Theoretical Statistics

Statistics 210A
Fall 2017

Professor Will Fithian
Scribe: Sinho Chewi

Contents

Lecture 1

August 24

1.1 Measure Theory Basics

Given a set $\mathcal{X}$, a measure μ maps subsets $A \subseteq \mathcal{X}$ to $[0, \infty]$.

Example 1.1. If $\mathcal{X}$ is countable (e.g. $\mathcal{X} = \mathbb{Z}$), the **counting measure** $\#(A)$ equals the number of points in A.

Example 1.2. If $\mathcal{X} = \mathbb{R}^n$, the **Lebesgue measure** is $\lambda(A) = \int \cdots \int_A \mathrm{d}x_1 \cdots \mathrm{d}x_n = \mathrm{Vol}(A)$.

Because of pathological sets, $\lambda(A)$ is only defined for some subsets $A \subseteq \mathbb{R}^n$. This leads to the idea of a σ-field (σ-algebra).

A **σ-field** $\mathcal{F}$ is a collection of sets on which μ is defined, satisfying certain closure properties.

Example 1.3. If $\mathcal{X}$ is countable, $\mathcal{F} = 2^{\mathcal{X}}$ (all subsets).

Example 1.4. If $\mathcal{X} = \mathbb{R}^n$, then $\mathcal{F}$ is the **Borel σ-field**, $\mathcal{B}$, the smallest σ-field containing all rectangles.

Given $(\mathcal{X}, \mathcal{F})$ (a **measurable space**), a **measure** is any map $\mu : \mathcal{F} \to [0, \infty]$ with $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ if $A_i \in \mathcal{F}$ are disjoint. If $\mu(\mathcal{X}) = 1$ (usually $\mathbb{P}$), then μ is a **probability measure**.

Measures let us define **integrals**, $\int f(x)\,\mathrm{d}\mu(x)$ or $\int f\,\mathrm{d}\mu$, that put weight $\mu(A)$ on A.

Counting: $\int f(x)\,\mathrm{d}\#(x) = \sum_{x \in \mathcal{X}} f(x)$.

Lebesgue: $\int f(x)\,\mathrm{d}\lambda(x) = \int \cdots \int f(x)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n$.

1.1.1 Densities

Given $(\mathcal{X}, \mathcal{F})$ and two measures μ, $\mathbb{P}$, we say that $\mathbb{P}$ is **absolutely continuous with respect to** μ if $\mathbb{P}(A) = 0$ whenever $\mu(A) = 0$ (if μ is the Lebesgue measure, we just say that $\mathbb{P}$ is **absolutely continuous**). Notate this as $\mathbb{P} \ll \mu$.

If $\mathbb{P} \ll \mu$, then we can define a **density function**

$$p = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}$$

with $\mathbb{P}(A) = \int_A p(x)\,\mathrm{d}\mu(x)$. Recall that $\mathbb{P}(A) = \int_A \mathrm{d}\mathbb{P}(x)$. Also, $\int f(x)\,\mathrm{d}\mathbb{P}(x) = \int f(x)p(x)\,\mathrm{d}\mu(x)$.

Let $\mathbb{P}$ be a probability measure. If $\mu = \#$, then p is a **probability mass function**. If $\mu = \lambda$, then p is a **probability density function**.

If $\mathrm{d}\mathbb{P} = p\,\mathrm{d}\lambda$, then $\mathbb{P}(A) = \int_A \mathrm{d}\mathbb{P}(x) = \int_A p(x)\,\mathrm{d}x$. If we redefine p at a single point, then we obtain another density, so density functions are not unique, but any two densities agree almost everywhere, so the distinction is not important.

1.1.2 Random Variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a **probability space**. $\omega \in \Omega$ is called an **outcome**. $A \in \mathcal{F}$ is called an **event**. $\mathbb{P}(A)$ is called the **probability of** A.

A **random variable** (**vector**) is a function $X : \Omega \to \mathbb{R}$ ($\mathbb{R}^n$). We say that X has **distribution** Q ($X \sim Q$) if $\mathbb{P}(X \in B) = \mathbb{P}(\{\omega : X(\omega) \in B\}) = Q(B)$ for $B \in \mathcal{B}$.

An **expectation** is an integral with respect to $\mathbb{P}$. $\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int \cdots \int_{\mathbb{R}^n} x\,\mathrm{d}Q(x)$.

