Statistics 210A

Introduction to Theoretical Statistics

Notes

Druv Pai

Contents

1 Measure Theory

Most statisticians learn probability at an intuitive level and go their whole career by working with the "common sense" rules. But this brings up weird contradictions; it doesn't make sense to talk about the probability of landing in some sets (some sets are non-measurable); also, the probability of an event conditioned on an event which happens with probability zero is not well-defined. *Measure theory* gives us a way to rigorously approach these problems.

Suppose we have a Gaussian $X \sim \mathcal{N}(0,1)$ and $\mathbb{E}[f(x)] = \int_X f(x)\phi(x)\,\mathrm{d}x$, where $\phi(x)$ is the Gaussian probability density function. But if $X \sim \mathrm{Binom}(10, 1/2)$, then $\mathbb{E}[f(x)] = \sum_{x=0}^{10} f(x)p(x)$, where $p(x)$ is the probability mass function of the binomial distribution. These conflicting notations are annoying; if $X \sim \mathcal{N}(0,1)$ and $Z = \max\{0, X\}$, then the expectation of Z requires both a sum and an integral, and more complex scenarios can add to the confusion.

The notation that we use is $\mathbb{E}[f(z)] = \int_Z f(z)\,\mathrm{d}\mathbb{P}(z)$., and we compute this integral via a Lebesgue integral or a sum. We'll talk about how to make sense of this notation.

Measure Theory Basics

Measure theory simplifies notation and clarifies a lot of concepts (integration, conditioning) for probability.

Definition 1.1 (Measure, Informal). Given a set X, a measure μ maps subsets $A \subseteq X$ to non-negative numbers $\mu(A) \in [0, \infty]$.

Example 1.2 (Counting Measure). If X is countable (for example $X = \mathbb{Z}$), we can define the counting measure $\#(A)$ as the number of points in A.

Example 1.3 (Lebesgue Measure). If $X = \mathbb{R}^n$, then the Lebesgue measure of a set is just the "usual" integral:

$$\lambda(A) = \int_A \mathrm{d}x = \mathrm{Vol}(A)$$

For all the sets that matter, essentially, the integral can be Riemann integral or Lebesgue integral. There are some sets that do not correspond to this definition, but they are not pathological.

Definition 1.4 (Standard Gaussian Distribution). The probability density function of the standard Gaussian distribution is

$$\phi(x) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_i x_i^2\right)$$

and so, if $Z \sim \mathcal{N}(0, I_n)$, then we can define a measure under the density $\mathcal{P}$ with

$$\mathcal{P}(A) = \int_A \phi(x)\,\mathrm{d}x = \mathbb{P}(Z \in A)$$

In general, the domain of a measure μ on a set X may not be all subsets of X. Define 2^X as the power set of X (the set of all subsets of X). Instead, the domain is a collection $\mathcal{F}$ of subsets ($\mathcal{F} \subseteq 2^X$) that is a **$\sigma$-field**, meaning that it satisfies certain closure properties (contains $\emptyset$ and closed under countable union and complement).

Example 1.5. If X is countable, then $\mathcal{F} = 2^X$.

Example 1.6. If $X = \mathbb{R}^n$, then $\mathcal{F}$ is the Borel σ-field $\mathcal{B}$.

Definition 1.7 (Borel σ-Field). The **Borel σ-field** is the smallest σ-field that contains all open rectangles $(a_1, b_1) \times \cdots \times (a_n, b_n)$ for all $a_i < b_i$.

Definition 1.8 (Measure). Given a **measurable space** $(X, \mathcal{F})$, a **measure** is a map $\mu\colon \mathcal{F} \to [0, \infty]$ which follows the property

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \text{ for disjoint } A_1, A_2, \ldots$$

Definition 1.9 (Probability Measure). μ is called a **probability measure** if $\mu(X) = 1$.

Measure let us define *integrals* that put "weight" $\mu(A)$ on $A \subseteq X$. Once we know what a measure is, we are close to knowing what an integral is.

We first define $\int_X 1\{x \in A\}\,\mathrm{d}\mu(x) = \mu(A)$. This tells us how to integrate $c1\{x \in A\}$ for $c \in \mathbb{R}$. and more generally $\sum_{i=1}^{\infty} 1\{x \in A_i\}\,\mathrm{d}\mu(x)$ for $A_i \in \mathcal{F}$. Indeed, for $A_i \in \mathcal{F}$ and $c_i \in \mathbb{R}$

$$\int_X \sum_{i=1}^{\infty} c_i 1\{x \in A_i\}\,\mathrm{d}\mu(x) = \sum_{i=1}^{\infty} c_i \mu(A_i).$$

For general (measurable) functions, we just approximate via these kinds of "simple" functions. Suppose $f\colon X \to \mathbb{R}$ is "nice enough" (measurable), we approximate f by taking limits and requiring the linearity of the integral, that is, $\int_X (af+g)\,\mathrm{d}\mu = a\int_X f\,\mathrm{d}\mu + \int_X g\,\mathrm{d}\mu$.

Example 1.10 (Counting Integral). If $\#$ is the counting measure, then

$$\int_X f\,\mathrm{d}\# = \sum_{x\in X} f(x).$$

Example 1.11 (Lebesgue Integral). If λ is the Lebesgue measure, then

$$\int_X f\,\mathrm{d}\lambda = \int_X f(x)\,\mathrm{d}x\,.$$

Example 1.12 (Gaussian Integral). If $\mathcal{P}$ is the probability measure associated with the Gaussian distribution, then

$$\int_X f\,\mathrm{d}\mathcal{P} = \int_X f(x)\phi(x)\,\mathrm{d}x\,.$$

This is all stuff we know how to compute, but this framework allows us to unify all these operations under a common notation.

Densities

The measures λ (Lebesgue) and $\mathcal{P}$ (Gaussian) are closely related to each other.

Definition 1.13 (Absolute Continuity of Measure). Given $(X,\mathcal{F})$ and two measures P,μ, we say P is **absolutely continuous** with respect to μ ($P \ll \mu$, or μ *dominates* P) if $P(A)=0$ whenever $\mu(A)=0$.

Definition 1.14 (Density Function). If $P\ll\mu$ then (under mild conditions) we can always define a **density function** $p\colon X\to(0,\infty)$ with

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

Sometimes it's written $p(x) = \frac{\mathrm{d}P}{\mathrm{d}\mu}(x)$, and called the **Radon-Nikodym derivative**.

For this density function we have

$$\int_X f(x)\,\mathrm{d}P(x) = \int_X f(x)p(x)\,\mathrm{d}\mu(x)\,.$$

Note that P cannot be counting measure and μ the Lebesgue measure, because then absolute continuity fails, but it can definitely be the other way around.

Generally, densities allow us to write the equivalence $\int f\,\mathrm{d}P = \int fp\,\mathrm{d}\mu$, where the latter is possible to compute.

Densities are helpful for two reasons:

- The main way we calculate $\int f\,\mathrm{d}P$ is to replace with either
  - $\int_X f(x)p(x)\,\mathrm{d}x$ if $P\ll\lambda$, then we say P is **absolutely continuous** and p is the **probability density function (p.d.f.)**
  - $\sum_{x\in X} f(x)p(x)$ if $p\ll\#$, then we say that P is a **discrete distribution**, p is the **probability mass function (p.m.f.)**
- The main way we define probability distributions is to start from a base measure like λ or $\#$ and add a density. For example, the standard Gaussian is defined as the distribution with p.d.f. $\phi(x)$ with respect to the Lebesgue measure.

Probability Spaces and Random Variables

Imagine some abstract *outcome* ω which encodes (in a way that's not specified) all of the random variables in the problem. Let $(\Omega,\mathcal{F},\mathbb{P})$ be a *probability space*, where

- $\omega\in\Omega$ is called an **outcome**
- $A\subseteq\mathcal{F}$ is called an **event**
- $\mathbb{P}(A)$ is called the **probability of** A.

Definition 1.15 (Random Variable). A random variable is a function $X : \Omega \to \mathcal{X}$. We say X has distribution Q ($X \sim Q$) if

$$\mathbb{P}(X \in B) = \mathbb{P}(\{\omega : X(\omega) \in B\}) = Q(B).$$

More generally, we can write events involving many random variables (i.e. $\mathbb{P}(X > Y) = \mathbb{P}(\{\omega : X(\omega) > Y(\omega)\})$).

Definition 1.16 (Expectation). The **expectation** is the integral with respect to $\mathbb{P}$:

$$\mathbb{E}[f(X)] = \int_\Omega f(X(\omega)) \, \mathrm{d}\mathbb{P}(\omega) \, .$$

We can also write

$$\mathbb{E}[f(X, Y)] = \int_\Omega f(X(\omega), Y(\omega)) \, \mathrm{d}\mathbb{P}(\omega) \, .$$

and extend to as many random variables we want.

Definition 1.17. If $\mathbb{P}(A) = 1$ we say A occurs **almost surely**.

2 Statistical Decision Theory

Definition 2.1 (Statistical Model). A **statistical model** is a family of candidate probability distributions $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ for some random variable $X \sim \mathbb{P}_\theta$. Then $X \in \mathcal{X}$ is called the **data** and is observed, and $\theta \in \Theta$ is called the **parameter** and is unobserved.

Note that in this case θ could be infinite dimensional (i.e. represents a density function). For now we consider θ fixed and known.

The goal of estimation is

$$\text{observe } X \sim \mathbb{P}_\theta \implies \text{guess the value of some } \textit{estimand } g(\theta)$$

Example 2.2 (3.1 in Keener). We flip a biased coin n heads with $\theta \in [0, 1]$ the probability of heads. Let X be the number of heads after n flips (assuming flips are independent). Then $X \sim \mathrm{Binom}(n, \theta)$. Then the probability mass function (the density w.r.t. $\#$ on the set $\mathcal{X} = \{0, 1, \dots, n\}$) is $p_\theta(x) = \theta^x (1 - \theta)^{n-x} \binom{n}{x}$. We want to find θ.

Definition 2.3 (Statistic). A **statistic** is any function $T(X)$ of the data X (not a function of X and θ).

Definition 2.4 (Estimator). An **estimator** $\delta(x)$ of $g(\theta)$ is a statistic which is intended to guess $g(\theta)$.

Example (3.1 in Keener, continued). A natural estimator is $\delta_0(X) = \frac{X}{n}$.

Question. Is this a good estimator?

Question. Is there a better estimator?

There are two notions of an estimator that are important to distinguish.

Loss and Risk

Definition 2.5 (Loss Function). A **loss function** $L(\theta, d)$ measures how bad an estimate is.

Example 2.6 (Squared Error Loss). The **squared error loss** is $L(\theta, d) = (d - g(\theta))^2$.

Typical properties we want the loss to have are

- $L(\theta, d) \geq 0$ for all θ, d – nonnegativity
- $L(\theta, g(\theta)) = 0$ for all θ – getting it right doesn't increase the loss.

Definition 2.7 (Risk Function). The **risk function** is the expected loss (**risk**) as a function of θ, for an estimator δ.

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

In this case, the subscript θ tells us which parameter is in effect, *not* "what randomness to integrate over."

Remark. The Bayes risk also integrates over the distribution of θ, but only once we've defined a distribution of θ.

Definition 2.8 (Mean-Squared Error). If L is the squared-error loss then R is the **mean-squared error** (MSE).

$$\mathrm{MSE}(\theta; \delta) = \mathbb{E}_\theta\left[(g(\theta) - \delta_0(X))^2\right]$$

Example (3.1 in Keener, continued). Let $\delta_0(X) = \frac{X}{n}$. Then $\mathbb{E}_\theta[\delta_0(X)] = \mathbb{E}_\theta\left[\frac{X}{n}\right] = \theta$. Then $\mathrm{MSE}(\theta; \delta_0) = \mathrm{Var}_\theta\left(\frac{X}{n}\right) = \frac{\theta(1-\theta)}{n}$.
On a plot with θ on the x-axis and the MSE on the y axis, the curve $R(\theta; \delta_0)$ is the downward-facing parabola with optimum $\left(\frac{1}{2}, \frac{1}{4n}\right)$.
Other (bad) choices of estimators would be $\delta_1(x) = \frac{x+3}{n}$, for which the curve $R(\theta; \delta_1)$ is a downward-facing parabola higher than $R(\theta; \delta_0)$ at each point, so that δ_1 is *inadmissible* (dominated by δ_0). And $\delta_2(x) = \frac{x+3}{n+6}$, which is a broader parabola. Indeed, if θ is close to $\frac{1}{2}$, then δ_2 is better estimator than δ_0, but otherwise δ_0 is better than δ_2.

Definition 2.9 (Inadmissible Estimator). An estimator δ is **inadmissable** if there exists another estimator δ^* with

- $R(\theta; \delta^*) \le R(\theta; \delta)$ for all $\theta \in \Theta$.
- there exists $\theta \in \Theta$ such that $R(\theta; \delta^*) < R(\theta; \delta)$.

Example (3.1 in Keener, continued). Suppose that $\delta_3(x) = \frac{2}{3}$ (no matter what X is). If $\theta = \frac{2}{3}$, then we can't possibly beat this.

We will discuss several strategies to resolve this ambiguity.

1. Summarize the risk function by a scalar.
   a) **Average-case risk**: minimize
   $$\int_\Theta R(\theta; \delta)\, \mathrm{d}\Lambda(\theta)$$
   with respect to some measure Θ. This is exactly the Bayes estimator, and Λ is called the prior.
   b) **Worst-case risk**: minimize
   $$\sup_{\theta \in \Theta} R(\theta; \delta).$$
   This is the minimax estimator, and is closely related t o the Bayes estimator.
2. Constrain the choice of estimator.
   a) Only consider unbiased estimators δ (such that $\mathbb{E}_\theta[\delta(X)] = g(\theta)$ for all $\theta \in \Theta$). This rules out the constant estimator, $\delta_1, \delta_2, \delta_3$, and δ_0 is the "best" unbiased estimator in some sense we will explore in the future.

3 Exponential Families

A lot of the models (Gaussian, Poisson, Binomial) we're interested in are known as *exponential families* and so if we prove something about exponential families it applies to many distributions. Also, exponential families are very simple objects that there are a lot of results for.

Definition 3.1 (Exponential Family). An **s-parameter exponential family** is a model $\mathcal{P} = \{P_\eta : \eta \in \Xi\}$ with densities p_η with respect to a common measure μ on $\mathcal{X}$ of the form
$$p_\eta(x) = \mathrm{e}^{\eta' T(x) - A(\eta)} h(x)$$
where

- $T : \mathcal{X} \to \mathbb{R}^s$ is a **sufficient statistic**
- $h : \mathcal{X} \to [0, \infty)$ is the **carrier or base density**
- $\eta \in \Xi \subseteq \mathbb{R}^s$ is the **natural parameter**
- $A : \mathbb{R}^s \to \mathbb{R}$ is the **cumulant-generating function (c.g.f.)** or **normalizing constant**

Remark 3.2. The c.g.f. A is totally determined by h and T, since we always have $\int_X p_\eta \, \mathrm{d}\mu = 1$ for all η. In particular,
$$A(\eta) = \log\left(\int_{\mathcal{X}} \mathrm{e}^{\eta' T(x)} h(x)\, \mathrm{d}\mu(x)\right).$$

Definition 3.3 (Normalizability, Natural Parameter Space, Canonical Form). p_η is **normalizable** if $A(\eta) < \infty$. The **natural parameter space** is the set of all allowable (normalizable) η:

$$\Xi_1 = \{\eta \colon A(\eta) < \infty\}.$$

We say $\mathcal{P}$ is in **canonical form** if $\Xi = \Xi_1$.

Since $A(\eta)$ is convex, Ξ_1 is convex (from Homework 1, Problem 2).
It is often convenient to use a different parameterization. Indeed, suppose $\eta = \eta(\theta)$ and $B(\theta) = A(\eta(\theta))$. Then

$$p_\theta(x) = \mathrm{e}^{\eta(\theta)'T(x) - B(\theta)} h(x).$$

Example 3.4. Suppose $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Let $\theta = \left(\mu, \sigma^2\right)$. Then

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \exp\left(\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log\left(2\pi\sigma^2\right)\right) \end{aligned}$$

and we can see that p_θ is an exponential family. We have

- $T(x) = \left(x, x^2\right)$,
- $h(x) = 1$,
- $\mu(\theta) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$,
- $B(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2} \log\left(2\pi\sigma^2\right)$.

In terms of η:

$$p_\eta(x) = \mathrm{e}^{\eta' \begin{bmatrix} x \\ x^2 \end{bmatrix} - A(\eta)}$$

where $A(\eta) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2} \log\left(-\frac{\pi}{\eta_2}\right)$.

It is clear that there are many ways to re-parameterize the members of an exponential family, i.e. by invertible matrices or just constant factors on T and η.
The set of exponential families is a subset of a hyperplane in the space of log-densities. Assume without loss of generality that $h(x) > 0$ (in fact by changing the base measure we can fix $h(x) = 1$ anywhere). Then $f_\eta(x) = \log(p_\eta(x)) = \log(h(x)) + \eta' T(x)$. Then $\mathrm{e}^{f_\eta(x)}$ is a density (maybe not normalizable). In this form we can construct a hyperplane by choosing basis vectors which are the T_i, i.e. the hyperplane is $\operatorname{span}(T_1, \ldots, T_s)$. The set of all normalizable densities in this hyperplane represents Ξ_1 in the sense that this set is $f_{\Xi_1} = \{f_\eta \colon \eta \in \Xi_1\}$..
The functional form of p_η is not unique. We can always

- Reformulate to $h(x) = 1$: set $\mu \to \widetilde{\mu}$ and $h \to \widetilde{h} \equiv 1$ with $\dot{\widetilde{\mu}} = h$.
- Reparameterize so $0 \in \Xi_1$: Take some $\eta_0 \in \Xi$ and for $eta \in \Xi$, set $\eta \to \widetilde{\eta} = \eta - \eta_0$, $h \to \widetilde{h} = p_{\eta_0} = p_{\widetilde{\eta}=0}$, and $A \to \widetilde{A}(\widetilde{\eta}) = A(\eta_0 + \widetilde{\eta}) - A(\eta_0)$, so without loss of generality $h(x) = p_0(x)$.
- For $c \in \mathbb{R}^s$ and invertible $D \in \mathbb{R}^{s \times s}$, set $T \to \widetilde{T}(x) = c + DT(x)$, $\eta \to \widetilde{\eta} = \left(D^{-1}\right)' \eta$, and $A \to \widetilde{A}$ as appropriate.

The Gaussian distribution is the *only* exponential form with base measure $h(x) = 1$ and sufficient statistics in $\operatorname{span}\left(x, x^2\right) \setminus \left\{\operatorname{span}(x) \cup \operatorname{span}\left(x^2\right)\right\}$ in functional space (any pair of nontrivial nondegenerate linear combinations in x and x^2 work).
We can interpret this in *exponential tilting*. This process is

- Start with carrier density $h(x)$
- Apply **exponential tilt**:
  1. multiply $h(x)$ by $\mathrm{e}^{\eta' T(x)}$
  2. re-normalize to get the probability density

The exponential family in canonical form is all normalizable exponential tilts of $h(x)$ by linear combinations of $T_1(x), \dots, T_s(x)$.

Example 3.5. Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. We claim that the distribution is still an exponential family and the number of sufficient statistics is still 2. Write

$$\begin{aligned} p_\theta(x) &= \prod_{i=1}^n p_\theta^{(i)}(x_i) \\ &= \exp\left(\sum_{i=1}^n \frac{\mu}{\sigma^2} x_i - \frac{1}{2\sigma^2} x_i^2 - \left(\frac{\mu}{\sigma^2} + \frac{1}{2}\log\left(2\pi\sigma^2\right)\right)\right) \\ &= \exp\left(\frac{\mu}{\sigma^2}\left(\sum_{i=1}^n x_i\right) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2\right) - \frac{n}{2}\log\left(2\pi\sigma^2\right)\right) \end{aligned}$$

is an exponential family, where

- $\eta(\theta) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$
- $T(x) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)$
- $B\left(\mu, \sigma^2\right) = nB^{(i)}\left(\mu, \sigma^2\right)$.

More generally, say $X_1, \dots, X_n \sim p_\eta^{(i)}(x) = \mathrm{e}^{\eta' T(x) - A(\nu)} h(x)$. Then

$$\begin{aligned} X &\sim \prod_{i=1}^n p_\eta^{(i)}(x_i) \\ &= \prod_{i=1}^n \mathrm{e}^{\eta' T(x_i) - A(\eta)} h(x_i) \\ &= \exp\left(\eta' \underbrace{\sum_{i=1}^n T(x_i)}_{\text{sufficient statistic}} - \underbrace{nA(\eta)}_{\text{c.g.f.}}\right) \underbrace{\prod_{i=1}^n h(x_i)}_{\text{carrier}} \end{aligned}$$

Example 3.6 (Binomial). Suppose $X \sim \mathrm{Binom}(n, \theta)$. Then, with respect to the counting measure $\#$ on $\{0, \dots, n\}$, the density of X is

$$\begin{aligned} p_\theta(x) &= \theta^x (1-\theta)^{n-x} \binom{n}{x} \\ &= \left(\frac{\theta}{1-\theta}\right)^x (1-\theta)^n \binom{n}{x} \\ &= \exp\left(\underbrace{x}_{\text{sufficient statistic}} \underbrace{\log\left(\frac{\theta}{1-\theta}\right)}_{\text{natural parameter}} - \underbrace{n\log(1-\theta)}_{\text{c.g.f.}}\right) \underbrace{\binom{n}{x}}_{\text{carrier}} \end{aligned}$$

Therefore $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$; this is usually called the "log-odds ratio."

All the functions on the exponential family are mutually absolutely continuous, so $\theta \in \{0, 1\}$ does not lead to densities in the exponential family.

Example 3.7 (Poisson). Suppose $X \sim \mathrm{Pois}(\lambda)$. Then, with respect to the counting measure $\#$ on $\mathbb{N}$, the density of X is

$$\begin{aligned} p_\lambda(x) &= \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} \\ &= \mathrm{e}^{\log(\lambda)x - \lambda} \cdot \frac{1}{x!} \\ \eta(\lambda) &= \log(\lambda) \end{aligned}$$

with $\lambda = \mathbb{E}_\lambda[x]$.

Differential Identities

Exponential families have many nice properties, some of which we can find via derivatives. Write

$$\mathrm{e}^{A(\eta)} = \int_{\mathcal{X}} \mathrm{e}^{\eta' T(x)} h(x) \,\mathrm{d}\mu(x) \tag{*}$$

Take the derivative on both sides and pass the derivative inside the integral, to obtain

$$\frac{\mathrm{d}}{\mathrm{d}\eta} \mathrm{e}^{A(\eta)} = \frac{\mathrm{d}^\eta}{\mathrm{d} \int^\eta_{\mathcal{X}}} \mathrm{e}^{\eta' T(x)} h(x) \,\mathrm{d}\mu(x)$$

Theorem 3.8 (Keener Theorem 2.4). For $f \colon \mathcal{X} \to \mathbb{R}$ let

$$\Xi_f = \left\{ \eta \in \mathbb{R}^s \colon \int_{\mathcal{X}} |f| \mathrm{e}^{\eta' T} h \,\mathrm{d}\mu < \infty \right\}.$$

Then the function

$$g(\eta) = \int_{\mathcal{X}} f \mathrm{e}^{\eta' T} h \,\mathrm{d}\mu$$

has continuous partial derivatives of all orders for $\eta \in \Xi_f^\circ$ and we can differentiating under the integral sign.

Corollary 3.9. On Ξ_1°, $A(\eta)$ has all partial derivatives.

Returning to our differentiation exercise, we obtain

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\eta} \mathrm{e}^{A(\eta)} &= \frac{\mathrm{d}^\eta}{\mathrm{d} \int^\eta_{\mathcal{X}}} \mathrm{e}^{\eta' T(x)} h(x) \,\mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \frac{\mathrm{d}}{\mathrm{d}\eta} \mathrm{e}^{\eta' T(x)} h(x) \,\mathrm{d}\mu(x) \\
\mathrm{e}^{A(\eta)} \frac{\partial A}{\partial \eta_i}(\eta) &= \int_{\mathcal{X}} T_j(x) \mathrm{e}^{\eta' T(x)} h(x) \,\mathrm{d}\mu(x) \\
\frac{\partial A}{\partial \eta_1}(\eta) &= \int_{\mathcal{X}} T_j \mathrm{e}^{\eta' T(x) - A(\eta)} h(x) \,\mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} T_j(x) \,\mathrm{d}\mathbb{P}(x) \\
&= \mathbb{E}_\eta[T_j(X)] \\
\boldsymbol{\nabla}_\eta A(\eta) &= \mathbb{E}_\eta[T(X)].
\end{aligned}$$

Differentiating again, we obtain

$$\begin{aligned}
\frac{\partial}{\partial \eta_i \partial \eta_k} \mathrm{e}^{A(\eta)} &= \frac{\partial}{\partial \eta_i \partial \eta_k} \int_{\mathcal{X}} \mathrm{e}^{\eta' T} h \,\mathrm{d}\mu \\
\mathrm{e}^{A(\eta)} \left(\frac{\partial^2 A}{\partial \eta_i \partial \eta_k} + \frac{\partial A}{\partial \eta_i} \frac{\partial A}{\partial \eta_k} \right) &= \int_{\mathcal{X}} T_j T_k \mathrm{e}^{\eta' T - A(\eta)} h \,\mathrm{d}\mu \\
\frac{\partial^2 A}{\partial \eta_i \partial \eta_k} + \mathbb{E}_\eta[T_j(X)] \mathbb{E}_\eta[T_k(X)] &= \int_{\mathcal{X}} T_j T_k \,\mathrm{d}\mathbb{P}(x) \\
\frac{\partial^2 A}{\partial \eta_i \partial \eta_k}(\eta) &= \mathbb{E}_\eta[T_j(X) T_k(X)] - \mathbb{E}_\eta[T_i(X)] \mathbb{E}_\eta[T_k(X)] \\
&= \operatorname{Cov}(T_j(X), T_k(X)) \\
\boldsymbol{\nabla}^2_\eta A(\eta) &= \operatorname{Var}_\eta(T(X)) \in \mathbb{R}^{s \times s}
\end{aligned}$$

Example 3.10. If $X \sim \operatorname{Pois}(\lambda)$, then setting $T(x) = x$ and $\eta(\lambda) = \log(\lambda)$, then $B(\lambda) = \lambda$ and $A(\eta) = \mathrm{e}^\eta$. Then $\mathbb{E}_\eta[X] = \frac{\mathrm{d}}{\mathrm{d}\eta} \mathrm{e}^\eta = \mathrm{e}^\eta = \lambda$, and $\operatorname{Var}_\eta(X) = \frac{\mathrm{d}^2}{\mathrm{d}\eta^2} \mathrm{e}^\eta = \mathrm{e}^\eta = \lambda$.

This is we we call $A(\eta)$ the cumulant generating function. But technically, it would be more accurate to call $c(\lambda) = A(\eta + \lambda) - A(\eta)$ to be the cumulant generating function (generated by p_η).

Moment-Generating Functions

Differentiating $(*)$ repeatedly, we get

$$\mathrm{e}^{-A(\eta)} \frac{\partial^{\sum_{i=1}^s k_i}}{\partial \eta_1^{k_1} \cdots \partial \eta_s^{k_s}} \left(\mathrm{e}^{A(\eta)}\right) = \mathbb{E}_\eta \left[T_1^{k_1}(X) \cdots T_s^{k_s}(X)\right].$$

That is because

$$M_\eta^T(u) = \mathrm{e}^{A(\eta+u)-A(\eta)}$$

is the **moment-generating function** (m.g.f.) of $T(X)$ when $X \sim p_\eta$. Indeed,

$$\begin{aligned} M_\eta^T(u) &= \mathbb{E}_\eta \left[\mathrm{e}^{u'T(X)}\right] \\ &= \int_{\mathcal{X}} \mathrm{e}^{u'T} \, \mathrm{d}\mathbb{P}_\eta(x) \\ &= \int_{\mathcal{X}} \mathrm{e}^{u'T} \mathrm{e}^{\eta'T - A(\eta)} h \, \mathrm{d}\mu \\ &= \mathrm{e}^{A(\eta+u)-A(\eta)} \int_{\mathcal{X}} \mathrm{e}^{(\eta+\mu)'T - A(\eta+\mu)} h \, \mathrm{d}\mu \\ &= \mathrm{e}^{A(\eta+u)-A(\eta)} \int_{\mathcal{X}} \mathrm{d}\mathbb{P}_{\eta+u}(x) \\ &= \mathrm{e}^{A(\eta+u)-A(\eta)} \end{aligned}$$

In particular,

$$\left[\frac{\mathrm{d}}{\mathrm{d}\eta} M_\eta^T(u)\right]_{u=0} = \frac{\mathrm{d}}{\mathrm{d}\eta}\left(\mathrm{e}^{A(\eta)}\right).$$

4 Properties of Statistics and Estimators

Sufficiency

The motivating example for this section is coin flipping.

Example 4.1. Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(\theta)$, and $X = (X_1, \ldots, X_n)$. Then $X \sim \prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i}$ on $\{0,1\}^n$. Then $T(X) = \sum_{i=1}^n X_i \sim \mathrm{Binom}(n, \theta) = \theta^t(1-\theta)^{n-t}\binom{n}{t}$ on $\{0, \ldots, n\}$.
The transformation $(X_1, \ldots, X_n) \mapsto T(X)$ is throwing away data (in particular the order of the results of the tosses). We wish to justify this.

Definition 4.2 (Sufficiency). Let $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ be a statistical model for data X. Then $T(X)$ is **sufficient** for $\mathcal{P}$ if $\mathbb{P}_\theta(X \mid T)$ does not depend on θ.

Example 4.3 (Example 4.1 Continued). We show that T is a sufficient statistic. In particular,

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

So given $T(x) = t$, the distribution of X does not depend on θ, and in this case it is uniform over all sequences with $\sum_i X_i = t$.
We can think of a concrete realization of the random variable X, as being generated in two stages:

1. Generate T, where the distribution of T depends on θ.
2. Generate $X \mid T$, where the distribution of $X \mid T$ *doesn't* depend on θ.

Theorem 4.4 (Sufficiency Principle). If $T(X)$ is sufficient for $\mathcal{P}$, then any statistical procedure should depend on X *only* through $T(X)$. This is called **sufficiency reduction**.

In fact, we could throw away X and generate a new dataset $\widetilde{X} \sim P(X \mid T(X))$ and it would be "just as good" as X.
The estimator $\delta\left(\widetilde{X}\right)$ has the same distribution (including the same risk function) as $\delta(X)$.

Theorem 4.5 (Factorization Theorem). Let $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ be a family of distributions dominated by μ (that is, $\mathbb{P}_\theta \ll \mu$ for all θ). Then T is sufficient for $\mathcal{P}$ if and only if there exists non-negative functions $g_\theta(x)$, $h(x)$ such that

$$p_\theta(x) = g_\theta(T(x))h(x) \ \mu\text{-a.e.}$$

A rigorous proof is in Keener 6.4.

