 < \nu(G_n) < \infty$ and $\bigcup_n G_n = G$. For any estimator δ, define

$$\bar{\delta}_n(x) = \frac{1}{\nu(G_n)} \int_{G_n} g^{-1}\delta(gx)\, d\nu(g).$$

Since $\nu_n := \nu(\cdot \cap G_n)/\nu(G_n)$ is a probability measure, convexity and Jensen's inequality give

$$L(\theta, \bar{\delta}_n(x)) \leqslant \frac{1}{\nu(G_n)} \int_{G_n} L(\theta, g^{-1}\delta(gx))\, d\nu(g).$$

Taking expectations and using invariance of the loss and equivariance of the model,

$$R(\theta, \bar{\delta}_n) \leqslant \frac{1}{\nu(G_n)} \int_{G_n} R(g\theta, \delta)\, d\nu(g) \leqslant \sup_{\theta'} R(\theta', \delta).$$

For abelian G, the sequence $\bar{\delta}_n$ converges to an equivariant estimator $\bar{\delta}$ satisfying the same risk bound (Exercise 2.19). Since $\bar{\delta}$ is equivariant, $R(\theta, \bar{\delta}) \geqslant r^*$. Hence $\sup_\theta R(\theta, \delta) \geqslant r^*$ for all δ, so δ^* is minimax. □

We now apply the Hunt-Stein theorem to determine the minimax risk in the Gaussian location model, and examine how this interacts with the James-Stein phenomenon from Section 2.3.1.

Example 2.48. Consider $X \sim N_d(\theta, \sigma^2 I_d)$ with $\theta \in \mathbb{R}^d$ under squared error loss $L(\theta, \delta) = \|\delta - \theta\|^2$. This is a location family: the translation group $G = (\mathbb{R}^d, +)$ acts on $\mathcal{X} = \mathbb{R}^d$ and $\Theta = \mathbb{R}^d$ by $g_c(x) = x + c$, the model is equivariant, the loss is invariant, and G is locally compact abelian.

By Example 2.32, the (UMREE) Pitman estimator is $\delta^*(X) = X$. It has constant risk $R(\theta, \delta^*) = \mathbb{E}_\theta\|X - \theta\|^2 = d\sigma^2$. By the Hunt-Stein theorem, δ^* is minimax, so the minimax risk for the estimation problem of estimating $\theta \in \mathbb{R}^d$ is $d\sigma^2$.

For $d \geqslant 3$, Theorem 2.40 shows that the James-Stein estimator satisfies

$$R(\theta, \delta_{\mathrm{JS}}) = d\sigma^2 - (d-2)^2\sigma^4\mathbb{E}_\theta[\|X\|^{-2}] < d\sigma^2 \quad \text{for all } \theta \in \mathbb{R}^d.$$

Thus the James-Stein estimator is also minimax. As $\|\theta\| \to \infty$, the correction term vanishes and $R(\theta, \delta_{\mathrm{JS}}) \to d\sigma^2$, so $\sup_\theta R(\theta, \delta_{\mathrm{JS}}) = d\sigma^2$.

The moral is that minimaxity and admissibility are complementary criteria. We now have two minimax estimators: the UMVUE/UMREE/MLE and the James-Stein estimator. In high dimensions, the minimax criterion alone does not distinguish between X and δ_{JS}—both achieve the same worst-case risk. Admissibility could break the tie: among minimax estimators, we might prefer those that are not dominated. Morover, for any bounded subset $\Theta' \subset \Theta$, $\sup_{\theta\in\Theta'} R(\theta, \delta_{\mathrm{JS}}) < \sup_{\theta\in\Theta} R(\theta, \delta)$: if we are even slightly more optimistic than worst case across the entire parameter space, we prefer the James-Stein estimator. ◊

Despite appearing to be opposing viewpoints, admissibility and minimaxity are not incompatible. In fact, minimaxity can imply admissibility under the right conditions.

Theorem 2.49 (Unique minimax implies admissible)**.** *If δ^* is the unique minimax estimator, then δ^* is admissible.*

Proof. Suppose δ^* is unique minimax. If δ' is any other estimator, then by uniqueness,

$$\sup_\theta R(\theta, \delta^*) < \sup_\theta R(\theta, \delta').$$

This implies there exists some $\theta_0 \in \Theta$ such that $R(\theta_0, \delta^*) < R(\theta_0, \delta')$, so δ' does not dominate δ^*. Since δ' was arbitrary, δ^* is admissible.

Alternatively, suppose for contradiction that δ^* is inadmissible, so some δ' dominates it: $R(\theta, \delta') \leqslant R(\theta, \delta^*)$ for all θ, with strict inequality for at least one θ. Then

$$\sup_\theta R(\theta, \delta') \leqslant \sup_\theta R(\theta, \delta^*) = R^*,$$

so δ' is also minimax—contradicting the uniqueness of δ^*. □

As Example 2.48 illustrates, uniqueness often fails in simple settings: both the MLE and James-Stein estimator are minimax for the Gaussian location model when $d \geqslant 3$. When multiple minimax estimators exist, admissibility provides a criterion for choosing among them.

We now turn to another tool for establishing minimaxity: the submodel flexibility noted in (2.3). If we can identify a submodel $\mathcal{P}_0 \subset \mathcal{P}$ that captures the "hardest" part of the problem, then finding the minimax estimator over $\mathcal{P}_0$ suffices.

Lemma 2.50. *If δ is minimax for θ under $P \in \mathcal{P}_0 \subset \mathcal{P}$ and*

$$\sup_{P \in \mathcal{P}_0} R(P, \delta) = \sup_{P \in \mathcal{P}} R(P, \delta),$$

then δ is minimax for θ under $P \in \mathcal{P}$.

Proof. For any other estimator δ',

$$\sup_{P \in \mathcal{P}} R(P, \delta') \geqslant \sup_{P \in \mathcal{P}_0} R(P, \delta') \geqslant \sup_{P \in \mathcal{P}_0} R(P, \delta) = \sup_{P \in \mathcal{P}} R(P, \delta).$$

□

We illustrate the power of this lemma by reducing a vast nonparametric problem to the Gaussian location model, where the Hunt-Stein theorem applies.

Example 2.51 (Population mean with bounded variance)**.** Consider the nonparametric model

$$\mathcal{P} = \{P^{\otimes n} : P \text{ probability measure on } (\mathbb{R}, \mathcal{B}(\mathbb{R})) \text{ with } \operatorname{Var}_P(X) \leqslant M\}$$

for some known $M > 0$. We wish to estimate $\phi(P) = \mathbb{E}_P[X]$ under squared error loss.

The sample mean has risk $R(P, \bar{X}) = \mathrm{Var}_P(X)/n \leqslant M/n$, with equality when $\mathrm{Var}_P(X) = M$. To show $\bar{X}$ is minimax, consider the Gaussian submodel $\mathcal{P}_0 = \{N(\theta, M)^{\otimes n} : \theta \in \mathbb{R}\}$. This is a location family, so by the Hunt-Stein theorem, $\bar{X}$ is minimax over $\mathcal{P}_0$ with constant risk M/n.

Since $\mathcal{P}_0 \subset \mathcal{P}$ and $\bar{X}$ achieves its maximum risk M/n on the submodel $\mathcal{P}_0$, Lemma 2.50 implies that $\bar{X}$ is minimax over all of $\mathcal{P}$. ♢

2.4.1 Minimax rates

For some models, like the Gaussian location model studied in Example 2.48, the minimax risk is relatively easy to compute exactly in terms of various problem characteristics, such as dimension, variance, or sample size. Let us summarize the findings thus far.

Example 2.52 (Minimax rate for the Gaussian location model)**.** Consider the family of Gaussian location estimation problems indexed by $i = (n, d, \sigma^2) \in \mathbb{N} \times \mathbb{N} \times (0, \infty)$. For each $i = (n, d, \sigma^2)$, we observe $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N_d(\theta, \sigma^2 I_d)$ and wish to estimate $\theta \in \mathbb{R}^d$ under squared error loss $L_i(\theta, \delta) = \|\delta - \theta\|^2$.

From Example 2.48, the minimax risk for a single observation ($n = 1$) is $d\sigma^2$. For n i.i.d. observations, it suffices to consider the sufficient statistic $\bar{X} \sim N_d(\theta, \sigma^2 I_d/n)$ (by Rao-Blackwell, Theorem 1.43), so rescaling gives minimax risk

$$R^*_{n,d,\sigma^2} = \frac{d\sigma^2}{n}.$$

We can extract rate information by examining how the risk scales with each characteristic; how much difficult does the problem become when we increase e.g. dimension, variance or how much easier it becomes when we increase the sample size. ♢

When the exact minimax risk is difficult to compute, we often settle for characterizing its *rate*—how the minimax risk scales with problem characteristics. This coarser lens is powerful: it allows us to compare the difficulty of different estimation problems and to identify which estimators are "rate-optimal" without pinning down exact constants.

Definition 2.53. Consider a collection of decision problems, indexed by $i \in I$, given by the tuple $(\mathcal{X}_i, \mathscr{X}_i, \mathcal{P}_i, \Theta_i, (\mathcal{D}_i, \mathscr{D}_i), L_i)$ with risk function $\mathcal{R}_i$. The *minimax rate* is a function $r : I \to \mathbb{R}$ such that

$$c_* r(i) \leqslant \inf_\delta \sup_{\theta \in \Theta_i} \mathcal{R}_i(\theta, \delta) \leqslant C_* r(i)$$

for some constants $c_*, C_* > 0$ and for all $i \in I$.

Knowledge of the exact minimax risk immediately yields the rate: in Example 2.52, the minimax rate is $r(n, d, \sigma^2) = d\sigma^2/n$ with constants $c_* = C_* = 1$. More often, exact constants are intractable but the rate remains accessible. Proving a minimax rate requires two ingredients: an *upper bound* (exhibiting an estimator achieving risk $O(r(i))$) and a *lower bound* (showing no estimator can do better than $\Omega(r(i))$).