1.2 Risk & Estimation

A **statistical model** is a family of candidate probability distributions. $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for some observed data $X \sim P_\theta$. θ is called the **parameter**.

Goal of Estimation: Observe $X \sim P_\theta$ and guess the value of $g(\theta)$ (**estimand**).

Example 1.5. Flip a biased coin n times. θ is the probability of landing heads and X is the number of heads after n flips. $\Theta = [0, 1]$. $X \sim \text{Binomial}(n, \theta)$, with $p_\theta(x) = \theta^x (1-\theta)^{n-x} \binom{n}{k}$ for $x \in \{0, \ldots, n\}$.

A **statistic** is any function $T(X)$ of data X. An **estimator** $\delta(X)$ **of** $g(\theta)$ is any statistic meant to guess $g(\theta)$.

In the example, a natural estimator is $\delta_0(X) = X/n$. Is this a good estimator?

A **loss function** $L(\theta, d)$ measures the "badness" of the guess.

Example 1.6. $L(\theta, d) = (d - g(\theta))^2$ is the **squared error**.

Typical properties:

- $L(\theta, d) \geq 0$ for all θ, d.
- $L(\theta, g(\theta)) = 0$ for all θ.

The **risk function** is $R(\theta, \delta(\cdot)) = \mathbb{E}_\theta[L(\theta, \delta(X))]$.

Example 1.7. If $L(\theta, d) = (d - g(\theta))^2$, then $R(\theta, \delta) = \mathbb{E}_\theta[(\delta(X) - g(\theta))^2]$ (the **MSE**).

Lecture 2

August 29

2.1 Review

2.1.1 Basic Measure Theory

A measure space is

$$(\underbrace{\mathcal{X}}_{\text{set}}, \underbrace{\mathcal{F}}_{\sigma\text{-field}}, \underbrace{\mu}_{\text{measure}})$$

where $\mu(A) \in [0, \infty]$ for $A \in \mathcal{F}$ is the "weight" on A. If $P \ll \mu$, then a density $p \left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)$ is a function such that $P(A) = \int_A \mathrm{d}P(x) = \int_A p(x)\,\mathrm{d}\mu(x)$ and $\int f\,\mathrm{d}P = \int fp\,\mathrm{d}\mu$.

2.1.2 Statistical Model

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

Estimation: We have

- an estimand, $g(\theta)$;
- an estimator, $\delta(X)$;
- loss $L(\theta, d)$, e.g. $(g(\theta) - d)^2$;
- risk, $R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(X))]$.

2.2 Comparing the Risk of Different Estimators

Example 2.1. $X \sim \text{Binomial}(n, \theta)$, so $p_\theta(x) = \theta^x(1-\theta)^{n-x}\binom{n}{x}$. An estimator for θ is $\delta_0(X) = X/n$. The expectation of the estimator is $\mathbb{E}_\theta[X/n] = \theta$ (it is unbiased). So, $R(\theta, \delta) = \text{var}_\theta(X/n) = \theta(1-\theta)/n$.

Other choices:

$$\delta_1(X) = \frac{X+3}{n},$$
$$\delta_2(X) = \frac{X+3}{n+6}.$$

$R(\theta, \delta_1)$ is always greater than $R(\theta, \delta_0)$ because δ_1 has the same variance as δ_0, but more bias. $R(\theta, \delta_2)$ is smaller than $R(\theta, \delta_0)$ when θ is close to $1/2$.

δ_1 is definitely bad, but the comparison between δ_0 and δ_2 is more ambiguous.

An estimator δ is **inadmissible** if there exists δ^* such that

(a) $R(\theta, \delta^*) \leq R(\theta, \delta)\ \forall \theta \in \Theta$,

(b) $R(\theta, \delta^*) < R(\theta, \delta)$ for some $\theta \in \Theta$.

Strategies to resolve ambiguity:

1. Summarize the risk function as a scalar.
   (a) Average-case risk: for some measure Λ, minimize $\int_\Theta R(\theta, \delta)\,\mathrm{d}\Lambda(\theta)$. This is called the **Bayes estimator**, and Λ is the **prior**.
   (b) Worst-case risk: minimize $\sup_{\theta \in \Theta} R(\theta, \delta)$ (over $\delta : \mathcal{X} \to \mathbb{R}$).
2. Constrain the choice of estimator.
   (a) Only consider unbiased δ. $\mathbb{E}_\theta[\delta(X)] = g(\theta)\ \forall \theta \in \Theta$.