Sketch. The proof is rigorous for discrete $\mathcal{X}$, but non-rigorous otherwise.
Suppose

$$p_\theta(x) = g_\theta(T(x))h(x)$$

μ-a.e. Then

$$\begin{aligned} p_\theta(x \mid T = t) &= 1\{T(x) = t\} \cdot \frac{g_\theta(t)h(x)}{\int_{T(z)=t} g_\theta(t)h(z)\,\mathrm{d}\mu(z)} \\ &= \frac{g_\theta(t)h(x)}{g_\theta(t)\int_{T(z)=t} h(z)\,\mathrm{d}\mu(z)} \\ &= \frac{h(z)}{\int_{T(z)=t} h(z)\,\mathrm{d}\mu(z)} \end{aligned}$$

does not depend on T.
Now suppose T is sufficient for $\mathcal{P}$ and take

$$g_\theta(t) = \int_{T(x)=t} p_\theta(x)\,\mathrm{d}\mu(x) = \mathbb{P}_\theta(T(x) = t)$$

and

$$h(x) = \frac{p_{\theta_0}(x)}{\int_{T(z)=t} p_{\theta_0}(z)\,\mathrm{d}\mu(z)} = \mathbb{P}_{\theta_0}(X = x \mid T(X) = T(x)).$$

Then

$$g_\theta(T(x))h(x) = \mathbb{P}_\theta(T = T(x))\mathbb{P}(X = x \mid T = t).$$

In some sense g_θ is the "step 1" distribution (distribution of contours of T) and h is the "step 2" distribution (distribution of X within contours of T). □

Example 4.6 (Exponential Families). Write

$$p_\theta(x) = \underbrace{\mathrm{e}^{\eta(\theta)'T(x) - B(\theta)}}_{g_\theta(T(x))} \underbrace{h(x)}_{h(x)}.$$

Example 4.7. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\theta^{(i)}$ for *any* model $\mathcal{P}^{(i)} = \left\{\mathbb{P}_\theta^{(i)} \,\middle|\, \theta \in \Theta\right\}$ on $X \subseteq \mathbb{R}$. The *order statistics* $\left(X_{(i)}\right)_{i=1}^n$ are sufficient, where $X_{(k)}$ is the k^{th} smallest value, i.e. $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$.
Obviously when converting to the order statistics, we lose the original order that the data points had. In general taking the sufficient statistic is a lossy compression of the data, even though it doesn't affect θ.
This phenomenon doesn't seem tied to $\mathbb{R}$. In fact, for more general $\mathcal{X}$ we can say that the **empirical distribution** $\widehat{P}_n(A) = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}(A)$ is sufficient where $\delta_{x_i}(A) = 1\{x_i \in A\}$.
It's a bit awkward to say that the sufficient statistic is a measure, but what we really mean is that we care about what data points we got and how many times we got them, but crucially not on the order that they came in.

The data set X is always sufficient, and we might find other $T(X)$ that are sufficient. We ask the question whether it's possible to find a minimal sufficient statistic, one that carries minimal information and compresses the data as much as possible while not losing any information about θ.

Example 4.8 (Uniform Location Family). If

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Uni}([\theta, \theta + 1]) = 1\{\theta \le x \le \theta + 1\},$$

then

$$p_\theta(X) = \prod_{i=1}^n 1\{\theta \le X_{(1)} \le \theta + 1\} = 1\{\theta \le X_{(1)}\}1\{X_{(n)} \le \theta + 1\}.$$

Minimal Sufficiency

Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$ (the Gaussian location family). Then $T(X) = \sum_{i=1}^n X_i$ is sufficient, $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is sufficient, $S(X) = \left(X_{(1)}, \dots, X_{(n)}\right)$ is also sufficient, and $X = (X_1, \dots, X_n)$ is also sufficient. Clearly $S(X)$ and X can be "compressed" further, but $T(X)$ and $\overline{X}$ cannot.
We can recover X only from itself, $S(X)$ can be recovered from X, $T(X)$ can be recovered from $\overline{X}$, $S(X)$, and X, and $\overline{X}$ can be recovered from $T(X)$, $S(X)$, and X.

Proposition 4.9. If $T(X)$ is sufficient and $T(X) = f(S(X))$ then $S(X)$ is sufficient.

Proof. The factorization theorem says

$$p_\theta(x) = g_\theta(T(x))h(x) = (g_\theta \circ f)(S(x))h(x).$$

□

Definition 4.10 (Minimal Sufficiency). $T(X)$ is **minimal sufficient** if

1. $T(X)$ is sufficient
2. For any other sufficient statistic $S(X)$, we can recover $T(X) = f(S(X))$ for some f (a.s. in $\mathcal{P}$, a.e. with respect to the common measure).

Definition 4.11 (Log Likelihood). Assume $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ has densities $p_\theta(x)$ with respect to μ.
The **(log)-likelihood function** is the (log) density reframed as a function of θ.

$$\text{Lik}(\theta; x) = p_\theta(x), \quad \ell(\theta; x) = \log(\text{Lik}(\theta; x)).$$

If $T(X)$ is sufficient then

$$\text{Lik}(\theta; x) = \underbrace{g_\theta(T(x))}_{\text{determines shape}} \underbrace{h(x)}_{\text{scaling}}.$$

Theorem 4.12 (Keener 3.11). Assume $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$, with densities p_θ with respect to some common measure, and $T(X)$ is sufficient to X. If

$$\text{Lik}(\theta; x) \propto_\theta \text{Lik}(\theta; y) \quad \text{implies} \quad T(x) = T(y)$$

then $T(X)$ is minimal sufficient.

Proof. Suppose S is sufficient and there does not exist f such that $f(S(X)) = T(X)$. Then there exist x, y with $S(x) = S(y)$, $T(x) \neq T(y)$. Then

$$\begin{aligned}\text{Lik}(\theta; x) &= g_\theta(S(x))h(x) \\ &\propto_\theta g_\theta(S(y))h(y) \\ &= \text{Lik}(\theta, y)\end{aligned}$$

which implies $T(x) = T(y)$ by assumption. □

This gives us a fairly easy minimal sufficiency check for exponential families.

Example 4.13. Suppose

$$p_\theta(x) = \exp(\eta(\theta)'T(x) - B(\theta))h(x).$$

Note that $B(\theta)$ doesn't change with x and $h(x)$ is constant in θ. Assume $\text{Lik}(\theta; x) \propto \text{Lik}(\theta; y)$, and we want to show that $T(x) = T(y)$. Then

$$\begin{aligned}\text{Lik}(\theta; x) &\propto \text{Lik}(\theta; y) \\ e^{\eta(\theta)'T(x)} &= e^{\eta(\theta)'T(y)}c(x, y) \\ \eta(\theta)'T(x) &= \eta(\theta)'T(y) + a(x, y) \\ (\eta(\theta_1) - \eta(\theta_2))'T(x) &= (\eta(\theta_1) - \eta(\theta_2))'T(y) \\ \eta(\theta_1) - \eta(\theta_2) &\perp T(x) - T(y) \\ T(x) - T(y) &\perp \text{span}(\{\eta(\theta_1) - \eta(\theta_2) \mid \theta_1, \theta_2 \in \Theta\})\end{aligned}$$

If $\operatorname{span}(\{\eta(\theta_1) - \eta(\theta_2) \mid \theta_1, \theta_2 \in \Theta\}) = \mathbb{R}^s$, then $T(x) - T(y) \perp \mathbb{R}^s$, so $T(x) - T(y) = 0$, so $T(x) = T(y)$. We do really need this condition. If $\eta(\theta) = \begin{bmatrix}\theta \\ 0\end{bmatrix}$ then $T_1(X)$ is sufficient, so $T(X)$ is not minimal.

The minimal sufficient statistic is not always one-dimensional. That is, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $T(X) = \begin{bmatrix}X \\ X^2\end{bmatrix}$ is minimal; if $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $T(X) = \begin{bmatrix}\sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i^2\end{bmatrix}$ is minimal.

Example 4.14. Suppose $X \sim \mathcal{N}(\mu(\theta), I_2)$ with $\theta \in \mathbb{R}$. Then

$$p_\theta(x) = \mathrm{e}^{\eta(\theta)'x - B(\theta)} e^{-\frac{1}{2}x'x}$$

If $\Theta = \mathbb{R}$, then $\mu(\theta) = a + \theta b$, $a, b \in \mathbb{R}^2$, and $\Xi = a + b\mathbb{R}$. Then

$$\begin{aligned} p_\theta(x) &= \mathrm{e}^{(a+\theta b)'x - B(\theta)} \mathrm{e}^{-\frac{1}{2}x'x} \\ &= \mathrm{e}^{\theta(b'x) - B(\theta)} \mathrm{e}^{-\frac{1}{2}(x-2a)'x}. \end{aligned}$$

Since $b'X$ is sufficient, X is not minimal.

Example 4.15 (Laplace Location Family). Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta^{(i)}(x) = \frac{1}{2}\mathrm{e}^{-|x-\theta|}$. Then $p_\theta(x) = \frac{1}{2}\exp(-\sum_{i=1}^n |x_i - \theta|)$ and

$$\begin{aligned} \ell(\theta; x) &= \log(p_\theta(x)) \\ &= -\sum_{i=1}^n |x_i - \theta| - n\log(2) \end{aligned}$$

This is piecewise linear in θ with "knots" at the $x_{(i)}$.
In this case $\ell(\theta; x) = \ell(\theta; y)$ plus a constant if and only if x, y have the same order statistics.

Completeness

Definition 4.16 (Completeness). $T(X)$ is **complete** for $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ if $\mathbb{E}_\theta[f(T(X))] = 0$ for all θ implies $f(T) = 0$ almost everywhere with respect to the measure $\mathbb{P}_\theta$ for all θ.

Example 4.17 (Laplace Continued). A minimal sufficient statistic is $S(X) = \left(X_{(1)}, \dots, X_{(n)}\right)$. We want to know if $S(X)$ is complete. We claim that the answer is "no" (since this statistic is so wasteful, in our minds). Take $f(S(X)) = \operatorname{median}(x) - \overline{\theta}$. Then $\mathbb{E}_\theta[f(S(x))] = \theta - \theta = 0$. So $S(X)$ is not complete.
Is there a way to find a complete minimal sufficient statistic? Not really, because $S(X)$ carries all the required information, and $f(S(X))$ only depends on the information carried by $S(X)$, so there's no way to rewrite $S(X)$ in a way that it's complete.

Example 4.18 (Uniform Scale Family). Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \operatorname{Uni}([0, \theta])$ for $\theta \in (0, \infty)$. We can show $T(X) = X_{(n)}$ is minimally sufficient. But is it complete?
We find the density of $T(X)$. Indeed,

$$\mathbb{P}_\theta(T \le t) = \left(\min \frac{t}{\theta}, 1\right)^n = \min\left(\frac{t}{\theta}\right)^n, 1.$$

Then

$$p_\theta(t) = n\frac{t^{n-1}}{\theta^n} 1\{t \le \theta\}.$$

Suppose that we can find a function $f(T)$ for which for all $\theta > 0$,

$$\begin{aligned} 0 &= \mathbb{E}_\theta[f(T)] \\ &= \frac{n}{\theta^n}\int_0^\theta f(t)t^{n-1}\,\mathrm{d}t \\ &= \int_0^\theta f(t)t^{n-1}\,\mathrm{d}t \\ &= f(t)t^{n-1} \text{ a.e., } t > 0 \end{aligned}$$

and so we have shown that $T(X) = X_{(n)}$ is complete.

Definition 4.19. Let $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ be an exponential family with densities

$$p_\theta(x) = \exp(\eta(\theta)'T(x) - B(\theta))h(x).$$

Assume without loss of generality that $\exists \alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^s$, with $\beta' T(X) = \alpha$. If so, replace $T(X)$ with a linearly independent basis for $\mathbb{R}^s$.

If $\Xi = \eta(\Theta) = \{\eta(\theta) \mid \theta \in \Theta\}$ contains an open set, we say $\mathcal{P}$ is **full rank**. Otherwise $\mathcal{P}$ is **curved**.

Theorem 4.20. If $\mathcal{P}$ is full rank, then any minimal sufficient statistic $T(X)$ on $\mathcal{P}$ is complete.

Theorem 4.21. If $T(X)$ is complete and sufficient for $\mathcal{P}$ then it is minimal and sufficient for $\mathcal{P}$.

Proof. Assume $S(X)$ is minimal sufficient (we have to assume that there is a meaningful minimal sufficient statistic, but this is almost always true). Then $S(X) = f(T(X))$ almost surely. Let $m(S(x)) = \mathbb{E}[T(X) \mid S(x)]$ and $g(t) = t - m(f(t))$. Then

$$\mathbb{E}_\theta[g(T(X))] = \mathbb{E}_\theta[T(X)] - \mathbb{E}_\theta[m(S(X))] = \mathbb{E}_\theta[T(X)] - \mathbb{E}_\theta[\mathbb{E}[T(X) \mid S(X)]] = 0$$

so $g(T(X)) \overset{\text{a.e.}}{=} 0$, and $T(X) \overset{\text{a.e.}}{=} m(S(X))$. We have recovered T from a minimal sufficient statistic. Suppose $R(X)$ is another statistic. We have assumed that $S(X)$ is minimal, so there exists measurable g such that $S(X) = g(R(X))$ and $T(X) = m(g(R(X)))$, so T can be recovered from any statistic R. □

Ancillarity

Definition 4.22 (Ancillary Statistic). $V(X)$ is **ancillary** for $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ if its distribution doesn't depend on θ.

The conditionality principle says that if we can find an ancillary statistic $V(X)$, then all analysis should condition on $V(X)$.

Theorem 4.23 (Basu). If $T(X)$ is complete sufficient and $V(X)$ is ancillary for $\mathcal{P}$, then $V(X)$ is independent of $T(X)$ for all $\theta \in \Theta$.

Proof. We want to show that for all θ, and measurable A, B that

$$\mathbb{P}_\theta(V \in A, T \in B) = \mathbb{P}_\theta(V \in A)\mathbb{P}_\theta(T \in B)$$
$$\mathbb{P}_\theta(V \in A \mid T \in B) = \mathbb{P}_\theta(V \in A) \text{ if } \mathbb{P}_\theta(T \in B) > 0$$

Let $q_A(T) = \mathbb{P}(V \in A \mid T)$ and let $p_A = \mathbb{P}(V \in A)$. Then

$$\mathbb{E}_\theta[q_A(T) - p_A] = p_A - p_A = 0 \,\forall\, \theta$$
$$q_A(T) \overset{\text{a.e.}}{=} p_A$$

where the measure is any $\mathbb{P}_\theta$. □

The reason this is so interesting is that we end up proving some fact about the distribution itself, regardless of the model, even though T and V are with respect to the model.

The most famous application of Basu's Theorem is as follows.

Example 4.24. Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Define $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean, and $S^2 = \frac{1}{2}\sum_{i=1}^n \left(X_i - \overline{X}_i^2\right)$ is the sample variance. Let $\mathcal{Q}_{\sigma^2} = \left\{\mathcal{N}(\mu, \sigma^2)^n \mid \mu \in \mathbb{R}\right\}$ (with σ^2 known). Clearly $\overline{X}$ is complete sufficient, as $\mathcal{Q}_{\sigma_2}$ is a full-rank exponential family. Suppose $Z_i = (X_i - \mu) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. Then $X_i - \overline{X} = Z_i - \overline{Z}$, so $S^2 = \frac{1}{n-1}\sum_{i=1}^n \left(Z_i - \overline{Z}\right)^2$. This value crucially doesn't rely on μ, so the distribution of S^2 is independent of μ, so S^2 is ancillary for $\mathcal{Q}_{\sigma^2}$ and $\overline{X} \perp\!\!\!\perp S^2$ for every distribution in $\mathcal{Q}_{\sigma^2}$.

5 Estimators

We will cover desirable properties of estimators, such as unbiasedness.

Convex Loss Function

Definition 5.1 (Convexity). $f(x)$ is **convex** if, for all $\gamma \in (0,1)$,

$$f(\gamma x + (1-\gamma)y) \le \gamma f(x) + (1-\gamma)f(y)$$

and **strictly convex** if

$$f(\gamma x + (1-\gamma)y) < \gamma f(x) + (1-\gamma)f(y).$$

Theorem 5.2 (Jensen's Inequality). If f is convex then

$$f(\mathbb{E}[X]) \le \mathbb{E}[f(X)]$$

for any probability measure $\mathbb{P}$. If f is strictly convex then, unless $X \overset{\text{a.e.}}{=} c$,

$$f(\mathbb{E}[X]) < \mathbb{E}[f(X)].$$

In particular, X can be a random vector.

If $L(\theta, d)$ is convex in d then it penalizes us for adding extra noise to an estimate. Suppose $\widetilde{\delta}(X) = \delta(X) + \varepsilon$, where ε is mean-zero noise and $\varepsilon \perp\!\!\!\perp X$. Then

$$\begin{aligned} R(\theta; \widetilde{\delta}) &= \mathbb{E}_\theta[\mathbb{E}[L(\theta, \delta(X) + \varepsilon) \mid X]] \\ &\ge \mathbb{E}_\theta[L(\theta, \delta(X))] \\ &= R(\theta; \delta) \end{aligned}$$

where the inequality is strict if L is strictly convex and $\varepsilon \overset{\text{a.e.}}{\neq} 0$.

Rao-Blackwell Theorem

Theorem 5.3 (Rao-Blackwell Theorem). Assume $T(X)$ is sufficient and $\delta(X)$ is an estimator. Let

$$\overline{\delta}(T(X)) = \mathbb{E}[\delta(X) \mid T(X)].$$

If $L(\theta, d)$ is convex then $R(\theta; \widetilde{\delta}) \le R(\theta; \delta)$ for all values of θ. If $L(\theta, d)$ is strictly convex then $R(\theta; \widetilde{\delta}) < R(\theta; \delta)$ for all values of θ unless $\delta(X) \overset{\text{a.e.}}{=} \overline{\delta}(T(X))$

Proof. The risk is just the expected loss, so

$$\begin{aligned} R(\theta; \overline{\delta}) &= \mathbb{E}_\theta[L(\theta, \mathbb{E}[\delta \mid T])] \\ &\le \mathbb{E}_\theta[\mathbb{E}[L(\theta; \delta) \mid T]] \\ &= \mathbb{E}_\theta[L(\theta; \delta)] \\ &= R(\theta; \delta). \end{aligned}$$

The inequality is strict if L is strictly convex, unless $\delta \overset{\text{a.e.}}{=} \overline{\delta}$. □

Definition 5.4. The estimator $\overline{\delta}(T(X))$ is called the **Rao-Blackwellization of $\delta(X)$.**

Remark 5.5. An estimator for θ cannot itself use θ, and $\overline{\delta}(T(X))$ correspondingly does not use θ; this is possible because $T(X)$ is sufficient.

Bias-Variance Decomposition

Definition 5.6 (Bias). The **bias** of $\delta(X)$ is $\mathbb{E}_\theta[\delta(X)] - g(\theta)$. The estimator $\delta(X)$ is **unbiased** if $\mathbb{E}_\theta[\delta(X)] = g(\theta)$ for all $\theta \in \Theta$.

The bias variance decomposition writes the MSE in terms of the bias and variance of the estimator.

$$\begin{aligned} \operatorname{MSE}(\theta; \delta) &= \mathbb{E}_\theta\left[(\delta(X) - g(\theta))^2\right] \\ &= \mathbb{E}_\theta\left[(\delta(X) - \mathbb{E}_\theta[\delta(X)] + \mathbb{E}_\theta[\delta(X)] - g(\theta))^2\right] \\ &= (\mathbb{E}_\theta[\delta(X)] - g(\theta))^2 + \mathbb{E}_\theta\left[(\delta(X) - \mathbb{E}_\theta[\delta(X)])^2\right] + 2\mathbb{E}_\theta\left[\underbrace{(\delta(X) - \mathbb{E}_\theta[\delta(X)])}_{\text{expectation } 0}\underbrace{(\mathbb{E}_\theta[\delta(X)] - g(\theta))}_{\text{constant}}\right] \end{aligned}$$

$$= (\mathbb{E}_\theta[\delta(X)] - g(\theta))^2 + \mathbb{E}_\theta\left[(\delta(X) - \mathbb{E}_\theta[\delta(X)])^2\right]$$
$$= \text{Bias}_\theta(\delta(X))^2 + \text{Var}_\theta(\delta(X)).$$

In this decomposition, if $\delta(X)$ is unbiased, then $\text{MSE}(\theta; \delta) = \text{Var}_\theta(\delta(X))$.
This decomposition gives us a tradeoff between an estimator's bias and variance. To choose between them isn't always easy. The next topic is how well can we do if we require $\text{Bias}_\theta(\delta(X)) = 0$. It's not always possible to get an unbiased estimator.

Definition 5.7 (Estimable). The function $g(\theta)$ is **$\mathcal{U}$-estimable** if there exists an estimator $\delta(X)$ with $\mathbb{E}_\theta[\delta(X)] = g(\theta)$ for all θ.

Definition 5.8 (UMVU). The estimator $\delta(X)$ is **uniformly minimum variance unbiased (UMVU)** if δ is unbiased and for any unbiased estimator $\widetilde{\delta}$,

$$\text{Var}_\theta(\delta(X)) \le \text{Var}_\theta\left(\widetilde{\delta}(X)\right)$$

for all θ.

Theorem 5.9 (Keener 4.4). Suppose $T(X)$ is a complete sufficient statistic for $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$. Then for any $\mathcal{U}$-estimable $g(\theta)$, there is a unique (up to $\overset{\text{a.e.}}{=}$) UMVU estimator of the form $\delta(T(X))$. More generally, for any convex loss L, there is a unique best estimator (uniformly minimum risk) $\delta(T(X))$.

Proof. Assume $\delta_0(X)$ is unbiased for $g(\theta)$. Then

$$\mathbb{E}_\theta[\delta] = \mathbb{E}_\theta[\mathbb{E}[\delta_0(X) \mid T(X)]]$$
$$= \mathbb{E}_\theta[g(\theta)]$$
$$= g(\theta),$$

so $\delta(T) = \mathbb{E}[\delta_0 \mid T]$ is unbiased. If $\widetilde{\delta}(T(X))$ is unbiased then

$$\mathbb{E}_\theta\left[\delta(T(X)) - \widetilde{\delta}(T(X))\right] = 0 \ \forall \theta \in \Theta \implies \delta(T(X)) \overset{\text{a.e.}}{=} \widetilde{\delta}(T(X)),$$

by completeness of $T(X)$.
If $\delta^*(X)$ is unbiased then the Rao-Blackwell theorem gives

$$R(\theta; \delta^*) = \mathbb{E}_\theta[L(\theta; \delta^*)] \ge \mathbb{E}_\theta[L(\theta; \delta)] = R(\theta; \delta)$$

and in particular

$$\text{MSE}(\theta; \delta^*) = \text{Var}_\theta(\delta^*) \ge \text{Var}_\theta(\delta) = \text{MSE}(\theta; \delta).$$

We can Rao-Blackwellize any estimator $\delta^*(X)$ to obtain an estimator $\delta(T(X))$ that depends only on $T(X)$ and is at least as good. Completeness is necessary for uniqueness of the Rao-Blackwellization of $\delta^*(X)$. □

The next point is how to find the UMVUE. There are two methods that the proof suggests.

1. Find any unbiased estimator that is only a function of T.
2. Find any unbiased estimator at all, then Rao-Blackwellize it.

Example 5.10. Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Uni}([0, \theta])$ for $\theta > 0$. Then $T(X) = X_{(n)}$ is complete sufficient. We use method 2 to find a UMVUE. Suppose $\delta(X) = 2X_1$. Then $X_1 \mid T \sim \begin{cases} T & \text{with probability } \frac{1}{n} \\ \text{Uni}([0, T]) & \text{with probability } \frac{n-1}{n} \end{cases}$. Then

$$\mathbb{E}[2X_1 \mid T] = 2T \cdot \frac{1}{n} + T \cdot \frac{n-1}{n} = \frac{n+1}{n} \cdot T.$$

If we use method 1 to find a UMVUE, we can use $T(X) = X_{(n)}$. Then $\mathbb{E}_\theta[T(X)] = \frac{n}{n+1} \cdot \theta$ and so $\mathbb{E}_\theta\left[\frac{n+1}{n} \cdot T\right] = \theta$, so $\frac{n+1}{n} \cdot T$ is the UMVUE.
Keener shows that among estimators $c \cdot T$, $\frac{n+2}{n+1} \cdot T$ has the best MSE. However, it's biased downward.

Example 5.11. Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\theta)$ for $\theta > 0$. Then $p_\theta^{(i)}(x) = \frac{\theta^x e^{-\theta}}{x!}$ for $x = 0, 1, \dots$ (where the density is taken with respect to the counting measure). The complete sufficient statistic $T(X) = \sum_{i=1}^n X_i \sim \text{Pois}(n\theta)$. This statistic is complete sufficient because the distribution family is an exponential family with natural parameter $\eta(\theta) = \frac{1}{\theta}$ and $\text{span}(\{\eta(\theta_1) - \eta(\theta_2) \mid \theta_1, \theta_2 > 0\}) = \mathbb{R}$, so the family is full rank.

We estimate $g(\theta) = \theta^2$. If $\delta(T)$ is unbiased, then for all $\theta > 0$,

$$\begin{aligned}\theta^2 &= \int_{T(\mathcal{X})} \delta(t)\,\mathrm{d}\mathbb{P}_\theta^T(t) \\ &= \sum_{t=0}^\infty \delta(t) p_\theta^T(t) \\ \mathrm{e}^{n\theta}\theta^2 &= \sum_{t=0}^\infty \delta(t)\cdot\frac{n^t\theta^t}{t!} \\ \sum_{k=0}^\infty \frac{n^k}{k!}\cdot\theta^{k+2} &= \sum_{t=0}^\infty \delta(t)\cdot\frac{n^t\theta^t}{t!}.\end{aligned}$$

Matching coefficients in the power series, we get $\delta(0) = \delta(1) = 0$, and for all $t \geq 2$,

$$\delta(t) = \frac{n^{t-2}}{(t-2)!}\cdot\frac{t!}{n^t} = \frac{t(t-1)}{n^2} \approx \left(\frac{T}{n}\right)^2.$$

We know T/n is the UMVUE for θ (and it happens to be the MLE for θ). However, $(T/n)^2$ is not the UMVUE for θ^2, although it happens to be the MLE for θ^2.

Log-Likelihood, Score

Assume $\mathcal{P}$ has densities p_θ with respect to μ. Let $\Theta \in \mathbb{R}^d$ and p_θ have a common support (that is, $\{x \mid p_\theta(x) = 0\}$ does not depend on θ).
Recall $\ell(\theta; X) = \log(p_\theta(X))$ is a random function of θ.

Definition 5.12 (Score). The **score** is $\nabla_\theta \ell(\theta; X)$.

The score can be thought of as the "local sufficient statistic". The family $\mathcal{P}$ at θ has a locally tangent family that approximates $\mathcal{P}$. In particular, for small values of η,

$$p_{\theta+\eta}(x) = \mathrm{e}^{\ell(\theta+\eta;x)} \approx \mathrm{e}^{\eta^\prime\nabla_\theta\ell(\theta;x)}p_\theta(x)$$

so we can approximate $\mathcal{P}$ by something that looks like a curved exponential family with natural parameter η and "sufficient statistic" $\nabla_\theta\ell(\theta; x)$; note that this is not a statistic because it depends on θ.
We can consider some differential identities for this score function:

$$\begin{aligned}1 &= \int_{\mathcal{X}} \mathrm{e}^{\ell(\theta;x)}\,\mathrm{d}\mu(x) \\ \frac{\partial}{\partial\theta_j}1 &= \frac{\partial}{\partial\theta_j}\int_{\mathcal{X}} \mathrm{e}^{\ell(\theta;x)}\,\mathrm{d}\mu(x) \\ 0 &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta_j}\ell(\theta;x)\right)\mathrm{e}^{\ell(\theta;x)}\,\mathrm{d}\mu(x) \\ &= \mathbb{E}_\theta[\nabla_\theta\ell(\theta;X)]\end{aligned}$$

Differentiating again, we obtain

$$\begin{aligned}0 &= \frac{\partial}{\partial\theta_k}\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_j}\ell(\theta;X)\right] \\ &= \frac{\partial}{\partial\theta_k}\int_{\mathcal{X}}\left(\frac{\partial}{\partial\theta_j}\ell(\theta;x)\right)\mathrm{e}^{\ell(\theta;x)}\,\mathrm{d}\mu(x) \\ &= \int_{\mathcal{X}}\left(\frac{\partial\ell(\theta;x)}{\partial\theta_j\theta_k} + \frac{\partial\ell(\theta;x)}{\partial\theta_j}\cdot\frac{\partial\ell(\theta;x)}{\partial\theta_k}\right)\mathrm{e}^{\ell(\theta;x)}\,\mathrm{d}\mu \\ &= \mathbb{E}_\theta\left[\frac{\partial^2\ell(\theta;X)}{\partial\theta_i\partial\theta_k}\right] + \mathbb{E}_\theta\left[\frac{\partial\ell(\theta;X)}{\partial\theta_j}\frac{\partial\ell(\theta;X)}{\partial\theta_k}\right] \\ \mathrm{Var}_\theta(\nabla_\theta\ell(\theta;X)) &= \mathbb{E}_\theta\left[-\nabla_\theta^2\ell(\theta;X)\right]\end{aligned}$$

This function $J(\theta) = \operatorname{Var}_\theta(\nabla_\theta \ell(\theta; X))$ is called the **Fisher information**.
Let's try with another statistic $\delta(X)$. Let $g(\theta) = \mathbb{E}_\theta[\delta(X)]$. Then

$$\begin{aligned} g(\theta) &= \mathbb{E}_\theta[\delta(X)] \\ &= \int_{\mathcal{X}} \delta(x) \mathrm{e}^{\ell(\theta;x)} \, \mathrm{d}\mu(x) \\ \nabla_\theta g(\theta) &= \int_{\mathcal{X}} \delta(x)(\nabla_\theta \ell(\theta; x)) \mathrm{e}^{\ell(\theta;x)} \, \mathrm{d}\mu(x) \\ &= \mathbb{E}_\theta[\delta(X) \nabla_\theta \ell(\theta; X)] \\ &= \operatorname{Cov}_\theta(\delta(X), \nabla_\theta \ell(\theta; X)). \end{aligned}$$

where the last line is given by homework problem 1.3b.

Cramër-Rao Lower bound

This is also called the information lower bound. In a one-parameter θ case, we obtain by the Cauchy-Schwarz inequality and the above differential equalities,

$$\operatorname{Var}_\theta(\delta(X)) \cdot \operatorname{Var}_\theta(\nabla_\theta \ell(\theta; X)) \geq \operatorname{Cov}_\theta(\delta, \nabla_\theta \ell(\theta; X))^2$$

$$\operatorname{Var}_\theta(\delta(X)) = \frac{(\nabla_\theta g(\theta))^2}{J(\theta)}.$$

In the general case, that is, $\theta \in \mathbb{R}^d$ and $g(\theta) \in \mathbb{R}$, we obtain

$$\operatorname{Var}_\theta(\delta(X)) \geq (\nabla_\theta g(\theta))' (J(\theta))^{-1} (\nabla_\theta g(\theta))$$

which can be obtained by a simple analysis on quadratic forms.

Example 5.13 (I.I.D. Sample). Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta^{(1)}(x)$, where $\theta \in \Theta$. Then

$$X \sim p_\theta(x) = \prod_{i=1}^n p_\theta^{(1)}(x_i).$$

Let $\ell_1(\theta; x_i) = \log\left(p_\theta^{(i)}(x_i)\right)$, and $\ell(\theta; x) = \sum_{i=1}^n \ell_1(\theta; x_i)$. Then

$$\begin{aligned} J(\theta) &= \operatorname{Var}_\theta(\nabla_\theta \ell(\theta; X)) \\ &= \operatorname{Var}_\theta\left(\sum_{i=1}^n \ell(\theta; X_i)\right) \\ &= n \cdot J_1(\theta) \end{aligned}$$

where J_1 is the univariate Fisher information. In this case the Cramer-Rao Lower Bound (CRLB) scales like n^{-1} (for "regular" families the standard deviation scales like $n^{-1/2}$).