In many nonparametric problems, achieving the optimal rate requires carefully balancing bias and variance—neither the unbiased estimator nor the lowest-variance estimator is rate-optimal. The minimax rate framework helps identify the correct trade-off, even when exact constants remain elusive. We illustrate with a classical nonparametric model where the exact minimax risk is unknown, but the rate can be determined.

Consider observing $X_1, \dots, X_n \overset{\text{iid}}{\sim} f$ where f is an unknown probability density on $[0, 1]$. Rather than estimating the entire density, we focus on a simpler target: evaluating f at a fixed point $x_0 \in (0, 1)$.

Without restrictions on f, this problem is hopeless—the density could have arbitrary local behavior near x_0. We therefore assume f belongs to a *Hölder smoothness class*. For $\beta > 0$ and $M > 0$, define

$$\mathcal{F}_\beta(M) = \left\{ f : [0,1] \to \mathbb{R}_+ \,:\, \int_0^1 f = 1,\ |f^{(k)}(x) - f^{(k)}(y)| \leqslant M|x-y|^\alpha \text{ for all } x, y \in [0,1] \right\},$$

where $k = \lfloor \beta \rfloor$ is the number of derivatives and $\alpha = \beta - k \in [0, 1)$ controls the smoothness of the kth derivative. The case $\beta = 1$ corresponds to Lipschitz densities; $\beta = 2$ requires a Lipschitz first derivative; and so on. Larger β means smoother densities, which should make estimation easier.

Formally, the statistical model is $\mathcal{P} = \{P_f^{\otimes n} : f \in \mathcal{F}_\beta(M)\}$, where P_f denotes the distribution on $[0, 1]$ wit Lebesgue density f.

Consider the family of estimation problems indexed by $i = (n, \beta) \in \mathbb{N} \times (0, \infty)$, with parameter space $\Theta_i = \mathcal{F}_\beta(M)$ for fixed $M > 0$, and loss $L_i(f, \delta) = (\delta - f(x_0))^2$.

We are interested in determining the minimax rate for the minimax risk

$$R^*_{n,\beta} = \inf_\delta \sup_{f \in \mathcal{F}_\beta(M)} \mathbb{E}_f[(\delta - f(x_0))^2].$$

First interesting observation: the problem has no unbiased estimator.

Proposition 2.54. *Consider the decision problem corresponding to estimating $f(x_0)$ for a fixed $x_0 \in (0, 1)$ on the basis of n i.i.d. observations $X_1, \dots, X_n \sim f(x)dx$ from $f \in \mathcal{F}_\beta(M)$. For any sample size $n \geqslant 1$, there is no unbiased estimator of $f(x_0)$.*

Proof. Exercise 2.22. □

Since no unbiased estimator exists, we must navigate the bias-variance trade-off. The minimax rate framework tells us how to do this optimally.

If f were constant in a neighborhood of x_0, then the probability that X_i falls in an interval $[x_0 - h, x_0 + h]$ would be approximately $2h \cdot f(x_0)$. Counting observations in this interval and dividing by $2nh$ would give an unbiased estimator. Of course, f is not constant, so this procedure introduces bias—but if f is smooth and h is small, the bias should be small.

This reasoning leads to the *kernel density estimator*

$$\hat{f}_h(x_0) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right),$$

where $K : \mathbb{R} \to \mathbb{R}$ is a *kernel function* satisfying $\int K = 1$, and $h > 0$ is the *bandwidth*. The simplest choice is the box kernel $K(u) = \frac{1}{2}\mathbb{1}_{\{|u| \leqslant 1\}}$, which recovers the histogram-style estimator described above. Smoother kernels (e.g., the 'Gaussian kernel' $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$) yield smoother estimates but the same asymptotic behavior.

The bandwidth h controls the bias-variance trade-off. A small h means we average over a narrow window, capturing local behavior but using few observations—low bias, high variance. A large h averages over many observations but blurs local structure—low variance, high bias.

To quantify this, we analyze the bias and variance separately. For the bias, Taylor expansion of f around x_0 combined with the Hölder condition yields (see Exercise 2.20)

$$|\mathbb{E}_f[\hat{f}_h(x_0)] - f(x_0)| \leqslant C_1 h^\beta$$

for a constant C_1 depending on M and K. The smoothness β determines how quickly the bias vanishes as $h \to 0$: smoother densities have smaller bias for the same bandwidth.

For the variance, each summand $\frac{1}{h} K(\frac{X_i - x_0}{h})$ has magnitude of order $1/h$ and is nonzero with probability of order h. This gives (see again Exercise 2.20)

$$\mathrm{Var}_f(\hat{f}_h(x_0)) \leqslant \frac{C_2}{nh}$$

for a constant C_2. The variance decreases with n (more observations) and increases as $h \to 0$ (narrower window).

By Lemma 2.7, we obtain

$$\mathbb{E}_f[(\hat{f}_h(x_0) - f(x_0))^2] \leqslant C_1^2 h^{2\beta} + \frac{C_2}{nh}.$$

This expression captures the bias-variance trade-off: as h decreases, the first term shrinks but the second grows. The optimal bandwidth h^* minimizes the sum by

balancing the two terms. Setting $h^{2\beta} \asymp 1/(nh)$ and solving yields

$$h^* \asymp n^{-1/(2\beta+1)}.$$

Substituting back, both the squared bias and the variance are of order $n^{-2\beta/(2\beta+1)}$, giving

$$\sup_{f \in \mathcal{F}_\beta(M)} \mathbb{E}_f[(\hat{f}_{h^*}(x_0) - f(x_0))^2] \lesssim n^{-2\beta/(2\beta+1)}.$$

This establishes an upper bound on the minimax risk: there exists an estimator achieving rate $n^{-2\beta/(2\beta+1)}$. But is this the best possible? Perhaps a cleverer construction—something other than kernel estimation—could achieve a faster rate. To rule this out, we need a *lower bound* showing that no estimator, however ingenious, can do better.

The constraint risk inequality (Lemma 2.43) is the key tool. Recall the intuition: if two parameter values f_0 and f_1 generate statistically similar distributions yet have well-separated values of the target functional $f(x_0)$, then no estimator can perform well at both. The product $|f_1(x_0) - f_0(x_0)| \cdot \rho(P_{f_0}^{\otimes n}, P_{f_1}^{\otimes n})$ measures this tension, and the constraint risk inequality converts it into a lower bound on the worst-case risk.

We construct a pair of densities that are hard to distinguish. Let $f_0 \equiv 1$ be the uniform density on $[0, 1]$, and let

$$f_1 = 1 + \epsilon\psi_h,$$

where ψ_h is a smooth bump function centered at x_0 with support of width h, normalized so that $\int \psi_h = 0$ (ensuring f_1 integrates to one) and scaled so that both f_0 and f_1 lie in $\mathcal{F}_\beta(M)$. The Hölder constraint forces the bump to have height at most of order h^β: a taller bump would violate the smoothness condition. Thus $|f_1(x_0) - f_0(x_0)| \asymp \epsilon h^\beta$.

How similar are the distributions $P_{f_0}^{\otimes n}$ and $P_{f_1}^{\otimes n}$? The Bhattacharyya coefficient satisfies (Exercise 2.21)

$$\rho(P_{f_0}^{\otimes n}, P_{f_1}^{\otimes n}) \geqslant (1 - c'\epsilon^2 h)^n$$

for a constant $c > 0$. For small enough peturbations, the distributions remain close (Bhattacharyya coefficient near 1) – provided $n\epsilon^2 h \lesssim 1$; they become distinguishable when $n\epsilon^2 h \gg 1$. This reflects the intuition that n observations, each falling in the bump region with probability h, provide roughly nh "effective observations" for detecting a perturbation of size ϵ.

Applying Lemma 2.43:

$$\sqrt{\mathbb{E}_{f_0}(\delta - f_0(x_0))^2} + \sqrt{\mathbb{E}_{f_1}(\delta - f_1(x_0))^2} \geqslant |f_1(x_0) - f_0(x_0)| \cdot \rho(P_{f_0}^{\otimes n}, P_{f_1}^{\otimes n}).$$

The left-hand side is bounded by $2\sqrt{\sup_f \mathbb{E}_f[(\delta - f(x_0))^2]}$. Consequently, the worst-case risk of any estimator δ is bounded below by the square of the right-hand side. The right-hand side is of order ϵh^β when $n\epsilon^2 h \lesssim 1$. Choosing ϵ as large as possible subject to this constraint gives $\epsilon \asymp (nh)^{-1/2}$, and hence the minimax risk satisfies

$$R^*_{n,\beta} \gtrsim \epsilon^2 h^{2\beta} \asymp \frac{h^{2\beta}}{nh} = \frac{h^{2\beta-1}}{n}.$$

This bound holds for any $h > 0$. Optimizing over h—choosing h to maximize the lower bound—yields $h \asymp n^{-1/(2\beta+1)}$, and substituting gives

$$\inf_{\delta} \sup_{f \in \mathcal{F}_\beta(M)} \mathbb{E}_f[(\delta - f(x_0))^2] \gtrsim n^{-2\beta/(2\beta+1)}.$$

Combining the upper and lower bounds, we conclude that the minimax rate for estimating $f(x_0)$ over the Hölder class $\mathcal{F}_\beta(M)$ is

$$R^*_{n,\beta} \asymp n^{-2\beta/(2\beta+1)}.$$

The kernel density estimator with optimally chosen bandwidth is rate-optimal: no estimator can achieve a faster rate, and the bias-variance trade-off we identified is indeed the correct one.

Remark 2.55 (CDF vs density estimation). The contrast with CDF estimation (Example 2.11) is instructive. The empirical CDF $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \leqslant t\}}$ is unbiased for $F(t)$ and achieves the parametric rate $n^{-1/2}$. Why is density estimation so much harder?