2.3 Exponential Families

An s-**parameter exponential family** is a family of probability densities $\{\rho_\eta : \eta \in \Xi\}$ with respect to a measure μ on $\mathcal{X}$ of the form

$$\rho_\eta(x) = \exp\{\eta^\intercal T(x) - A(\eta)\} h(x)$$

where $T : \mathcal{X} \to \mathbb{R}^s$ is a **sufficient statistic**, $h : \mathcal{X} \to \mathbb{R}$ is the **carrier/base density**, $\eta \in \Xi \subseteq \mathbb{R}^s$ is the **natural parameter**, and $A : \Xi \to \mathbb{R}$ is the **cumulant generating function** (CGF). The CGF A is totally determined by T, h since we have $\int_\mathcal{X} \rho_\eta \,\mathrm{d}\mu = 1\ \forall \eta$. So,

$$A(\eta) = \log \int_\mathcal{X} \mathrm{e}^{\eta^\intercal T(x)} h(x)\,\mathrm{d}\mu(x).$$

ρ_η is only normalizable if $A(\eta) < \infty$. The **natural parameter space** is the set of all "allowable" η,

$$\Xi = \left\{\eta : \int e^{\eta^\intercal T} h\,\mathrm{d}\mu < \infty\right\}.$$

If Ξ is the natural parameter space, $\{\rho_\eta : \eta \in \Xi\}$ is in **canonical form**. ρ_η is convex in η, so Ξ is convex. Note that we have the same exponential family if:

- we change $\mu \rightsquigarrow \tilde{\mu}$, where
  $$\frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\mu} = h,$$
  and then $h \rightsquigarrow \tilde{h} = 1$.
- Or, (if $0 \in \Xi$), take $h \rightsquigarrow \tilde{h} = \rho_0$, and $A(\eta) \rightsquigarrow \tilde{A}(\eta) = A(\eta) - A(0)$.

Interpretation of Exponential Families:

- Start with a base density ρ_0.
- Apply an "**exponential tilt**":
  1. multiply by $\mathrm{e}^{\eta^\intercal T}$
  2. renormalize (if possible)

An exponential family in canonical form is all possible tilts of h (or any ρ_η) using any linear combination of T.

Example 2.2. Let $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Let $\theta = (\mu, \sigma^2)$.

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

Then:

$$\begin{aligned} \eta(\theta) &= \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right) \\ T(x) &= \left(x, x^2\right) \\ h(x) &= \frac{1}{\sqrt{2\pi}} \\ B(\theta) &= A(\eta(\theta)) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2} \log\left(\sigma^2\right) \end{aligned}$$

In canonical form:

$$\begin{aligned} \rho_\eta(x) &= \mathrm{e}^{\eta_1 x - \eta_2 x^2 - A(\eta)}, \\ A(\eta) &= \frac{\eta_1^2}{4\eta_2} - \frac{1}{2} \log(2\eta_2) + \log\left(\sqrt{2\pi}\right) \end{aligned}$$

Example 2.3. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

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

Lecture 3

August 31

3.1 Integrals

The integral $\int f \, \mathrm{d}\mu$ is generally abstract.

If $\dfrac{\mathrm{d}\mu}{\mathrm{d}\lambda_{\mathbb{R}^n}} = p$, then $\displaystyle\int f \, \mathrm{d}\mu = \int_{x \in \mathbb{R}^n} f(x) p(x) \, \mathrm{d}x$.

If $\dfrac{\mathrm{d}\mu}{\mathrm{d}\#_{\mathcal{X}}} = p$, then $\displaystyle\int f \, \mathrm{d}\mu = \sum_{x \in \mathcal{X}} f(x) p(x)$.

Note that if $X \sim \mathcal{N}(0,1)$, then X_+, the positive part of X, does not have a density with respect to Lebesgue measure or counting measure.