Hammersley-Chapman-Robbins Inequality

Hammersley-Chapman-Robbins Inequality (HCR) is a more general bound, replacing $\nabla_\theta \ell(\theta; X)$ with a finite difference. We obtain

$$\frac{p_{\theta+\varepsilon}(x)}{p_\theta(x)} - 1 = \mathrm{e}^{\ell(\theta+\varepsilon;x) - \ell(\theta;x)} - 1 \approx \varepsilon'(\nabla_\theta(\theta; x))$$

for small ε. Then

$$\mathbb{E}_\theta\left[\frac{p_{\theta+\varepsilon}(X)}{p_\theta(X)} - 1\right] = \int_{\mathcal{X}} \left(\frac{p_{\theta+\varepsilon}(x)}{p_\theta(x)} - 1\right) \mathrm{d}\mathbb{P}_\theta(x) = \int_{\mathcal{X}} \left(\frac{p_{\theta+\varepsilon}(x)}{p_\theta(x)} - 1\right) p_\theta(x) \, \mathrm{d}\mu(x) = 1 - 1 = 0$$

assuming that $\mathbb{P}_{\theta+\varepsilon} \ll \mathbb{P}_\theta$, and

$$\operatorname{Cov}_\theta\left(\delta(X), \frac{p_{\theta+\varepsilon}(X)}{p_\theta(X)} - 1\right) = \int_{\mathcal{X}} \delta(x) \left(\frac{p_{\theta+\varepsilon}(x)}{p_\theta(x)} - 1\right) \mathrm{d}\mathbb{P}_\theta(x)$$

$$\begin{aligned}
&= \int_{\mathcal{X}} \delta(x)\left(\frac{p_{\theta+\varepsilon}(x)}{p_\theta(x)} - 1\right) p_\theta(x)\,\mathrm{d}\mu(x) \\
&= \mathbb{E}_{\theta+\varepsilon}[\delta(X)] - \mathbb{E}_\theta[\delta(X)] \\
&= g(\theta+\varepsilon) - g(\theta) \\
\mathrm{Var}_\theta(\delta(X)) &\geq \frac{(g(\theta+\varepsilon)-g(\theta))^2}{\mathbb{E}_\theta\left[\left(\frac{p_{\theta+\varepsilon}(X)}{p_\theta(X)} - 1\right)^2\right]}
\end{aligned}$$

If $\varepsilon \to 0$, and derivatives work out, then this approaches the CRLB. Since ε isn't on the left hand side, side, we can take the supremum to obtain

$$\mathrm{Var}_\theta(\delta(X)) \geq \sup_\varepsilon \frac{(g(\theta+\varepsilon)-g(\theta))^2}{\mathbb{E}_\theta\left[\left(\frac{p_{\theta+\varepsilon}(X)}{p_\theta(X)} - 1\right)^2\right]}.$$

The CRLB isn't always achievable.

Definition 5.14 (Efficiency). The efficiency of an estimator $\delta(X)$ is defined to be

$$\mathrm{eff}_\theta(\delta) = \frac{\mathrm{CRLB}}{J(\theta)}.$$

if $g(\theta) = \theta \in \mathbb{R}$. We have $\mathrm{eff}_\theta(\delta) \leq 1$. We call a estimator **efficient** if $\mathrm{eff}_\theta(\delta) = 1$ for all θ.

If $g(\theta) = \theta$, then

$$\mathrm{eff}_\theta(\delta) = \frac{1}{\mathrm{Var}_\theta(\delta(X)) \cdot J(\theta)}$$

We note that $\mathrm{eff}_\theta(\delta)$ is related to $\mathrm{Corr}_\theta(\delta(X), \nabla_\theta \ell(\theta; X))$; in fact

$$\begin{aligned}
\mathrm{eff}_\theta(\delta) &= \frac{\mathrm{Cov}_\theta(\delta(X), \nabla_\theta \ell(\theta;X))^2}{\mathrm{Var}_\theta(\delta(X)) \cdot \mathrm{Var}_\theta(\nabla_\theta \ell(\theta;X))} \\
&= \mathrm{Corr}_\theta(\delta(X), \nabla_\theta \ell(\theta;X))^2 \\
&\leq 1.
\end{aligned}$$

Thus $\delta(X)$ is efficient if and only if $\mathrm{Corr}_\theta(\delta(X), \nabla_\theta \ell(\theta;X))^2 = 1$. This very rarely happens in the finite n case, but can be approached as $n \to \infty$.

Example 5.15 (Curved Exponential Family). Suppose for $\theta \in \mathbb{R}$,

$$p_\theta(x) = \mathrm{e}^{\eta(\theta)'T(x) - B(\theta)} h(x),$$

so then

$$\begin{aligned}
\ell(\theta;x) &= \eta(\theta)'T(x) - B(\theta) + \log(h(x)) \\
\nabla_\theta \ell(\theta;x) &= \nabla_\theta \eta(\theta)' T(x) - \nabla_\theta B(\theta) \\
&= (\nabla_\theta \eta(\theta))'(T(x) - \nabla_\theta A(\eta(\theta))) \\
&= (\nabla_\theta \eta(\theta))'(T(x) - \mathbb{E}_\theta[T(X)]).
\end{aligned}$$

In this way the score family stitches together all the locally linear families at each θ. The curved exponential family is a prototypical example of a stitched-together locally-linear family. In the asymptotic setting, we can restrict our attention to a little neighborhood of θ, at which point the family behaves linearly.

6 Bayesian Decision Theory

Doubts about UMVUE

The UMVUE might me inefficient, inadmissible, or dumb, in cases where another approach makes more sense.

Example 6.1. Suppose $X \sim \mathrm{Binom}(n, \theta)$ where n is large. We estimate $g(\theta) = \mathbb{P}_\theta(X \geq n/2)$ and the UMVUE is $1\{X \geq n/2\}$. Then $X = (n/2) + 1$ gives $g(\theta) = 1$; $X = (n/2) - 1$ gives $g(\theta) = 0$. This is very unreasonable.

The UMVUE can sometimes be ridiculous. It's not always ridiculous, but sometimes it's better to just leave "unbiasedness" as a side goal when trying to find estimators.
Another goal of estimation design is to minimize some function of the risk. Sometimes this function is the worst-case risk; sometimes this function is the average-case risk.

Frequentist Motivation

Suppose the model $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \mathbb{R}\}$ is for data X. Let Θ be a random variable. As a reminder, the loss is $L(\theta, \delta(X))$ and the risk is $R(\theta; d) = \mathbb{E}_\theta[L(\theta; \delta(X))]$.

Definition 6.2 (Bayes Risk). The **Bayes risk** is the average-case risk, integrated with respect to some measure Λ called the **prior**. For now we set $\Lambda(\Omega) = 1$ and if Λ is a finite measure then we can normalize by $\Lambda(\Omega)$.

Then the Bayes risk is

$$\begin{aligned} R_{\text{Bayes}}(\Lambda, \delta) &= \int_\Omega R(\theta, d)\,\mathrm{d}\Lambda(\theta) \\ &= \mathbb{E}_{\Theta\sim\Lambda}[R(\Theta; d)] \\ &= \mathbb{E}_{\Theta\sim\Lambda, X\mid\Theta=\theta\sim\mathbb{P}_\theta}[L(\Theta; d(X))]. \end{aligned}$$

An estimator δ minimizing $R_{\text{Bayes}}(\Lambda, \cdot)$ is called **Bayes** (and is a **Bayes estimator**).

Theorem 6.3. Suppose $\Theta \sim \Lambda$ and $X \mid \Theta = \theta \sim \mathbb{P}_\theta$. Suppose $L(\theta, \delta) \geq 0$ for all θ, δ and there is some $\delta_0(X)$ for which $R_{\text{Bayes}}(\Lambda; \delta_0) < \infty$ for some $\delta_0(X)$. Then $\delta_\Lambda(X) \in \operatorname{argmin}_\delta \mathbb{E}[L(\Theta, d(X)) \mid X = X]$ $\mathbb{P}_\theta$-a.e. if and only if $\delta_\Lambda(X)$ is Bayes with $R_{\text{Bayes}}(\Lambda, \delta_\Lambda(X)) < \infty$.

Proof. Let δ be any other estimator. Then if $\delta = \delta_0$,

$$\begin{aligned} \infty > R_{\text{Bayes}}(\Lambda, \delta) &= \mathbb{E}[L(\Theta, \delta(X))] \\ &= \mathbb{E}[\mathbb{E}[L(\theta, \delta(X)) \mid X = x]] \\ &\geq \mathbb{E}[\mathbb{E}[L(\theta, \delta_\Lambda(X)) \mid X = x]] \\ &\geq R_{\text{Bayes}}(\Lambda, \delta_\Lambda). \end{aligned}$$

For the other direction, let $\mathbb{E}_x[\delta] = \mathbb{E}[L(\Theta, d) \mid X = x]$. Define

$$\delta^*(x) = \begin{cases} \delta_\Lambda(x) & \text{if } \delta_\Lambda(x) \in \operatorname{argmin}_\delta E_x \\ \delta_0(x) & \text{if } E_x(\delta_0(x)) < E_x(\delta_\Lambda(x)) \\ \delta^*(x) & \text{otherwise} \end{cases}$$

where $E_x(\delta^*) < E_x(\delta_\Lambda(x))$. Then $E_x(\delta^*(x)) \overset{\mathbb{P}_\theta-\text{a.e.}}{\leq} E_x(\delta_0(x))$ and $E_x(\delta^*(x)) \overset{\mathbb{P}_\theta-\text{a.e.}}{\leq} E_x(\delta_\Lambda(x))$, with the inequality strict on a set of positive measure. □

Prior, Posterior

The usual interpretation is that Λ reflects a "prior belief" about Θ before we see data.

Definition 6.4 (Posterior Definition). The conditional expectation of Θ given X (that is, $\mathcal{L}(\Theta \mid X)$) is called the **posterior distribution**.

Suppose the densities related are

- prior $\lambda(\theta)$
- likelihood $p_\theta(x)$
- marginal density of X, mixture density of X distributed according to the priors $q(x) = \int_\Omega \lambda(\theta) p_\theta(x)\,\mathrm{d}\theta$
- **posterior density** $\lambda(\theta \mid x) = \frac{\lambda(\theta)p_\theta(x)}{q(x)}$

Then the Bayes estimator is $\delta_\Lambda = \operatorname{argmin}_d \int_\Omega L(\theta, d)\lambda(\theta \mid x)\,\mathrm{d}\theta$.

Posterior Mean

If $L(\theta, d) = (g(\theta) - d)^2$ then δ_Λ is the **posterior mean** of $g(\theta)$.

$$\begin{aligned}\mathbb{E}\left[(g(\Theta) - d)^2 \,\middle|\, X\right] &= \mathbb{E}\left[(g(\Theta) - \mathbb{E}[g(\Theta) \mid X] + \mathbb{E}[g(\Theta) \mid X] - d)^2 \,\middle|\, X\right] \\ &= \operatorname{Var}(g(\Theta) \mid X) + (\mathbb{E}[g(\Theta) \mid X] - d)^2 \\ &\geq \operatorname{Var}(g(\Theta) \mid X)\end{aligned}$$

with equality if $d = \mathbb{E}[g(\Theta) \mid X]$.
In the case of the weighted squared error $L(\theta, d) = w(\theta)(g(\theta) - d)^2$, then

$$\mathbb{E}\left[(d - g(\Theta))^2 w(\Theta) \,\middle|\, X\right] = d^2 \mathbb{E}[w(\Theta) \mid X] - 2d\mathbb{E}[w(\Theta)g(\Theta) \mid X] + \underbrace{\mathbb{E}\left[w(\Theta)g(\Theta)^2 \mid X\right]}_{\text{no dependence on } d}$$

and this is just a quadratic function in d, so solving obtains

$$d = \frac{\mathbb{E}[w(\Theta)g(\Theta) \mid X]}{\mathbb{E}[w(\Theta) \mid X]} = \delta_\Lambda(X).$$

Example 6.5 (Beta-Binomial). Suppose $X \mid \Theta = \theta \sim \operatorname{Binom}(n, \theta)$ and the density is $p_\theta(x) = \theta^x(1-\theta)^{n-x}\binom{n}{x}$. One popular choice for the prior distribution of Θ is the Beta distribution $\Theta \sim \operatorname{Beta}(\alpha, \beta)$ with $\lambda(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Then the marginal distribution of X is called the **Beta-binomial**. Computing the posterior distribution of θ is straightforward :

$$\begin{aligned}\lambda(\theta \mid x) &= \frac{\lambda(\theta)p_\theta(x)}{q(x)} \\ &\propto_\theta \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^x(1-\theta)^{n-x} \\ &= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \\ &\sim_\theta \operatorname{Beta}(x+\alpha, n-x+\beta).\end{aligned}$$

Thus we can directly conclude that $\Theta \mid X = x \sim \operatorname{Beta}(x+\alpha, n-x+\beta)$ and

$$\begin{aligned}\mathbb{E}[\Theta \mid X] &= \frac{x+\alpha}{n+\alpha+\beta} \\ &= \underbrace{\frac{x}{n}}_{\text{UMVU}}\frac{n}{n+\alpha+\beta} + \underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{prior mean}}\frac{\alpha+\beta}{n+\alpha+\beta}.\end{aligned}$$

The interpretation of this is that there are $k = \alpha + \beta$ "pseudo-trials" with α successes.

Example 6.6 (Gaussian). If $X \mid \Theta = \theta \sim \mathcal{N}(\theta, \sigma^2)$ with density $p_\theta(x) \sim_\theta \mathrm{e}^{-\frac{(x-\theta)^2}{2\sigma^2}}$ and $\Theta \sim \mathcal{N}(\mu, \tau^2)$ with density $\lambda(\theta) \propto_\theta \mathrm{e}^{-\frac{(\theta-\mu)^2}{2\tau^2}}$, then

$$\begin{aligned}\lambda(\theta \mid x) &\propto_\theta \exp\left(-\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu)^2}{2\tau^2}\right) \\ &\propto_\theta \exp\left(\theta\left(\frac{X}{\sigma^2} + \frac{\mu}{\tau^2}\right) - \theta^2\left(\frac{\sigma^2+\tau^2}{2}\right)\right) \\ &= \exp\left(\theta b - \theta^2 a^2\right)\end{aligned}$$

which is a normal distribution with no carrier distribution. In particular

$$\Theta \mid X \sim \mathcal{N}\left(\underbrace{\frac{X\sigma^{-2} + \mu\tau^{-2}}{\sigma^{-2}+\tau^{-2}}}_{\text{precision-weighted average of } x,\, \mu}, \underbrace{\frac{1}{\sigma^{-2}+\tau^{-2}}}_{\text{harmonic mean of } \sigma^2,\, \tau^2}\right),$$

We can again view the posterior mean as a convex combination of x and μ:

$$\mathbb{E}[\Theta \mid X] = X\frac{\sigma^{-2}}{\sigma^{-2}+\tau^{-2}} + \mu\frac{\tau^{-2}}{\sigma^{-2}+\tau^{-2}}.$$

Example 6.7 (Gaussian). Suppose that $\Theta \sim \mathcal{N}(\mu, \tau^2)$ and each $X_i \mid \Theta = \theta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$ for $i = 1, \dots, n$. Then $\overline{X} \mid \Theta = \theta \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)$ and

$$\mathbb{E}[\Theta \mid X] = \overline{X}\frac{n\sigma^{-2}}{n\sigma^{-2}+\tau^{-2}} + \mu\frac{\tau^{-2}}{n\sigma^{-2}+\tau^{-2}} = \overline{X}\frac{n}{n+\frac{\sigma^2}{\tau^2}} + \frac{\sigma^2/\tau^2}{n+\frac{\sigma^2}{\tau^2}}.$$

The interpretation of this is that $k = \frac{\sigma^2}{\tau^2}$ is the number of pseudo-observations. If $n \gg k$ then the "data swamps the prior," i.e., the posterior is highly dependent on the data and not the prior. Likewise, if $n \ll k$ then the "prior swamps the data," i.e., the posterior is highly dependent on the prior and not the data.

Bias and Bayes

Bayes estimators are usually biased, and unbiased estimators don't minimize the Bayes risk (if the loss is the squared-error loss).

Theorem 6.8. The posterior mean is biased unless $\delta_\Lambda(x) \overset{\text{a.e.}}{=} g(\Theta)$.

Proof. Suppose δ_Λ is unbiased. Then $g(\Theta) = \mathbb{E}[\delta(X) \mid \Theta]$ and $\delta_\Lambda(X) = \mathbb{E}[g(\Theta) \mid X]$. Then conditioning on X, we obtain

$$\begin{aligned}\mathbb{E}[\delta_\Lambda(X)g(\Theta) \mid X] &= \delta_\Lambda(X)\mathbb{E}[g(\Theta) \mid X] \\ &= \delta_\Lambda(X)^2.\end{aligned}$$

Conditioning on Θ,

$$\begin{aligned}\mathbb{E}[\delta_\Lambda(X)g(\Theta) \mid \Theta] &= \mathbb{E}[\delta_\Lambda(X) \mid \Theta]g(\Theta) \\ &= g(\Theta)^2.\end{aligned}$$

Thus

$$\mathbb{E}[\delta_\Lambda(X)g(\Theta)] = \mathbb{E}\left[\delta_\Lambda(X)^2\right] = \mathbb{E}\left[g(\Theta)^2\right]$$

and so

$$\begin{aligned}\mathbb{E}\left[(\delta_\Lambda - g(\Theta))^2\right] &= \mathbb{E}\left[\delta_\Lambda(X)^2\right] + \mathbb{E}\left[g(\Theta)^2\right] - 2\mathbb{E}[\delta_\Lambda(X)g(\Theta)] \\ &= 0.\end{aligned}$$

□

Conjugate Prior

Definition 6.9 (Conjugate Prior). If the posterior is always from the same family as the prior, we say that the prior is **conjugate** to the likelihood.

Example 6.10 (Exponential Family). Suppose $X_i \mid \eta \overset{\text{i.i.d.}}{\sim} p_\eta(x) = \mathrm{e}^{\eta' T(x) - A(\eta)}h(x)$ for $\eta \in \Xi \subseteq \mathbb{R}^s$ and $i = 1, \dots, n$. For a carrier distribution $\lambda_0(\eta)$, we define the $(s+1)$-dimensional family

$$\lambda_{k\mu,k}(\eta) = \mathrm{e}^{k\mu'\eta - kA(\eta) - B(k\mu,k)}\lambda_0(\eta)$$

with sufficient statistic $(\eta, -A(\eta)) \in \mathbb{R}^{s+1}$ and natural parameter $(k\mu, k)$. Then

$$\begin{aligned}\lambda(\eta \mid x_1, \dots, x_n) &= \left(\prod_{i=1}^n \mathrm{e}^{\eta' T(x_i) - A(\eta)}h(x_i)\right) \cdot \mathrm{e}^{k\mu'\eta - kA(\eta) - B(k\mu,k)}\lambda_0(\eta) \\ &\propto_\eta \mathrm{e}^{\left(k\mu + \sum_{i=1}^n T(x_i)\right)'\eta - (k+n)A(\eta)}\lambda_0(\eta) \\ &\propto \lambda_{k\mu + n\overline{T}, k+n}(\eta)\end{aligned}$$

where $\overline{T} = \frac{1}{n}\sum_{i=1}^n T(x_i)$. There are two interpretations:

- Prior $\lambda_{k\mu,k}$, observe the average sufficient statistic $\overline{T}$ on sample size n

- Prior λ_0, observe average sufficient statistic
  - μ on sample size k
  - $\overline{T}$ on sample size n

Here are a list of conjugate priors:

Table 1: Conjugate Priors

Likelihood	Prior	Posterior
$X \mid \Theta = \theta \sim \mathrm{Binom}(n, \theta)$	$\Theta \sim \mathrm{Beta}(\alpha, \beta)$	$\Theta \mid X \sim \mathrm{Beta})(\alpha + X, \beta + n - X)$
$X_i \mid \Theta = \theta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$	$\Theta \sim \mathcal{N}(\mu, \tau^2)$	$\Theta \mid X \sim \mathcal{N}\left(\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)\left(\frac{\mu}{\tau^2} + \frac{\sum_{i=1}^n X_i}{\sigma^2}\right), \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$
$X_i \mid \Theta = \theta \overset{\text{i.i.d.}}{\sim} \mathrm{Pois}(\theta)$	$\Theta \sim \mathrm{Gamma}(\nu, s)$	$\Theta \mid X \sim \mathrm{Gamma}\left(\nu + \sum_{i=1}^n X_i, \frac{s}{ns+1}\right)$

Philosophy

Bayesian statistics gives us a lot of utility, in that we can calculate any weird property of the posterior distribution once we actually have the posterior distribution. But for that, we require the prior distribution. We want a principled method to choose the prior. The main ways to get a prior are

- Previous or parallel experiences (A/B testing). This is relatively non-controversial, we can fit this prior from data, and we can test the validity of the prior because we have many draws of the prior.
- Subjective beliefs. The prior represents epistemic uncertainty and the posterior represents a rational update of beliefs. We can't be wrong about the prior assignment (as it's our own opinion) and we can use hard-to-formalize knowledge from outside data. This represents a philosophical conundrum: we're discussing uncertainty in our own head and applying it to the real world. It's also generally impossible to truly write down beliefs about the prior distribution. But this is the most philosophically coherent way to do statistics.
- It can be generally very hard to compute the normalizing constant $\int_\Omega \lambda(\theta) p_\theta(x) \, \mathrm{d}\theta$. If $\dim(\Omega) \gg 0$, then the posterior is approximately 0 for most of Ω. It helps to use conjugate priors.
- "Objective" prior. Suppose $X_i \mid \Theta = \theta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$. The natural choice is the "flat prior", not biasing us towards any particular value of θ. In particular $\lambda(\theta) \propto_\theta 1$. This prior is improper but it's fine: $\lambda(\theta \mid x) \propto_\theta \mathrm{e}^{\theta \sum_{i=1}^n x_i - n\theta^2/2} \propto_\theta \mathcal{N}\left(\overline{x}, n^{-1}\right)$. The posterior mean is $\overline{X}$. This prior arises naturally as the limit of $\Theta \; sim \mathcal{N}\left(0, \tau^2\right)$ as $\tau^2 \to \infty$. One issue with the flat prior is that the "flat" prior depends on the parameterization of Θ; in particular if we reparameterize θ nonlinearly then the flat prior also becomes non-constant.
- Jeffreys proposed using $\lambda(\theta) \propto_\theta |J(\theta)|^{1/2}$. The binomial Jeffreys prior is $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$.

The only real uncertainty it makes sense to consider is our epistemic uncertainty. Rejecting this makes doing any statistics very untenable.
The advantages of using Bayes methods are

1. It's very easy and simple to define a "common sense," interpretable estimator, and in a general case you can even define the estimator pointwise.
2. It has appealing frequentist properties.
3. It has detailed output (i.e. you get a joint distribution back instead of an expectation).

The disadvantages of using Bayes methods are

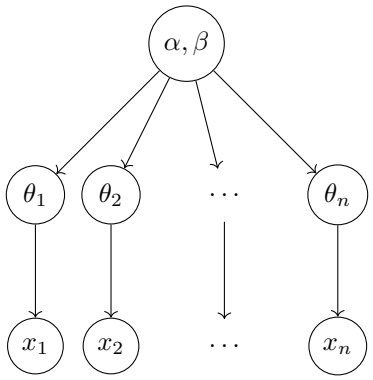
1. It's difficult to choose a good prior, and if you choose the wrong prior then your estimator is efficacy-capped by the efficacy of your prior.

2. We have to have a coherent opinion at everything. Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P$ where P is any probability measure. Suppose $g(P) = \mathbb{E}_P[X]$. We have shown that $\overline{X}$ is a UMVU estimator. This is a straightforward computation. However, if we want to use Bayesian methods, we need to put a prior on P, which is an (infinite-dimensional) set of probability measures. It's really hard to determine such a prior, and then the posterior will be very complicated, and even if we somehow estimate the posterior mean then the UMVU will probably be better anyways.

Hierarchical Bayes

The strongest argument for using Bayes is when we encounter many similar instances of the same problem and want to learn from previous predictions.

Example 6.11. Suppose $X \sim \text{Binom}(n, \theta)$. We want to use $\text{Beta}(\alpha, \beta)$ as a prior distribution for θ, and want to choose α, β based on pooling information from past instantiations of X, say $X_1, \dots, X_m$. Suppose $\alpha, \beta \sim \lambda_{\alpha,\beta}$, so that $\theta_i \mid \alpha, \beta \overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, \beta)$ and $X_i \mid \theta_i \overset{\text{i.i.d.}}{\sim} \text{Binom}(n_i, \theta_i)$. Using this information we can find the "true value" of α, β and thereby estimate θ. In this case θ_i are the **parameters**, conditionally independent on α, β, and α, β are hyperparameters.
The **directed graphical model** for this scenario is



In our model,

$$p(\alpha, \beta, \theta_1, \dots, \theta_m, x_1, \dots, x_m) = p(\alpha, \beta)\left(\prod_{i=1}^{m} p(\theta_i \mid \alpha, \beta)\right)\left(\prod_{i=1}^{m} p(x_i \mid \theta_i, \alpha, \beta)\right)$$

Definition 6.12 (Directed Graphical Model). The graph, a directed acyclic graph, tells us how the joint distribution can be factorized. If $G = (V, E)$ then

$$p(z_1, \dots, z_{|V|}) = \prod_{i=1}^{|V|} p_i(z_i \mid \text{Parents}_G(z_i)).$$

We can turn any model into a DAG, by making the in-edges to Z_i from $Z_1, \dots, Z_{i-1}$.
A common situation in hierarchical Bayes modeling is that we have one draw from the hyperparameter, $\zeta \sim \lambda(\zeta)$, n draws from the parameters $\theta_i \mid \zeta \overset{\text{i.i.d.}}{\sim} \pi_\zeta(\theta)$, and n draws from data $X_i \mid \zeta, \theta \overset{\text{i.i.d.}}{\sim} p_{\theta_i}(X_i)$. Using the **empirical Bayes** method, we may estimate ζ as $\widehat{\zeta}$ based on all the data, then plug in $\widehat{\zeta}$ as if it were known.

James-Stein

Suppose $\tau^2 \sim \lambda(\tau^2)$ (for example $\frac{1}{\tau^2} \sim \text{Gamma}(k, s)$) and $\theta_i \mid \tau^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2)$, with $X_i \mid \tau^2, \theta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_i, 1)$, for $i = 1, \dots, n$. Then the posterior mean of θ_i is

$$\begin{aligned}\delta_i(X) &= \mathbb{E}[\theta_i \mid X] \\ &= \mathbb{E}\left[\mathbb{E}\left[\theta_i \,\middle|\, X, \tau^2\right] \mid X\right] \\ &= \mathbb{E}\left[\frac{\tau^2}{1+\tau^2} X_i \,\middle|\, X\right]\end{aligned}$$

$$= \mathbb{E}\left[\frac{\tau^2}{1+\tau^2}\,\middle|\, X\right] \cdot X_i$$

Define $\zeta = \frac{1}{1+\tau^2}$. (If $\zeta = 0$ then there is no shrinkage, if $\zeta = 1$ then there is full shrinkage). Since

$$X_i \mid \tau^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, 1+\tau^2\big), \quad X \mid \zeta \sim \mathcal{N}\big(0, \zeta^{-1} I_n\big) \quad \text{and} \quad \mathsf{P}(X \mid \zeta) = \left(\frac{\zeta}{2\pi}\right)^{n/2} \mathrm{e}^{-\frac{\zeta}{2}\|X\|_2^2}$$

and so by the LLN,

$$X \mid \zeta \sim \operatorname{Gamma}\left(1 + \frac{n}{2}, \frac{2}{\|X\|_2^2}\right) \approx \mathcal{N}\left(\underbrace{\frac{n+2}{\sum_{i=1}^n X_i^2}}_{\approx \zeta}, \underbrace{2\frac{n+2}{\left(\sum_{i=1}^n X_i\right)^2}}_{\approx 0}\right).$$

The likelihood thus has a sharp peak at

$$\frac{n+2}{\|X\|_2^2} \approx \zeta,$$

so that

$$\mathbb{E}[\zeta \mid X] \approx \zeta,$$

and finally

$$\delta_i(X) = (1-\zeta) X_i.$$

James and Stein proposed a different estimator

$$\delta_{\mathrm{JS},i}(X) = \left(1 - \frac{d-2}{\|X\|_2^2}\right) X_i.$$

Proposition 6.13. If $Y \sim \chi_d^2 = \operatorname{Gamma}(\frac{d}{2}, 2)$ for $d > 3$ then

$$\mathbb{E}\left[\frac{1}{Y}\right] = \frac{1}{d-2}.$$

Proof. A straight computation gives

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{Y}\right] &= \int_0^\infty \frac{1}{y} \frac{1}{2^{d/2}\Gamma(\frac{d}{2})} y^{\frac{d}{2}-1} \mathrm{e}^{-y/2}\,\mathrm{d}y \\
&= \frac{2^{(d-2)/2}\Gamma(\frac{d-2}{2})}{2^{d/2}\Gamma(\frac{d}{2})} \underbrace{\int_0^\infty \frac{1}{2^{(d-2)/2}\Gamma(\frac{d-2}{2})} y^{\frac{d-2}{2}-1} \mathrm{e}^{-y/2}\,\mathrm{d}y}_{1} \\
&= \frac{2^{(d/2)/2}\Gamma(\frac{d-2}{2})}{2^{d/2}\Gamma(\frac{d}{2})} \\
&= \frac{1}{d-2}.
\end{aligned}$$

□

Thus $\frac{d-2}{\|X\|_2^2}$ is the UMVU of $\frac{1}{1+\tau^2}$ (as it's unbiased and a function of a complete sufficient statistic).

James-Stein Paradox

Back to the non-Bayesian Gaussian sequence model. Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(\theta, \sigma^2 I\big)$. For $d \geq 3$, X (which is UMVU, objective Bayes, minimax estimator, of θ) is *inadmissible* as an estimator of θ under squared error loss: if

$$\delta_{\mathrm{JS}}(X) = \left(1 - \frac{(d-2)\sigma^2}{n\|X\|_2^2}\right),$$

then

$$\mathrm{MSE}(\theta, \delta_{\mathrm{JS}}) < \mathrm{MSE}(\theta, X) \text{ for all } \theta \in \mathbb{R}^d$$

The reason this is better is because the JS estimator uses adaptive shrinking which uses the value of every coordinate, as opposed to forming d separate estimation problems which don't cross-inform. Going forward, we assume without loss of generality that $n = \sigma^2 = 1$. There's nothing special about 0 – we can also shrink towards any value:

$$\delta_{\mathrm{JS};\theta_0}(X) = \theta_0 + \left(1 - \frac{d-2}{\|X-\theta_0\|_2^2}\right)(X - \theta_0)$$

but the surprising fact is that this *also* dominates X for all values of θ (and θ_0):

$$\mathrm{MSE}(\theta, \delta_{\mathrm{JS};\theta_0}) < \mathrm{MSE}(\theta, X) \text{ for all } \theta, \theta_0 \in \mathbb{R}^d.$$

The deeper meaning here is that we are gaining by shrinkage because it reduces variance (at the cost of increasing bias) and the James-Stein estimator ensures that we don't over-shrink. A corollary for this is that admissibility isn't everything; the estimator $\delta(X) = X$ is still "good" in some sense, and UMVU, and minimax.

Lemma 6.14 (Stein's Lemma). Suppose $X \sim \mathcal{N}\big(\theta, \sigma^2\big)$, with $h(x) \colon \mathbb{R} \to \mathbb{R}$ differentiable. Further suppose

$$\mathrm{Cov}(X, h(X)), \mathbb{E}\big[\big|\dot{h}j(X)\big|\big] < \infty.$$

Then

$$\mathrm{Cov}(X, h(X)) = \mathbb{E}[(X-\theta)h(X)] = \sigma^2 \mathbb{E}\big[\dot{h}(X)\big].$$

Proof. We assume without loss of generality that $h(0) = 0$. First assume $\theta = 0$ and $\sigma^2 = 1$. Then

$$\begin{aligned}
\mathbb{E}[Xh(X)] &= \int_0^\infty xh(x)\phi(x)\,\mathrm{d}x + \int_{-\infty}^0 xh(x)\phi(x)\,\mathrm{d}x \\
\int_0^\infty xh(x)\phi(x)\,\mathrm{d}x &= \int_0^\infty x\left(\int_0^x \dot{h}(y)\,\mathrm{d}y\right)\phi(x)\,\mathrm{d}x \\
&= \mathit{int}_0^\infty \int_0^\infty \mathbb{1}(y < x)x\dot{h}(y)\phi(x)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^\infty \dot{h}(y)\left(\int_y^\infty x\phi(x)\,\mathrm{d}x\right)\mathrm{d}y \\
&= \int_0^\infty \dot{h}(y)\left(\int_y^\infty -\dot{\phi}(x)\,\mathrm{d}x\right)\mathrm{d}y \\
&= \int_0^\infty \dot{h}(y)\phi(y)\,\mathrm{d}y \\
\int_{-\infty}^0 xh(x)\phi(x)\,\mathrm{d}x &= \int_{-\infty}^0 \dot{h}(y)\phi(y)\,\mathrm{d}y \\
\mathbb{E}[Xh(X)] &= \mathbb{E}\big[\dot{h}(X)\big].
\end{aligned}$$

Now assume general θ, σ^2 and write $X = \theta + \sigma Z$, where $Z \sim \mathcal{N}(0,1)$. Then

$$\begin{aligned}
\mathbb{E}[(X-\theta)h(X)] &= \sigma\mathbb{E}[Zh(\theta + \sigma Z)] \\
&= \sigma\mathbb{E}\left[\frac{\mathrm{d}h(\theta+\sigma Z)}{\mathrm{d}Z}\right] \\
&= \sigma^2\mathbb{E}\left[\frac{\mathrm{d}h(\theta+\sigma Z)}{\mathrm{d}\theta + \sigma Z}\right] \\
&= \sigma^2\mathbb{E}\big[\dot{h}(X)j\big]
\end{aligned}$$

as desired. □

Suppose, if $h \colon \mathbb{R}^d \to \mathbb{R}^d$ is differentiable, we define the differential operator D by $Dh \in \mathbb{R}^{d\times d}$ for which $[Dh(x)]_{ij} = \frac{\partial h_i(x)}{\partial x_j}$.