Geometrically, the CDF integrates the density up to t, averaging over a macroscopic region. This averaging stabilizes estimation: whether or not X_i falls below t is informative about $F(t)$ regardless of the local shape of f. The density at a point, however, describes infinitesimal behavior—how much probability mass is packed into an arbitrarily small neighborhood of x_0. No finite sample can resolve infinitesimal structure without assumptions, which is why smoothness (the Hölder condition) is essential and why the rate $n^{-2\beta/(2\beta+1)}$ is slower than $n^{-1/2}$ for any finite β.

Exercises

Exercise 2.1 (UMVUE for the mean). Consider the (nonparametric) model corresponding to observing $X_1, \dots, X_n$ i.i.d. from an unknown distribution P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with finite variance. We wish to estimate the population mean $\phi(P) = \mathbb{E}_P[X_1]$.

(a) Show that the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the UMVUE for $\phi(P) = \mathbb{E}_P[X_1]$. You may use the result from Exercise 1.17 that the order statistics are a complete sufficient statistic for this model.

(b) Show that the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is the UMVUE for the population variance $\sigma^2(P) = \mathrm{Var}_P(X_1)$.

Exercise 2.2 (UMVUE for the CDF in the normal mean model). Consider a statistical model corresponding to $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$, $\theta \in \mathbb{R}$ unknown, known variance $\sigma^2 > 0$. We wish to estimate the CDF at a fixed point t, i.e., $\phi(\theta) = \Phi((t-\theta)/\sigma)$.

(a) Show that the UMVUE for $\phi(\theta)$ is given by

$$\delta(X) = \Phi\left(\frac{t - \bar{X}}{\sigma\sqrt{1 - 1/n}}\right).$$

(b) Compare the variance of $\delta(X)$ with the variance of the empirical CDF $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i \leqslant t\}}$ of Example 2.11. Which one is smaller and why?

Exercise 2.3 (Linear model). Let $Y \sim N_n(X\beta, \sigma^2 I_n)$ for unknown $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$, where $X \in \mathbb{R}^{n \times p}$ has full column rank.

(a) Show that $(\hat{\beta}_{\text{OLS}}, s^2)$, where $\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top Y$ and $s^2 = \frac{1}{n-p}\|Y - X\hat{\beta}_{\text{OLS}}\|^2$, is a complete sufficient statistic for (β, σ^2).

(b) Conclude that $\hat{\beta}_{\text{OLS}}$ is the UMVUE for β.

(c) Recall that the Gauss-Markov theorem states that $\hat{\beta}_{\text{OLS}}$ is the Best Linear Unbiased Estimator (BLUE) regardless of the distribution of Y, as long as it has mean $X\beta$ and covariance $\sigma^2 I_n$. How does the UMVUE property under normality relate to the BLUE property?

Exercise 2.4 (Bias-Variance Decomposition). Prove Lemma 2.7.

Exercise 2.5 (Consistency and Bias). Let $\hat{\theta}_1, \hat{\theta}_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^d$ with finite covariance matrix Σ and mean vector μ. Consider 'the estimator' $\bar{\theta}_m = \frac{1}{m}\sum_{j=1}^{m} \hat{\theta}_j$ of $\theta \in \mathbb{R}^d$. Show that $\mathbb{E}[\|\bar{\theta}_m - \theta\|^2] \to 0$ if and only if $\mathbb{E}[\hat{\theta}_1] = \theta$.

Exercise 2.6 (Uncorrelated with 0-unbiased estimators). Let $\delta(X)$ have finite variance. Show that a necessary and sufficient condition for δ to be the UMVUE of its expectation $g(\theta) = \mathbb{E}_\theta[\delta(X)]$ is that $\mathrm{Cov}_\theta(\delta(X), U(X)) = 0$ for all $\theta \in \Theta$ and all statistics U such that $\mathbb{E}_\theta[U(X)] = 0$ for all $\theta \in \Theta$ (i.e., U is an unbiased estimator of zero).

Exercise 2.7 (Cramer-Rao Lower Bounds). (a) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$ with known variance $\sigma^2 > 0$. Compute the Cramer-Rao lower bound for the variance of any unbiased estimator of $\theta \in \mathbb{R}$.

(b) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ for $p \in (0, 1)$. Derive the Cramer-Rao lower bound for the variance of any unbiased estimator of p. What is the noticeable difference compared to the normal distribution case? What is the worst-case lower bound (over $p \in (0, 1)$)?

Exercise 2.8 (DQM implications). The aim is to prove Lemma 2.13. Throughout, let μ be a dominating measure for the model with densities $p_\theta = dP_\theta/d\mu$.

(a) Show that $\sqrt{p_\theta}, \sqrt{p_{\theta+h}} \in L^2(\mu)$, and conclude that $A_h := \sqrt{p_{\theta+h}} - \sqrt{p_\theta} \in L^2(\mu)$.

(b) Using part (a) and the fact that $L^2(\mu)$ is a vector space, conclude that $h^\top S_\theta \sqrt{p_\theta} \in L^2(\mu)$ for all sufficiently small h and hence $I(\theta) = E_\theta[S_\theta S_\theta^\top]$ has all entries finite.

(c) Using the algebraic identity $a - b = (\sqrt{a} - \sqrt{b})(\sqrt{a} + \sqrt{b})$ and the DQM expansion, show that

$$p_{\theta+h} - p_\theta = h^\top S_\theta \, p_\theta + \tilde{r}_h,$$

where $\tilde{r}_h$ is a remainder term that you should specify explicitly in terms of $r_h := \sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^\top S_\theta \sqrt{p_\theta}$.

(d) Let $T : \mathcal{X} \to \mathbb{R}$ satisfy $E_\theta[T^2] < \infty$. Show that the contribution of the remainder term to $\psi(\theta + h) - \psi(\theta)$ is negligible:

$$\int T(x) \tilde{r}_h(x) \, d\mu(x) = o(\|h\|).$$

Hint: Use the Cauchy-Schwarz inequality and the bound $\|r_h\|_{L^2(\mu)} = o(\|h\|)$.

(e) Combine the results of parts (c) and (d) to conclude that if $E_{\theta'}[T^2] < \infty$ for all θ' in a neighborhood of θ, then $\psi(\theta') = E_{\theta'}[T]$ is differentiable at θ with $\nabla\psi(\theta) = E_\theta[T \cdot S_\theta]$.

Exercise 2.9 (Covariance matrix estimation as an invariant decision problem). Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N_p(0, \Sigma)$ with Σ positive definite, so the sufficient statistic is $S = \sum_{i=1}^n X_i X_i^\top \sim \text{Wishart}_p(n, \Sigma)$. The parameter space is $\Theta = \mathcal{S}_p^{++}$, the set of $p \times p$ positive definite matrices.

Consider the general linear group $G = \text{GL}(p)$ acting on data by $g_A(X_1, \dots, X_n)$ as $(A, X_i) \mapsto AX_i$ (equivalently, $g_A S = ASA^\top$) and on parameters by $g_A \Sigma = A\Sigma A^\top$.

(a) Show that the model is equivariant under this group action.

(b) The squared Frobenius loss $L(\Sigma, \delta) = \|\delta - \Sigma\|_F^2$, where $\|M\|_F = \sqrt{\text{Tr}(M^\top M)}$ is the Frobenius norm, is *not* invariant under this action. Verify this by finding matrices A, Σ, and δ such that $L(A\Sigma A^\top, A\delta A^\top) \neq L(\Sigma, \delta)$.

(c) The *Stein loss* is defined as

$$L_S(\Sigma, \delta) = \text{Tr}(\delta\Sigma^{-1}) - \log|\delta\Sigma^{-1}| - p.$$

Show that Stein loss is invariant under the action of $\text{GL}(p)$.

Exercise 2.10 (Pitman estimator for a scale family). Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\sigma)$ with density $f(x \mid \sigma) = \frac{1}{\sigma} e^{-x/\sigma}$ for $x > 0$ and $\sigma > 0$. We wish to estimate σ under the scale-invariant loss

$$L(\sigma, \delta) = \left(\frac{\delta}{\sigma} - 1\right)^2.$$

(a) Verify that this loss is invariant under the multiplicative group $G = (\mathbb{R}_{>0}, \times)$ acting by $g_c\sigma = c\sigma$ and $g_c\delta = c\delta$.

(b) Show that an UMREE estimator is of the form $\delta^*(x) = a\sum_{i=1}^n x_i$ for some constant $a > 0$. *Hint*: Consider an equivariant estimator $\delta(cx) = c\delta(x)$ and its Rao-Blackwellization $\delta^* = \mathbb{E}_\sigma[\delta(X) \mid \bar{X}]$, where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$.

(c) Find the UMREE by minimizing $\mathcal{R}(1, \delta^*)$ in $a > 0$. *Hint:* If $X_i \overset{\text{iid}}{\sim} \text{Exp}(1)$, then $\sum_{i=1}^n X_i \sim \text{Gamma}(n, 1)$.

Exercise 2.11 (Pitman estimator for the Cauchy location family). Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Cauchy}(\theta, 1)$ with density

$$f(x - \theta) = \frac{1}{\pi(1 + (x - \theta)^2)}, \quad x \in \mathbb{R},\ \theta \in \mathbb{R}.$$

(a) Write down the Pitman estimator for θ under squared error loss.

(b) For $n = 2$, show that the Pitman estimator can be written as

$$\delta^*(X_1, X_2) = \frac{X_1 + X_2}{2} + \frac{1}{2}(X_1 - X_2) \cdot g\left(\frac{X_1 - X_2}{2}\right)$$

for some odd function $g : \mathbb{R} \to \mathbb{R}$, and reason that $g \equiv 0$.

Hint: Use the substitution $\eta = \theta - \frac{X_1+X_2}{2}$ and let $u = \frac{X_1-X_2}{2}$.

(c) For $n = 3$, show that as $x_3 \to \infty$ with x_1, x_2 fixed, the Pitman estimator satisfies

$$\delta^*(x_1, x_2, x_3) \to \frac{x_1 + x_2}{2}.$$

Interpret this result.