3.2 Exponential Family Examples

Example 3.1. If $X \sim \mathcal{N}(\mu, \sigma^2)$, with density

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)},$$

then

$$\eta = \begin{bmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{bmatrix}, \qquad T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}, \qquad A\big(\eta(\mu, \sigma^2)\big) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2} \log \sigma^2.$$

Example 3.2. If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

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

and

$$T(x) = \begin{bmatrix} \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2 \end{bmatrix}, \qquad \eta = \begin{bmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{bmatrix}, \qquad A(\eta) = nA^{(1)}(\eta).$$

Generally, suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} e^{\eta^\intercal T - A(\eta)} h$. Then,

$$\begin{aligned} X \sim p_\eta(x) &= \prod_{i=1}^n e^{\eta^\intercal T(x_i) - A(\eta)} h(x_i) \\ &= e^{\eta^\intercal \sum_{i=1}^n T(x_i) - nA(\eta)} \prod_{i=1}^n h(x_i). \end{aligned}$$

$T(X)$ also follows an exponential family. If $X \sim p_\eta^X(x) = e^{\eta^\intercal T(x) - A(\eta)} h^X(x)$, then (informally)

$$\mathbb{P}_\eta\big(T(X) = t\big) = \int_{\{x : T(x) = t\}} e^{\eta^\intercal t - A(\eta)} h^X(x) \, d\mu(x)$$

so

$$p_\eta^T(t) = e^{\eta^\intercal t - A(\eta)} \underbrace{\int_{\{x : T(x) = t\}} h^X(x) \, d\mu(x)}_{h^T(t)}.$$

3.2.1 Binomial

If $X \sim \text{Binomial}(n, \theta)$,

$$\begin{aligned} p_\theta(x) &= \theta^x (1-\theta)^{n-x} \binom{n}{x} \\ &= \Big(\frac{\theta}{1-\theta}\Big)^x (1-\theta)^n \binom{n}{x} \\ &= e^{x \log(\theta/(1-\theta)) + n \log(1-\theta)} \binom{n}{x}, \end{aligned}$$

with natural parameter

$$\begin{aligned} \eta(\theta) &= \log \frac{\theta}{1-\theta}, \\ A\big(\eta(\theta)\big) &= -n \log(1-\theta). \end{aligned}$$

3.2.2 Poisson

If $X \sim \text{Poisson}(\lambda)$, then

$$\begin{aligned} p_\lambda(x) &= \frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, \ldots \\ &= \exp\{(\log \lambda) x - \lambda\} \frac{1}{x!}, \end{aligned}$$

with natural parameter

$$\eta(\lambda) = \log \lambda.$$

3.3 Differential Identities

Theorem 3.3 (Keener Theorem 2.4). *For $f : \mathcal{X} \to \mathbb{R}$, let*

$$\Xi_f = \Big\{ \eta \in \mathbb{R}^s : \int |f| e^{\eta^\intercal T} h \, d\mu < \infty \Big\}.$$

(Ξ_1 is the natural parameter space.) Then, $g(\eta) = \int f(x) e^{\eta^\intercal T(x)} h(x) \, d\mu(x)$ has continuous partial

derivatives of all orders for $\eta \in \Xi_f^\circ$, which can be computed by differentiating under the integral.

This implies

$$e^{A(\eta)} = \int e^{\eta^\top T(x)} h(x)\, d\mu(x) \tag{3.1}$$

has partial derivatives of all orders.

Differentiate (3.1) *once*:

$$\begin{aligned}
\frac{\partial}{\partial \eta_j} e^{A(\eta)} &= \frac{\partial}{\partial \eta_j} \int e^{\eta^\top T(x)} h(x)\, d\mu(x) \\
&= \int \frac{\partial}{\partial \eta_j} e^{\eta^\top T} h\, d\mu \\
\frac{\partial}{\partial \eta_j} A(\eta) &= \int T_j e^{\eta^\top T - A(\eta)} h\, d\mu \\
\frac{\partial}{\partial \eta_j} A(\eta) &= \mathbb{E}_\eta[T_j(X)]
\end{aligned}$$

so $\nabla A(\eta) = \mathbb{E}_\eta[T(X)]$.