Theorem 6.15 (Multivariate Stein's Lemma). Suppose $X \sim \mathcal{N}\big(\theta, \sigma^2 I\big)$, where $\theta \in \mathbb{R}^d$, $h \colon \mathbb{R}^d \to \mathbb{R}^d$ is differentiable and

$\mathbb{E}[\|Dh(X)\|_F] < \infty$. Then

$$\mathbb{E}\big[(X-\theta)'h(X)\big] = \sigma^2\mathbb{E}[\operatorname{tr}(Dh(x))] = \sum_{i=1}^d \mathbb{E}\left[\frac{\partial h_i(x)}{\partial x_i}\right].$$

Proof. The expectation is the sum of d terms of the form $\mathbb{E}[(X_i-\theta_i)h_i(X)]$. We use

$$\begin{aligned}\mathbb{E}[(X_i-\theta_i)h_i(X)] &= \mathbb{E}[\mathbb{E}[(X_i-\theta_i)h_i(X) \mid X_j : j \neq i]] \\ &= \mathbb{E}\left[\mathbb{E}\left[\sigma^2\frac{\partial h_i(X)}{\partial x_i} \,\middle|\, X_j : j \neq i\right]\right] \\ &= \sigma^2\mathbb{E}\left[\frac{\partial h_i(X)}{\partial x_i}\right]\end{aligned}$$

from which the claim follows. Note here the second equality follows because $X_i \mid \{X_j : j \neq i\} \sim \mathcal{N}\big(\theta_i, \sigma^2\big)$. □

Stein's Unbiased Risk Estimator (SURE)

We can use Stein's Lemma to estimate the MSE of any $\delta(X)$, by applying it to $h(X) = X - \delta(X)$. Assume $\sigma^2 = 1$. Then

$$\begin{aligned}\operatorname{MSE}(\theta,\delta) &= \mathbb{E}_\theta\Big[\|\delta(X)-\theta\|_2^2\Big] \\ &= \mathbb{E}_\theta\Big[\|X-\theta-h(X)\|_2^2\Big] \\ &= \mathbb{E}_\theta\Big[\|X-\theta\|_2^2\Big] + \mathbb{E}\Big[\|h(X)\|_2^2\Big] - 2\mathbb{E}_\theta\big[(X-\theta)'h(X)\big] \\ &= d + \mathbb{E}_\theta[\|h(X)\|]_2^2 - 2\mathbb{E}_\theta[\operatorname{tr}(Dh(x))]\end{aligned}$$

If we define

$$\widehat{\operatorname{MSE}}(X) = d + \|h(X)\|_2^2 \operatorname{tr}(Dh(X)),$$

then

$$\mathbb{E}_\theta\Big[\widehat{\operatorname{MSE}}(X)\Big] = \operatorname{MSE}(\theta,\delta) \text{ for all } \theta$$

and $\widehat{\operatorname{MSE}}(X)$ is an unbiased estimator for $\operatorname{MSE}(\theta, d)$. We can also use this to calculate the MSE via $\operatorname{MSE}(\theta,\delta) = \mathbb{E}_\theta\Big[\widehat{\operatorname{MSE}}(X)\Big]$.

Example 6.16. If $\delta(X) = X$ and $h(X) = 0$ and $Dh(X) = 0$, then $\widehat{\operatorname{MSE}}(X) = \operatorname{MSE}(\theta,\delta) = d$.

Example 6.17. Suppose $\delta_\zeta(X) = (1-\zeta)X$ for fixed ζ. Then $h(X) = \zeta X$ so $Dh(X) = \zeta I$. Then

$$\widehat{\operatorname{MSE}}(X) = d + \zeta^2\|X\|_2^2 - 2\zeta d = (1-2\zeta)d + \zeta^2\|X\|_2^2,$$

while we can obtain by taking the expectation that

$$\operatorname{MSE}(\theta,\delta) = \big(1 - 2\zeta = \zeta^2\big)d + \zeta^2\|\theta\|_2^2 = \underbrace{(1-\zeta)^2 d}_{\operatorname{Var}_\theta(\delta)} + \underbrace{\zeta^2\|\theta\|_2^2}_{\operatorname{Bias}_\theta(\delta)^2}.$$

To find the optimal ζ, we obtain

$$\begin{aligned}\frac{\partial\zeta}{\partial \operatorname{MSE}(\theta,\delta_\zeta)\zeta} &= 2\zeta\|\theta\|_2^2 - 2(1-\zeta)d \\ \zeta^*(\theta) &= \operatorname*{argmin}_\zeta R(\theta;\delta_\zeta) \\ &= \frac{d}{d+\|\theta\|_2^2}.\end{aligned}$$

which is a term in the James-Stein estimator.

Let $h(x) = \dfrac{d-2}{\|X\|_2^2}X$. Then

$$\|h(X)\|_2^2 = \frac{(d-2)^2}{\|X\|_2^4}\|X\|_2^2$$

$$
\begin{aligned}
&= \frac{d-2}{\|X\|_2^2} \\
\frac{\partial h_i(x)}{\partial x_i} &= \frac{\partial}{\partial x_i} \frac{(d-2)X_i}{\sum_{j=1}^d X_j^2} \\
&= (d-2)\frac{\|X\|_2^2 - 2X_i^2}{\|X\|_2^4} \\
\operatorname{tr}(Dh(x)) &= \frac{d-2}{\|X\|_2^4} \sum_{i=1}^d \left(\|X\|^2 - 2X_i^2\right) \\
&= \frac{(d-2)^2}{\|X\|_2^2} \\
\widehat{R} &= \mathbb{E}\left[\widehat{\operatorname{MSE}}(\theta)\right] \\
&= d + \frac{(d-2)^2}{\|X\|_2^2} - 2\frac{(d-2)^2}{\|X\|_2^2} \\
&= d - \frac{(d-2)^2}{\|X\|_2^2} \\
R(\theta; \delta_{\mathrm{JS}}) &= d - \underbrace{(d-2)^2 \mathbb{E}_\theta\left[\frac{1}{\|X\|_2^2}\right]}_{>0} \\
&< d \\
&= R(\theta; X).
\end{aligned}
$$

If $\theta = 0$ then

$$\mathbb{E}_\theta\left[\frac{1}{\|X\|_2^2}\right] = d-2,$$

so

$$R(\theta, \delta_{\mathrm{JS}}) = d - (d-2) = 2 \ll d.$$

If θ is large then $\mathbb{E}_\theta\left[\frac{1}{\|X\|_2^2}\right] \approx \frac{1}{\|\theta\|_2^2}$, so

$$R(\theta, \delta_{\mathrm{JS}}) \approx d - \frac{(d-2)^2}{\|\theta\|_2^2} < d,$$

which always has a smaller risk than $\delta(X) = X$, but it's always better.
The James-Stein estimator is also inadmissible; taking the positive part is better:

$$\delta_{\mathrm{JS}+}(x) = \left(1 - \frac{d-2}{\|X\|_2^2}\right)_+ X$$

which is better because the shrinkage factor should never be negative. A practically more useful version of the estimator is where we shrink to $\overline{X}$:

$$\delta_{\mathrm{JS};2}(X) = \overline{X} + \left(1 - \frac{d-2}{\left\|X - \overline{X}1\right\|_2^2}\right)\left(X - \overline{X}1\right),$$

where $\overline{X} = \frac{1}{d}\sum_{i=1}^d X_i$. This dominates $\delta(X) = X$ for $\delta \geq 4$.
It's important to note that the James-Stein estimator improves $\mathbb{E}\left[\|\delta - \theta\|_2^2\right]$, not necessarily $\mathbb{E}\left[(\delta_i - \theta_i)^2\right]$ for any given i, so that pooling isn't always useful.
The James-Stein estimator can be viewed as an empirical Bayes estimator for the hierarchical distribution $\theta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\overline{X}, \tau^2\right)$, and $X_i \mid \theta \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, 1)$.

7 Minimax Estimation

The idea is to choose an estimator by minimizing the worst-case risk, which is $\sup_\theta R(\theta;\delta)$. The minimum achievable worst-case risk is called the minimax risk of the estimation function,

$$r^* = \inf_\delta \sup_\theta R(\theta;\delta).$$

δ^* is called **minimax** if it achieves minimax risk:

$$\sup_\theta R(\theta,\delta^*) = r^*.$$

There is a game theoretic interpretation. For Bayesian estimation, we assume nature is "nice," giving us predictable results and fixed θ – we go second, in this two-player game. For minimax estimation, we assume nature is "mean," picking adversarial θ for our estimator δ. – we go first, in this two-player game.
The minimax estimator is closely related to the Bayes estimator. In all cases,

$$\mathbb{E}[R(\theta,\delta)] \le \sup_\theta R(\theta,\delta),$$

and for any proper prior Λ, the Bayes risk is

$$\begin{aligned} r_\Lambda &= \inf_\delta \mathbb{E}_\Lambda[R(\theta;\delta)] \\ &\le \inf_\delta \sup_\theta R(\theta,\delta) \\ &= r^*. \end{aligned}$$

If δ_Λ is a Bayes estimator for Λ, then

$$r_\Lambda = \int_\Omega R(\theta,\delta_\Lambda)\,\mathrm{d}\Lambda(\theta)\,,$$

so that any Bayes risk lower bounds r^*. The **least favorable prior** Λ^* gives the best lower bound $r_{\Lambda^*} = \sup_\Lambda r_\Lambda$. Thus

$$\underbrace{\sup_\theta R(\theta,\delta)}_{\text{for all } \delta} \ge r^* \ge r_{\Lambda^*} \ge \underbrace{r_\Lambda}_{\text{for all } \Lambda}.$$

Thus if we can find a pair (δ,Λ) for which $\sup_\theta R(\theta,\delta) = r_\Lambda$, then all inequalities become equalities, and the Bayes estimator is equal to the minimax estimator.

Theorem 7.1. If Λ is a prior, with Bayes estimator δ_Λ, if

$$r_\Lambda = \sup_\theta R(\theta,\delta_\Lambda)$$

then

(i) δ_Λ is minimax,

(ii) If δ_Λ is the unique Bayes estimator (up to equality a.e.) for Λ, then it is the unique minimax estimator.

(iii) Λ is a least favorable prior.

Proof. For (a): for any other δ,

$$\begin{aligned} \sup_\theta R(\theta,\delta) &\ge \int_\Omega R(\theta,\delta)\,\mathrm{d}\Lambda(\theta) \\ &\ge \int_\Omega R(\theta,\delta_\Lambda)\,\mathrm{d}\Lambda(\theta) \\ &= r_\Lambda \\ &= \sup_\theta R(\theta,\delta_\Lambda) \text{ by assumption} \end{aligned}$$

so r_Λ is minimax-risk and δ_Λ is minimax. For (b): replace "$\ge$" by "$>$".
For (c): take another prior $\widetilde{\Lambda}$. Then

$$r_{\widetilde{\Lambda}} = \inf_\delta \int_\Omega R(\theta,\delta)\,\mathrm{d}\widetilde{\Lambda}(\theta)$$

$$\begin{aligned}
&\leq \int_\Omega R(\theta, \delta_\Lambda)\,\mathrm{d}\widetilde{\Lambda}(\theta) \\
&\leq \sup_\theta R(\theta, \delta_\Lambda) \\
&= r_\Lambda \text{ by assumption.}
\end{aligned}$$

Thus Λ is least favorable. □

Off the top of my head, if the parameter space is compact then we can find a least favorable prior.
This theorem gives us a checkable condition:

$$\text{does } \mathbb{E}[R(\theta, \delta)] = \sup_\theta R(\theta, \delta)?$$

This is true for a Bayes estimator δ_Λ if

1. $R(\theta; \delta_\Lambda)$ is constant (on the support of Λ).
2. $$\Lambda\left(\left\{\theta \,\middle|\, R(\theta, \delta_\Lambda) = \max_\zeta R(\zeta, \delta_\Lambda))\right\}\right) = 1.$$

Example 7.2. Suppose $X \sim \mathrm{Binom}(n, \theta)$, and we want to estimate θ with squared error loss. We guess a prior $\Lambda = \mathrm{Beta}(\alpha, \beta)$ and let

$$\delta_{\alpha,\beta}(X) = \frac{\alpha + X}{\alpha + \beta + n}.$$

Then

$$\begin{aligned}
R(\theta; \delta_{\alpha,\beta}(X)) &= \mathbb{E}_\theta\left[\left(\frac{\alpha + X}{\alpha + \beta + n} - \theta\right)^2\right] \\
&= \mathrm{Var}_\theta\left(\frac{X}{\alpha + \beta + n}\right) + \left(\frac{\alpha + n\theta}{\alpha + \beta + n} - \theta\right)^2 \\
&= \frac{n\theta(1-\theta)}{(\alpha + \beta + n)^2} + \left(\frac{\alpha + n\theta}{\alpha + \beta + n} - \theta\right)^2 \\
&\propto_\theta \Big((\alpha + \beta)^2 - n\Big)\theta^2 + (n - 2\alpha(\alpha + \beta))\theta + \alpha^2.
\end{aligned}$$

The solution here is for $\alpha = \beta = \frac{\sqrt{n}}{2}$, $\mathrm{Beta}(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2})$ is least favorable prior. Thus the minimax estimator is

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

Another example: If $X \sim \mathcal{N}(\theta, 1)$ then the least favorable prior "should" spread measure everywhere.
The least favorable prior is a supremum, which is not always achievable.

Definition 7.3. A sequence $\{\Lambda_n\}$ is least favorable if

$$\lim_{n\to\infty} r_{\Lambda_n} \to \sup_\Lambda r_\Lambda.$$

Theorem 7.4. Suppose $\{\Lambda_n\}$ is a prior sequence and δ satisfies $\sup_\theta R(\theta, \delta) = \lim_{n\to\infty} r_{\Lambda_n}$. Then

1. δ is minimax, and
2. $\{\Lambda_n\}$ is least favorable.

Proof. Take any other estimator $\widetilde{\delta}$. Then for all n,

$$\begin{aligned}
\sup_\theta R(\theta, \widetilde{\delta}) &\geq \int_\Omega R(\theta, \widetilde{\delta})\,\mathrm{d}\Lambda_n(\theta) \\
&\geq r_{\Lambda_n} \\
\sup_\theta R(\theta, \widetilde{\delta}) &\geq \sup_n r_{\Lambda_n}
\end{aligned}$$

$$\begin{aligned}
&\geq \lim_{n\to\infty} r_{\Lambda_n} \\
&= \sup_\theta R(\theta, \delta).
\end{aligned}$$

Now for any least favorable prior Λ,

$$\begin{aligned}
r_\Lambda &= \inf_\zeta \int_\Omega R(\theta, \zeta)\,\mathrm{d}\Lambda(\theta) \\
&\leq \int_\Omega R(\theta, \delta)\,\mathrm{d}\Lambda(\theta) \\
&\leq \sup_\theta R(\theta, \delta) \\
&= \lim_{n\to\infty} r_{\Lambda_n}.
\end{aligned}$$

□

8 Sampling Methods

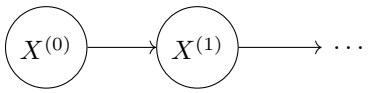
It's very easy to turn graphical models very complicated and unsuitable for direct computation. Let's consider what we can actually compute, for the posterior distribution. If

$$\lambda(\theta \mid x) = \frac{p_\theta(x)\lambda(\theta)}{\int_\Omega p_\zeta(x)\lambda(\zeta)\,\mathrm{d}\zeta}$$

then $p_\theta(x)\lambda(\theta)$ is usually easy to compute, but $\int_\Omega p_\zeta(x)\lambda(\zeta)\,\mathrm{d}\zeta$ is usually a high dimensional integral and completely intractable. We want to computationally estimate this integral. Our strategy is to set up a Markov chain with stationary distribution $\propto p_\theta(x)\lambda(\theta)$ and run it to get approximate samples from $\lambda(\theta \mid x)$.

Definition 8.1 (Markov Chain). A (stationary) Markov chain with transition kernel (or transition probabilities) $Q(y \mid x)$ and initial distribution $\pi_0(x)$ is a sequence of random variables $X^{(0)}, X^{(1)}, \ldots$, where $X^{(0)} \sim \pi_0$ and $X^{(t+1)} \mid X^{(0)}, \ldots, X^{(t)} \sim Q(\cdot \mid X^{(t)})$.

We can "think" of $Q(y \mid x) = \mathbb{P}\big(X^{(t+1)} = y \mid X^{(t)} = x\big)$, which is precise in the discrete case but not the continuous case. This gives a directed graphical model



Then we define the distribution of $X^{(t)}$, π_t, recursively as follows:

$$\begin{aligned}
\pi_t(y) &= \mathbb{P}\Big(X^{(t)} = y\Big) \\
&= \int_{\mathcal{X}} \mathbb{P}\Big(X^{(t)} = y \,\Big|\, X^{(t-1)} = x\Big)\pi_{t-1}(x)\,\mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} Q(y \mid x)\pi_{t-1}(x)\,\mathrm{d}\mu(x)
\end{aligned}$$

Definition 8.2 (Stationary Distribution). If $\pi(y) = \int_{\mathcal{X}} Q(y \mid x)\pi(x)\,\mathrm{d}\mu(x)$ we say π is a **stationary distribution** for Q.

A sufficient condition for stationarity of a distribution π is the **detailed balance equations**:

Proposition 8.3. If

$$\pi(x)Q(y \mid x) = \pi(y)Q(x \mid y)\ \forall x, y$$

then π is stationary.

Proof. If

$$\pi(x)Q(y \mid x) = \pi(y)Q(x \mid y)\ \forall x, y$$

then for all y,

$$\begin{aligned}\int_{\mathcal{X}} Q(y \mid x)\pi(x)\,\mathrm{d}\mu(x) &= \pi(y)\int_{\mathcal{X}} Q(x \mid y)\,\mathrm{d}\mu(x)\\ &= \pi(y).\end{aligned}$$

□

Definition 8.4 (Reversible Markov Chain). A Markov chain with the detailed balance property is called **reversible**. That is, $p\big(X^{(0)},\dots,X^{(t)}\big) \overset{\text{a.e.}}{=} p\big(X^{(t)},\dots,X^{(0)}\big)$.

Theorem 8.5. If a Markov chain with a stationary distribution π is

1. **irreducible** (for all x, y there exists n for which $p(X^{(n)} = y \mid X^{(0)} = x) = 0$),
2. and **aperiodic** (for all x, $\gcd\big(\big\{n \geq 0 \,\big|\, p(X^{(n)} = x \mid X^{(0)} = x)\big\}\big) = 1$).

Then $\pi_t \to \pi$ (under the total variation norm) regardless of π_0.

Thus if we have irreducibility and aperiodicity, and run many iterations of the Markov chain update, then we get close to the stationary distribution. One downside of this estimate is that we're not sure how many iterations it'll take to get a "good" approximation.

Our strategy is to find Q with stationary distribution proportional to $\lambda(\theta \mid x)$, prove irreducibility and aperiodicity, start at any X, run the chain for a long time, and pick $X^{(t)}$ a random sample from $\lambda(\theta \mid x)$.

Gibbs Sampling

One particular method we use is Gibbs sampling. Suppose we have a parameter vector $\theta = (\theta_1, \dots, \theta_d)$.

Algorithm 1 Gibbs Sampling

Input: $\theta^{(0)} \in \mathbb{R}^d$
Output: Sampled $\theta \sim \lambda(\theta \mid X)$
  $\theta \leftarrow \theta^{(0)}$
  for $t = 1, \dots, T$ **do**
    for $j = 1, \dots, d$ **do**
      $\theta_j \sim \lambda(\theta_j \mid \{\theta_i \,|\, i \neq j\}, X)$
    $\theta^{(t)} \leftarrow \theta$
  return $\theta^{(T)}$

Variations include updating coordinates at random and in random orders.

For hierarchical priors there is an advantage:

$$\lambda(\theta_j \mid \theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots \theta_n, X) \propto p(\theta_j \mid \{\theta_i \,|\, i \in \text{Parents}(j)\}) \prod_{i:\, j \in \text{Parents}(i)} p(\theta_i \mid \{\theta_k \colon k \in \text{Parents}(i)\})$$

In theory we can take any initialization $\theta^{(0)}$ and a valid (irreducible, aperiodic) kernel Q, sample long enough (T steps, where T is large), and $\theta^{(T)} \equiv \lambda(\theta \mid X)$.

Trace plots show how the Markov chain is mixing. In particular we plot $\theta_j^{(t)}$ against t (trace plot) and see when we approximately reach equilibrium. Good plots converge fast; bad plots converge extremely slowly or not at all. In addition, some Markov chains mix pathologically, in that it can swap between arbitrary modes for long times before finally converging on the correct value.

If we get to a region of the trace plot which indicates convergence of $\theta_j^{(t)}$, then the time leading up to that is called the "burn-in" period (say B is the index of the end of the burn-in period). Sampling at a lower frequency (say once every s time-step) is called "thinning". Then the posterior mean is given by

$$\lim_{N\to\infty} \frac{1}{N+1} \sum_{k=0}^{N} \theta_j^{(B+ks)} \mathbb{E}[\theta_j \,|\, X].$$

We want to find out why the posterior density is stationary for the Gibbs sampling mechanism. We consider updating a fixed coordinate j and hold the other coordinates fixed, resampling θ_j from the conditional distribution. We assume that

$$\theta^{(t)} \sim \lambda(\theta \mid X)$$

and define

$$\theta^{(t+1)} = \left(\theta_1^{(t)}, \dots, \theta_{j-1}^{(t)}, \theta_j^{(t+1)}, \theta_j^{(t)}, \dots, \theta_d^{(t)}\right),$$

so that

$$\theta_{\text{not } j}^{(t+1)} = \theta_{\text{not } j}^{(t)} \sim \lambda(\theta_{\text{not } j} \mid X)$$

so that the transition kernel Q for the whole Markov chain is $Q_1, \dots, Q_d$ in order.

Example 8.6. Suppose $\theta \sim \mathcal{N}(0, I_2)$ and $X_i \mid \theta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_1 + \theta_2, 1)$, for $i = 1, \dots, n$. Then

$$\begin{bmatrix} \theta \\ \overline{X} \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} I & 1 \\ 1 & 2 + \frac{1}{n} \end{bmatrix}\right).$$

Then

$$\begin{aligned}
\theta \mid \overline{X} &\sim \mathcal{N}\big(\mu(\overline{X}), \sigma(\overline{X})\big) \\
\mu(\overline{X}) &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(2 + \frac{1}{n}\right)^{-1} \overline{X} = \frac{n\overline{X}}{2n+1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
\Sigma(\overline{X}) &= I - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(2 + \frac{1}{n}\right)^{-1} \begin{bmatrix} 1 & 1 \end{bmatrix} \\
&= \frac{n+1}{2n+1} \begin{bmatrix} 1 & -n/(n+1) \\ -n/(n+1) & 1 \end{bmatrix}
\end{aligned}$$

Because of the high correlations, Gibbs sampling takes a long time to mix. A better parameterization would be to take $\beta_1 = \theta_1 + \theta_2$, and $\beta_2 = \theta_1 - \theta_2$, which are marginally independent and also conditionally independent given X.

9 Hypothesis Testing

In **hypothesis testing**, we use data X to infer which of two submodels generated X. Suppose we have the model $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. The **null hypothesis** is $H_0 : \theta \in \Theta_0$; the **alternate hypothesis** is $H_1 : \theta \in \Theta_1$. Whenever H_1 is unspecified, we assume $\Theta_1 = \Theta \setminus \Theta_0$.
H_0 is the default choice: we either

1. accept H_0 (more accurately, fail to reject H_0, no definite conclusion), or
2. reject H_0 (conclude that Θ_0 is false, and Θ_1 is true).

Example 9.1. Suppose $X \sim \mathcal{N}(\theta, 1)$. Then two alternate hypotheses are:

$$H_0 : \theta \le 0 \text{ vs } H_1 : \theta > 0$$

and

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta \ne 0.$$

Example 9.2. If $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$, then $H_0 : P = Q$ and $H_1 : P \ne Q$.

A common conceptual objection is that we "know" $\theta \ne 0$ or $P \ne Q$ already, since it's very unlikely these quantities are exactly equal. We will return to this objection later.

Power Functions

We can describe a test formally by its **critical function** (or **test function**):

$$\phi(x) = \begin{cases} 0 & \text{accept } H_0 \\ \pi \in (0,1) & \text{accept } H_0 \text{ with probability } \pi \\ 1 & \text{reject } H_0 \end{cases}.$$

In practice, randomization is rarely used, so that $\phi(\mathcal{X}) = \{0, 1\}$. For non-randomized ϕ, the **rejection region** is $R = \{x \colon \phi(x) = 1\}$, and $A = \mathcal{X} \setminus R$ is called the **acceptance region**.
The power function is

$$\beta_\phi(\theta) = \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta(\text{reject } H_0)$$

and it fully summarizes the behavior of the test.
The **significance level** of ϕ is $\sup_{\theta \in \Theta_0} \beta_\phi(\theta)$. We say ϕ is a **level-α test** if its significance level is $\leq \alpha \in [0, 1]$. The ubiquitous choice is $\alpha = 0.05$. The goal is to maximize $\beta_\phi(\theta)$ on Θ, subject to level-α constraint.

Likelihood Ratio Test

A **simple** hypothesis is a single distribution:

$$\Theta_0 = \{\theta_0\} \text{ or } \Theta_1 = \{\theta_1\}.$$

When the null and alternative hypotheses are both simple, there exists a unique best test which rejects for large values of the likelihood ratio:

$$\phi^*(x) = \begin{cases} 1 & \frac{p_1(x)}{p_0(x)} > c \\ \gamma & \frac{p_1(x)}{p_0(x)} = c \\ 0 & \frac{p_1(x)}{p_0(x)} < c \end{cases}$$

where p_0, p_1 are null and alternative densities. A dominating measure always exists, for example $P_0 + P_1$. In this case c and γ are chosen to make $\mathbb{E}_{\theta_0}[\phi^*(x)] = \alpha$. ϕ^* is called the **likelihood ratio test** (LRT).

Neyman-Pearson

Proposition 9.3 (Keener 12.1). Suppose $c \geq 0$ and ϕ^* maximizes the Lagrangian:

$$\mathbb{E}_{\theta_1}[\phi(X)] - c\mathbb{E}_{\theta_0}[\phi(X)]$$

among all critical functions ϕ. If $\mathbb{E}_{\theta_0}[\phi^*(X)] = \alpha$ then ϕ^* maximizes $\mathbb{E}_{\theta_1}[\phi(X)]$ among all level-α tests ϕ.

Proof. Suppose $\mathbb{E}_{\theta_0}[\phi(X)] \leq \alpha$. Then

$$\begin{aligned} \mathbb{E}_{\theta_1}[\phi(X)] &\leq \mathbb{E}_{\theta_1}[\phi(X)] + c(\alpha - \mathbb{E}_{\theta_0}[\phi(X)]) \\ &\leq \mathbb{E}_{\theta_1}[\phi^*(X)] - c\mathbb{E}_{\theta_1}[\phi^*(X)] + c\gamma \\ &= \mathbb{E}_{\theta_1}[\phi^*(X)]. \end{aligned}$$

□

Theorem 9.4 (Neyman-Pearson Lemma). The likelihood ratio test with significance level α is optimal for testing $H_0 \colon X \sim p_0$ vs $H_1 \colon X \sim p_1$.

Proof. Maximize the Lagrangian:

$$\begin{aligned} \mathbb{E}_{p_1}[\phi(X)] - c\mathbb{E}_{p_0}[\phi(X)] &= \int_{\mathcal{X}} (p_1(x) - cp_0(x))\phi(x) \, \mathrm{d}\mu(x) \\ &= \int_{p_1 > cp_0} |p_1 - cp_0|\phi \, \mathrm{d}\mu - \int_{p_1 < cp_0} |p_1 - cp_0|\phi \, \mathrm{d}\mu \end{aligned}$$

Maximizing the first term, we get $\phi(x) = 1$ when $p_1(x) > cp_0(x)$. Minimizing the second term, we get $\phi(x) = 0$ when $p_1(x) < cp_0(x)$. We choose the minimum c such that

$$\mathbb{P}_{p_0}\left(\frac{p_1(x)}{p_0(x)} > c\right) \leq \alpha \leq \mathbb{P}_{p_0}\left(\frac{p_1(x)}{p_2(x)} \geq c\right)$$

and choose γ to "top up" the significance level:

$$\mathbb{P}_{p_0}\left(\frac{p_1}{p_0} > c\right) + \gamma\mathbb{P}_{p_0}\left(\frac{p_1}{p_0} = c\right) = \alpha.$$

□

Keener gives the converse up to wiggle room if $\frac{p_1(x)}{p_0(x)} = c$ for more than one value of X.

Proposition 9.5 (Keener 12.4). If $p_0 \overset{\mathbb{P}\text{-a.e.}}{\neq} p_1$ and ϕ is a likelihood ratio test with level $\alpha \in (0,1)$, then $\mathbb{E}_{p_1}[\phi(X)] > \alpha$.

Proof. By assumption $\mu(\{x \,|\, p_1(x) > p_0(x)\}) > 0$ and $\mu(\{x \,|\, p_1(x) < p_0(x)\}) > 0$. If $c \geq 1$ then

$$\text{power} - \alpha = \mathbb{E}_{p_1}[\phi^*(X)] - \mathbb{E}_{p_2}[\phi^*(X)] = \int_{p_1 > p_0} |p_1 - p_0| \phi^* \, \mathrm{d}\mu - \int_{p_1 < p_0} |p_1 - p_0| \phi^* \, \mathrm{d}\mu = \int_{p_1 > p_0} |p_1 - p_0| \phi^* \, \mathrm{d}\mu > 0.$$

If $c < 1$ then

$$(1 - \text{power}) - (1 - \alpha) = \mathbb{E}_{p_1}[1 - \phi^*(X)] - \mathbb{E}_{p_0}[1 - \phi^*(X)] = -\int_{p_1 < p_0} |p_1 - p_0| (1 - \phi^*) \, \mathrm{d}\mu < 0.$$

□

Example 9.6 (One-Parameter Exponential Family). Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\eta(x) = \mathrm{e}^{\eta T(x) - A(\eta)} h(x)$. Let $H_0 \colon \eta = \eta_0$ vs $H_1 \colon \eta = \eta_1 > \eta_0$. Then

$$\begin{aligned} \frac{p_1(x)}{p_0(x)} &= \frac{\prod_{i=1}^n p_{\eta_1}(x_i)}{\prod_{i=1}^n p_{\eta_0}(x_i)} \\ &= \frac{\mathrm{e}^{\eta_1 \sum_{i=1}^n x_i - nA(\eta_1)}}{\mathrm{e}^{\eta_0 \sum_{i=1}^n x_i - nA(\eta_0)}} \\ &= \mathrm{e}^{(\eta_1 - \eta_0) \sum_{i=1}^n x_i - n(A(\eta_1) - A(\eta_0))}. \end{aligned}$$

ϕ^* rejects for large $\sum_{i=1}^n T(x_i)$:

$$\phi^*(x) = \begin{cases} 0 & \sum_{i=1}^n T(x_i) < c \\ \gamma & \sum_{i=1}^n T(x_i) = c \\ 1 & \sum_{i=1}^n T(x_i) > c \end{cases}.$$

We choose c, γ for which

$$\mathbb{P}_{\eta_0}\left(\sum_{i=1}^N T(x_i) > c\right) + \gamma \mathbb{P}_{\eta_0}\left(\sum_{i=1}^n T(x_i) = c\right) = \alpha.$$

Surprise: $\phi^*(x)$ depends only on η_0 and $\operatorname{sign}(\eta_1 - \eta_0)$, not on η_1.