Exercise 2.12 (Admissibility of the MLE in the normal mean model for $d = 1$). For $d = 1$ and $d = 2$ the MLE X is admissible in the model $X \sim N_d(\theta, I)$. This exercise deals with the case $d = 1$. So we assume that $X \sim N(\theta, 1)$. The goal is to prove that there exists no other estimator $\hat{\theta}$ such that $E_\theta(\hat{\theta} - \theta)^2 \leqslant E_\theta(X - \theta)^2$ for all $\theta \in \mathbb{R}$, with strict inequality for some $\theta \in \mathbb{R}$.

For $\tau > 0$, consider the $N(0, \tau)$ prior on the parameter θ. Denote the corresponding prior density by π_τ.

(i) Show that if an estimator $\hat{\theta}$ as described above would exist, then there would exist an $\varepsilon > 0$ and $\theta_0 < \theta_1$ such that

$$1 - \int E_\theta(\hat{\theta} - \theta)^2 \pi_\tau(\theta)\, d\theta \geqslant \varepsilon \int_{\theta_0}^{\theta_1} \pi_\tau(\theta)\, d\theta.$$

(ii) Let $\tilde{\theta}_\tau$ be the posterior mean corresponding to the prior π_τ. Compute the corresponding Bayes risk

$$\int E_\theta(\tilde{\theta}_\tau - \theta)^2 \pi_\tau(\theta)\, d\theta.$$

You may use without proof that the posterior mean minimizes this integrated risk among all estimators.

(iii) Using the results of (i) and (ii), show that if an estimator $\hat{\theta}$ as described above would exist, then

$$\frac{1 - \int E_\theta(\hat{\theta} - \theta)^2 \pi_\tau(\theta)\, d\theta}{1 - \int E_\theta(\tilde{\theta}_\tau - \theta)^2 \pi_\tau(\theta)\, d\theta} \to \infty$$

as $\tau \to \infty$. Derive a contradiction.

Remark 2.56. Admissibility of the MLE in the case $d = 2$ can also be proved using this approach via the Bayes risk. The analysis is more involved however, since using conjugate Gaussian priors as in the case $d = 1$ does not work. See Problem 4.5 on p. 398 of Lehmann and Casella (1998).

Exercise 2.13 (Negative moments of the multivariate Gaussian). Let $X \sim N_d(0, I)$. Show that $E(1/\|X\|^p) < \infty$ if and only if $d > p$.

Exercise 2.14 (Proof of the James-Stein lemma). Prove Lemma 2.41.

Exercise 2.15 (Shrinking towards another point). Let $X \sim N_d(\theta, I)$ and $v \in \mathbb{R}^d$. Define the estimator

$$\tilde{\theta}_{\mathrm{JS}} = v + \left(1 - \frac{d-2}{\|X - v\|^2}\right)(X - v).$$

Prove that for $d \geqslant 3$, this estimator also satisfies $E_\theta\|\tilde{\theta}_{\mathrm{JS}} - \theta\|^2 < E_\theta\|\hat{\theta}_{\mathrm{MLE}} - \theta\|^2$ for all $\theta \in \mathbb{R}^d$.

Exercise 2.16 (Oracle version of James-Stein). Use the expression for the risk of the James-Stein estimator to prove that if $X \sim N(\theta, \sigma^2 I)$, then for every $\theta \in \mathbb{R}^d$ and $d \geqslant 3$,

$$E_\theta\|\hat{\theta}_{\mathrm{JS}} - \theta\|^2 \leqslant 4\sigma^2 + \inf_{c \in \mathbb{R}} E_\theta\|cX - \theta\|^2.$$

This is a so-called oracle inequality that asserts that up to a constant, the risk of the James-Stein estimator is as good as the risk that could be achieved by an oracle that may use its knowledge of the true parameter θ to choose the degree of shrinking.

Exercise 2.17 (Estimating the distribution is hard). Let $\mathcal{P}$ be the set of all probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with variance bounded by σ^2. This exercise shows that estimating the distribution P itself in total variation distance is impossible with uniform control over $\mathcal{P}$.

Let $P_0 = N(0, 1)$. For $M > 1$, define the mixture distribution

$$P_M = \left(1 - \frac{1}{M}\right) N(0, 1) + \frac{1}{M} N(M^2, 1).$$

(a) Verify that $P_0, P_M \in \mathcal{P}$ for all $M > 1$.

(b) Show that $d_{TV}(P_0, P_M) \to 1$ as $M \to \infty$.

Hint: Consider the event $A_M = \{x : x > M^2/2\}$ *and compute* $P_0(A_M)$ *and* $P_M(A_M)$.

(c) Let $P^{\otimes n}$ denote the n-fold product measure corresponding to n i.i.d. draws from P. Show that for any fixed n,

$$d_{TV}(P_0^{\otimes n}, P_M^{\otimes n}) \to 0 \quad \text{as } M \to \infty.$$

Hint: Let N be the number of samples from the $N(M^2, 1)$ component. Show that $P_M^{\otimes n}(N = 0) \to 1$ as $M \to \infty$, and that conditionally on $N = 0$, the two product measures coincide.

(d) Conclude that for any estimator $\hat{P}_n : \mathbb{R}^n \to \mathcal{P}$ and any sample size n,

$$\sup_{P \in \mathcal{P}} E_P[d_{TV}(\hat{P}_n, P)] \geqslant \frac{1}{2}.$$

Hint: Use Le Cam's method: for any estimator and any pair of distributions P, Q,

$$E_P[d_{TV}(\hat{P}_n, P)] + E_Q[d_{TV}(\hat{P}_n, Q)] \geqslant d_{TV}(P, Q)(1 - d_{TV}(P^{\otimes n}, Q^{\otimes n})).$$

Exercise 2.18. Let $P_\theta = N(\theta, \sigma^2)$ and $P_{\theta'} = N(\theta', \sigma^2)$ be two univariate normal distributions with the same variance. Show that

$$d_{TV}(P_\theta, P_{\theta'}) \leqslant \frac{1}{2\sigma}|\theta - \theta'|.$$

Hint: Use Pinsker's inequality, which relates total variation distance to Kullback-Leibler divergence: $d_{TV}(P, Q) \leqslant \sqrt{\frac{1}{2} D_{KL}(P \| Q)}$.

Exercise 2.19 (♠). This exercise completes the proof of the Hunt-Stein theorem. Let G be a locally compact abelian group with Følner sequence $\{G_n\}$, and let $\bar{\delta}_n$ be the partial group averages defined in the proof of Theorem 2.47.

(a) Show that $\bar{\delta}_n$ is asymptotically equivariant: for all $h \in G$,

$$\mathbb{E}_\theta \|\bar{\delta}_n(hX) - h\bar{\delta}_n(X)\|^2 \to 0 \quad \text{as } n \to \infty.$$

Hint: Express the difference as integrals over $G_n \triangle hG_n$ and use the Følner property.

(b) Let r^* be the constant risk of the UMREE. Show that $\liminf_n R(\theta, \bar{\delta}_n) \geqslant r^*$.

Hint: If $R(\theta, \bar{\delta}_{n_k}) < r^* - \epsilon$ along a subsequence, use a compactness argument to extract a limit that is equivariant with risk strictly below r^*, contradicting the

definition of the UMREE.

Exercise 2.20 (Kernel density estimation bounds). Consider the kernel density estimator $\hat{f}_h(x_0) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right)$ for a density $f \in \mathcal{F}_\beta(L)$ at a point x_0. Assume the kernel K satisfies $\int K(u)\,du = 1$, $\int |u|^\beta |K(u)|\,du < \infty$, and $\int u^j K(u)\,du = 0$ for all integers $1 \leqslant j < \beta$.

(a) Show that the bias is bounded by

$$|\mathbb{E}_f[\hat{f}_h(x_0)] - f(x_0)| \leqslant C_1 h^\beta,$$

where C_1 depends on L and K.

(b) Show that the variance is bounded by

$$\mathrm{Var}_f(\hat{f}_h(x_0)) \leqslant \frac{C_2}{nh},$$

where C_2 depends on $\|K\|_\infty$ (or $\|K\|_2^2$) and $f(x_0)$ (or $\|f\|_\infty$).

Exercise 2.21 (Bhattacharyya affinity under perturbation). Let P_{f_0} be the uniform distribution on $[0,1]$ (density $f_0 \equiv 1$) and let P_{f_1} have density $f_1(x) = 1 + \epsilon\psi_h(x)$, where $\psi_h(x) = \psi((x - x_0)/h)$ for a function ψ supported on $[-1/2, 1/2]$ with $\int \psi = 0$ and bounded magnitude. Assume $|\epsilon\psi_h(x)| \leqslant 1/2$ so that f_1 is a valid density.

Show that the Bhattacharyya affinity between the product measures satisfies

$$\rho(P_{f_0}^{\otimes n}, P_{f_1}^{\otimes n}) \geqslant 1 - cn\epsilon^2 h$$

for some constant $c > 0$ depending on ψ.

Exercise 2.22 (Non-existence of unbiased density estimators). Let $\mathcal{F}$ be the class of Lipschitz densities on $[0,1]$. We wish to show that for any fixed $x_0 \in (0,1)$ and any sample size $n \geqslant 1$, there is no unbiased estimator of $f(x_0)$.

Proceed by contradiction: Suppose $\delta(X_1, \ldots, X_n)$ is an unbiased estimator, i.e., $\mathbb{E}_f[\delta] = f(x_0)$ for all $f \in \mathcal{F}$.

(a) Fix $f_0 \in \mathcal{F}$. Consider perturbations $f_\epsilon = f_0 + \epsilon g$ where g is Lipschitz, supported on an interval $[a,b] \subset [0,1]$ not containing x_0, and satisfies $\int g = 0$. Show that the unbiasedness condition implies

$$\int_0^1 g(t) h_{f_0}(t)\,dt = 0,$$

where $h_{f_0}(t) = \int_{[0,1]^{n-1}} \delta(t, x_2, \ldots, x_n) \prod_{j=2}^{n} f_0(x_j)\, dx_j$.