Differentiate (3.1) *twice*:

$$\begin{aligned}
\frac{\partial^2}{\partial \eta_j \partial \eta_k} e^{A(\eta)} &= \frac{\partial^2}{\partial \eta_j \partial \eta_k} \int e^{\eta^\top T} h\, d\mu \\
\Big(\frac{\partial^2}{\partial \eta_j \partial \eta_k} A(\eta) + \frac{\partial}{\partial \eta_j} A(\eta) \frac{\partial}{\partial \eta_k} A(\eta)\Big) e^{A(\eta)} &= \int T_j T_k e^{\eta^\top T} h\, d\mu \\
\frac{\partial^2}{\partial \eta_j \partial \eta_k} A(\eta) + \mathbb{E}_\eta[T_j(X)]\, \mathbb{E}_\eta[T_k(X)] &= \mathbb{E}_\eta[T_j(X) T_k(X)] \\
\frac{\partial^2}{\partial \eta_j \partial \eta_k} A(\eta) &= \operatorname{cov}_\eta\big(T_j(X), T_k(X)\big)
\end{aligned}$$

so $\nabla^2 A(\eta) = \operatorname{var}_\eta T(X) \in \mathbb{R}^{s \times s}$.

3.3.1 Moment Generating Function

$$e^{-A(\eta)} \frac{\partial^{k_1 + \cdots + k_s}}{\partial \eta_1^{k_1} \cdots \partial \eta_s^{k_s}} e^{A(\eta)} = \mathbb{E}_\eta[T_1^{k_1} \cdots T_s^{k_s}].$$

In fact, $e^{A(\eta+u) - A(\eta)}$ is the MGF of $T(X)$ if $X \sim p_\eta$.

$$\begin{aligned}
M_{T(X)}(u) &= \mathbb{E}_\eta[e^{u^\top T(X)}] \\
&= \int e^{u^\top T + \eta^\top T - A(\eta)} h\, d\mu \\
&= e^{A(\eta+u) - A(\eta)} \int e^{(\eta+u)^\top T - A(\eta+u)} h\, d\mu \\
&= e^{A(\eta+u) - A(\eta)}.
\end{aligned}$$

The cumulant generating function is $K_{T(X)}(u) = \log M_{T(X)}(u) = A(\eta + u) - A(\eta)$.

3.4 Sufficiency

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim}$ Bernoulli(θ), then $T(X) = \sum_{i=1}^n X_i \sim$ Binomial(n, θ). How do we justify throwing away information?

Definition 3.4. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a model for $X \in \mathcal{X}$. We say $T(X)$ is **sufficient for $\mathcal{P}$** if $P_\theta(X \mid T)$ does not depend on θ.

Example 3.5. If $T(X) = \sum_{i=1}^n X_i = t \in \{0, \ldots, n\}$, then conditionally, $X \in \{0,1\}^n$ is uniformly distributed on all sequences with $\sum_{i=1}^n x_i = t$.

$$\begin{aligned}
\mathbb{P}_\theta(X = x \mid T = t) &= \mathbb{1}\left\{\sum_{i=1}^n x_i = t\right\} \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(T = t)} \\
&= \mathbb{1}\left\{\sum_{i=1}^n x_i = t\right\} \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\theta^t (1-\theta)^{n-t}\binom{n}{t}} \\
&= \mathbb{1}\left\{\sum_{i=1}^n x_i = t\right\} \frac{1}{\binom{n}{t}}.
\end{aligned}$$

3.4.1 Sufficiency Principle

If $T(X)$ is sufficient, any statistical procedure should depend only on $T(X)$.

Suppose $\delta(X)$ is an estimator of θ which is not a function of $T(X)$. Then, $\delta(X)$ and $\delta(\tilde{X})$ have the same distribution, where $\tilde{X}$ is "made up" given $T(X)$.

Bayesian interpretation: If θ is random, $\theta \sim \Lambda$, $X \mid \theta \sim P_\theta$, then $\theta \to T(X) \to X$ is a Markov chain if T is sufficient. Then, we could generate fake data $\tilde{X}$ from $T(X)$.

3.4.2 Minimal Sufficiency

X and $T(X)$ are both sufficient in the binomial example, but $T(X)$ is "more compressed" than X.

Definition 3.6. $T(X)$ is **minimal sufficient** if

1. $T(X)$ is sufficient,
2. for any sufficient $S(X)$, $T(X) = f(S(X))$ for some f.

Lecture 4

September 5

4.1 Sufficiency

$T(X)$ is sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if $P_\theta(X \mid T)$ does not depend on θ.

Interpretation: Nature generates data in two steps.