Uniformly Most Powerful Tests

Now we will talk about uniformly most powerful tests. Fix $\mathcal{P}_1$, Θ_0, Θ_1.

Definition 9.7 (Uniformly Most Powerful). If ϕ^* has significance level α, and for any other level-α test ϕ we have

$$\mathbb{E}_\theta[\phi^*] \geq \mathbb{E}_\theta[\phi]$$

for all $\theta \in \Theta$, then ϕ^* is **uniformly most powerful (UMP)**.

Typically UMP tests only exist for 1-sided testing in certain 1-parameter families.

Definition 9.8 (Identifiable). A model $\mathcal{P}$ is identifiable if

$$\theta_1 \neq \theta_2 \implies \mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}.$$

Definition 9.9 (Monotone Likelihood Ratios). Assume $\mathcal{P} = \{\mathbb{P}_\theta \colon \theta \in \Theta \subseteq \mathbb{R}\}$ has densities p_θ and is identifiable. We say $\mathcal{P}$ has **monotone likelihood ratios** (MLR) if there is some statistic $T(X)$ for which

$$\theta_1 < \theta_2 \implies \frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} \text{ is a nondecreasing function of } T(X).$$

We define $\frac{c}{0} = \infty$ if $c > 0$, and $\frac{0}{0}$ undefined.

The one-parameter exponential family has monotone likelihood ratios in $\sum_{i=1}^n T(x)$.

We define $\frac{c}{0} = \infty$ if $c > 0$, and $\frac{0}{0}$ undefined.

Theorem 9.10. Assume $\mathcal{P}$ has MLR, test $H_0\colon \theta = \theta_0$ vs $H_1\colon \theta > \theta_0$ at level $\alpha = (0,1)$. Let $\phi^*(x) = \begin{cases} 0 & T(x) < c \\ \gamma & T(x) = c \\ 1 & T(x) > c \end{cases}$ with c, γ chosen so that $\mathbb{E}_{\theta_0}[\phi^*(X)] = \alpha \in (0,1)$. Then

(a) ϕ^* is a UMP level-α test.

(b) $\mathbb{E}_\theta[\phi^*(X)]$ is non-decreasing in θ and strictly increasing wherever $\mathbb{E}_\theta[\phi^*(X)] \in (0,1)$.

(c) If $\theta_1 < \theta_0$ then ϕ^* minimizes $\mathbb{E}_{\theta_1}[\phi(X)]$ among all tests ϕ with $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$.

Proof. For (b): Suppose $\theta_1 < \theta_2$, then $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}$ is a non-decreasing function of $T(x)$. Thus ϕ^* is a likelihood ratio test for $H_0\colon \theta = \theta_1$ vs $H_1\colon \theta = \theta_2$ with level $\widetilde{\alpha} = \mathbb{E}_{\theta_1}[\phi^*(X)]$. By Corollary 12.4, $\mathbb{E}_{\theta_2}[\phi(X)] \geq \mathbb{E}_{\theta_1}[\phi(X)]$, a strict inequality unless both sides are 0 or 1.
For (a): Suppose $\phi_1 > \phi_0$ and $\widetilde{\phi}$ has level $\leq \alpha$. Then $\mathbb{E}_{\theta_1}[\phi^*(X)] \geq \mathbb{E}_{\theta_1}\left[\widetilde{\phi}(X)\right]$ since ϕ^* is a likelihood ratio test for $H_0\colon \theta = \theta_0$ vs $H_1\colon \theta = \theta_1$.
For (c): Suppose $\theta_1 < \theta_0$, assume $\mathbb{E}_{\theta_0}\left[\widetilde{\phi}(X)\right] = \mathbb{E}_{\theta_0}[\phi^*(X)] = \alpha$. Both $1 - \phi^*$ and $1 - \widetilde{\phi}$ are tests of $H_0\colon \theta = \theta_0$ vs $H_1\colon \theta = \theta_1$. Both have significance level $1 - \alpha$. The test $1 - \phi^*$ is a likelihood ratio test since $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)}$ is non-increasing in $T(x)$. Thus

$$\mathbb{E}_{\theta_1}\left[1 - \widetilde{\phi}\right] \leq \mathbb{E}_{\theta_1}[1 - \phi^*] = 1 - \alpha$$

as desired. □

The intuition is that ϕ^* is a likelihood ratio test for $H_0\colon \theta = \theta_0$ vs $H_1\colon \theta = \theta_1$ for any pair $\theta_0 < \theta_1$ (where the significance level depends on θ_0). This lets us extend our simple vs simple result to a very special case of composite vs composite.
We now move onto two-sided alternatives.
The setup is that $\mathcal{P} = \{\mathbb{P}_\theta\colon \theta \in \Theta \subseteq \mathbb{R}\}$, $\theta_0 \in \Theta^o$, and the test is $H_0\colon \theta = \theta_0$ vs $H_1\colon \theta \neq \theta_0$. It can be generalized naturally to $H_0\colon \theta \in [\theta_1, \theta_2]$.

Definition 9.11 (Stochastically Increasing). A real-valued statistic $T(x)$ is **stochastically increasing** in θ if $\mathbb{P}_\theta(T(x) \leq t)$ is non-increasing in θ, for all t.

Assume $T(x)$ is a stochastically increasing summary test statistic.

Example 9.12. If $X_i \overset{\text{i.i.d.}}{\sim} p(x - \theta)$ (location family) then $T(X) = \overline{X}$ or $T(X) = \text{Median}(X)$.

Example 9.13. If $X_i \overset{\text{i.i.d.}}{\sim} \frac{1}{\theta} p\left(\frac{X}{\theta}\right)$ (scale family), then $T(x) = \sum_{i=1}^n X_i^2$ or $T(X) = \text{Median}(|X|)$.

A two-tailed test rejects when $T(x)$ is extreme:

$$\phi(x) = \begin{cases} 1 & T(x) > c_2 \text{ or } T(x) < c_1 \\ \gamma_i & T(x) = c_i \\ 0 & T(x) \in (c_1, c_2) \end{cases}.$$

Let $\alpha_1 = \mathbb{P}_{\theta_2}(T(X) < c_1) + \gamma_1 \mathbb{P}_{\theta_0}(T(X) = c_1), \alpha_2 = \mathbb{P}_{\theta_2}(T(X) > c_2) + \gamma_2 \mathbb{P}_{\theta_0}(T(X) = c_2)$. We need $\alpha_1 + \alpha_2 = \alpha$, and we have some choice to set them.
The first idea is a **equal-tailed test**, that is, constrain α_1, α_2 to $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$.

Definition 9.14 (Unbiased Test). $\phi(x)$ is **unbiased** if $\inf_{\theta \in \Theta_1} \beta(\theta) \geq \alpha$.

The second idea is an unbiased test we ensure $\inf_\theta \beta(\theta) = \alpha$. Usually this is true if and only if $\dot{\beta}(\theta_0) = 0$.

Theorem 9.15. Assume $X_i \overset{\text{i.i.d.}}{\sim} e^{\theta T(x) - A(\eta)} h(x)$ with the test $H_0\colon \theta \in [\theta_1, \theta_2]$ vs $H_1\colon \theta \in \mathbb{R} \setminus [\theta_1, \theta_2]$. Then

(a) The unbiased test based on (can be written as a function of) $\sum_{i=1}^n T(X_i)$ with significance level α is UMP among all unbiased tests (UMPU).

(b) If $\theta_1 < \theta_2$ the UMPU test can be found by solving for c_i, γ_i such that $\beta(\theta_1) = \beta(\theta_2) = \alpha$.

(c) If $\theta_1 = \theta_2 = \theta_0$ the UMPU test can be found solving for c_i, γ_i such that $\beta(\theta_0) = \alpha$ and

$$\dot{\beta}(\theta_0) = \mathbb{E}_{\theta_0}\left[\sum_{i=1}^n T(X_i)(\phi(X) - \alpha)\right] = 0.$$

The proof is in Keener.

Definition 9.16 (p-Values, Informal). Suppose $\phi(X)$ rejects for large values of $T(X)$. Then

$$p(x) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X) > T(x)).$$

Example 9.17. If $X \sim \mathcal{N}(\theta_1)$ with $H_0 \colon \theta = 0$ and $H_1 \colon \theta \neq 0$, the two-sided test rejects for large $T(X) = |X|$ (and so $\phi_\alpha(X) = \mathbb{1}\left(|X| > z_{\alpha/2}\right)$). The two-sided p-value is $p(X)$ where

$$p(x) = \inf_{\theta \in \Theta_0} \mathbb{P}_\theta(|X| > |x|) = \mathbb{P}_0(|X| > |x|) = 2(1 - \Phi(|x|)).$$

For $H_0 \colon |\theta| < \delta$ vs $H_1 \colon |\theta| > \delta$,

$$p(x) = \mathbb{P}_\delta(|X| > |x|) = 1 - \Phi(|x| - \delta) + \Phi(-|x| - \delta).$$

Definition 9.18 (p-Values). Suppose we have a test ϕ_α for each significance level, $\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi_\alpha(X)] \leq \alpha$. (The non-randomized case is: $\phi_\alpha = \mathbb{1}(x \in R_\alpha)$). We assume tests are monotone in α:

$$\text{if } \alpha \leq \alpha_2 \text{ then } \phi_{\alpha_1}(x) \leq \phi_{\alpha_2}(x)$$

from which it follows that in the non-randomized case $R_{\alpha_1} \subseteq R_{\alpha_2}$. Then

$$p(x) = \inf\{\alpha \colon \phi_\alpha(x) = 1\} = \inf\{\alpha \colon x \in R_\alpha\}.$$

It's possible to define a randomized p-value, but not worth it.
Note that $p(x) \leq \alpha$ if and only if $\phi^\sim_\alpha(x) = 1$ for all $\widetilde{\alpha} > \alpha$.

$$\mathbb{P}_\theta(p(X) \leq \alpha) \leq \inf_{\widetilde{\alpha} > \alpha} \underbrace{\mathbb{P}_\theta\left(\phi^\sim_\alpha(X) = 1\right)}_{\leq \widetilde{\alpha}} \leq \alpha.$$

This implies that $p(x)$ stochastically dominates $\mathrm{Uni}([0,1])$.
If ϕ_α rejects for large $T(x)$, it coincides with the informal definition.
The p-value depends on

- the model,
- the null hypothesis,
- the data, and
- the choice of test.

Example 9.19. If $X \sim \mathcal{N}(\theta, I)$ and $H_0 \colon \theta = 0$ and $H_1 \colon \theta \neq 0$, we can use $T_1(X) = \|X\|_2^2$ (in the χ^2 test) or $T_2(X) = \|X\|_\infty = \max_{1 \leq i \leq d} |X_i|$ (in the max test). We get very different p-values and power if d is large. The choice affects the model.

Confidence Sets

Testing that $\theta \neq \theta_0$ is not really interesting. So we want to estimate our degree of certainty in our test.

Definition 9.20 (Confidence). $C(X)$ is a $1 - \alpha$ **confidence set** for $g(\theta)$ if

$$\mathbb{P}_\theta(g(\theta) \in C(X)) \geq 1 - \alpha \text{ for all } \theta \in \Theta.$$

We say $C(X)$ **covers** $g(\theta)$ if $g(\theta) \in c(X)$. The quantity $\mathbb{P}_\theta(g(\theta) \in C(X))$ is called the **coverage probability**, and $\inf_\theta \mathbb{P}_\theta(g(\theta) \in C(X))$ is the **confidence level**.

In some sense confidence sets are dual to hypothesis testing. Suppose we have a level-α test $\phi(x; a)$ of $H_0 \colon g(\theta) = a$ vs $H_1 \colon g(\theta) \neq \alpha$. We can use this to make a confidence set $C(x) = \{a \colon \phi(x; a) < 1\}$, all the non-rejected values of $g(\theta)$ for continuous distributions. Then

$$\begin{aligned} \mathbb{P}_\theta(C(X) \neq g(\theta)) &= \mathbb{P}_\theta(\phi(X; g(\theta)) = 1) \\ &\leq \alpha. \end{aligned}$$

Alternatively, suppose $C(x)$ is a $1 - \alpha$ confidence set for $g(\theta)$. We can use $C(x)$ to construct a test $\phi(x)$ of $H_0 \colon g(\theta) \in A$ vs $H_1 \colon g(\theta) \notin A$. Then $\phi(x) = \mathbb{1}(C(x) \cap A = \emptyset)$. Then for θ such that $g(\theta) \in A$,

$$\mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta(C(X) \cap A = \emptyset)$$

$$\begin{aligned}&\leq \mathbb{P}_\theta(C(X) \neq g(\theta))\\&\leq \alpha.\end{aligned}$$

This is called **inverting** the test or confidence set.
If $C(X) = [C_1(X), C_2(X)]$ we call it a **confidence interval** (CI). If $C(X) = [C_1(X), \infty)$ a **lower confidence bound** (LCB). If $C(X) = (-\infty, C_2(X)]$ a **upper confidence bound** (UCB). We get confidence intervals from two-sided tests; we get confidence bounds from one-sided tests. A confidence interval is called **uniformly most accurate** (UMA) if the corresponding test is UMP.

Example 9.21. Suppose $X \sim \operatorname{Exp}\left(\theta^{-1}\right) = \frac{1}{\theta}\mathrm{e}^{-X/\theta}$ for $x > 0$ and $\theta > 0$. Then $\mathbb{P}_\theta(X \leq x) = 1 - \mathrm{e}^{-x/\theta}$. To find the LCB we invert the test for $H_0 \colon \theta \leq \theta_0$. We solve $\alpha = \mathbb{P}_{\theta_0}(X > c(\theta_0)) = \mathrm{e}^{-c(\theta_0)/\theta_0}$. Then $c(\theta_0 = -\theta_0 \log(\alpha))$. If $X \leq c(\theta_0)$ then $\theta_0 \geq -\frac{X}{\log(\alpha)}$, so $C(X) = \left[-\frac{X}{\log(\alpha)}, \infty\right)$. Correspondingly the UCB is $C(X) = \left(0, -\frac{X}{\log(1-\alpha)}\right]$. The equal-tailed CI is obtained by inverting the equal-tailed test of $H_0 \colon \theta = \theta_0$ vs $H_1 \colon \theta \neq \theta_0$. Thus $\phi_\alpha^{\mathrm{2T}}(X) = \phi_{\alpha/2}^{\geq \theta_0}(X) + \phi_{\alpha/2}^{\leq \theta_0}$, so $C(X) = \left[-\frac{X}{\log\left(\frac{\alpha}{2}\right)}, -\frac{X}{\log\left(1-\frac{\alpha}{2}\right)}\right]$.

Common misinterpretations of hypothesis tests are:

1. If $p < 0.05$ it does *not* mean that there "is an effect" or "the effect size is exactly equal to the estimate.".
2. An even worse idea is that $p > 0.05$ means that "there is no effect" – but absence of evidence does not imply evidence of absence.
3. If p is very small that does not mean "the effect is huge". It's also a function of the amount of data collected.
4. Extrapolation is bad.

Nuisance Parameters

Usually we have extra unknown parameters which are not of direct interest.

Definition 9.22. Suppose our family is $\mathcal{P} = \{P_{\theta,\lambda} \colon (\theta, \lambda) \in \Omega\}$, where $H_0 \colon \theta \in \Theta_0$ vs $H_1 \colon \theta \in \Theta_1$. Then θ is the **parameter of interest** and λ is the **nuisance parameter**.

The issue is that λ is unknown but may affect type I error or power of a given test.

Example 9.23. Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, with $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\nu, \sigma^2\right)$, where μ, ν, σ^2 are all unknown. Suppose $H_0 \colon \mu = \nu$ and $H_1 \colon \mu \neq \nu$. Then $\theta = \mu - \nu$ and $\lambda = (\mu + \nu, \sigma^2)$ or $\lambda = (\mu, \sigma^2)$.

Example 9.24. Suppose $X_1 \sim \operatorname{Binom}(n_1, \pi_i)$ and $X_2 \sim \operatorname{Binom}(n_2, \pi_2)$, where n_1, n_2 are known constants (and thus *not* nuisance parameters). The test can be $H_0 \colon \pi_1 \leq \pi_2$ vs $H_1 \colon \pi_1 > \pi_2$.

Multiparameter Exponential Families

Suppose

$$X \sim p_{\theta,\lambda}(x) = \exp(\theta' T(x) + \lambda' U(x) - A(\theta, \lambda))h(x)$$

where $\theta \in \mathbb{R}^s$, $\lambda \in \mathbb{R}^r$, and both are unknown. The idea is to condition on the function $U(X)$ which acts as a sufficient statistic for λ, to eliminate dependency on θ. The steps are:

1. Let
$$(T(X), U(X)) \sim q_{\theta,\lambda}(t, u) = \mathrm{e}^{\theta' t + \lambda' u - A(\theta,\lambda)} g(t, u),$$
where the density is taken with respect to the product measure μ on $\mathbb{R}^{s+r}$.
2. Condition on $U(X)$:
$$\begin{aligned}q_\theta(t \mid u) &= \frac{q_{\theta,\lambda}(t, u)}{\int_{\mathbb{R}^s} q_{\theta,\lambda}(z, u)\,\mathrm{d}z}\\&= \frac{\mathrm{e}^{\theta' t + \lambda' u - A(\theta,\lambda)} g(t, u)}{\int_{\mathbb{R}^s} \mathrm{e}^{\theta' z + \lambda' u - A(\theta,\lambda)} g(z, u)\,\mathrm{d}z}\end{aligned}$$

$$= \frac{\mathrm{e}^{\theta' t} g(t,u)}{\int_{\mathbb{R}^s} \mathrm{e}^{\theta' z} g(z,u)\,\mathrm{d}z}$$
$$= \mathrm{e}^{\theta' t - B_u(\theta)} g(t,u)$$

where $B_u(\theta) = \log\left(\int_{\mathbb{R}^s} \mathrm{e}^{\theta' z} g(z,u)\,\mathrm{d}z\right)$.

3. Conditional test: test $H_0 \colon \theta \in \Theta_0$ vs $H_1 \colon \theta \in \Theta_1$ in the s-parameter model $\mathcal{Q}_u = \{q_\theta(t \mid u) \colon \theta \in \Theta\}$.

Note: If $s = 1$, this family has MLR in $T(x)$. Even if $s > 1$, we have still gotten rid of λ.

Theorem 9.25. If $\mathcal{P}$ is a full rank exponential family with densities

$$p_{\theta,\lambda}(x) = \mathrm{e}^{\theta T(x) + \lambda' U(x) - A(\theta,\lambda)} h(x),$$

with $\theta \in \mathbb{R}$, $\lambda \in \mathbb{R}^r$, $(\theta, \lambda) \in \Omega$, Ω is an open set, and θ_0 is achievable, then

(a) To test $H_0 \colon \theta \le \theta_0$ vs $H_1 \colon \theta > \theta_0$, there is a UMPU test $\phi^*(x) = \psi(T(x); U(x))$ where

$$\psi(t,u) = \begin{cases} 1, & t > c(u) \\ \gamma(u), & t = c(u) \\ 0, & t < c(u) \end{cases}$$

with $c(u), \gamma(u)$ chosen to make

$$\mathbb{E}_{\theta_0}[\phi^*(x) \,|\, U(x) = u] = \alpha.$$

(b) To test $H_0 \colon \theta = \theta_0$ vs $H_1 \colon \theta \ne \theta_0$, there is a UMPU test $\phi^*(x) = \psi(T(x); U(x))$ where

$$\psi(t,u) = \begin{cases} 1, & t < c_1(u) \text{ or } t > c_2(u) \\ \gamma_i(u), & t = c_i(u) \\ 0, & t \in (c_1(u), c_2(u)) \end{cases}$$

with $c_i(u), \gamma_i(u)$ chosen to make

$$\mathbb{E}_{\theta_0}[\phi^*(x) \,|\, U(x) = u] = \alpha \text{ and } \mathbb{E}_{\theta_0}[T(x)(\phi^*(x) - \alpha) \,|\, U(x) = u] = 0.$$

Solution Sketch. Any unbiased test has $\beta(\theta_0, \lambda) = \alpha$ for all λ. Since the power is α on the boundary, $\mathbb{E}_{\theta_0}[\phi \,|\, u] \overset{\text{a.e.}}{=} \alpha$ ($U(X)$ is complete sufficient on the boundary submodel). Thus we can show ϕ^* is optimal among all tests with conditional level α (by reduction to the univariate model). □

Note that λ has disappeared from the problem.

Proof. Assume ϕ is any unbiased test. Since $\mathbb{E}_{\theta,\lambda}[|\phi(X)| \le 1]$ for all $(\theta, \lambda) \in \Omega$, then $\mathbb{E}_{\theta,\lambda}[\phi(X)]$ is infinitely differentiable on Ω, and by Dominated Convergence Theorem we can differentiate under the integral sign.
We now look at the boundary submodel $\mathcal{P}_{\theta_0} = \{P_{\theta_0,\lambda} \colon (\theta_0, \lambda) \in \Omega\}$. Then

$$p_{\theta_0,\lambda}(x) = \mathrm{e}^{\lambda' U(x) - A(\theta_0,\lambda)} \mathrm{e}^{\theta_0 T(x)} h(x),$$

and since $\mathcal{P}_{\theta_0}$ is full-rank, s-parameter exponential family, then $U(x)$ is complete sufficient. Let $f(u) = \mathbb{E}_{\theta_0}[\phi(X) \,|\, U(x) = u] - \alpha$. Then

$$\mathbb{E}_{\theta_0,\lambda}[f(U(X))] = \mathbb{E}_{\theta_0,\lambda}[\phi(X)] - \alpha = 0$$

for all λ. Thus $f(u) \overset{\text{a.e.}}{=} 0$, so $\mathbb{E}_{\theta_0}[\phi(X) \,|\, U(X) = u] = 0$ for all u. In the two-sided case,

$$g(u) = \frac{\partial}{\partial\theta} \mathbb{E}_{\theta_0}[\phi(X) \,|\, U(X) = u] = \mathbb{E}_{\theta_0}[(T(X) - \mathbb{E}_{\theta_0}[T(X) \,|\, U(X) = u])\phi(X) \,|\, U(X) = u] = \mathbb{E}_{\theta_0}[T(X)(\phi(X) - \alpha) \,|\, U(X) = u].$$

Thus

$$\mathbb{E}_{\theta_0,\lambda}[g(U(X))] = \mathbb{E}_{\theta_0,\lambda}[T(X)(\phi(X) - \alpha)] = \frac{\partial}{\partial\theta} B_\phi(\theta_0) = 0$$

for all λ. Thus $\frac{\partial}{\partial\theta} \mathbb{E}_{\theta_0}[\phi(X) \,|\, U(X) = u] \overset{\text{a.e.}}{=} 0$.
The final step is that, for any u, the conditional model is

$$q_\theta(t \mid u) = \mathrm{e}^{\theta t - B_u(\theta)} g(t,u),$$

a one-parameter exponential family. In the one/two-sided case, we have shown that $\psi(t;u)$ is UMP(U) in $\mathcal{Q}_u$. Let $\overline{\phi}(t;u) = \mathbb{E}[\phi(X) \,|\, T(X) = t, U(X) = u]$. Then

$$\mathbb{E}_\theta\left[\overline{\phi}(T(X);u) \,\middle|\, U(X) = u\right] = \mathbb{E}_\theta[\phi(X) \,|\, U(X) = u] = \alpha \text{ if } \theta = \theta_0.$$

Thus $\overline{\phi}(\cdot;u)$ is a conditional test of H_0 vs. H_1 in $\mathcal{Q}_u$ with power α at the boundary. In the one-sided case, $\phi(t;u)$ is the UMP test of $\theta = \theta_0$ or $\theta \le \theta_0$ vs $\theta > \theta_0$ in $\mathcal{Q}_u$, with is a one-parameter exponential family. In the two-sided case, $\phi(t,u)$ is the UMP test of $\theta = \theta_0$ vs $\theta \ne \theta_0$ among tests φ with power α, and $\frac{\partial}{\partial\theta}\beta_\varphi(\theta_0) = 0$. In either case ψ has higher conditional power than $\overline{\phi}$, almost everywhere. For $(\theta,\lambda) \in \Omega$,

$$\begin{aligned}\mathbb{E}_{\theta,\lambda}[\phi(X)] &= \mathbb{E}_{\theta,\lambda}\left[\mathbb{E}_\theta\left[\overline{\phi}(T(X);u) \,\middle|\, U(X) = u\right]\right] \\ &\le \mathbb{E}_{\theta,\lambda}[\mathbb{E}_\theta[\psi(T(X);u) \,|\, U(X) = u]] \\ &= \mathbb{E}_{\theta,\lambda}[\phi^*(X)].\end{aligned}$$

This completes the proof. □

Example 9.26. Suppose $X_i \overset{\text{ind}}{\sim} \text{Pois}(\mu_i)$, for $i = 1, 2$. Let $H_0 \colon \mu_1 < \mu_2$ vs $H_1 \colon \mu_1 > \mu_2$. Then

$$p_\mu(x) = \prod_{i=1}^2 \frac{\mu_i^{X_i}\mathrm{e}^{-\mu_i}}{X_i!} = \exp(X_1\eta_1 + X_2\eta_2 - (\mathrm{e}^{\eta_1} + \mathrm{e}^{\eta_2}))\frac{1}{X_1!X_2!}$$

where $\eta_i = \log(\mu_i)$. Then $H_0 \colon \eta_1 \le \eta_2$ and $H_1 \colon \eta_1 > \eta_2$. Then

$$\begin{aligned}p_\mu(x) &= \exp(X_1\eta_1 + X_2\eta_2 - (\mathrm{e}^{\eta_1} + \mathrm{e}^{\eta_2})) \\ &= \exp\left(\underbrace{(X_1 + X_2)}_{T(X)}\underbrace{\left(\frac{\eta_1 - \eta_2}{2}\right)}_{\theta} + \underbrace{(X_1 + X_2)}_{U(X)}\underbrace{\left(\frac{\eta_1 + \eta_2}{2}\right)}_{\lambda} - A(\eta)\right)\frac{1}{X_1!X_2!}\end{aligned}$$

We reject for **conditionally** large values of $X_1 - X_2$, given $X_1 + X_2 = u$, which is the same as saying we reject for conditionally large values of X_1, given $X_1 + X_2 = u$. In particular

$$\begin{aligned}\mathbb{P}_\theta(X_1 = x \,|\, U = u) &\propto_x \exp((2x - u)\theta + u\lambda - A(\theta,\lambda))\frac{1}{x!(u-x)!} \\ &\propto_x \exp(2\theta x)\frac{u!}{x!(u-x)!} \\ &= \text{Binom}\left(u, \frac{\mathrm{e}^{2\theta}}{1 + \mathrm{e}^{2\theta}}\right) \\ &= \text{Binom}\left(u, \frac{\mu_1}{\mu_1 + \mu_2}\right)\end{aligned}$$

so in the end we do a binomial test.

Gaussian-Adjacent Distributions

If $Z_1, \dots, Z_d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ then

$$V = \|Z\|_2^2 = \sum_{i=1}^d Z_i^2 \sim \chi_d^2 = \text{Gamma}\left(\frac{d}{2}, 2\right),$$

for which

$$\mathbb{E}[V] = d \text{ and } \text{Var}(V) = 2d.$$

In particular

$$\frac{V}{d} \xrightarrow{p} 1 \text{ and } V \approx \mathcal{N}(d, 2d).$$

If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi_d^2$, with $Z \perp\!\!\!\perp V$, then

$$\frac{Z}{\sqrt{V/d}} \sim t_d \xrightarrow{d} \mathcal{N}(0,1) \text{ as } d \to \infty.$$

If $V_1 \sim \chi^2_{d_1}$ and $V_2 \sim \chi^2_{d_2}$ with $V_1 \perp\!\!\!\perp V_2$, then

$$\frac{V_1/d_1}{V_2/d_2} \sim F_{d_1,d_2} \xrightarrow{d} \frac{1}{d_1}\chi^2_{d_1} \text{ as } d_2 \to \infty.$$

If $T \sim t_d$ then $T^2 \sim F_{1,d}$.
Recall that if $Z \sim \mathcal{N}(\mu, \Sigma)$ then

$$AZ + b \sim \mathcal{N}(A\mu + b, A\Sigma A').$$

Example 9.27. Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(\mu, \sigma^2\big)$ with μ, σ^2 both unknown. Suppose $H_0 \colon \mu = 0$ vs $H_1 \colon \mu \neq 0$. Then

$$p_{\mu,\sigma^2}(x) = \exp\left(\underbrace{\frac{\mu}{\sigma^2}}_{\theta}\underbrace{\sum_{i=1}^n X_i}_{T(X)} - \underbrace{\frac{1}{2\sigma^2}}_{\lambda}\underbrace{\sum_{i=1}^n X_i^2}_{U(X)} - \frac{n\mu}{2\sigma^2}\right)\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}.$$

We condition on $U(X) = \|X\|_2^2$. In particular if $\mu = 0$ then

$$X \mid \|X\|_2^2 \overset{H_0}{\sim} \operatorname{Uni}\big(\|X\|_2 S^{n-1}\big) \text{ or } \frac{X}{\|X\|_2} \overset{H_0}{\sim} \operatorname{Uni}\big(S^{n-1}\big) \text{ independent of } \|X\|_2.$$

In this case, S^{n-1} is the unit sphere in n dimensions. The optimal test rejects when $\sum_{i=1}^n X_i$ is extreme given $\|X\|_2$. Also $\sqrt{(n-1)S^2} \sim \sqrt{\sigma^2\chi^2_{n-1}}$.
Let

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^n \big(X_i - \overline{X}\big)^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - 2\overline{X}\sum_{i=1}^n X_i^2 + n\overline{X}^2\right) \\ &= \frac{1}{n-1}\Big(\|X\|_2^2 - n\overline{X}^2\Big) \\ (n-1)S^2 &= \|X\|_2^2 - \left(\frac{1}{\sqrt{n}}1'X\right)^2 \\ &= \Big\|\operatorname{proj}_{\operatorname{span}(1)^\perp}(X)\Big\|_2^2 \end{aligned}$$

so we reject when

$$\frac{\sqrt{n}\overline{X}}{\sqrt{\|X\|_2^2 - n\overline{X}^2}} = \frac{\sqrt{n}\big(\overline{X}/\|X\|_2\big)}{\sqrt{1 - n\big(\overline{X}/\|X\|_2\big)^2}}$$

is extreme, and since $X/\|X\|_2$ is independent of $\|X\|_2$ under the null hypothesis there is no dependence on $\|X\|_2$. Thus we reject when $\frac{\sqrt{n}\overline{X}}{\sqrt{S^2}}$ is high. In particular $\frac{\sqrt{n}\overline{X}}{\sqrt{S^2}} \overset{H_0}{\sim} t_{n-1}$.
Why is this a t distribution? It has a lot to do with the orthogonal projections. Let $q_1 = \frac{1}{\sqrt{n}}$, and $q_2, \dots, q_n$ orthonormally for which $\{q_1, \dots, q_n\}$ is an orthonormal basis for $\mathbb{R}^n$. Let $Q = \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix} = \begin{bmatrix} q_1 & Q_r \end{bmatrix}$, and $X \sim \mathcal{N}\big(\mu 1, \sigma^2 I\big)$. Then in a new basis

$$Z = Q'X = \begin{bmatrix} q_1'X \\ Q_r'X \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \sqrt{n}\mu \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \sigma^2 I\right).$$

Let $Z_1 = \sqrt{n}\overline{X}$. If $Z_r = Q_r'X$, then $Z_r \perp\!\!\!\perp Z_1$ and $Z_r \sim \mathcal{N}\big(0, \sigma^2 I\big)$. Thus

$$\begin{aligned} (n-1)S^2 &= \|X\|_2^2 - \big(\sqrt{n}\overline{X}\big)^2 \\ &= \|Z\|_2^2 - (Z_1)^2 \end{aligned}$$

$$= \sum_{i=2}^{n} Z_i^2$$
$$\sim \sigma^2 \chi_{n-1}^2.$$

Thus

$$S^2 \sim \frac{\sigma^2}{n-1}\chi_{n-1}^2$$

and

$$\frac{\sqrt{n}\overline{X}}{\sqrt{S^2}} = \frac{Z_1}{\|Z_r\|_2^2/(n-1)} \sim \frac{\mathcal{N}\left(\sqrt{n}\mu, \sigma^2\right)}{\sqrt{\sigma^2\chi_{n-1}^2/(n-1)}} \sim \operatorname{nct}_{n-1}\left(\frac{\sqrt{n}\mu}{\sigma^2}\right).$$

Even if we don't get a UMPU test at the end, conditioning on null sufficient statistic still helps.