(b) Use the result from (a) to show that $h_{f_0}(t)$ must be constant for almost every $t \in [0,1] \backslash \{x_0\}$.

(c) Show that this constant must be $f_0(x_0)$.

(d) Deduce that for any fixed $t \neq x_0$, the function $\delta_t(x_2, \ldots, x_n) = \delta(t, x_2, \ldots, x_n)$ is an unbiased estimator of $f(x_0)$ based on $n-1$ observations. Explain why this leads to a contradiction.

Part II

Asymptotic Statistics

Part III

Appendix

A Metric Spaces

A.1 Metrics

Definition A.1 (Metric)**.** Let X be a set. A *metric* on X is a function $\mathsf{d} : X \times X \to \mathbb{R}$ such that for all $x, y, z \in X$:

(i) $\mathsf{d}(x, y) \geqslant 0$ (non-negativity);

(ii) $\mathsf{d}(x, y) = 0$ if and only if $x = y$ (identity of indiscernibles);

(iii) $\mathsf{d}(x, y) = \mathsf{d}(y, x)$ (symmetry);

(iv) $\mathsf{d}(x, z) \leqslant \mathsf{d}(x, y) + \mathsf{d}(y, z)$ (triangle inequality).

Definition A.2 (Metric Space)**.** A *metric space* is a pair (X, d), where X is a set and d is a metric on X.

Example A.3 (Euclidean Space)**.** Let $X = \mathbb{R}^n$. The Euclidean metric is defined by

$$\mathsf{d}(x, y) = \|x - y\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Then $(\mathbb{R}^n, \mathsf{d})$ is a metric space. ◊

Example A.4 (Function Space)**.** Let $X = C[0, 1]$, the set of continuous real-valued functions on the interval $[0, 1]$. The supremum metric (or uniform metric) is defined by

$$\mathsf{d}(f, g) = \sup_{t \in [0,1]} |f(t) - g(t)|.$$

Then $(C[0, 1], \mathsf{d})$ is a metric space. ◊

A.2 Topology

Definition A.5 (Open Ball)**.** Let (X, d) be a metric space. The *open ball* of radius $r > 0$ centered at $x \in X$ is the set

$$B_r(x) = \{y \in X : \mathsf{d}(x, y) < r\}.$$

Definition A.6 (Open Set in Metric Spaces)**.** A subset $U \subseteq X$ is called *open* if for every $x \in U$, there exists an $\epsilon > 0$ such that $B_\epsilon(x) \subseteq U$.

Definition A.7 (Closed Set)**.** A subset $F \subseteq X$ is *closed* if its complement $X \backslash F$ is open.

Definition A.8 (Closure)**.** The *closure* of a subset $A \subseteq X$, denoted $\overline{A}$, is the intersection of all closed sets containing A. It is the smallest closed set containing A.

Definition A.9 (Neighborhood)**.** A subset $N \subseteq X$ is called a *neighborhood* of a point $x \in X$ if there exists an open set U such that $x \in U \subseteq N$. Equivalently, N is a neighborhood of x if there exists an $\epsilon > 0$ such that $B_\epsilon(x) \subseteq N$.

Proposition A.10. *Let (X, d) be a metric space. The collection $\mathcal{T}$ of open sets in X (as defined in Definition A.6) satisfies the following properties:*

(i) $\varnothing \in \mathcal{T}$ and $X \in \mathcal{T}$;

(ii) The union of any collection of open sets is open;

(iii) The intersection of any finite collection of open sets is open.

Definition A.11 (Continuous Function)**.** Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is *continuous* at a point $x \in X$ if for every $\epsilon > 0$, there exists a $\delta > 0$ such that $\mathsf{d}_X(x, y) < \delta$ implies $\mathsf{d}_Y(f(x), f(y)) < \epsilon$. The function f is *continuous* if it is continuous at every point in X.

Proposition A.12. *Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is continuous if and only if for every open set $V \subseteq Y$, the preimage $f^{-1}(V)$ is open in X.*

Proposition A.12 reveals that continuity can be characterized entirely in terms of open sets, without explicit reference to the underlying metric.

Definition A.13 (Topology Generated by a Metric)**.** The collection of all open sets in a metric space (X, d) forms a topology on X, called the *topology induced by the metric* d.

This motivates the generalization of continuity in metric spaces to spaces where only the notion of "openness" is defined, which leads to the definition of a topological space. It turns out that the properties of Proposition A.10 are precisely the properties needed to have things function the way they do for metrics.

Definition A.14 (Topology)**.** A *topology* on a set X is a collection $\mathcal{T}$ of subsets of X satisfying:

(i) $\varnothing \in \mathcal{T}$ and $X \in \mathcal{T}$;

(ii) The union of any collection of sets in $\mathcal{T}$ is in $\mathcal{T}$;

(iii) The intersection of any finite collection of sets in $\mathcal{T}$ is in $\mathcal{T}$.

The pair $(X, \mathcal{T})$ is called a *topological space*. The elements of $\mathcal{T}$ are called *open sets*.

Remark A.15. Every metric induces a topology, but not every topology arises from a metric. A topological space whose topology is induced by a metric is called *metrizable*.

Definition A.16 (Separable Space)**.** A topological space X is called *separable* if it contains a countable dense subset. That is, there exists a countable set $D \subseteq X$ such that $\overline{D} = X$.

Definition A.17 (Polish Space)**.** A topological space X is called a *Polish space* if it is separable and completely metrizable. That is, there exists a metric d on X which induces the topology of X such that (X, d) is a complete metric space.

A.3 Compactness

Definition A.18 (Bounded Set)**.** A subset A of a metric space (X, d) is *bounded* if there exists $x \in X$ and $R > 0$ such that $A \subseteq B_R(x)$.

Definition A.19 (Compactness)**.** A subset K of a topological space X is *compact* if every open cover of K has a finite subcover. That is, if $K \subseteq \bigcup_{i \in I} U_i$ where each U_i is open, then there exists a finite subset $J \subseteq I$ such that $K \subseteq \bigcup_{j \in J} U_j$.

Definition A.20 (Sequential Compactness)**.** A subset K of a metric space is *sequentially compact* if every sequence in K has a convergent subsequence whose limit belongs to K.

In metric spaces, compactness and sequential compactness are equivalent.

Definition A.21 (Locally Compact Space)**.** A topological space X is *locally compact* if every point $x \in X$ has a compact neighborhood. That is, for every $x \in X$, there exists an open set U containing x such that $\overline{U}$ is compact.

Example A.22 (Euclidean space is locally compact)**.** The space $\mathbb{R}^d$ with the Euclidean topology is locally compact. For any $x \in \mathbb{R}^d$, the open ball $B_1(x)$ has closure $\overline{B_1(x)}$ equal to the closed ball $\{y : \|y - x\| \leqslant 1\}$, which is compact by the Heine–Borel theorem.

More generally, any open or closed subset of $\mathbb{R}^d$ is locally compact. Compact spaces are trivially locally compact. ♢

Definition A.23 (Limit Point)**.** A point $x \in X$ is a *limit point* (or accumulation point) of a set A if every open neighborhood of x contains a point of A distinct from x.

Definition A.24 (Generated Topology)**.** Let X be a set and $\mathcal{S}$ be a collection of subsets of X. The *topology generated by* $\mathcal{S}$ is the smallest topology on X containing $\mathcal{S}$. It consists of all arbitrary unions of finite intersections of elements of $\mathcal{S}$. The elements of $\mathcal{S}$ are called a *subbasis* for the topology.

Example A.25 (Standard Topology on $\mathbb{R}$)**.** Let $X = \mathbb{R}$. The standard topology on $\mathbb{R}$ is the topology generated by the collection of all open intervals (a, b). In fact, this is the same as the topology induced by the Euclidean metric $\mathsf{d}(x, y) = |x - y|$. ◊

Example A.26 (Topology of Pointwise Convergence)**.** Let X be the set of all functions $f : [0, 1] \to \mathbb{R}$. The topology of pointwise convergence is the topology generated by sets of the form

$$S_{t,(a,b)} = \{f \in X : a < f(t) < b\}$$

where $t \in [0, 1]$ and $a < b$ are real numbers. Convergence in this topology corresponds exactly to pointwise convergence: a sequence $f_n \to f$ if and only if $f_n(t) \to f(t)$ for all $t \in [0, 1]$. ◊

B Measure Theory

B.1 Measure and Probability

The foundational concept in measure theory is the sigma-algebra, which defines the collection of subsets to which we can assign a measure.

Definition B.1. A σ-algebra $\mathcal{F}$ on a set Ω is a collection of subsets of Ω that satisfies the following properties:

(i) $\varnothing \in \mathcal{F}$

(ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$

(iii) If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Once we have a σ-algebra, we can define a measure, which generalizes the concepts of length, area, and probability.

Definition B.2. Consider a measurable space $(\Omega, \mathcal{F})$. A *measure* μ on a σ-algebra $\mathcal{F}$ is a function that assigns a non-negative real number to each set in $\mathcal{F}$ and satisfies the following properties:

1. $\mu(\varnothing) = 0$
2. If $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$

If $\mu(\Omega) < \infty$, then μ is called a *finite measure*. If in addition $\mu(\Omega) = 1$, then μ is a *probability measure*.

These components form the standard objects of study in measure theory.

Definition B.3. A pair $(\Omega, \mathcal{F})$ consisting of a set Ω and a σ-algebra $\mathcal{F}$ is called a *measurable space*. A triple $(\Omega, \mathcal{F}, \mu)$ consisting of a measurable space and a measure μ is called a *measure space*. If μ is a probability measure, the triple is called a *probability space*.

Many important measures are not finite, but satisfy a weaker condition called σ-finiteness.

Definition B.4. A measure μ on $(\Omega, \mathcal{F})$ is called *σ-finite* if there exists a sequence of sets $A_1, A_2, \ldots \in \mathcal{F}$ such that $\bigcup_{i=1}^{\infty} A_i = \Omega$ and $\mu(A_i) < \infty$ for all i.