1. Generate T (uses θ).
2. Generate X given $T(X) = T$ (does not use θ).

4.2 Factorization Theorem

Theorem 4.1 (Factorization)**.** *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of distributions dominated by μ ($P_\theta \ll \mu$, $\forall\theta$). T is sufficient for $\mathcal{P}$ iff there exists functions $g_\theta, h \geq 0$ such that $p_\theta(x) = g_\theta(T(x))h(x)$ (for a.e. x under μ).*

"Proof" (rigorous proof in Keener 6.4). ($\Longleftarrow$)

$$p_\theta\big(x \bigm| T(x) = t\big) = \frac{\cancel{g_\theta(t)}h(x)\,\mathbb{1}\{T(x) = t\}}{\int_{\{T(s)=t\}} \cancel{g_\theta(t)}h(s)\,\mathrm{d}\mu(s)}.$$

($\Longrightarrow$) Take

$$\begin{aligned} g_\theta(t) &= P_\theta\big(T(X) = t\big) \\ &= \int_{\{T(x)=t\}} p_\theta(x)\,\mathrm{d}\mu(x), \\ h(x) &= \frac{p_\theta(x)}{\int_{\{T(s)=t\}} p_\theta(s)\,\mathrm{d}\mu(s)} = P_\theta\big(X = x \bigm| T(X) = T(x)\big). \end{aligned}$$

□

Example 4.2 (Exponential Families)**.**

$$p_\theta(x) = \underbrace{\mathrm{e}^{\eta(\theta)^\intercal T(x) - B(\theta)}}_{g_\theta(T(x))} h(x).$$

Example 4.3. If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} U[\theta, \theta+1]$, then the density is $p_\theta(x) = \mathbb{1}\{\theta \le x \le \theta+1\}$. So,

$$p_\theta(x) = \prod_{i=1}^n \mathbb{1}\{\theta \le x_i \le \theta+1\} = \mathbb{1}\{\theta \le x_{(1)}, x_{(n)} \le \theta+1\},$$

and $(X_{(1)}, X_{(n)})$ is sufficient.

Example 4.4. Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta^{(1)}$, where $\mathcal{P}^{(1)} = \{P_\theta^{(1)} : \theta \in \Theta\}$ is any univariate model on $\mathcal{X} \subseteq \mathbb{R}$. P_θ is invariant to permutations of the vector $X = (X_1, \ldots, X_n)$. Therefore, the order statistics $(X_{(1)}, \ldots, X_{(n)})$ (where $X_{(1)} \le \cdots \le X_{(n)}$) are sufficient. More generally, the **empirical distribution** $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is sufficient.

4.3 Minimal Sufficiency

When $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$, then $T(X) \sim \text{Binomial}(n, \theta)$ is sufficient.

Definition: $T(X)$ is **minimal sufficient** for $\mathcal{P}$ if

- $T(X)$ is sufficient,
- for any sufficient $S(X)$ there exists f with $T(X) = f(S(X))$ a.s. in $\mathcal{P}$.

Suppose S, T are both minimal. Then, $S(x) = f(T(x))$ and $T(x) = g(S(x))$, so they can be recovered from each other.

Theorem 4.5 (Keener 3.1)**.** *Assume $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ is a family of densities w.r.t. μ and $T(X)$ is sufficient. If $p_\theta(x) \propto_\theta p_\theta(y)$ implies $T(x) = T(y)$, then $T(X)$ is minimal sufficient. [The log-likelihood satisfies $\ell(\theta; x) = \ell(\theta; y) + $ constant, where $\ell(\theta; x) = \log p_\theta(x)$.]*

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

$$\begin{aligned} p_\theta(x) &= g_\theta\big(S(x)\big)h(x) \\ &\propto_\theta g_\theta\big(S(y)\big)h(y) \\ &= p_\theta(y) \end{aligned}$$

so $T(x) = T(y)$, which is a contradiction. □

Example 4.6. If $p_\theta(x) = \mathrm{e}^{\eta(\theta)^\mathsf{T} T(x) - B(\theta)} h(x)$, is $T(x)$ minimal? We want to show that if $p_\theta(x) \propto_\theta p_\theta(y)$, then $T(x) = T(y)$.