Example 9.28. Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P, Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} Q$, with $H_0 \colon P = Q, H_1 \colon P \neq Q$. Under H_0, $X_1, \dots, X_n, Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} P$. Let $(Z_1, \dots, Z_{n+m}) = (X_1, \dots, X_n, Y_1, \dots, Y_m)$. Then under H_0, $U(Z) = \left(Z_{(1)}, \dots, Z_{(n+m)}\right)$ is complete sufficient. Let Π_{n+m} be the set of permutations on $n+m$ elements. Then

$$(X, Y) \mid U \overset{H_0}{\sim} \operatorname{Uni}(\{\pi U \colon \pi \in \Pi_{n+m}\}).$$

Thus for *any* test statistic T, if $P = Q$, then

$$\mathbb{P}_{P,Q}(T(Z) \geq t \mid U(Z) = u) = \frac{1}{(n+m)!} \sum_{\pi \in \Pi_{n+m}} \mathbb{1}(T(\pi Z) \geq t).$$

In practice, we sample $\pi_1, \dots, \pi_B \overset{\text{i.i.d.}}{\sim} \Pi_{n+m}$ for i.e., $B = 1000$. Then $Z, \pi_1 Z, \dots, \pi_B Z \overset{\text{i.i.d.}}{\sim} \operatorname{Uni}(\Pi_{n+m} U)$ under H_0. Then the Monte-Carlo p value is

$$p \geq \frac{1}{1+B}\sum_{b=1}^{B} \mathbb{1}(T(Z) \leq T(\pi_b Z)) \overset{H_0}{\sim} \operatorname{Uni}\left(\left\{\frac{1}{1+B}, \dots, \frac{B-1}{1+B}, 1\right\}\right)$$

with equality if and only if all the Z_i are distinct.

Canonical Linear Model

Assume $Z = (Z_0, Z_1, Z_r)$ where $Z_0 \in \mathbb{R}^{d_0}$, $Z_1 \in \mathbb{R}^{d_1}$, and $Z_r \in \mathbb{R}^{d_r} = \mathbb{R}^{n-d_0-d_1}$. Suppose $Z \sim \mathcal{N}\left((\mu_0, \mu_1, 0), \sigma^2 I\right)$, where $\mu_0 \in \mathbb{R}^{d_0}$, $\mu_1 \in \mathbb{R}^{d_1}$, and $\sigma^2 > 0$. The test is $H_0 \colon \mu_1 = 0$ vs $H_1 \colon \mu_1 \neq 0$ (or possibly one-sided if $d_1 = 1$). The model is an exponential family:

$$p_{\mu,\sigma^2}(Z) \propto_Z \exp\left(\frac{2\mu_1' Z_1 + 2\mu_0' Z_0 - \|Z\|_2^2}{2\sigma^2}\right).$$

1. If σ^2 is known and $d_1 = 1$, we "condition on Z_0", so as to reject for large/small/extreme values of $Z_1 \mid Z_0$. Since $Z_0 \perp\!\!\!\perp Z_1$, it's equivalent to reject for large/small/extreme values of Z_1. The test statistic is $Z_1 \sim \mathcal{N}\left(\mu_1, \sigma^2\right)$, so that $\frac{Z_1}{\sigma} \overset{H_0}{\sim} \mathcal{N}(0,1)$ (this is the z-**test**).

2. If σ^2 is known and $d_1 \geq 1$ we reject for large values of $\|Z_1\|_2^2$. In particular $\frac{\|Z_1\|_2^2}{\sigma^2} \sim \chi_{d_1}^2$ (this is the χ^2-**test**).

3. If σ^2 is unknown and $d_1 = 1$, we condition on Z_0 and
$$\|Z\|_2^2 = \|Z_0\|_2^2 + \|Z_1\|_2^2 + \|Z_r\|_2^2,$$
so we can reject for large/small/extreme $Z_1 \mid Z_0$. Since $Z_0 \perp\!\!\!\perp Z_1$, this is the same as rejecting for large/small/extreme values of Z_1, or $\frac{Z_1}{\|Z\|_2}$, or $\frac{Z_1}{\|Z_r\|_2}$, or $\frac{Z_1}{\sqrt{\|Z_r\|_2^2/d_r}} \overset{H_0}{\sim} t_{d_r}$. This is the t-**test**.

4. If σ^2 is unknown and $d_1 \geq 1$, we condition on Z_0 and reject for conditionally large $\|Z_1\|_2^2$, equivalently rejecting for large $\frac{\|Z_1\|_2^2/d_1}{\|Z_r\|_2^2/d_r} \overset{H_0}{\sim} F_{d_1,d_r}$. This is the F-**test**.

Here $\frac{\|Z_r\|_2^2}{d_r} \sim \frac{\sigma^2}{d_r}\chi^2_{d_r}$ is an unbiased estimator of σ^2 as d_r is large. In particular $\mathbb{E}_{\sigma^2}\left[\widehat{\sigma}^2\right] = \sigma^2$ and $\mathrm{Var}_{\sigma^2}\left(\widehat{\sigma}^2\right) = \frac{2\sigma^2}{d_r}$. Accordingly let $\widehat{\sigma}^2 = \frac{\|Z_r\|_2^2}{d_r}$. If σ^2 is unknown then we just use $\widehat{\sigma}^2$ in place of σ but the tests are the same.

General Linear Model

Assume $Y \sim \mathcal{N}\left(\theta, \sigma^2 I\right)$ where $\sigma^2 > 0$ is known or unknown. We want to test $H_0 \colon \theta \in \Theta_0$ vs $\theta \in \Theta \setminus \theta_0$, where $\Theta_0 \subset \Theta$ is a linear subspace of $\mathbb{R}^n$. In particular $\dim(\Theta_0) = d_0$ and $\dim(\Theta) = d = d_0 + d_1$.
The idea is to change basis. Let $Q = \begin{bmatrix} Q_0 & Q_1 & Q_r \end{bmatrix}$, where $Q_0 \in \mathbb{R}^{n \times d_0}$ is an orthonormal basis for Θ_0, $Q_1 \in \mathbb{R}^{n \times d_1}$ is an orthonormal basis for $\Theta \cap \Theta_0^\perp$, and $Q_r \in \mathbb{R}^{n \times (n-d)}$ is an orthonormal basis for $\Theta^\perp$. Then $Q'Q = I$. Let

$$Z = Q'Y \sim \mathcal{N}\left(\begin{bmatrix} Q_0'\theta \\ Q_1'\theta \\ 0 \end{bmatrix}, \sigma^2 I\right),$$

so that the null hypothesis is $H_0 \colon Q_1'\theta = 0$ vs $H_1 \colon Q_1'\theta \neq 0$. Then we use the Z, t, χ^2, or F test as appropriate.

Example 9.29 (Linear Regression). Suppose $Y = X\beta + \varepsilon$, where X is fixed and $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$. Then $Y \sim \mathcal{N}\left(X\beta, \sigma^2 I\right)$. Let's assume X has full column rank. Let $\theta = X\beta \in \Theta = \mathrm{span}(X_{:,1}, \ldots, X_{:,d})$. The null hypothesis is $H_0 \colon \beta_1 = \cdots = \beta_{d_1} = 0$ for $1 \leq d_1 \leq d$, which happens if and only if $\theta \in \mathrm{span}(X_{:,d+1}, \ldots, X_{:,d})$ (or $\theta = 0$ if $d_1 = d$). Then $\|Z_r\|_2^2 = \|Y - \mathrm{proj}_\Theta(Y)\|_2^2$. It's well known that $\widehat{\beta}_{\text{OLS}} = (X'X)^{-1}X'Y$, so

$$\|Z_1\|_2^2 + \|Z_r\|_2^2 = \left\|Y - \mathrm{proj}_{\theta_0}(Y)\right\|_2^2 = \mathrm{RSS}_0\left(\widehat{\beta}_{\text{OLS}}\right).$$

The F-statistic is

$$\frac{\|Z_1\|_2^2/(d-d_0)}{\|Z_r\|_2^2/(n-d)} = \frac{(\mathrm{RSS}_0\left(\widehat{\beta}_{\text{OLS}}\right) - \mathrm{RSS}_\theta\left(\widehat{\beta}_{\text{OLS}}\right))/(d-d_0)}{\mathrm{RSS}_\theta\left(\widehat{\beta}_{\text{OLS}}\right)/(n-d)} \sim F_{d-d_0, n-d}.$$

The term $n - d$ is called the **residual degrees of freedom**.

Example 9.30 (One-Way ANOVA). Suppose $Y_{ki} \overset{\text{ind}}{\sim} \mu_k + \varepsilon_{ki}$, where $\varepsilon_{ki} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$, for $k = 1, \ldots, m$ and $i = 1, \ldots, n$. Then $H_0 \colon \mu_1 = \cdots = \mu_m = \mu$ (in this case $d_0 = 1$) vs $H_1 \colon \mu_1, \ldots, \mu_m$ anything else ($d > m$). Define

$$\overline{Y}_k = \frac{1}{n}\sum_{i=1}^n Y_{ki} \quad \text{and} \quad S_k^2 = \frac{1}{n-1}\sum_{i=1}^n \left(Y_{ki} - \overline{Y}_k\right)^2 \quad \text{and} \quad \overline{Y} = \frac{1}{mn}\sum_{k=1}^m \sum_{i=1}^n Y_{ki}$$

Then

$$d_0 = 1, d = m, d_r = mn - d = m(n-1).$$

We have

$$\mathrm{RSS}_{H_1}(\{\mu_k\}) = \sum_{k=1}^m \sum_{i=1}^n \left(Y_{ki} - \overline{Y}_k\right)^2 = \|Y\|_F^2 - n\sum_{k=1}^m \overline{Y}_k^2 \quad \text{and} \quad \mathrm{RSS}_{H_0}(\{\mu_k\}) = \sum_{k=1}^m \sum_{i=1}^n \left(Y_{ki} - \overline{Y}\right)^2 = \|Y\|_F^2 - mn\overline{Y}^2.$$

Then

$$\mathrm{RSS}_{H_1}(\{\mu_k\}) - \mathrm{RSS}_{H_0}(\{\mu_k\}) = n\left(\sum_{k=1}^m \overline{Y}_k^2 - m\overline{Y}^2\right) = n\sum_{k=1}^m \left(\overline{Y}_k - \overline{Y}\right)^2.$$

The F-statistic is

$$F\text{-stat} = \frac{\frac{n}{m-1}\sum_{k=1}^m \left(\overline{Y}_k - \overline{Y}\right)^2}{\frac{1}{m(n-1)}\sum_{k=1}^m \sum_{i=1}^n \left(Y_{ki} - \overline{Y}_k\right)^2} = \frac{\text{between-groups variance}}{\text{within-groups variance}}.$$

10 Asymptotics

Let $\{X_n\} \in \mathbb{R}^d$ be a sequence of random vectors.

Definition 10.1 (Convergence in Probability). $\{X_n\}$ **converges in probability** to X (that is, $X_n \xrightarrow{p} X$) if for every $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(\|X_n - X\|_p \geq \varepsilon\right) = 0.$$

In particular since all ℓ^p norms are equivalent we can take any convenient value for p (usually $p = 1, 2, \infty$).

Definition 10.2 (Convergence in Distribution). $\{X_n\}$ converges in distribution to X (that is, $X_n \xrightarrow{d} X$) if for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)].$$

Theorem 10.3. If $F_n(x) = \mathbb{P}(X_n \le x)$ and $F(x) = \mathbb{P}(X \le x)$ then $X_n \xrightarrow{d} X$ if and only if $F_n(x) \to F(x)$ for all continuity points x of F.

Example 10.4. If $X_n \sim \delta_{1/n}$, then $X \sim \delta_0$, then $X_n \xrightarrow{d} X$, since $F_n(x) = \mathbb{1}_{\ge \frac{1}{n}} \to \mathbb{1}_{\ge 0} = F(x)$ except at $x = 0$.

Proposition 10.5. $X_n \xrightarrow{p} c$ if and only if $X_n \xrightarrow{d} c$.

Definition 10.6 (Consistency). For a sequence of statistical models $\mathcal{P}_n = \{P_{n,\theta} : \theta \in \Theta\}$, if $X_n \sim P_{n,\theta}$, we say $\delta_n(X_n)$ is consistent for $g(\theta)$ if $\delta_n(X_n) \xrightarrow{\mathbb{P}_\theta} g(\theta)$, meaning

$$\lim_{n\to\infty} \mathbb{P}_\theta(\|\delta_n(X_n) - g(\theta)\| > \varepsilon) = 0 \text{ for all } \varepsilon > 0.$$

We can say an estimator is consistent for some θ and not others, but without specifying we mean that the estimator is consistent for all θ.

Limit Theorems

Let $\{X_n\}$ be a sequence of i.i.d. random vectors and define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Theorem 10.7 (Law of Large Numbers (LLN)). If $\mathbb{E}[|X_i|] < \infty$ and $\mathbb{E}[X_i] = \mu$ then

$$\overline{X}_n \xrightarrow{\text{a.e.}} \mu \implies \overline{X}_n \xrightarrow{p} \mu, \overline{X}_n \xrightarrow{d} \mu.$$

Theorem 10.8 (Central Limit Theorem). If $\mathbb{E}[|X_i|] < \infty$ and $\mathbb{E}[X_i] = \mu$, $\operatorname{Var}(X_i) = \Sigma$, then

$$\overline{X}_n \xrightarrow{d} \mathcal{N}(\mu, \Sigma).$$

Theorem 10.9 (Continuous Mapping Theorem). If g is continuous and $X_n \xrightarrow{d} X$ then $g(X_n) \xrightarrow{d} g(X)$. Similarly if $X_n \xrightarrow{p} X$ then $g(X_n) \xrightarrow{p} g(X)$.

Corollary 10.10 (Slutsky's Theorem). If c is a constant and $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ then $X_n + Y_n \xrightarrow{d} X + c$ and $X_n Y_n \xrightarrow{d} cX$; if c is nonzero then $X_n / Y_n \xrightarrow{d} X/c$; if $X_n \xrightarrow{p} X$ then $X_n + Y_n \xrightarrow{p} X + c$ $X_n Y_n \xrightarrow{p} cX$; if c is nonzero then $X_n / Y_n \xrightarrow{p} X/c$.

Delta Method

Theorem 10.11 (δ-Method for Approximation). If $\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and $f(x)$ is differentiable at $x = \mu$ then $\sqrt{n}(f(X_n) - f(u)) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \left(\dot{f}(\mu)\right)^2\right)$.

Intuitively, if $X_n \sim \mathcal{N}(\mu, \sigma^2/n)$ and f is locally linear, then $f(X_n) \sim \mathcal{N}\left(f(\mu), \sigma^2 \left(\dot{f}(\mu)\right)^2 / n\right)$.

Proof. First order Taylor approximation gives

$$f(X_n) = f(\mu) + \dot{f}(\mu)(X_n - \mu) + o(X_n - \mu).$$

Thus

$$\sqrt{n}(f(X_n) - f(\mu)) = \dot{f}(\mu)\sqrt{n}(X_n - \mu) + \underbrace{\sqrt{n} \cdot o(X_n - \mu)}_{\xrightarrow{p} 0} \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \left(\dot{f}(\mu)\right)^2\right).$$

□

The multivariate analogue is that if $f : \mathbb{R}^d \to \mathbb{R}^k$ is differentiable then

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma) \implies \sqrt{n}(f(X_n) - f(\mu)) \xrightarrow{d} \mathcal{N}(0, (Df(\mu))'\Sigma(Df(\mu))).$$

Example 10.12. Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} (\mu, \sigma^2)$ (a distribution with mean μ and σ^2), and $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} (\nu, \tau^2)$, with $X \perp\!\!\!\perp Y$. By LLN $\overline{X} \xrightarrow{p} \mu$ and $\overline{Y} \xrightarrow{p} \nu$ so $\left(\overline{X} + \overline{Y}\right)^2 \xrightarrow{p} (\mu + \nu)^2$. We want to know the rate of convergence. We know $\sqrt{n}(\overline{X} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$ and $\sqrt{n}(\overline{Y} - \nu) \xrightarrow{d} \mathcal{N}\left(0, \tau^2\right)$. Then by the delta method,

$$\left(\overline{X} + \overline{Y}\right)^2 \approx \mathcal{N}\left((\mu + \nu)^2, \begin{bmatrix} 2(\mu + \nu) \\ 2(\mu + \nu) \end{bmatrix} \begin{bmatrix} \sigma^2 & \\ & \tau^2 \end{bmatrix} \begin{bmatrix} 2(\mu + \nu) \\ 2(\mu + \nu) \end{bmatrix}\right) = \mathcal{N}\left((\mu + \nu)^2, \frac{4(\mu + \nu)^2(\sigma^2 + \tau^2)}{n}\right).$$

This gives the estimator $\delta(X, Y) = (\overline{X} + \overline{Y})^2$ as an estimator of $(\mu + \nu)^2$. Then δ is consistent, asymptotically normal and unbiased.

One major point is that if $(\mu + \nu)^2 = 0$ then the conclusion still holds, that is, $\sqrt{\operatorname{Var}(\delta(X, Y))}\delta(X, Y) \xrightarrow{p} 0$. Note $\sqrt{n}\overline{X} + \sqrt{n}\overline{Y} \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 + \tau^2\right)$ by the continuous mapping theorem and the delta method, so $n(\overline{X} + \overline{Y})^2 \xrightarrow{d} (\sigma^2 + \tau^2)\chi_1^2$.

In general we can do higher-order Taylor derivatives if some derivatives are equal to 0. In particular

$$f(X_n) \approx \underbrace{f(\mu)}_{O(1)} + \underbrace{\dot{f}(\mu)(X_n - \mu)}_{O_p(n^{-1/2})} + \underbrace{\frac{\left(\ddot{f}(\mu)\right)^2}{2}(X_n - \mu)^2}_{O_P(n^{-1})} + \cdots.$$

If $\dot{f}(\mu) = 0$, we can use the second order term:

$$n(f(X_n) - f(\mu)) \approx \frac{\ddot{f}(\mu)^2}{2}\left(\sqrt{n}(X_n - \mu)\right)^2 \approx \frac{\ddot{f}(\mu)^2}{2}\sigma^2\chi_1^2.$$

MLE

Suppose $\mathcal{P}$ has densities p_θ. A simple estimator for θ is

$$\widetilde{\theta}_{\text{MLE}}(x) = \operatorname*{argmax}_{\theta \in \Theta} p_\theta(x) = \operatorname*{argmax}_{\theta \in \Theta} \log(p_\theta(x)).$$

A non-parametric estimator could be

$$\widetilde{p}_{\text{MLE}}(x) = \operatorname*{argmax}_{p \in \mathcal{P}} p(x).$$

Remark 10.13. The set of θ which achieve the maximum could have multiple elements, or no elements, or not be easily computable.

Remark 10.14. The maximum likelihood estimator doesn't depend on parameterization, or base measure. In particular the MLE for $g(\theta)$ is $g(\widehat{\theta}_{\text{MLE}})$.

Example 10.15. For exponential families of the form

$$p_\eta(x) = \mathrm{e}^{\eta' T(x) - A(\eta)} h(x),$$

so

$$\ell(\eta; x) = \log(p_\eta(x)) = \eta' T(x) - A(\eta) + \log(h(x)).$$

Thus

$$\boldsymbol{\nabla}_\eta \ell(\eta; x) = T(x) - \mathbb{E}_\eta[T(X)],$$

so

$$\widehat{\eta}_{[MLE]} \quad \text{solves} \quad T(x) = \mathbb{E}_\eta[T(X)].$$

There may not be an MLE if we take the parameter space too small. We have uniqueness because $\boldsymbol{\nabla}_\eta^2 \ell(\eta; x) = -\boldsymbol{\nabla}_\eta^2 A(\eta) = -\operatorname{Var}_\eta(T(X))$ is negative definite unless $\nu' T(X) \overset{\text{a.e.}}{=} c$ for some ν, c, in which case there is a redundant natural parameter. Thus at most one solution exists. If ψ is the inverse of $\mu(\eta) = \boldsymbol{\nabla}_\eta A(\eta) = \mathbb{E}_\eta[T(X)]$ then $\widehat{\eta}_{\text{MLE}}(x) = \psi(T(x))$.

Example 10.16. Suppose $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{e}^{\eta T(x) - A(\eta)} h(x)$ for $\eta \in \Xi \subseteq \mathbb{R}$. Then $\widehat{\eta}_{\text{MLE}} = \psi(\overline{T})$ where $\overline{T}(X) = \sum_{i=1}^n T(X_i)$.
Assume $\eta \in \Xi^o$, so that $\dot{\mu}(\eta) = \ddot{A}(\eta) > 0$ for all $\eta \in \Xi^o$, so if ψ is continuous then

$$\dot{\psi}(\mu(\eta)) = (\dot{\mu}(\eta))^{-1} = \left(\ddot{A}(\eta)\right)^{-1}.$$

For consistency, we know by LLN that $\overline{T} \xrightarrow{p_\eta} \mu(\eta)$, so

$$\widehat{\eta}_{\text{MLE}}(x) = \psi(\overline{T}(x)) \xrightarrow{p_\eta} \psi(\mu(\eta)) = \eta$$

by the continuous mapping theorem. Then

$$\sqrt{n}(\overline{T} - \mu(\eta)) \xrightarrow{d} \mathcal{N}(0, \operatorname{Var}_\eta(T(X))) = \mathcal{N}\big(0, \ddot{A}(\eta)\big).$$

The delta method gives

$$\sqrt{n}(\widehat{\eta}_{\text{MLE}} - \eta) = \sqrt{n}(\psi(\overline{T}) - \psi(\mu(\eta))) \xrightarrow{d} \mathcal{N}\Big(0, \big(\ddot{A}(\eta)\big)^{-1}\Big).$$

The Fisher information is $J(\eta) = \operatorname{Var}_\eta(T(X)) = \ddot{A}(\eta)$. Thus $\widehat{\eta}_{\text{MLE}}(X) \approx \mathcal{N}\Big(\eta, \frac{1}{nT(\theta)}\Big)$. In the asymptotic distribution $\widehat{\eta}$ is unbiased and achieves the Cramer-Rao lower bound.

Example 10.17. If $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \operatorname{Pois}(\theta)$, and $\eta = \log(\theta)$. Then $\widehat{\eta}(X) = \log\big(\overline{X}\big)$, we know that $\sqrt{n}(\overline{X} - \theta) \xrightarrow{d} \mathcal{N}(0, \theta)$, or more generally $\overline{X} \approx \mathcal{N}\big(\theta, \frac{\theta}{n}\big)$. Then

$$\widehat{\eta}_{\text{MLE}}(X) = \log\big(\overline{X}\big) \approx \mathcal{N}\left(\log(\theta), \frac{\theta}{n\theta^2}\right) = \mathcal{N}\left(\log(\theta), \frac{1}{\theta n}\right).$$

Then

$$\sqrt{\widehat{\eta}_{\text{MLE}} - \eta} \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\theta}\right).$$

Finally,

$$\mathbb{P}_\theta(\widehat{\eta}_{\text{MLE}}(X) = -\infty) = \mathbb{P}_\theta(X_1 = 0)^n = \mathrm{e}^{-\theta n} > 0.$$

Thus

$$\mathbb{E}_\eta[\widehat{\eta}_{\text{MLE}}(X)] = -\infty \quad \text{and} \quad \operatorname{Var}_\eta(\widehat{\eta}(X)) = \infty,$$

but wait, what gives? This apparently achieves the Cramer-Rao lower bound, but the estimator sucks. However, in the end there is no contradiction – convergence in distribution doesn't imply convergence in moments. Most of the time the estimator is sensible but in the end there is some nontrivial probability of getting $-\infty$.

Proposition 10.18. If $\mathbb{P}(B_n) \to 0$ and $X_n \xrightarrow{p} X$, with Z_n arbitrary, then

$$X_n \mathbb{1}_{B_n^c} + Z_n \mathbb{1}_{B_n} \xrightarrow{p} X.$$

(And similarly for convergence in distribution).

Proof. We start with

$$\mathbb{P}(\|Z_n \mathbb{1}_{B_n}\|_2 > \varepsilon) \le \mathbb{P}(B_n) \to 0.$$

Thus $Z_n \mathbb{1}_{B_n} \xrightarrow{p} 0$. Also

$$\mathbb{1}_{B_n^c} \xrightarrow{p} 1.$$

Applying Slutsky's theorem gets

$$X_n \mathbb{1}_{B_n^c} + Z_n \mathbb{1}_{B_n} \xrightarrow{p} 1.$$

□

Asymptotic Efficiency

The nice behavior of MLE we found in the exponential family case generalizes to a much broader case of models.

Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x)$, for $\theta \in \mathbb{R}^d$. Suppose p_θ has two continuous integrable derivatives in θ. Let $\ell_1(\theta; X_i) = \log(p_\theta(X_i))$ and $\ell_n(\theta; X) = \sum_{i=1}^n \ell_1(\theta; X_i)$. Then

$$J_1(\theta) = \operatorname{Var}_\theta(\nabla_\theta \ell_1(\theta; X_i)) = -\mathbb{E}_\theta\big[\nabla_\theta^2 \ell_1(\theta; X_i)\big] \quad \text{and} \quad J_n(\theta) = \operatorname{Var}_\theta(\nabla_\theta \ell_n(\theta; X)) = nJ_1(\theta).$$

Definition 10.19. We say an estimator $\widehat{\theta}_n$ is **asymptotically efficient** if

$$\sqrt{n}\Big(\widehat{\theta}_n - \theta\Big) \xrightarrow{\mathbb{P}_\theta} \mathcal{N}\big(0, J_1(\theta)^{-1}\big).$$

The delta method for the differentiable estimand $g(\theta)$ is

$$\sqrt{n}\Big(g(\widehat{\theta}_n) - g(\theta)\Big) \xrightarrow{\mathbb{P}_\theta} \mathcal{N}\big(0, (\nabla_\theta g(\theta))' J_1(\theta)^{-1} (\nabla_\theta g(\theta))\big).$$

This also achieves the CRLB if $\widehat{\theta}_n$ does.

Asymptotic Distribution of MLE

Under mild conditions, $\widehat{\theta}_{\text{MLE}}$ is asymptotically Gaussian and asymptotically efficient. We will be interested in $\ell_n(\theta; X)$ as a function of θ. Notate the "true" value as θ_0, that is, $X \sim \mathbb{P}_{\theta_0}$.
The derivatives of ℓ_n at θ_0 are:

$$\begin{aligned}
\nabla_\theta \ell_1(\theta_0; X_i) &\overset{\text{i.i.d.}}{\sim} (0, J_1(\theta_0)) \\
\frac{1}{\sqrt{n}} \nabla_\theta \ell_n(\theta_0; X) = \sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^n \nabla_\theta \ell_1(\theta_0; X_i) &\xrightarrow{\mathbb{P}_{\theta_0}} \mathcal{N}(0, J_1(\theta_0)) \\
\frac{1}{n} \nabla_\theta^2 \ell_n(\theta_0; X) &\xrightarrow{\mathbb{P}_{\theta_0}} \mathbb{E}_{\theta_0}\big[\nabla_\theta^2 \ell_1(\theta_0; X_i)\big] = -J_1(\theta_0).
\end{aligned}$$

Proposition 10.20. $\sqrt{n}\Big(\widehat{\theta}_n - \theta_0\Big) \xrightarrow{d} \mathcal{N}\big(0, J(\theta_0)^{-1}\big)$.

Solution Sketch. We have

$$\begin{aligned}
0 = \nabla_\theta \ell_n(\widehat{\theta}_n; X) &\approx \nabla_\theta \ell_n(\theta_0) + \nabla_\theta^2 \ell_n(\theta_0)(\widehat{\theta}_n - \theta_0) \\
\sqrt{n}\Big(\widehat{\theta}_n - \theta_0\Big) &\approx \underbrace{-\left(\frac{1}{n} \nabla_\theta^2 \ell_n(\theta_0)\right)^{-1}}_{\xrightarrow{p} J(\theta_0)^{-1}} \underbrace{\frac{1}{\sqrt{n}} \nabla_\theta \ell_n(\theta_0)}_{\xrightarrow{d} \mathcal{N}(0, J(\theta_0).)} \\
&\xrightarrow{d} \mathcal{N}\big(0, J(\theta_0)^{-1}\big).
\end{aligned}$$

We will give a full proof later, but note we need consistency of $\widehat{\theta}_n$ first to even justify the Taylor expansion. □

We want to find when $\widehat{\theta}_n \xrightarrow{\mathbb{P}_\theta} \theta_0$, assuming an identifiable model, that is, $\mathbb{P}_\theta \neq \mathbb{P}_\theta$ for $\theta \neq \theta_0$. Computing the KL Divergence,,

$$\begin{aligned}
D_{\text{KL}}(\theta_0 \parallel \theta) &= \mathbb{E}_{\theta_0}\left[\log\left(\frac{p_{\theta_0}(X)}{p_\theta(X)}\right)\right] \\
-D_{\text{KL}}(\theta_0 \parallel \theta) &= \mathbb{E}_{\theta_0}\left[\log\left(\frac{p_\theta(X)}{p_{\theta_0}(X)}\right)\right] \\
&\leq \log\left(\mathbb{E}_{\theta_0}\left[\frac{p_\theta(X)}{p_{\theta_0}(X)}\right]\right) \\
&= \log\left(\int_{x\colon p_{\theta_0}(x) > 0} \frac{p_\theta(x)}{p_{\theta_0}(x)} \,\mathrm{d}\mathbb{P}_{\theta_0}\big(X^{-1}(x)\big)\right) \\
&= \log\left(\int_{x\colon p_{\theta_0}(x) > 0} \frac{p_\theta(x)}{p_{\theta_0}(x)} p_{\theta_0}(x) \,\mathrm{d}\mu(x)\right) \\
&= \log\left(\int_{x\colon p_{\theta_0}(x) > 0} p_\theta(x) \,\mathrm{d}\mu(x)\right) \\
&\leq \log(1) \\
&\leq 0.
\end{aligned}$$

This inequality is strict unless $\frac{p_\theta}{p_{\theta_0}}$ is constant (and thus 1), so unless $\mathbb{P}_\theta = \mathbb{P}_{\theta_0}$.
Let $W_i(\theta) = \ell_1(\theta; X_i) - \ell(\theta_0; X_i)$ and $\overline{W}_n = \frac{1}{n}\sum_{i=1}^n W_i$. Note

$$\mathbb{E}_{\theta_0}[W_i(\theta)] = -D_{\text{KL}}(\theta_0 \parallel \theta)$$

so $\widehat{\theta}_n \in \operatorname{argmax}_{\theta\in\Theta} \overline{W}_n(\theta)$ too. Then

$$\overline{W}_n(\theta) \xrightarrow{p} \mathbb{E}_{\theta_0}[W_n(\theta)] = -D_{\text{KL}}(\theta_0 \parallel \theta) \le 0 \quad \text{with equality if and only if } \theta = \theta_0.$$

Definition 10.21. For compact K, let $C(K) = \{f\colon K \to \mathbb{R}, f \text{ is continuous}\}$. For $f \in C(K)$ let $\|f\|_\infty = \sup_{t\in K} f(t)$. Then $f_n \to f$ in this norm if $\|f_n - f\|_\infty \to 0$.