A simple example of a measure that can be finite or σ-finite is the counting measure.

Example B.5 (Counting Measure)**.** Let Ω be a countable set and $\mathcal{F} = 2^\Omega$. The counting measure μ is defined by $\mu(A) = |A|$ (the number of elements in A) for any $A \subseteq \Omega$. This measure is σ-finite since Ω is countable (take $A_i = \{\omega_i\}$). $\lozenge$

Measures are often defined on a smaller class of sets (like intervals in $\mathbb{R}$) and then extended to the full σ-algebra. Carathéodory's Extension Theorem guarantees that this extension is unique for σ-finite measures.

Theorem B.6 (Uniqueness of Measure Extension)**.** *Let $\mathcal{A}$ be a collection of subsets of Ω that is closed under finite intersections (a π-system) and generates the σ-algebra $\mathcal{F} = \sigma(\mathcal{A})$. If two measures μ and ν on $(\Omega, \mathcal{F})$ agree on $\mathcal{A}$ (i.e., $\mu(A) = \nu(A)$ for all $A \in \mathcal{A}$), and they are σ-finite on $\mathcal{A}$, then $\mu = \nu$ on $\mathcal{F}$.*

Measures also behave continuously with respect to increasing or decreasing sequences of sets.

Proposition B.7. *Let μ be a measure on $(\Omega, \mathcal{F})$.*

1. *(**Continuity from below**) If $A_1 \subseteq A_2 \subseteq \cdots$ is an increasing sequence of sets in $\mathcal{F}$ and $A = \bigcup_{n=1}^{\infty} A_n$, then*

$$\mu(A) = \lim_{n\to\infty} \mu(A_n).$$

2. *(**Continuity from above**) If $A_1 \supseteq A_2 \supseteq \cdots$ is a decreasing sequence of sets in $\mathcal{F}$ with $\mu(A_1) < \infty$ and $A = \bigcap_{n=1}^{\infty} A_n$, then*

$$\mu(A) = \lim_{n\to\infty} \mu(A_n).$$

We now turn to the functions between measurable spaces, which must preserve the measurable structure.

Definition B.8. Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{G})$ be measurable spaces. A function $f : \Omega \to S$ is *measurable* (or $\mathcal{F}/\mathcal{G}$-measurable) if for every $B \in \mathcal{G}$, the preimage $f^{-1}(B) \in \mathcal{F}$.

That is, f is measurable if

$$f^{-1}(B) = \{\omega \in \Omega : f(\omega) \in B\} \in \mathcal{F} \quad \text{for all } B \in \mathcal{G}.$$

Conversely, any function induces a σ-algebra on its domain.

Definition B.9. Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{G})$ be measurable spaces, and let $f : \Omega \to S$ be a measurable function. The *σ-algebra generated by f*, denoted by $\sigma(f)$, is the collection of all preimages of sets in $\mathcal{G}$:

$$\sigma(f) = \{f^{-1}(B) : B \in \mathcal{G}\}.$$

This is the smallest σ-algebra on Ω with respect to which f is measurable. Note that $\sigma(f) \subseteq \mathcal{F}$ since f is measurable.

Definition B.10. The *Borel σ-algebra* on a topological space $(X, \mathcal{T})$, denoted by $\mathcal{B}(X)$, is the σ-algebra generated by the open sets $\mathcal{T}$. In particular, if (X, d) is a metric space, $\mathcal{B}(X)$ is generated by the open balls.

If $(X, \mathcal{B}(X))$ and $(Y, \mathcal{B}(Y))$ are two measurable spaces equipped with their Borel σ-algebras, then any continuous function $f : X \to Y$ is measurable.

For $X = \mathbb{R}$, $\mathcal{B}(\mathbb{R})$ is the σ-algebra generated by the collection of all open intervals in $\mathbb{R}$. Sets in $\mathcal{B}(\mathbb{R})$ are called *Borel sets*. This is the standard σ-algebra used when the sample space is $\mathbb{R}$ (or $\mathbb{R}^d$).

On the real line, the most important measure is the one that assigns lengths to intervals.

Definition B.11 (Lebesgue Measure)**.** The Lebesgue measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is the unique measure satisfying $\lambda((a, b]) = b - a$ for all intervals $(a, b]$.

The Lebesgue measure is σ-finite since $\mathbb{R} = \bigcup_{n=1}^{\infty}(-n, n]$.

Measurable functions are closed under various operations.

Proposition B.12. *Let $(\Omega, \mathcal{F})$, $(S, \mathcal{G})$, and $(T, \mathcal{H})$ be measurable spaces.*

1. *(**Composition**) If $f : \Omega \to S$ is $\mathcal{F}/\mathcal{G}$-measurable and $g : S \to T$ is $\mathcal{G}/\mathcal{H}$-measurable, then the composition $g \circ f : \Omega \to T$ is $\mathcal{F}/\mathcal{H}$-measurable.*

A measurable function can be used to transport a measure from its domain to its codomain.

Definition B.13 (Push-forward Measure)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, $(S, \mathcal{G})$ a measurable space, and $T : \Omega \to S$ a measurable function. The *push-forward measure* of μ by T, denoted μ^T (or sometimes $T_{\#}\mu$ or $\mu \circ T^{-1}$), is the measure on $(S, \mathcal{G})$ defined by

$$\mu^T(B) = \mu(T^{-1}(B)) \quad \text{for all } B \in \mathcal{G}.$$

Intuitively, μ^T describes the distribution of the random element $T(\omega)$ when ω is distributed according to μ.

The relationship between integrals under the original and push-forward measures is given by the change of variables formula.

Theorem B.14 (Change of Variables Formula)**.** *Let $T : (\Omega, \mathcal{F}, \mu) \to (S, \mathcal{G})$ be measurable. For any measurable function $g : S \to \mathbb{R}$, g is integrable with respect to μ^T if and only if $g \circ T$ is integrable with respect to μ, and*

$$\int_S g(y)\, d\mu^T(y) = \int_\Omega g(T(\omega))\, d\mu(\omega).$$

Definition B.15 (Equivalence Relation)**.** An *equivalence relation* $\sim$ on a set X is a binary relation that satisfies three properties for all $a, b, c \in X$:

1. **Reflexivity:** $a \sim a$.
2. **Symmetry:** If $a \sim b$, then $b \sim a$.
3. **Transitivity:** If $a \sim b$ and $b \sim c$, then $a \sim c$.

Given an equivalence relation $\sim$ on a set X, the *equivalence class* of an element $x \in X$, denoted $[x]$, is the set of all elements in X equivalent to x:

$$[x] = \{y \in X : y \sim x\}.$$

The set of all equivalence classes is called the *quotient set* and denoted by $X/\sim$.

Equivalence relations allow us to define measurable structures on quotient spaces.

Definition B.16 (Quotient σ-algebra)**.** Let (X, Σ) be a measurable space and $\sim$ an equivalence relation on X. The *quotient σ-algebra* on the quotient space $X/\sim$, denoted by $\Sigma/\sim$, is defined as

$$\Sigma/\sim = \left\{B \subseteq X/\sim \;\middle|\; \pi^{-1}(B) \in \Sigma\right\},$$

where $\pi : X \to X/\sim$ is the canonical projection map $\pi(x) = [x]$.

This is the largest σ-algebra on $X/\sim$ making the projection π measurable.

B.2 Integration

B.2.1 The Standard Machinery

A common strategy in measure theory to prove a property $\mathfrak{p}$ for all measurable functions is the so-called "standard machine" or "approximation by simple functions". The steps are typically:

1. **Indicator Functions:** Prove that $\mathfrak{p}$ holds for indicator functions $\mathbb{1}_A$ for all measurable sets A.
2. **Simple Functions:** Extend the result to non-negative simple functions $s = \sum_{i=1}^{n} c_i \mathbb{1}_{A_i}$ by linearity.
3. **Non-negative Measurable Functions:** Use the fact that any non-negative measurable function f is the limit of an increasing sequence of non-negative simple functions $s_n \uparrow f$. Prove that $\mathfrak{p}$ is preserved under this limit (often using the Monotone Convergence Theorem).

4. **General Measurable Functions:** For a general measurable function f, write $f = f^+ - f^-$ where $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$. Extend the result by linearity, provided integrability conditions are met.

Key theorems supporting this machinery include:

Theorem B.17 (Monotone Class Theorem)**.** *Let $\mathcal{A}$ be an algebra of sets generating a σ-algebra $\mathcal{F}$. Let $\mathcal{M}$ be a collection of subsets of Ω that is a monotone class (i.e., closed under countable increasing unions and countable decreasing intersections). If $\mathcal{A} \subseteq \mathcal{M}$, then $\mathcal{F} \subseteq \mathcal{M}$.*

Theorem B.18 (Monotone Convergence Theorem)**.** *If $\{f_n\}$ is a sequence of non-negative measurable functions such that $f_n \uparrow f$ pointwise, then*

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

Lemma B.19 (Fatou's Lemma)**.** *If $\{f_n\}$ is a sequence of non-negative measurable functions, then*

$$\int \liminf_{n\to\infty} f_n \, d\mu \leqslant \liminf_{n\to\infty} \int f_n \, d\mu.$$

Theorem B.20 (Dominated Convergence Theorem)**.** *Let $\{f_n\}$ be a sequence of measurable functions converging pointwise to f. If there exists an integrable function g such that $|f_n| \leqslant g$ for all n, then f is integrable and*

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

The condition of a single dominating function in the DCT can be relaxed to uniform integrability, which controls the integrals of the sequence uniformly over sets of small measure.

Definition B.21 (Uniform Integrability)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. A collection of measurable functions $\{f_i\}_{i\in I}$ is called *uniformly integrable* if

$$\lim_{M\to\infty} \sup_{i\in I} \int_{\{|f_i|>M\}} |f_i| \, d\mu = 0.$$

Equivalently, for every $\varepsilon > 0$, there exists $M > 0$ such that

$$\sup_{i\in I} \int_{\{|f_i|>M\}} |f_i| \, d\mu < \varepsilon.$$

When μ is a finite measure, uniform integrability admits an equivalent characterization in terms of sets of small measure.