Theorem 10.22 (LLN for Random Functions). Suppose K is compact, and $W_1, \cdots \in C(K)$ are i.i.d., with $\mathbb{E}[\|W_n\|_\infty] < \infty$ and $\mu(t) = \mathbb{E}[W_i]$. Then $\mu(t) \in C(K)$ and $\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n W_i - \mu\right\|_\infty > \varepsilon\right) \to 0$. That is, $\overline{W}_n \xrightarrow{p} \mu$ in the ℓ^∞ norm, or $\left\|\overline{W}_n - \mu\right\|_\infty \xrightarrow{p} 0$.

Theorem 10.23 (Keener 9.4). Let $G_1, \dots$ be random functions in $C(K)$ where K is compact. Assume $\|G_n - g\|_\infty \xrightarrow{p} 0$ for some fixed $g \in C(K)$. Then

(i) If $t_n \xrightarrow{p} t^* \in K$ (t^* fixed) then $G_n(t_n) \xrightarrow{p} g(t^*)$.

(ii) If g is maximized at a unique value t^* and $G_n(t_N) = \sup_n G_n(t)$ then $t_n \xrightarrow{p} t^*$.

(iii) If $K \subseteq \mathbb{R}$ and $g(t) = 0$ has a unique solution t^* and t_n solves $G_n(t_n) = 0$ then $t_n \xrightarrow{p} t^*$.

Proof. By triangle inequality,

$$|G_n(t) - g(t^*)| \le \|G_n(t_n) - g(t_n)\| + |g(t_n) - g(t^*)| \qquad \le \underbrace{\|G_n - g\|_\infty}_{\xrightarrow{p}0} + \underbrace{|g(t_n) - g(t^*)|}_{\xrightarrow{p}0}.$$

Thus (i) is proved.
Fix $\varepsilon > 0$. Let $B_\varepsilon(t^*) = \{t\colon \|t - t^*\|_2 < \varepsilon\}$. Let

$$K_\varepsilon = K \setminus B_\varepsilon(t^*) = K \cap B_\varepsilon(t^*)^c.$$

Define $\delta = g(t^*) - \max_{t\in K_\varepsilon} g(t) > 0$. If $t_n \in K_\varepsilon$ then

$$G_n(t_n) \le g(t^*) - \delta + \|G_n - g\|_\infty \quad \text{and} \quad G_n(t_n) \ge G_n(t^*) \ge g(t^*) - \|G_n - g\|_\infty$$

Then $2\|G_n - g\|_\infty \ge \delta$. Thus

$$\mathbb{P}(\|t_n - t^*\| \ge \varepsilon) \le \mathbb{P}\left(\|G_n - g\|_\infty \ge \frac{\delta}{2}\right) \to 0.$$

Thus (ii) is proved.
The proof for (iii) is analogous to (ii). □

Theorem 10.24 (Consistency of MLE for Compact θ). If $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$, and $\mathcal{P}$ has continuous densities p_θ for $\theta \in \Theta$ compact, and $\mathcal{P}$ is identifiable, and

$$\mathbb{E}_{\theta_0}[\|W_i\|_\infty] = \mathbb{E}_{\theta_0}[\|\ell_1(\theta; X_i) - \ell_1(\theta_0; X_i)\|_\infty] < \infty,$$

then if $\widehat{\theta}_n \in \operatorname{argmax}_\theta \ell_n(\theta; X)$, then $\widehat{\theta}_n \xrightarrow{p} 0$.

Proof. If $W_i \in C(\Theta)$ are i.i.d. with mean $\mu(\theta) = -D_{\text{KL}}(\theta_0 \parallel \theta)$, $\mu(\theta_0) = 0$, $\mu(\theta) < 0$ for all $\theta \ne \theta_0$. By definition $\widehat{\theta}_n$ maximizes $\overline{W}_n$, and by LLN $\left\|\overline{W}_n - \mu\right\|_\infty \xrightarrow{p} 0$. Applying Theorem 9.4 (ii) of Keener obtains the conclusion. □

We usually care about non-compact parameter spaces, but we need some extra assumptions to get us there.

Theorem 10.25 (Keener 9.11 With Stronger Conditions). If $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$, $\mathcal{P}$ has continuous densities p_θ, and $\theta \in \Theta = \mathbb{R}^d$, $\mathcal{P}$ is identifiable, for all compact $K \subseteq \mathbb{R}^d$ $\mathbb{E}[\sup_{\theta\in K} |W_i(\theta)|] < \infty$, and there exists $r > 0$ such that $\mathbb{E}\left[\sup_{\|\theta-\theta_0\|_2 > r} W_i(\theta)\right] < 0$, then $\widehat{\theta}_n \xrightarrow{p} \theta_0$ if $\widehat{\theta}_n \in \operatorname{argmax}_\theta \ell_n(\theta; X)$.

Proof. Let $A = B_r(\theta_0)$ and $\alpha = \mathbb{E}[\sup_{\theta\in A} W_i(\theta)] < 0$. Then

$$\sup_{\theta\in A} \overline{W}_n(\theta) \le \frac{1}{n}\sum_{i=1}^n \sup_{\theta\in A} W_i(\theta) \to \alpha < 0.$$

Hence

$$\mathbb{P}\Big(\widehat{\theta}_n \in A\Big) \le \mathbb{P}\left(\overline{W}_n(\theta_0) < \sup_{\theta \in A} \overline{W}_n(\theta)\right) \to 0.$$

Let $\widetilde{\theta}_n = \widehat{\theta}_n \mathbb{1}\Big(\widehat{\theta}_n \in A^c\Big) + \theta_0 \mathbb{1}\Big(\widehat{\theta}_n \in A\Big)$. Then $\widetilde{\theta}_n \xrightarrow{p} 0$ by the previous theorem, so $\widehat{\theta}_n \xrightarrow{p} \theta_0$. □

Theorem 10.26 (Asymptotic Distribution of MLE). Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$ for $\theta_0 \in \Theta^0 \subseteq \mathbb{R}^d$. Assume

- $\widehat{\theta}_n(X) \in \operatorname{argmax}_{\theta \in \Theta} \ell_n(\theta; X)$ and $\widehat{\theta}_n \xrightarrow{p} \theta_0$.
- In a neighborhood $B_\varepsilon(\theta_0) \subseteq \Theta^0$:
  - $\ell_1(\theta; X)$ has two continuous derivatives on $B_\varepsilon(\theta_0)$
  - $\mathbb{E}_{\theta_0}\left[\sup_{\theta \in B_\varepsilon} \left\|\nabla_\theta^2 \ell_1(\theta; X_i)\right\|\right] < \infty$.
- Fisher information:
  - $\mathbb{E}_{\theta_0}[\nabla_\theta \ell_1(\theta_0; X_i)] = 0$.
  - $J_1(\theta_0) = \operatorname{Var}_{\theta_0}(\nabla_\theta \ell_1(\theta_0; X_i)) = -\mathbb{E}_{\theta_0}\left[\nabla_\theta^2 \ell_1(\theta_0, X_i)\right] \succ 0$. (It's enough to have a 3rd derivative of ℓ_1 that's bounded in $B_\varepsilon(\theta_0)$).

Then $\sqrt{n}\Big(\widehat{\theta}_n - \theta_0\Big) \xrightarrow{d} \mathcal{N}\big(0, J_1(\theta_0)^{-1}\big)$, or informally $\widehat{\theta}_n \approx \mathcal{N}\big(\theta_0, J_n(\theta_0)^{-1}\big)$.

Proof. Let $A_n = \left\{\left\|\widehat{\theta}_n - \theta_0\right\| \ge \varepsilon\right\}$. By assumption $\mathbb{P}_{\theta_0}(A_n) \to 0$. On A_n^c, $\widehat{\theta}_n \in B_\varepsilon(\theta_0)$. Thus by Mean Value Theorem,

$$\begin{aligned} 0 &= \nabla_\theta \ell_n(\widehat{\theta}_n; X) \\ &= \nabla_\theta \ell_n(\theta_0; X) + \left(\nabla_\theta^2 \ell_n\left(\underbrace{\widetilde{\theta}_n}_{\widetilde{\theta}_n \text{ between } \theta_0, \widehat{\theta}_n}; X\right)\Big(\widehat{\theta}_n - \theta_0\Big)\right). \end{aligned}$$

Thus since $\widetilde{\theta}_n$ is sandwiched between a sequence $\widehat{\theta}_n$ and its limit in probability θ_0, $\widehat{\theta}_n \xrightarrow{\mathbb{P}_{\theta_0}} \theta_0$. By the continuous mapping theorem and Theorem 9.14 (i) of Keener,

$$\begin{aligned} \sqrt{n}\Big(\widehat{\theta}_n - \theta_0\Big) &= \underbrace{\left(-\frac{1}{n}\nabla_\theta^2 \ell_n\Big(\widetilde{\theta}_n; X\Big)\right)^{-1}}_{\xrightarrow{\mathbb{P}[\theta_0]} J_1(\theta_0)^{-1}} \underbrace{\left(\frac{1}{\sqrt{n}}\nabla_\theta \ell_n(\theta_0; X)\right)}_{\mathcal{N}(0, J_1(\theta))} \\ &\xrightarrow{d} \mathcal{N}\big(0, J_1(\theta_0)^{-1}\big) \end{aligned}$$

where the last step is by Slutsky's Theorem.
On A_n we can just replace $\widehat{\theta}_n$ by θ_0 or anything else. The behavior on A_n doesn't affect anything since $\mathbb{P}_{\theta_0}(A_n) \to 0$. □

Likelihood-Based Interference

Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x)$, where $p_\theta(x)$ is "smooth" in θ. We assume

- $\mathbb{E}_\theta[\nabla_\theta \ell_1(\theta; X_i)] = 0$
- $\operatorname{Var}_\theta(\ell_1(\theta; X_i)) = -\mathbb{E}_\theta\left[\nabla_\theta^2 \ell_1(\theta; X)\right] = J_1(\theta) \succ 0$.
- $\widehat{\theta}_{\text{MLE}} \xrightarrow{\mathbb{P}_\theta} \theta$.

Then if $\theta = \theta_0$,

$$\frac{1}{n}\nabla_\theta \ell_n(\theta_0; X) \xrightarrow{d} \mathcal{N}(0, J_1(\theta_0)) \quad \text{and} \quad \frac{1}{n}\nabla_\theta^2 \ell_n(\theta_0; X) \xrightarrow{\mathbb{P}[\theta_0]} J_1(\theta_0),$$

and

$$\sqrt{n}\Big(\widehat{\theta}_n - \theta_0\Big) \approx -\left(\frac{1}{n}\nabla_\theta^2 \ell_n(\theta_0)\right)\big(\sqrt{n}\nabla_\theta \ell(\theta_0)\big) \approx \mathcal{N}\big(0, J_1(\theta_0)^{-1}\big).$$

We can use this for inference on θ_0.

Wald-Type Confidence Regions

Assume we have some estimator $\widehat{J}_n \succeq 0$ such that $\frac{1}{n}\widehat{J}_n \overset{\mathbb{P}_{\theta_0}}{\to} J_1(\theta_0) \succ 0$. If

$$\sqrt{n}\Big(\widehat{\theta}_n - \theta_0\Big) \overset{d}{\to} \mathcal{N}\big(0, J_1(\theta_0)^{-1}\big) \implies \Big(J_1(\theta_0)^{-1/2}\Big)\sqrt{n}\Big(\widehat{\theta}_n - \theta_0\Big) \overset{d}{\to} \mathcal{N}(0, I_d) \implies \left(\frac{1}{n}\widehat{J}_n\right)^{-1/2}\sqrt{n}\Big(\widehat{\theta}_n - \theta_0\Big) \overset{d}{\to} \mathcal{N}(0, I_d),$$

by Slutsky's theorem. This leads to a test of $H_0 = \theta = \theta_0$ vs $H_1 \colon \theta \neq \theta_0$:

$$\text{reject when} \qquad \left\|\widehat{J}_n^{-1/2}\Big(\widehat{\theta}_n - \theta_0\Big)\right\|_2^2 \overset{d}{\to} \chi_d^2 \qquad \text{is large.}$$

So

$$\mathbb{P}_{\theta_0}\left(\left\|\widehat{J}_n^{-1/2}\Big(\widehat{\theta}_n - \theta_0\Big)\right\|_2^2 \geq \underbrace{\chi_d^2(\alpha)}_{\text{upper } \alpha \text{ quantile}}\right) \to \alpha.$$

Suppose we do this for all $\theta_0 \in \Theta$:

$$\text{reject } \theta_0 \text{ if and only if} \qquad \left\|\widehat{J}_n^{-1/2}\Big(\widehat{\theta}_n - \theta_0\Big)\right\|_2^2 > \chi_d^2(\alpha),$$

then it's equivalent to

$$\text{reject } \theta_0 \text{ if and only if} \qquad \theta_0 \notin \widehat{\theta}_n + \widehat{J}_n^{-1/2} B_{\sqrt{\chi_d^2(\alpha)}}(0).$$

This gives a Wald confidence ellipsoid. The larger $\widehat{J}_n$ is, the smaller the ellipsoid gets. The radii shrink like $n^{-1/2}$. By definition this test is asymptotically valid.
There are some options for $\widehat{J}_n$:

- MLE:
$$\widehat{J}_n = J_n(\widehat{\theta}_n) = \big[\mathrm{Var}_\theta(\boldsymbol{\nabla}_\theta \ell_n(\theta; X))\big]_{\theta=\widehat{\theta}_n} \neq \mathrm{Var}_{\widehat{\theta}_n}\Big(\boldsymbol{\nabla}_\theta \ell_n(\widehat{\theta}_n(X); X)\Big) = 0.$$
- Observed Fisher information:
$$\widehat{J}_n = -\boldsymbol{\nabla}_\theta^2 \ell_n(\widehat{\theta}_n; X).$$

Both have

$$\frac{1}{n}\widehat{J}_n \overset{\theta_0}{\to} J_1(\theta_0) \quad \text{in the "smooth" i.i.d. setting.}$$

Now we want to find a Wald interval for θ_j. Indeed, if $\widehat{\theta}_n \approx \mathcal{N}\big(\theta_0, J_n(\theta_0)^{-1}\big)$, then

$$\widehat{\theta}_{n,j} \approx \mathcal{N}\left(\theta_{0,j}, \underbrace{\big(J_n(\theta_0)^{-1}\big)_{j,j}}_{\widehat{\text{s.e.}}\left(\widehat{\theta}_{n,j}\right)^2}\right).$$

Thus the confidence interval is

$$C_j = \widehat{\theta}_{n,j} \pm \widehat{\text{s.e.}}\Big(\widehat{\theta}_{n,j}\Big) \cdot z_{\alpha/2} = \widehat{\theta}_{n,j} \pm \sqrt{\Big(\widehat{J}_n^{-1}\Big)_{j,j}} \cdot z_{\alpha/2}.$$

If $S \subseteq \{1, \dots, d\}$, the confidence ellipsoid for $\theta_{0,S} = (\theta_{0,j})_{j \in S}$ is

$$\widehat{\theta}_{n,s} \approx \mathcal{N}\Big(\theta_{0,S}, \big(J_n(\theta_0)^{-1}\big)_{S,S}\Big) \implies C_S = \widehat{\theta}_{n,S} + \left(\Big(\widehat{J}_n^{-1}\Big)_{S,S}\right)^{1/2} B_{z_{\alpha/2}}(0).$$

More generally if $\frac{1}{n}\widehat{\Sigma}_n \overset{\mathbb{P}_{\theta_0}}{\to}$ and $\sqrt{n}\Big(\widehat{\theta}_n - \theta_0\Big) \approx \mathcal{N}\big(0, \Sigma(\theta_0)^{-1}\big)$ then we can do all the same things.

Example 10.27 (General Linear Model). Suppose $X_1, \dots, X_n \in \mathbb{R}^d$ are fixed. Suppose

$$Y_i \overset{\text{ind}}{\sim} p_{\eta_i}(Y) = \mathrm{e}^{\eta_i Y_i - A(\eta_i)} h(Y_i).$$

Let $\eta_i = \beta' X_i$ (the **canonical form**). Let

$$\mu_i(\beta) = \mathbb{E}_\beta[Y_i] = \mu(\eta_i(\beta)).$$

More generally $f(\mu_i) = \beta' X_i$ for some **link function** f.
The most common examples of this are

- Logistic regression: $Y_i \overset{\text{ind}}{\sim} \text{Bern}\left(\frac{\mathrm{e}^{X_i'\beta}}{1+\mathrm{e}^{X_i'\beta}}\right)$
- Poisson log-linear model: $Y_i \sim \text{Pois}\left(\mathrm{e}^{X_i'\beta}\right)$.

Then

$$\begin{aligned}
\ell_n(\beta, Y) &= \sum_i (X_i'\beta)Y_i - A(X_i'\beta) - \log(h(Y_i)) \\
\nabla_\theta \ell_n(\beta; Y) &= \sum_i Y_i X_i - \underbrace{\nabla A(X_i'\beta)}_{\mu(\beta)} X_i \\
&= \sum_i (Y_i - \mu_i(\beta)) X_i. \\
-\nabla_\beta^2 \ell_n(\beta; Y) &= \sum_i \nabla^2 A(X_i'\beta) X_i X_i' \\
&= \sum_i \text{Var}_\beta(Y_i) X_i X_i' \\
&= \text{Var}_\beta(\nabla_\beta \ell_n(\beta; Y)).
\end{aligned}$$

Thus

$$\left(-\nabla_\beta^2 \ell_n(\beta_0)\right)^{-1/2}(\nabla_\beta \ell_n(\beta_0)) \sim (0, I_d) \quad \text{and under regularity conditions on } X \quad \left(-\nabla_\beta^2 \ell_n(\beta_0)\right)^{-1/2}(\nabla_\beta \ell_n(\beta_0)) \xrightarrow{d} \mathcal{N}(0, I_d).$$

By Taylor expansion of ℓ_n,

$$\widehat{J}_n\left(\widehat{\beta}_n - \beta\right)^{1/2} \xrightarrow{d} \mathcal{N}(0, I_d).$$

The advantages of Wald inference are

- Simple, easy confidence regions.
- Asymptotically correct.

and the disadvantages are

- We have to calculate θ_{MLE}.
- The interval depends on the parameterization.
- Assumes that the $\nabla_\theta \ell_n$ is approximately normal, which may not happen until ridiculously large n.
- We require the MLE to be consistent.
- The confidence ellipsoid or any interval could actually go outside the parameter space.

10.1 Score Test

The simple **score test** for $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ uses the score as a test statistic. In particular

$$J_n(\theta_0)^{-1/2}(\nabla_\theta \ell_n(\theta_0)) \xrightarrow{\mathbb{P}_{\theta_0}} \mathcal{N}(0, I_d).$$

We can reject $H_0 : \theta = \theta_0$ if

$$\left\| J_n(\theta_0)^{-1/2}(\nabla_\theta \ell_n(\theta_0, x)) \right\|_2^2 \geq \chi_d^2(\alpha).$$

In the case $d = 1$ this simplifies to

$$\frac{\dot{\ell}_n(\theta_0)}{\sqrt{J_n(\theta_0)}} \xrightarrow{d} \mathcal{N}(0,1) \implies \quad \text{reject if} \quad \left\| \frac{\dot{\ell}_n(\theta_0)}{\sqrt{J_n(\theta_0)}} \right\|_2^2 \geq \chi_1^2(\alpha).$$

Remark 10.28.

1. There is no quadratic approximation and no MLE referenced.
2. There is no need to estimate the Fisher information at θ_0; we can just evaluate it.
3. We don't even need the normal approximation either if we are really only testing the simple null hypothesis. (We can simulate the distribution under H_0 and get arbitrarily close to the exact CDF).

Remark 10.29. The score test is invariant to (smooth) reparameterizations. Assume $d = 1$. Let $\theta = g(\zeta)$ with $\dot{g}(\zeta) > 0$ for all ζ. Let $q_\zeta(x) = p_{g(\zeta)}(x)$. Then

$$\dot{\ell}(\zeta; x) = \frac{\mathrm{d}}{\mathrm{d}\zeta} \log\big(p_{g(\zeta)}(x)\big) = \dot{\ell}(g(\zeta); x)\dot{g}(\zeta)$$

and

$$J(\zeta) = J(g(\zeta))\dot{g}(\zeta)^2,$$

so if $\theta_0 = g(\zeta_0)$,

$$\frac{\dot{\ell}(\zeta_0; x)}{\sqrt{J(\zeta_0)}} \overset{\text{a.e.}}{=} \frac{\dot{\ell}(\theta_0; x)}{\sqrt{J(\theta_0)}}.$$

Example 10.30. Suppose we have an s-parameter exponential family, with $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{e}^{\eta' T(x) - A(\eta)} h(x)$. Then

$$\nabla_\eta \ell_n(\eta; x) = \sum_{i=1}^s T(x_i) - n(\nabla_\eta A(\eta)) \implies \left\| J_n(\eta_0)^{-1/2} \left(\sum_{i=1}^n T(x_i) - n\nabla_\eta A(\eta_0) \right) \right\|_2^2 \xrightarrow{d} \chi_d^2.$$

Thus in the case $d = 1$,

$$\frac{\sum_{i=1}^n T(X_i) - n\nabla_\eta A(\eta_0)}{\sqrt{n \operatorname{Var}_{\eta_0}(T(X_1))}} \xrightarrow{d} \mathcal{N}(0, 1),$$

though in some cases we can get an exact distribution (the **exact score test**).

In the general parametric model, we can do a local approximation:

$$p_{\theta_0 + \eta}(x) = p_{\theta_0}(x) \mathrm{e}^{\ell(\theta_0 + \eta; x) - \ell(\theta_0; x)} \approx p_{\theta_0}(x) \mathrm{e}^{\eta' (\nabla_\theta \ell(\theta_0; x))}.$$

Example 10.31 (Pearson's χ^2 Goodness-Of-Fit Test). Suppose

$$N = (N_1, \dots, N_d) \sim \operatorname{Multinom}(n, (\pi_1, \dots, \pi_d)) = \frac{n! \, \mathbb{1}\left(\sum_{j=1}^d N_j = n\right)}{\prod_{i=1}^d N_i!} \sum_{i=1}^d \pi_i^{N_i}.$$

Note $\sum_{i=1}^d \pi_i = 1$ so this is a full-rank $(d-1)$-parameter exponential family. We have the parameterization $\eta_2, \dots, \eta_d$ (implicitly saying $\eta_1 = 0$ and $\pi_j \propto \mathrm{e}^{\eta_j}$), in particular

$$\pi_j = \begin{cases} \frac{1}{1 + \sum_{k=2}^d \mathrm{e}^{\eta_k}} & j = 1 \\ \frac{\mathrm{e}^{\eta_j}}{1 + \sum_{k=2}^d \mathrm{e}^{\eta_k}} & j > 1 \end{cases}.$$

In this parameterization

$$p_\eta(N) = \frac{n!}{\prod_{i=1}^d N_i!} \left(\prod_{j=2}^d \exp\left(N_j \log\left(\frac{\mathrm{e}^{\eta_j}}{1 + \sum_{k=2}^d \mathrm{e}^{\eta_k}} \right) \right) \right) \exp\left(N_1 \log\left(\frac{1}{1 + \sum_{k=2}^d \mathrm{e}^{\eta_k}} \right) \right) = h(N) \exp\left(\sum_{j=2}^d N_j(\eta_j - \log(1 + \varepsilon)) - \right.$$

Thus

$$\nabla_\eta \ell(\eta; N) = (N_2, \dots, N_d) - (n\pi_2, \dots, n\pi_d),$$

and, if $\pi_{-1} = (\pi_2, \dots, \pi_d)$

$$J_n(\eta) = \operatorname{Var}_\eta(\nabla_\eta \ell(\eta)) = n\big(\operatorname{diag}(\pi_{-1}) - \pi_{-1}\pi_{-1}'\big).$$

To compute the inverse, we use a formula which gives

$$(A - uv')^{-1} = A^{-1} - \frac{A^{-1}uv'A^{-1}}{1 - v'A^{-1}u}.$$

Thus

$$J_n(\eta)^{-1} = \big(n\big(\operatorname{diag}(\pi_{-1}) - \pi_{-1}\pi_{-1}'\big)\big)^{-1} = \frac{1}{n}\left(\operatorname{diag}(\pi_{-1})^{-1} - \frac{11'}{\pi_1}\right).$$

The score test for $H_0 : \pi = \pi_0$ vs $H_1 : \pi = \pi_0$ is

$$(\nabla_\eta \ell(\eta_0; X))'(J_n(\eta_0))^{-1}(\nabla_\eta \ell(\eta_0; X)) = \sum_{j=1}^{d} \frac{(N_j - \pi_{0,j})^2}{n\pi_{0,j}} \overset{\mathbb{P}_{\pi_0}}{\to} \chi^2_{d-1}.$$

Generalized Likelihood Ratio Test

We want to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. If we take the Taylor expansion around $\widehat{\theta}_n$,

$$\begin{aligned}
\ell_n(\theta_0) - \ell_n(\widehat{\theta}_n) &\approx \left(\nabla_\theta \ell(\widehat{\theta}_n)\right)\left(\theta_0 - \widehat{\theta}_n\right) + \frac{1}{2}\left(\theta_0 - \widehat{\theta}_n\right)'\left(\nabla^2_\theta \ell_n(\widehat{\theta}_n)\right)\left(\theta_0 - \widehat{\theta}_n\right) \\
&= -\frac{1}{2}\left\| \left(\underbrace{\nabla^2_\theta \ell_n(\widehat{\theta}_n)}_{\overset{p}{\to} -nJ_1(\theta_0)}\right)^{1/2} \left(\underbrace{\theta_0 - \widehat{\theta}_n}_{\overset{d}{\to}\mathcal{N}(0, J_1(\theta_0)/n)}\right) \right\| \\
&\overset{d}{\to} -\frac{1}{2}\chi^2_d.
\end{aligned}$$

The **generalized LRT statistic** is

$$2\Big(\ell_n(\widehat{\theta}_n; x) - \ell_n(\theta_0; x)\Big) \overset{\mathbb{P}_{\theta_0}}{\to} \chi^2_d.$$

Of course if we can calculate the left hand side or approximate it, we don't need to make an approximation.
Let's consider a composite test, that is $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta \setminus \Theta_0$. We make the following assumptions:

- $\Theta = \mathbb{R}^d$ and Θ_0 is a d_0-dimensional manifold.
- $\theta_0 \in \Theta_0^\circ$.
- $\widehat{\theta}_n \overset{\mathbb{P}_{\theta_0}}{\to} \theta_0$.
- The likelihood is a smooth function.

Then

$$2\Big(\ell_n(\widehat{\theta}_n) - \ell_n(\widehat{\theta}_0)\Big) \overset{d}{\to} \chi^2_{d-d_0} \quad \text{where} \quad \widehat{\theta}_0 = \operatorname*{argmax}_{\theta \in \Theta_0} \ell_n(\theta; x).$$

To see why, let's assume $\theta_0 = 0$ and $J_1(0) = I_d$ (and reparameterize otherwise). Then locally

$$\ell_n(\theta) \approx \ell_n(\widehat{\theta}_n) - \frac{n}{2}\left\|\theta - \widehat{\theta}_n\right\|_2^2 \quad \text{and} \quad \widehat{\theta}_0 = \operatorname*{argmax}_{\theta \in \Theta_0} \ell_n(\theta) \approx \operatorname{proj}_{\Theta_0}\left(\widehat{\theta}_n\right),$$

so

$$2\Big(\ell_n(\widehat{\theta}_n; x) - \ell_n(\widehat{\theta}_0)\Big) \approx n\left\|\widehat{\theta}_n - \operatorname{proj}_{\Theta_0}\left(\widehat{\theta}_n\right)\right\|_2^2 = n\left\|\operatorname{proj}_{\tan\left(\Theta_0, \widehat{\theta}_0\right)^\perp}\left(\widehat{\theta}_n\right)\right\|_2^2 \overset{d}{\to} \chi^2_{d-d_0}.$$

where $\tan\left(\Theta_0, \widehat{\theta}_0\right)$ is the tangent space to Θ_0 at $\widehat{\theta}_0$.

Asymptotic Equivalence and Relative Efficiency

Recall the quadratic approximation picture:

$$\ell_n(\theta) - \ell_n(\theta_0) \approx (\nabla_\theta \ell_n(\theta_0))(\theta - eta a_0) + \frac{1}{2}(\theta - \theta_0)^{\mathsf{T}}(J_n(\theta_0))(\theta - \theta_0)^2.$$

For large n,

$$\left\|J_n(\theta_0)^{1/2}\left(\widehat{\theta}_n - \theta_0\right)\right\|_2^2 \approx \underbrace{\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)}_{\text{(GRLT)}}$$

$$\approx \underbrace{\left\| \widehat{J}_n^{1/2}\left(\widehat{\theta}_n - \theta_0\right)\right\|_2^2}_{\text{(Wald)}}$$

$$\approx \underbrace{\left\| J_n(\theta_0)^{-1/2}(\nabla_\theta \ell_n(\theta_0))\right\|_2^2}_{\text{(Score)}}.$$

Now suppose $\widehat{\theta}_n^{(i)}$ for $i = 1, 2$ are two asymptotically normal estimators of $\theta_0 \in \mathbb{R}$ with

$$\sqrt{n}\left(\widehat{\theta}_n^{(i)} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right),$$

then the **asymptotic relative efficiency** of $\widehat{\theta}^{(0)}$ with respect to $\widehat{\theta}^{(i)}$ is $\frac{\sigma_1^2}{\sigma_2^2}$, e.g. if $\sigma_2^2 = 2\sigma_1^2$, then $\widehat{\theta}^{(2)}$ is 50% as efficient.

Suppose $\frac{\sigma_1^2}{\sigma_2^2} = \gamma \in (0, 1)$. Then for large n,

$$\theta_{\lfloor \gamma n \rfloor}^{(1)}(X_1, \dots, X_{\lfloor \gamma n \rfloor}) \overset{d}{=} \widehat{\theta}_n^{(2)}(X_1, \dots, X_n) \approx \mathcal{N}\left(\theta, \frac{\sigma_2^2}{n}\right).$$

Using $\widehat{\theta}^{(2)}$ is like throwing away $1 - \gamma$ of the data and then using $\widehat{\theta}^{(1)}$.

11 Nonparametric Estimation

Suppose we sample $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\text{NP}}$, where $\mathbb{P}_{\text{NP}}$ is an unknown probability measure over $\mathcal{X}$. We want to do inference on some parameter functional $\theta(\mathbb{P}_{\text{NP}})$, for example:

- $\theta(\mathbb{P}_{\text{NP}}) = \text{median}(\mathbb{P}_{\text{NP}})$ (if $\mathcal{X} \subseteq \mathbb{R}$)
- $\theta(\mathbb{P}_{\text{NP}}) = \lambda_{\max}(\text{Var}_{\mathbb{P}_{\text{NP}}}(X))$ (if $\mathcal{X} \subseteq \mathbb{R}^d$)
- $\theta(\mathbb{P}_{\text{NP}}) = \text{argmin}_{\theta \in X} \mathbb{E}_{\mathbb{P}_{\text{NP}}}\left[(Y - \langle X, \theta\rangle)^2\right]$ (if $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\text{NP}}$)
- $\theta(\mathbb{P}_{\text{NP}}) = \text{argmin}_{\theta \in \Theta} D_{\text{KL}}(\mathbb{P}_{\text{NP}} \,\|\, \mathbb{P}_\theta) = \text{argmax}_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{\text{NP}}}[\ell_1(\theta; X)]$.