Proposition B.22. *Let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space. A collection $\{f_i\}_{i\in I}$ of integrable functions is uniformly integrable if and only if:*

1. $\sup_{i\in I} \int |f_i|\, d\mu < \infty$, *and*
2. *for every $\varepsilon > 0$, there exists $\delta > 0$ such that for all $A \in \mathcal{F}$ with $\mu(A) < \delta$,*

$$\sup_{i\in I} \int_A |f_i|\, d\mu < \varepsilon.$$

Uniform integrability provides a necessary and sufficient condition for L^1 convergence.

Theorem B.23 (Vitali Convergence Theorem)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a finite measure space and let $\{f_n\}$ be a sequence of integrable functions converging in measure to f. Then $f_n \to f$ in $L^1(\mu)$ if and only if $\{f_n\}$ is uniformly integrable.*

B.2.2 Function spaces

Definition B.24 ($\mathcal{L}^p$ spaces)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. For $1 \leqslant p < \infty$, let $\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$ denote the set of all measurable functions $f : \Omega \to \mathbb{R}$ such that

$$\|f\|_p := \left(\int_\Omega |f|^p \, d\mu \right)^{1/p} < \infty.$$

Similarly, $\mathcal{L}^\infty(\Omega, \mathcal{F}, \mu)$ consists of all essentially bounded measurable functions, i.e., those for which there exists a constant C such that $|f(\omega)| \leqslant C$ for almost all ω. The essential supremum is defined as:

$$\|f\|_\infty := \inf\{C \geqslant 0 : |f(\omega)| \leqslant C \text{ for } \mu\text{-almost all } \omega\}.$$

The quantity $\|\cdot\|_p$ satisfies most properties of a norm (non-negativity, homogeneity, triangle inequality), but it is only a *semi-norm* on $\mathcal{L}^p$, because $\|f\|_p = 0$ implies $f = 0$ only almost everywhere (not everywhere). To obtain a Banach space, we must identify functions that are equal almost everywhere.

Definition B.25 (L^p spaces)**.** We define an equivalence relation $\sim$ on $\mathcal{L}^p$ by $f \sim g$ if and only if $f = g$ μ-almost everywhere. The L^p *space* is the quotient space of equivalence classes:

$$L^p(\Omega, \mathcal{F}, \mu) := \mathcal{L}^p(\Omega, \mathcal{F}, \mu)/\sim .$$

Elements of L^p are equivalence classes $[f]$, but it is standard practice to abuse notation and refer to them as functions f.

Equipped with the norm $\|[f]\|_p := \|f\|_p$, the space L^p becomes a Banach space (a complete normed vector space).

Important special cases include:

- $L^p(\mathbb{R}^d)$: When $\Omega = \mathbb{R}^d$ equipped with the Lebesgue measure.
- $L^p([0,1])$: The space of functions on the unit interval square-integrable with respect to Lebesgue measure. This is a standard setting for functional analysis.
- ℓ^p: When μ is the counting measure on $\mathbb{N}$, the space is the set of sequences (x_n) with $\sum |x_n|^p < \infty$.
- $L^2(\mu)$: For $p = 2$, the space is a Hilbert space with inner product $\langle f, g \rangle = \int fg \, d\mu$.

B.2.3 Change of measure

Let μ be a measure on $(\Omega, \mathcal{F})$ and let $f : \Omega \to [0, \infty]$ be a non-negative measurable function. We can define a new measure ν on $(\Omega, \mathcal{F})$ by setting

$$\nu(A) = \int_A f \, d\mu \quad \text{for all } A \in \mathcal{F}.$$

It is a standard exercise in measure theory to verify that ν indeed satisfies the properties of a measure.

Definition B.26 (Probability Density)**.** If the function f is non-negative and the induced measure ν satisfies $\nu(\Omega) = 1$ (i.e., ν is a probability measure), then f is called a *probability density* of ν with respect to the reference measure μ.

The relationship between ν and μ constructed above implies a specific property called absolute continuity.

Definition B.27 (Absolute Continuity)**.** Let ν and μ be two measures on a measurable space $(\Omega, \mathcal{F})$. We say ν is *absolutely continuous* with respect to μ (denoted $\nu \ll \mu$) if for all $A \in \mathcal{F}$,

$$\mu(A) = 0 \implies \nu(A) = 0.$$

The fundamental result connecting these concepts is the Radon-Nikodym theorem, which states that under mild conditions, absolute continuity is sufficient to guarantee the existence of a density.

Theorem B.28 (Radon-Nikodym Theorem)**.** *Let ν and μ be two measures on a measurable space $(\Omega, \mathcal{F})$, and assume that μ is σ-finite. If $\nu \ll \mu$, then there exists a*

non-negative measurable function $f : \Omega \to [0, \infty)$ *such that for all* $A \in \mathcal{F}$,

$$\nu(A) = \int_A f \, d\mu.$$

The function f *is unique up to a set of* μ*-measure zero. We call* f *the* Radon-Nikodym derivative *or* density *of* ν *with respect to* μ*, and denote it by* $f = \frac{d\nu}{d\mu}$.

The next theorem provides a characterization of sufficient statistics (Definition 1.9). The theorem provides the measure-theoretic foundation for the Factorization Theorem (Theorem 1.12) encountered in the main text. Its proof is quite involved and is omitted here, but one can find it in Halmos and Savage 1949.

Theorem B.29 (Halmos–Savage)**.** *Let* $\mathcal{P}$ *be a family of probability measures dominated by a* σ*-finite measure. A statistic* T *is sufficient for* $\mathcal{P}$ *if and only if for all* $P, Q \in \mathcal{P}$*, the likelihood ratio* dP/dQ *admits a* $\sigma(T)$*-measurable version.*

B.3 Joint distributions

B.3.1 Product measures and independence

Given two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, the *product* σ*-algebra*, denoted $\mathcal{F}_1 \otimes \mathcal{F}_2$, is the σ-algebra on $\Omega_1 \times \Omega_2$ generated by the collection of measurable rectangles $\{A \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$.

If μ_1 and μ_2 are σ-finite measures on $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ respectively, there exists a unique measure $\mu = \mu_1 \otimes \mu_2$ on the product space such that

$$\mu(A \times B) = \mu_1(A)\mu_2(B) \quad \text{for all } A \in \mathcal{F}_1, B \in \mathcal{F}_2.$$

Definition B.30 (Independence)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. Two events $A, B \in \mathcal{F}$ are *independent* if $P(A \cap B) = P(A)P(B)$. Two random variables $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ are *independent* if for all $A \in \mathscr{X}$ and $B \in \mathscr{Y}$, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent.

In terms of joint distributions, independence means the joint distribution is the product measure of the marginals. That is, the joint law of (X, Y) is $P_{(X,Y)} = P_X \otimes P_Y$.

Definition B.31 (i.i.d.)**.** A sequence of random variables $X_1, X_2, \ldots, X_n$ is *independent and identically distributed (i.i.d.)* if they are mutually independent and all have the same marginal distribution.

If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$, their joint distribution on the product space $(\mathcal{X}^n, \mathscr{X}^{\otimes n})$ is the product measure $P^{\otimes n}$, defined inductively by $P^{\otimes 1} = P$ and $P^{\otimes(n+1)} = P^{\otimes n} \otimes P$.

B.3.2 Conditional probability and expectation

The definition of conditional probability is based on the concept of conditional expectation.

Definition B.32 (Conditional Expectation)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra, and X an integrable random variable (i.e., $E|X| < \infty$). The *conditional expectation* of X given $\mathcal{G}$, denoted $E[X \mid \mathcal{G}]$, is the equivalence class of $\mathcal{G}$-measurable random variables Z such that

$$\int_G Z \, dP = \int_G X \, dP \quad \text{for all } G \in \mathcal{G}.$$

The existence and uniqueness (up to almost sure equivalence) of Z are guaranteed by the Radon-Nikodym theorem.

Theorem B.33 (Existence and Uniqueness of Conditional Expectation)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra, and X an integrable random variable. Then there exists a unique (up to almost sure equivalence) $\mathcal{G}$-measurable random variable Z such that*

$$\int_G Z \, dP = \int_G X \, dP \quad \textit{for all } G \in \mathcal{G}.$$

With this tool, we can rigorously define the probability of an event given partial information.

Definition B.34 (Conditional Probability)**.** The *conditional probability* of an event $A \in \mathcal{F}$ given a sub-σ-algebra $\mathcal{G}$, denoted $P(A \mid \mathcal{G})$, is defined as the conditional expectation of the indicator function of A:

$$P(A \mid \mathcal{G}) := E[\mathbb{1}_A \mid \mathcal{G}].$$

When conditioning on a random variable Y, we mean conditioning on the σ-algebra generated by Y, i.e., $E[X \mid Y] := E[X \mid \sigma(Y)]$.

Conditional expectations satisfy a generalized version of Bayes' theorem.

Theorem B.35 (Abstract Bayes Formula)**.** *Let P and Q be probability measures on $(\Omega, \mathcal{F})$ such that $P \ll Q$, and let $L = dP/dQ$ be the Radon-Nikodym derivative. For any sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ and any P-integrable random variable f,*

$$E_P[f \mid \mathcal{G}] = \frac{E_Q[fL \mid \mathcal{G}]}{E_Q[L \mid \mathcal{G}]} \quad P\text{-a.s.}$$

Often, we want to view the conditional probability $P(\cdot \mid \mathcal{G})(\omega)$ as a probability measure on $(\Omega, \mathcal{F})$ for each fixed ω. This is not guaranteed by the general definition (due to null sets for each A). However, it is possible in sufficiently "nice" spaces.

A crucial property relating measurability with respect to a random variable and functions of that random variable is given by the Doob-Dynkin Lemma.

Lemma B.36 (Doob–Dynkin Lemma)**.** *Let $(S, \mathcal{S})$ be a standard Borel space and let $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ be measurable. Then a random variable $Y : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is $\sigma(X)$-measurable if and only if there exists a measurable function $g : S \to \mathbb{R}$ such that $Y = g(X)$.*

This lemma implies that $E[Z \mid X] = g(X)$ for some measurable function g. Specifically, if Y is $\sigma(X)$-measurable, it is a function of X.

Under certain conditions, conditional probabilities can be realized as a kernel that is a measure for each fixed ω.

Definition B.37 (Regular Conditional Probability)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. A *regular conditional probability* is a function $\kappa : \Omega \times \mathcal{F} \to [0, 1]$ such that:

1. For each $\omega \in \Omega$, $\kappa(\omega, \cdot)$ is a probability measure on $(\Omega, \mathcal{F})$.
2. For each $A \in \mathcal{F}$, $\omega \mapsto \kappa(\omega, A)$ is a version of $P(A \mid \mathcal{G})$.

Regular conditional probabilities are guaranteed to exist when Ω is a standard Borel space (e.g. a Polish space (see Definition A.17 in Appendix A) equipped with its Borel σ-algebra).

Theorem B.38 (Existence of Regular Conditional Probabilities)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space where Ω is a Polish space and $\mathcal{F} = \mathcal{B}(\Omega)$ is its Borel σ-algebra. For any sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, there exists a regular conditional probability given $\mathcal{G}$.*

A related concept is the Markov kernel, which generalizes the idea of a transition matrix.

Definition B.39 (Markov Kernel)**.** Let $(X, \mathscr{X})$ and $(Y, \mathscr{Y})$ be measurable spaces. A *Markov kernel* (or probability kernel) from $(X, \mathscr{X})$ to $(Y, \mathscr{Y})$ is a function $K : X \times \mathscr{Y} \to [0, 1]$ such that:

1. For each $x \in X$, the map $B \mapsto K(x, B)$ is a probability measure on $(Y, \mathscr{Y})$.
2. For each $B \in \mathscr{Y}$, the map $x \mapsto K(x, B)$ is $\mathscr{X}$-measurable.

Markov kernels are used to model random mappings where the output distribution depends on the input, such as in conditional distributions $P(Y \in B \mid X = x)$.

Finally, we state the version of Bayes' rule for densities, which is the most common form used in statistical inference.

Theorem B.40 (Bayes' Rule for Densities)**.** *Let Θ and $\mathcal{X}$ be random variables taking values in measurable spaces $(\Omega_\Theta, \mathscr{F}_\Theta)$ and $(\Omega_\mathcal{X}, \mathscr{F}_\mathcal{X})$, respectively. Suppose the joint distribution of $(\Theta, \mathcal{X})$ is dominated by a product measure $\nu \otimes \mu$, with joint density $p(\theta, x)$. Then the conditional distribution of Θ given $\mathcal{X} = x$ has density (with respect to ν):*

$$p(\theta \mid x) = \frac{p(\theta, x)}{\int_{\Omega_\Theta} p(\vartheta, x)\, d\nu(\vartheta)},$$

provided the denominator is positive and finite. In the common case where $p(\theta, x) = p(x \mid \theta)\pi(\theta)$ (likelihood $\times$ prior), this becomes the familiar form:

$$p(\theta \mid x) = \frac{p(x \mid \theta)\pi(\theta)}{\int p(x \mid \vartheta)\pi(\vartheta)\, d\nu(\vartheta)}.$$

B.4 Concentration of measure

Lemma B.41 (Jensen's Inequality)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $X \in L^1(P)$ be real-valued, and let $\varphi : \mathbb{R} \to \mathbb{R}$ be convex such that $\mathbb{E}|\varphi(X)| < \infty$. Then*

$$\varphi(\mathbb{E}[X]) \leqslant \mathbb{E}[\varphi(X)].$$

If φ is strictly convex, then equality holds if and only if X is constant P-a.s. Moreover, for any sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$,

$$\varphi(\mathbb{E}[X \mid \mathcal{G}]) \leqslant \mathbb{E}[\varphi(X) \mid \mathcal{G}] \qquad P\text{-a.s.}$$

Lemma B.42 (Markov's inequality)**.** *If $X \geqslant 0$, then for any $a > 0$,*

$$\mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a}.$$

Proof. Note that $a \cdot \mathbb{1}X \geqslant a \leqslant X$. Taking expectations gives $a \cdot \mathbb{P}(X \geqslant a) \leqslant \mathbb{E}[X]$. □

The following concentration inequalities are immediate consequences.

Lemma B.43 (Chebyshev's inequality)**.** *If $\mathrm{Var}(X) < \infty$, then for any $k > 0$,*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant k) \leqslant \frac{\mathrm{Var}(X)}{k^2}.$$

Proof. Apply Markov's inequality to $(X - \mathbb{E}[X])^2$ with threshold k^2. □

Lemma B.44 (Chernoff's bound)**.** *For any random variable X and any $a \in \mathbb{R}$,*

$$\mathbb{P}(X \geqslant a) \leqslant \inf_{t>0} e^{-ta}\mathbb{E}[e^{tX}].$$

Proof. For any $t > 0$, the event $\{X \geqslant a\}$ implies $\{e^{tX} \geqslant e^{ta}\}$. Apply Markov's inequality to e^{tX} and take the infimum over $t > 0$. □

B.5 Transforms

Definition B.45 (Laplace Transform)**.** Let μ be a finite measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. The *Laplace transform* of μ is the function $\psi : \mathbb{R}^k \to \mathbb{R}$ defined by

$$\psi(t) = \int_{\mathbb{R}^k} e^{\langle t,x \rangle} \, d\mu(x),$$

provided the integral exists.

The Laplace transform is a powerful tool for characterizing measures. A key property is its uniqueness:

Theorem B.46 (Uniqueness of Laplace Transform)**.** *Let μ and ν be two finite measures on $\mathbb{R}^k$. If their Laplace transforms agree on an open set containing the origin, then $\mu = \nu$.*

Proof Sketch. We sketch the argument for $k = 1$ and compact support. Suppose μ and ν are supported on a compact interval $[a, b]$. The Laplace transform condition implies

$$\int_a^b e^{tx} \, d\mu(x) = \int_a^b e^{tx} \, d\nu(x)$$

for all t in a neighborhood of 0. By analyticity, this equality extends to all $t \in \mathbb{R}$. By linearity,

$$\int_a^b P(e^x) \, d\mu(x) = \int_a^b P(e^x) \, d\nu(x)$$

for any polynomial P. The algebra of functions of the form $x \mapsto P(e^x)$ separates points on $[a, b]$ and vanishes at no point. By the Stone-Weierstrass theorem, such functions are dense in the space of continuous functions $C([a, b])$ with respect to the uniform norm.

Thus, for any continuous function f, $\int f \, d\mu = \int f \, d\nu$. Since measures on Borel σ-algebras are determined by their integrals against continuous functions (Riesz Representation Theorem), we conclude $\mu = \nu$. The extension to non-compact support requires more careful analysis involving truncation or compactification, but the core idea remains the density of exponential families in function spaces. □

This uniqueness property extends to signed measures. If μ is a signed measure with $\int e^{\langle t,x \rangle} \, d\mu(x) = 0$ for all t in an open set, then μ is the zero measure. This fact is crucial for proving completeness of exponential families.

Another important transform is the characteristic function, which similarly provides a powerful tool for characterizing measures.

Definition B.47 (Characteristic Function)**.** Let μ be a finite measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. The *characteristic function* of μ is the function $\phi : \mathbb{R}^k \to \mathbb{C}$ defined by

$$\phi(t) = \int_{\mathbb{R}^k} e^{i\langle t,x\rangle}\, d\mu(x),$$

where $i = \sqrt{-1}$.

Unlike the Laplace transform, the characteristic function is always defined for any finite measure (since $|e^{i\langle t,x\rangle}| = 1$ is bounded). It also uniquely determines the measure.

Theorem B.48 (Uniqueness of Characteristic Functions)**.** *Let μ and ν be two finite measures on $\mathbb{R}^k$. If their characteristic functions agree, i.e., $\phi_\mu(t) = \phi_\nu(t)$ for all $t \in \mathbb{R}^k$, then $\mu = \nu$.*

This theorem is a direct consequence of the Fourier Inversion Theorem. Since the characteristic function is essentially the Fourier transform of the measure, and the Fourier transform is injective, the measure is uniquely determined.

Bibliography

Berger, James O (2013). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Ferguson, Thomas S. (1967). *Mathematical statistics: A decision theoretic approach*. Vol. 1. Probability and Mathematical Statistics. New York: Academic Press.

Folland, Gerald B. (2016). *A Course in Abstract Harmonic Analysis*. 2nd. Textbooks in Mathematics. Boca Raton, FL: CRC Press. ISBN: 978-1-4987-2713-6.

Halmos, Paul R. and Leonard J. Savage (1949). "Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics". In: *Annals of Mathematical Statistics* 20.2, pp. 225–241. DOI: 10.1214/aoms/1177730032.

Le Cam, Lucien and Grace Lo Yang (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. New York, NY: Springer-Verlag. ISBN: 978-0-387-96307-5.

Lehmann, E. L. and Joseph P. Romano (2005). *Testing Statistical Hypotheses*. 3rd. Springer Texts in Statistics. New York: Springer.

Lehmann, Erich L and George Casella (2006). *Theory of Point Estimation*. Springer Science & Business Media.

Regazzini, Eugenio (2013). "The Origins of de Finetti's Critique of Countable Additivity". In: *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*. Ed. by Galin Jones and Xiaotong Shen. Vol. 10. IMS Collections. Institute of Mathematical Statistics, pp. 63–82. DOI: 10.1214/12-IMSCOLL1204. URL: https://doi.org/10.1214/12-IMSCOLL1204.