Recall the **empirical distribution** of $X_1, \dots, X_n$ is given by the probability measure

$$\widehat{\mathbb{P}}_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i \in A}.$$

The **plug-in estimator** of $\theta(\mathbb{P}_{\text{NP}})$ is $\widehat{\theta} = \theta(\widehat{\mathbb{P}}_n)$:

- sample median
- $\lambda_{\max}$(sample variance)
- OLS estimator
- MLE for $\{\mathbb{P}_\theta : \theta \in \Theta\}$

We want to find whether the plug-in estimator works, i.e., under which notions of convergence $\widehat{\mathbb{P}}_n \to \mathbb{P}_{\text{NP}}$. It turns out that $\widehat{\mathbb{P}}_n(A) \xrightarrow{p} \mathbb{P}_{\text{NP}}(A)$ for all A, but $\sup_A \left|\widehat{\mathbb{P}}_n(A) - \mathbb{P}_{\text{NP}}(A)\right| \not\xrightarrow{p} 0$ even when $\mathbb{P}_{\text{NP}}$ is continuous (for example in the total variation metric). However if $\mathcal{X} \subseteq \mathbb{R}$ then $\sup_x \left|\widehat{\mathbb{P}}_n((-\infty, x]) - \mathbb{P}_{\text{NP}}((-\infty, x])\right| \xrightarrow{p} 0$. In the end we want $\theta(\mathbb{P}_{\text{NP}})$ to be continuous with respect to some topology in which $\widehat{\mathbb{P}}_n \xrightarrow{p} \mathbb{P}_{\text{NP}}$. Then $\theta(\widehat{\mathbb{P}}_n) \xrightarrow{p} \theta(\mathbb{P}_{\text{NP}})$.

Bootstrap

Bootstrapping is when, given a sample $X_1, \dots, X_n$, we sample $X^* \sim \widehat{\mathbb{P}}_n$.
Suppose $\widehat{\theta}_n(X)$ is an estimator for $\theta(\mathbb{P})$. Then

$$\widehat{\text{s.e.}}(\widehat{\theta}_n) = \sqrt{\text{Var}_{\widehat{\mathbb{P}}_n}\left(\widehat{\theta}_n^*\right)} = \sqrt{\text{Var}_{X_1^*,\dots,X_n^* \overset{\text{i.i.d}}{\sim} \widehat{\mathbb{P}}_n}\left(\widehat{\theta}_n(X_1^*, \dots, X_n^*)\right)}.$$

We compute this via Monte-Carlo.

Algorithm 2 Estimates $\text{s. e.}\left(\widehat{\theta}_n\right)$ by approximating $\widehat{\text{s. e.}}(\widehat{\theta}_n)$ via Monte-Carlo.

Input: A sample dataset $X_1, \dots, X_n$, an estimator $\widehat{\theta}_n$, a stopping time B
Output: An estimate $\widehat{\text{s.e.}}(\widehat{\theta}_n)$
  for $b = 1, \dots, B$ **do**
    $X_1^{*b}, \dots, X_n^{*b} \overset{\text{i.i.d.}}{\sim} \widehat{\mathbb{P}}_n$
    $\widehat{\theta}^{*b} \leftarrow \widehat{\theta}_n(X_1^{*b}, \dots, X_n^{*b})$
  $\overline{\theta}^* \leftarrow \frac{1}{B}\sum_{b=1}^B \widehat{\theta}^{*b}$
  return $\sqrt{\frac{1}{B}\sum_{b=1}^B \left(\widehat{\theta}^{*b} - \overline{\theta}^*\right)^2}$

This is a Monte Carlo numerical approximation to the idealized bootstrap estimator, which we could compute by iterating through all n^n possible X^* vectors.
We also want to correct for the bias of $\widehat{\theta}_n$. We have

$$\text{Bias}_P\left(\widehat{\theta}_n\right) = \mathbb{E}_{\mathbb{P}_{\text{NP}}}\left[\widehat{\theta}_n - \theta(\mathbb{P}_{\text{NP}})\right],$$

and so the idea is to plug in $\widehat{\mathbb{P}}_n$ for $\mathbb{P}_{\text{NP}}$:

$$\text{Bias}_{\widehat{\mathbb{P}}_n}\left(\widehat{\theta}_n^*\right) = \mathbb{E}_{\widehat{\mathbb{P}}_n}\left[\widehat{\theta}_n^* - \theta(\widehat{\mathbb{P}}_n)\right],$$

which we can again estimate via Monte-Carlo.

Algorithm 3 Estimates $\text{Bias}\left(\widehat{\theta}_n\right)$ by approximating $\widehat{\text{Bias}}\left(\widehat{\theta}_n\right)$ via Monte Carlo.

Input: A sample dataset $X_1, \dots, X_n$, an estimator $\widehat{\theta}_n(X)$, a stopping time B
Output: An estimate $\widehat{\text{Bias}}(\widehat{\theta}_n)$
  for $b = 1, \dots, B$ **do**
    $X_1^{*b}, \dots, X_n^{*b} \overset{\text{i.i.d.}}{\sim} \widehat{\mathbb{P}}_n$
    $\widehat{\theta}^{*b} \leftarrow \widehat{\theta}_n(X_1^{*b}, \dots, X_n^{*b})$
  $\overline{\theta}^* = \frac{1}{B}\sum_{b=1}^B \widehat{\theta}^{*b}$
  return $\overline{\theta}^* - \theta(\widehat{\mathbb{P}}_n)$

We can use this to correct bias:

$$\widehat{\theta}_n^{\text{BC}} = \widehat{\theta}_n - \widehat{\text{Bias}}(\widehat{\theta}_n).$$

While $\widehat{\theta}_n - \text{Bias}\left(\widehat{\theta}_n\right)$ is always better than $\widehat{\theta}_n$, $\widehat{\theta}_n - \widehat{\text{Bias}}(\widehat{\theta}_n)$ may not be; we might be adding more variance.
Our next goal is to get a confidence interval for $\theta(\mathbb{P}_{\text{NP}})$. If we know the distribution of

$$R_n(X, \mathbb{P}_{\text{NP}}) = \widehat{\theta}_n(X) - \theta(\mathbb{P}_{\text{NP}}),$$

then we can define the cumulative density function

$$G_{n,\mathbb{P}_{\text{NP}}}(r) = \mathbb{P}_{\mathbb{P}_{\text{NP}}}\left(\widehat{\theta}(X) - \theta(\mathbb{P}_{\text{NP}}) \le r\right).$$

The lower $\alpha/2$ quantile is $r_1 = G^{-1}_{n,\mathbb{P}_{\text{NP}}}(\alpha/2)$; the upper $\alpha/2$ quantile is $r_2 = G^{-1}_{n,\mathbb{P}_{\text{NP}}}(1-\alpha/2)$. Then $\left[\widehat{\theta}_n - r_2, \widehat{\theta}_n - r_1\right]$ is a $1-\alpha$ confidence interval for θ. If we don't know $G_{n,\mathbb{P}_{\text{NP}}}$ then we bootstrap using the approximation

$$G_{n,\widehat{\mathbb{P}}_n}(r) = \mathbb{P}_{\widehat{\mathbb{P}}_n}\left(\widehat{\theta}(X^*) - \theta(\widehat{\mathbb{P}}_n) \le r\right).$$

We can use the confidence interval

$$C_{n,\alpha}(X) = \left[\widehat{\theta}_n - \widehat{r}_2, \widehat{\theta}_n - \widehat{r}_1\right] \quad \text{with} \quad \widehat{r}_1 = G^{-1}_{n,\widehat{\mathbb{P}}_n}(\alpha/2) \quad \text{and} \quad \widehat{r}_2 = G_{n,\widehat{\mathbb{P}}_n}(1-\alpha/2).$$

The bootstrap algorithm for this is simple:

Algorithm 4 Estimates $G_{n,\mathbb{P}_{\text{NP}}}$ by approximating $G_{n,\widehat{\mathbb{P}}_n}$ via Monte-Carlo.

Input: A sample dataset $X_1, \dots, X_n$, estimators θ and $\widehat{\theta}_n$, a stopping time B
Output: An estimate $G_{n,\widehat{\mathbb{P}}_n}$
  for $b = 1, \dots, B$ **do**
    $X_1^{*b}, \dots, X_n^{*b} \overset{\text{i.i.d.}}{\sim} \widehat{\mathbb{P}}_n$
    $R_n^{*b} \leftarrow \widehat{\theta}_n(X^{*b}) - \theta(\widehat{\mathbb{P}}_n)$
  return $\frac{1}{B}\sum_{b=1}^B \mathbb{1}_{(-\infty, R_n^{*b}]}$

The quantity $R_n(X, \mathbb{P}_{\text{NP}}) = \widehat{\theta}_n(X) - \theta(\mathbb{P}_{\text{NP}})$ is called a **root** (a function of the data and distribution, and used to make confidence intervals). If $\widehat{\sigma}(X)$ is an estimate of $\text{s.e.}(\widehat{\theta}_n)$ then other examples of roots are

$$R_n(X, \mathbb{P}_{\text{NP}}) = \frac{\widehat{\theta}_n(X) - \theta(\mathbb{P}_{\text{NP}})}{\widehat{\theta}(X)} \quad \text{and} \quad R_n(X, \mathbb{P}_{\text{NP}}) = \frac{\widehat{\theta}_n(X)}{\theta(\mathbb{P}_{\text{NP}})}.$$

We want to choose R_n so that its sampling distribution $G_{n,\widehat{\mathbb{P}}_n}$ changes slowly with $\mathbb{P}_{\text{NP}}$ (so $G_{n,\widehat{\mathbb{P}}_n} \approx G_{n,\mathbb{P}_{\text{NP}}}$). The *Studentized* root $\frac{\widehat{\theta}_n(X) - \theta(\mathbb{P}_{\text{NP}})}{\widehat{\sigma}(X)}$ usually works better than $\widehat{\theta}_n(X) - \theta(\mathbb{P}_{\text{NP}})$, then we get

$$C_{n,\alpha}(X) = \left[\widehat{\theta}_n(X) - \widehat{r}_2\widehat{\sigma}(X), \widehat{\theta}_n - \widehat{r}_1\widehat{\sigma}(X)\right].$$

We might have a theory that tells us that, for example,

$$\sup_{a<b}\left|G_{n,\widehat{\mathbb{P}}_n}([a,b]) - G_{n,\mathbb{P}_{\text{NP}}}([a,b])\right| \xrightarrow{p} 0$$

but still we worry about finite sample coverage. Let

$$\gamma_{n,\mathbb{P}_{\text{NP}}}(\alpha) = \mathbb{P}_{\mathbb{P}_{\text{NP}}}(\theta(\mathbb{P}_{\text{NP}}) \in C_{n,\alpha}) \to 1-\alpha \quad \text{if } C_{n,\alpha} \text{ has asymptotic coverage.}$$

But in finite samples we might have $\gamma_{n,\mathbb{P}_{\text{NP}}}(\alpha) < 1-\alpha$. The solution is to use *double bootstrap*: the algorithm is to

1. Estimate $\gamma_{n,\mathbb{P}_{\text{NP}}}$ via plug-in estimator $\gamma_{n,\widehat{\mathbb{P}}_n}$.
2. Use $C_{n,\widehat{\alpha}}(X)$ where $\widehat{\gamma}(\widehat{\alpha}) = 1-\alpha$.

The algorithm for step 1 is

Algorithm 5 Estimates $\gamma_{n,\mathbb{P}_{\text{NP}}}$ by approximating $\gamma_{n,\widehat{\mathbb{P}}_n}$ via Monte-Carlo.

Input: A sample dataset $X_1, \dots, X_n$; estimators θ, $\widehat{\theta}_n$, and $\widehat{\sigma}$; a set Λ of α values; Monte-Carlo stopping times A, B
Output: Estimates $\widehat{\gamma}_{n,\widehat{\mathbb{P}}_n}$ and $\widehat{\alpha}$.

for $a \in [A]$ **do**
  $X_1^{*a}, \dots, X_n^{*a} \overset{\text{i.i.d.}}{\sim} \widehat{\mathbb{P}}_n$
  $\widehat{\mathbb{P}}_n^{*a} \leftarrow (S \mapsto \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i^{*a} \in S))$
  for $b \in [B]$ **do**
    $X_1^{**a,b}, \dots, X_n^{**a,b} \overset{\text{i.i.d.}}{\sim} \widehat{\mathbb{P}}_n^{*a}$
    $R_n^{**a,b} \leftarrow \left(\widehat{\theta}_n(X^{**a,b}) - \theta(\widehat{\mathbb{P}}_n^{*a})\right)/\widehat{\sigma}(X^{**a,b})$
  $\widehat{G}_n^{*a} \leftarrow \frac{1}{B}\sum_{b=1}^n \mathbb{1}_{\left(-\infty, R_n^{**a,b}\right]}$
  for $\alpha \in \Lambda$ **do**
    $C_{n,\alpha}^{*a} \leftarrow \left[\widehat{\theta}_n^{*a} - \widehat{\sigma}^{*a} r_2(\widehat{G}_n^{*a}), \widehat{\theta}_n^{*a} - \widehat{\sigma}^{*a} r_1(\widehat{G}_n^{*a})\right]$
for $\alpha \in \Lambda$ **do**
  $\widehat{\gamma}(\alpha) \leftarrow \frac{1}{A}\sum_{a=1}^A \mathbb{1}\left(\theta(\widehat{\mathbb{P}}_n) \in C_{n,\alpha}^{*a}\right)$
$\widehat{\alpha} \leftarrow \widehat{\gamma}^{-1}(1-\alpha)$

and the rest can be done as already described.

12 Multiple Testing

In many testing problems, we want to test many hypotheses at a time, e.g.,

- Test $H_{0j}\colon \beta_j = 0$ for $j \in [d]$ in linear regression.
- Test whether each of $2 \cdot 10^6$ SNPs is associated with a given phenotype (e.g. diabetes/schizophrenia).
- Test whether each of $2 \cdot 10^3$ tweaks affects unit engagement.

The setup is $X \sim \mathbb{P}_\theta \in \mathcal{P} = \{\mathbb{P}_\theta\colon \theta \in \Theta\}$, where $\mathcal{P}$ can be non-parametric (and then θ represents, i.e., a density). Let $H_{0i}\colon \theta \in \Theta_{0i}$ for $i \in [m]$. (The null hypothesis $H_{0i}\colon \theta_i = 0$ is common).
The goal is to return an accept/reject decision for each i. Let

$$\mathcal{R}(X) = \{i\colon H_{0i} \text{ rejected}\} \subseteq [m], \quad R(X) = |\mathcal{R}(X)| \quad \text{and} \quad \mathcal{H}_0(\theta) = \{i\colon H_{0i} \text{ are true}\}, \quad m_0 = |\mathcal{H}_0|.$$

Family-Wise Error Rate

The naive option to testing all H_{0i} at the same time is that, for large m, we might have

$$\mathbb{P}(\text{any } H_{0i} \text{ rejected}) \gg \alpha$$

just by random variation, and so the false rejection rate of this combined test is higher than α.

Example 12.1. If $X_i \overset{\text{ind}}{\sim} \mathcal{N}(\theta_i, 1)$ and $H_{0i}\colon \theta_i = 0$ for $i \in [m]$, then

$$\lim_{m\to\infty} \mathbb{P}_0(\text{any } H_{0i} \text{ rejected}) = \lim_{m\to\infty} (1 - (1-\alpha)^m) = 1.$$

This is a problem, since all attention will be focused on the false rejections and not on the correct non-rejections.
The classical solution is to control the **family-wise error rate** (FWER):

$$\text{FWER}(\theta) = \mathbb{P}_\theta(\text{any false rejections}) = \mathbb{P}_\theta(\mathcal{R}(X) \cap \mathcal{H}_0(\theta) \neq \emptyset).$$

We want, mirroring the original error rate consideration,

$$\sup_{\theta\in\Theta} \text{FWER}(\theta) \leq \alpha.$$

It's typically achieved by "correcting" the marginal error rate to a quantity $\widetilde{\alpha}_m$. Suppose $p_i(X)$ is the p-value for the m^{th} experiment. Then $p_i \overset{H_{0i}}{\sim} \operatorname{Uni}([0,1])$.

Šidák's Correction

Assume the $p_i(X)$ are mutually independent. And since $p_i \overset{H_{0i}}{\sim} \operatorname{Uni}([0,1])$, $\mathbb{P}_\theta(p_i(X) \le t) \le t$ for $\theta \in \Theta_{0i}$. Then

$$\mathbb{P}_\theta(\text{no false rejections}) = \mathbb{P}_\theta(p_i(X) \ge \widetilde{\alpha}_m \text{ for all } i \in \mathcal{H}_0) \ge (1-\widetilde{\alpha}_m)^{m_0} \ge (1-\widetilde{\alpha}_m)^m.$$

Then

$$\mathbb{P}_\theta(\text{no false rejections}) \ge 1-\alpha \implies \widetilde{\alpha}_m \ge 1-(1-\alpha)^{1/m} \approx \frac{\alpha}{m} \quad \text{for small } m.$$

This gives us tight control of FWER whenever $p_i \overset{H_{0i}}{\sim} \operatorname{Uni}([0,1])$ and are independent. This is **Šidák's correction**.

Bonferroni Correction

Now we don't have to assume that the $p_i(X)$ are mutually independent. Then

$$\mathbb{P}_\theta(\text{any false rejections}) = \mathbb{P}_\theta\left(\bigcup_{i\in\mathcal{H}_0}\{H_{0i} \text{ rejected}\}\right) \le \sum_{i\in\mathcal{H}_0}\mathbb{P}_\theta(H_{0i} \text{ rejected}) = m_0\widetilde{\alpha}_m \le m\widetilde{\alpha}_m,$$

so

$$\mathbb{P}_\theta(\text{any false rejections}) \le \alpha \implies \widetilde{\alpha}_m \ge \frac{\alpha}{m}.$$

This is **Bonferroni Correction**. There are no assumptions on interdependence of the $p_i(X)$ and the performance is not much worse than Šidák.

Holm's Procedure

We can directly improve on Bonferroni by using a **step-down procedure**, or **Holm's procedure**. First, we order p-values $p_{(1)} \le p_{(2)} \le \cdots p_{(m)}$ and hypotheses $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$. Then

$$\text{let} \quad R(X) = \sup\left\{r \colon p_{(i)}(X) \le \frac{\alpha}{m-i+1} \text{ for all } i \le r\right\} \quad \text{and reject} \quad H_{0(1)}, \ldots, H_{0(R)}.$$

Proposition 12.2. Holm's procedure controls FWER at level α.

Proof. Let $p_0^* = \inf\{p_i \colon i \in \mathcal{H}_0\}$. Then

$$\mathbb{P}\left(p_0^* \le \frac{\alpha}{m_0}\right) \le \sum_{i\in\mathcal{H}_0}\mathbb{P}\left(p_i \le \frac{\alpha}{m_0}\right) \le \sum_{i\in\mathcal{H}_0}\frac{\alpha}{m_0} \le m_0\frac{\alpha}{m_0} \le \alpha.$$

Suppose $p_0^* > \frac{\alpha}{m_0}$ and let

$$k = |\{i \colon p_i \le p_0^*\}| \le m - m_0 + 1.$$

Then

$$p_{(k)} = p_0^* > \frac{\alpha}{m_0} \ge \frac{\alpha}{m-k+1}.$$

So $t < k$, and $p_0^* > p_{(t)}$. Thus there are no false rejections, so the FWER is controlled at level α. □

Holm's procedure strictly dominates the Bonferroni correction. A different step-down procedure dominates Šidák's procedure,

$$\text{reject } H_{0(i)} \text{ when } p_{(i)} \le \frac{\alpha}{m-i+1}.$$

There is a general framework for making such improvements called the **closure principle**.

Closure Principle

Assume we can construct a (marginal) level α test for every **intersection null** hypothesis:

$$\text{for every } S \subseteq [m] \text{ let} \quad H_{0S} \colon \theta \in \bigcap_{i \in S} \Theta_{0i}$$

for example, rejecting H_{0S} if $\min_{i \in S} p_i \le \frac{\alpha}{|S|}$.
There is a two-step general procedure to improve a marginal test (such as testing every hypothesis individually using Bonferroni's or Šidák's correction, as was detailed above).

1. Provisionally reject H_{0S} if the marginal test rejects H_{0i} for all $i \in S$.
2. Reject H_{0i} in the multiple test if H_{0S} is rejected for all $S \ni i$.

Proposition 12.3. This two-step procedure controls FWER.

Proof. Simple,

$$\mathbb{P}(\text{any false rejections}) \le \mathbb{P}(H_{0\mathcal{H}_0} \text{ rejected in Step 1}) = \mathbb{P}\left(\bigcup_{i \in \mathcal{H}_0} \{H_{0i} \text{ rejected in marginal test}\} \right) \le \alpha.$$

□

Testing with Dependence

In cases with independence, Bonferroni's correction isn't much worse than Šidák's correction. For example, if $\alpha = 1/20$ and $m = 20$, then Bonferroni's correction gives $\widetilde{\alpha}_m = 1/400$ and Šidák's correction gives $\widetilde{\alpha}_m = 1 - \sqrt[20]{19/20}$, very close together. But when tests are highly dependent, we can often do *much* better.

Example 12.4 (Scheffe's S-Method). Suppose $X \sim \mathcal{N}(\theta, I_d)$ with $\theta \in \Theta = \mathbb{R}^d$. Let $H_{0,\lambda} \colon \theta'\lambda = 0$ for $\theta \in S^{d-1}$. In this case $m_0 = \aleph_1$ and thus it is impossible to use either correction. In this case the procedure is

$$\text{reject } H_{0,\lambda} \text{ if} \quad \|X'\lambda\|_2^2 \ge \chi_d^2(\alpha) \approx d + 3\sqrt{d}$$

for $\alpha = 1/20$. This controls the FWER:

$$\sup_{\lambda \colon \theta'\lambda = 0} \|X'\lambda\|_2^2 \le \sup_{\lambda} \|(X - \theta)'\lambda\|_2^2 \sim \chi_d^2(\alpha).$$

We can view this as a **deduction** from the confidence region

$$C(X) = \left\{ \theta \colon \|\theta - X\|_2^2 \le \chi_d^2(\alpha) \right\}.$$

Deduced Inference

Given any joint confidence region $C(X)$ for $\theta \in \Theta$, we may freely assume $\theta \in C(X)$ and "deduce" any and all implied conclusions without any FWER inflation. Indeed, simply,

$$\mathbb{P}_\theta(\text{any deduced inference is wrong}) \le \mathbb{P}_\theta(\theta \notin C(X)) \le \alpha.$$

Deduction is also a good paradigm for deriving simulatneous intervals.
We say $C_1(X), \dots, C_k(X)$ are simultaneous $1 - \alpha$ confidence intervals for $g_1(\theta), \dots, g_k(\theta)$ if

$$\mathbb{P}_\theta(g_i(\theta) \in C_i(X) \text{ for } i \in [k]) \ge 1 - \alpha.$$

Example 12.5 (Simultaneous Confidence Intervals for Multivariate Gaussian). Assume $X \sim \mathcal{N}(\theta, \Sigma)$, for Σ known with $\Sigma_{ii} = 1$ for all i. Let t_α be the upper-α quantile of $\|X - \theta\|_\infty$. Then

$$C(X) = \{\theta \colon |\theta_i - X_i| \le c_\alpha \text{ for } i \in [d]\} = \bigtimes_{i=1}^d (X_i - t_\alpha, X_i + t_\alpha) = \bigtimes_{i=1}^d C_i(X_i).$$

Then

$$\mathbb{P}_\theta(\theta_i \notin C_i(X) \text{ for any } i) = \mathbb{P}_\theta(\theta \notin C(X)) = \alpha.$$

We could have instead constructed an elliptical confidence region, but then the intervals would be conservative.

Example 12.6 (Linear Regression). Suppose $X \in \mathbb{R}^{n \times d}$ is a design matrix, and $Y \in \mathbb{R}^n$ is the response vector, with

$$Y = X\beta + \varepsilon \quad \text{where} \quad \varepsilon \sim \mathcal{N}\big(0, \sigma^2 I\big).$$

Then

$$\widehat{\beta} \sim \mathcal{N}\big(\beta, \sigma^2 (X'X)^{-1}\big).$$

We estimate

$$\widehat{\sigma}^2 = \frac{\mathrm{RSS}(\beta)}{n-d} \perp\!\!\!\perp \widehat{\beta}.$$

Then

$$\frac{\widehat{\beta} - \beta}{\widehat{\sigma}} = \frac{Z}{\sqrt{V/(n-d)}} \quad \text{where} \quad Z = \frac{\widehat{\beta} - \beta}{\sigma} \sim \mathcal{N}\big(0, (X'X)^{-1}\big), \quad V = \frac{\mathrm{RSS}(\beta)}{\sigma^2} \sim \chi^2_{n-d}.$$

Since $Z \perp\!\!\!\perp V$, the distribution of $\frac{\widehat{\beta}-\beta}{\widehat{\sigma}}$ is fully known.

Assume without loss of generality (via scaling) that $((X'X)^{-1})_{jj} = 1$ for $j \in [n]$. Let t_α denote the upper-α quantile of $\left\| \frac{\widehat{\beta}-\beta}{\widehat{\sigma}} \right\|_\infty$. Then $C_i = \widehat{\beta}_i \pm \widehat{\sigma} t_\alpha$ are simultaneous confidence intervals for $\widehat{\beta}_i$, $i \in [d]$. We can compute t_α by simulation. Then

$$\mathbb{P}(\beta_i \in C_i(X) \text{ for } i \in [d]) = \mathbb{P}\Big(\Big|\widehat{\beta}_i - \beta_i\Big| \le \widehat{\sigma} t_\alpha \text{ for } i \in [d]\Big) = 1 - \alpha.$$

False Discovery Rate

Suppose we test 10^4 hypotheses with independent test statistics, all at level $\alpha = 10^{-3}$. We expect 10 rejections by chance. What if we get 50? Probably only 20% of them are fake rejections. Can we accept 10 false rejections as long as most rejections are valid?

Benjamini and Hochberg (in 1995) proposed a more liberal error control criterion called FDR. We define

$$R(X) = |\mathcal{R}(X)| = \#\text{ of rejections/“discoveries”} \quad \text{and} \quad V(X;\theta) = |\mathcal{R}(X) \cap \mathcal{H}_0(\theta)| = \#\text{ of false discoveries}.$$

The **false discovery proportion** (FDP) is

$$\mathrm{FDP}(X, \theta) = \frac{V(X;\theta)}{\min\{R(X), 1\}}$$

and the FDR is given by

$$\mathrm{FDR}(X, \theta) = \mathbb{E}_\theta[\mathrm{FDP}(X,\theta)] = \mathbb{E}_\theta\left[\frac{V(X;\theta)}{\min\{R(X),1\}}\right]$$

Benjamini-Hochberg Procedure

Benjamini and Hochberg (in 1995) also proposed a **step-up** method to control FDR. Given ordered p-values $p_{(1)} \le \cdots \le p_{(m)}$,

$$\text{let} \quad R(X) = \sup\left\{r \colon p_{(r)} \le \frac{\alpha r}{m}\right\} \quad \text{and reject} \quad H_{0(1)}, \ldots, H_{0(R)}.$$

This is *much* more liberal than Holm's procedure when $1 \ll R \ll m$. The Benjamini-Hochberg procedure rejects at least r null hypotheses if $p_{(r)} \le \frac{\alpha r}{m}$; Holm's procedure needs $p_{(r)} \le \frac{\alpha}{m-r+1} \approx \frac{\alpha}{m}(1 + \frac{r}{m})$.

Benjamini-Hochberg as Empirical Bayes

An equivalent formulation of Benjamini-Hochberg is

$$\text{let} \quad R_t(X) = |\{i \colon p_i(X) \le t\}|, \quad \widehat{\mathrm{FDP}_t}(X) = \frac{mt}{\min\{R_t(X), 1\}}, \quad t^*(X) = \sup\left\{t \colon \widehat{\mathrm{FDP}_t}(X) \le \alpha\right\},$$

$$\text{and reject } H_{0i} \text{ when} \quad p_i(X) \le t^*(X).$$

This works because $\widehat{\mathrm{FDP}_t}$ continuously increass at t, except at $p_{(i)}$ values where it has a downward jump discontinuity. The only values of t that matter for the algorithm are

$$t = p_{(i)} \implies \widehat{\mathrm{FDP}_t}(X) = \frac{m p_i(X)}{i},$$

and so equivalence is shown by

$$\frac{mp_{(i)}(X)}{i} \le \alpha \iff p_{(i)}(X) \le \frac{\alpha i}{m}.$$

FDR Control

Proposition 12.7. Suppose $p_i(X)$ are independent, so that $p_i \sim \mathrm{Uni}([0,1])$ for $i \in \mathcal{H}_0$. Then the Benjamini-Hochberg procedure controls the FDR.

Proof. The proof is due to Storey, Taylor, and Siegmund.
Let $V_t(X;\theta) = |\{i \in \mathcal{H}_0(\theta) \colon p_i(X) \le t\}|$, and FDP_t and Q_t be defined by the following

$$\mathrm{FDP}_t(X,\theta) = \frac{V_t(X;\theta)}{\min\{R_t(X),1\}} = \widehat{\mathrm{FDP}_t}(X,\theta) \cdot \frac{V_t(X;\theta)}{mt} = \widehat{\mathrm{FDP}_t}(X,\theta) Q_t(X,\theta).$$

Then we know $\widehat{\mathrm{FDP}_{t^*}}(X,\theta) \overset{\text{a.s.}}{=} \alpha$, so

$$\mathrm{FDR}(X,\theta) = \mathbb{E}_\theta[\mathrm{FDP}_{t^*}(X,\theta)] = \mathbb{E}_\theta\left[\widehat{\mathrm{FDP}_{t^*}}(X,\theta) \cdot Q_{t^*}(X,\theta)\right] = \alpha \mathbb{E}_\theta[Q_{t^*}(X,\theta)].$$

Then Q_t is a backwards martingale from $t=1$ to $t=0$ with respect to the filtration defined by, again from $t=1$ to $t=0$, $\Sigma_t = \sigma(\min\{p_1(X),t\},\dots,\min\{p_m(X),t\})$: if $s<t$ then

$$\mathbb{E}_\theta[V_s(X,\theta) \mid V_t(X,\theta)] = \mathbb{E}_\theta[|\{i \colon p_i(X) \le s\}| \mid |\{i \colon p_i(X) \le t\}|] = |\{i \colon p_i(X) \le t\}| \frac{s}{t} = V_t(X,\theta)\frac{s}{t}.$$

And note that

$$\mathbb{E}_\theta\left[\frac{V_s(X,\theta)}{ms} \,\middle|\, V_t(X,\theta)\right] = \frac{1}{ms}\left(\frac{s}{t}V_t(X,\theta)\right) = \frac{V_t(X,\theta)}{mt}.$$

Thus $Q_t(X,\theta)$ is a martingale. And t^* is a stopping time with respect to the same filtration, since for $s \ge t$, if

$$R_s(X) = |\{i \colon p_i(X) \le s\}| = |\{i \colon \min\{p_i(X),t\} \le s\}|$$

then

$$\widehat{\mathrm{FDP}_s}(X) = \frac{ns}{R_s(X)}.$$

Thus by the optional stopping theorem,

$$\mathrm{FDR}(X,\theta) = \alpha \mathbb{E}\left[\frac{V_{t^*}}{mt^*}\right] = \alpha \mathbb{E}\left[\frac{V_1}{m}\right] = \alpha \frac{m_0}{m} \le \alpha.$$

□

This proof only really works if the p-values are independent and the p-values corresponding to the null hypotheses are exactly uniform. A more robust proof shows that the FDR is controlled when the null p-values are conservative, and this idea can be extended to "positive dependence." The FDR can be controlled under general dependence if we use the corrected level $\frac{\alpha}{H_m} \approx \frac{\alpha}{\log(m)}$.