$$\begin{aligned} p_\theta(x) \propto_\theta p_\theta(y) &\iff \mathrm{e}^{\eta(\theta)^\mathsf{T} T(x)} \propto_\theta \mathrm{e}^{\eta(\theta)^\mathsf{T} T(y)} \\ &\iff \eta(\theta)^\mathsf{T} T(x) = \eta(\theta)^\mathsf{T} T(y) + \text{constant} \\ &\iff \big(\eta(\theta_1) - \eta(\theta_2)\big)^\mathsf{T}\big(T(x) - T(y)\big) = 0, \qquad \forall \theta_1, \theta_2 \\ &\iff T(x) - T(y) \perp \text{span}\{\eta(\theta_1) - \eta(\theta_2) : \theta_1, \theta_2 \in \Theta\}. \end{aligned}$$

So, if $\text{span}\{\eta(\theta_1) - \eta(\theta_2) : \theta_1, \theta_2 \in \Theta\} = \mathbb{R}^s$, then $T(X)$ is minimal.

Example 4.7. Suppose $X \sim \mathcal{N}_2(\mu(\theta), I_2)$. The density is $p_\theta(x) = \mathrm{e}^{\mu(\theta)^\top x - B(\theta)} \mathrm{e}^{-x^\top x/2}$. If $\Theta = \mathbb{R}$, $\mu(\theta) = a + b\theta$ for some $a, b \in \mathbb{R}^2$, then X is *not* minimal ($b^\top X$ is). If

$$\mu(\theta) = \begin{bmatrix} \theta \\ \theta^2 \end{bmatrix}$$

then X is minimal.

Example 4.8. Let

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta^{(1)}(x) = \frac{1}{2} \mathrm{e}^{-|x-\theta|}.$$

Then,

$$p_\theta(x) = \frac{1}{2^n} \exp\Big\{-\sum_{i=1}^n |x_i - \theta|\Big\},$$

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

The function $\ell(\theta; x)$ is piecewise linear with knots at the x_i. The maximum likelihood estimator is the median. When is $\ell(\theta; x) = \ell(\theta; y) + \text{constant}$? This occurs if and only if x and y have the same order statistics. Therefore, $(X_{(i)})_{i=1}^n$ is *minimal sufficient*.

4.4 Completeness

Definition 4.9. $T(X)$ is **complete** for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if $\mathbb{E}_\theta[f(T(x))] = 0 \ \forall \theta$ implies

$$f\big(T(X)\big) \overset{\text{a.s.}}{=} 0 \qquad \forall \theta.$$

Example 4.10. If $X_i \overset{\text{i.i.d.}}{\sim} U[0, \theta]$, where $\theta \in (0, \infty)$, one can show that $T(X) = X_{(n)}$ is minimal sufficient. The density of $T(X)$ with respect to $\lambda([0, \infty))$ is:

$$P_\theta(T \le t) = \Big(\frac{t}{\theta} \vee 1\Big)^n = \Big(\frac{t}{\theta}\Big)^n \vee 1,$$

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

Suppose

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

then

$$0 = \int_0^\infty f(t) t^{n-1} \, \mathrm{d}t$$

which implies $f(t) t^{n-1} = 0$ for a.e. $t > 0$, and so $f(T(X)) \overset{\text{a.s.}}{=} 0$.

Lecture 5

September 7

5.1 Completeness

Definition: $T(X)$ is complete for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if $\mathbb{E}_\theta[f(T)] = 0\ \forall\theta$ implies $f(T) \overset{\text{a.s.}}{=} 0\ \forall\theta$.

Definition 5.1. Let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ be an exponential family of densities (with respect to μ),

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

Assume WLOG that there does not exist $v \in \mathbb{R}^s$, $c \in \mathbb{R}$ with $v^\intercal T(X) \overset{\text{a.s.}}{=} c$, $\forall\theta$. If

$$\Xi = \eta(\Theta) = \{\eta(\theta) : \theta \in \Theta\}$$

contains an open set, we say that $\mathcal{P}$ is **full-rank**. Otherwise, $\mathcal{P}$ is **curved**.

Theorem 5.2. *If $\mathcal{P}$ is full-rank, then $T(X)$ is complete sufficient for $\mathcal{P}$.*

Proof. The proof is in Lehmann & Romano, Theorem 4.3.1. □

Example 5.3. If $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\eta = \begin{bmatrix} \mu/\sigma^2 \\ 1/(2\sigma^2) \end{bmatrix}, \qquad T(x) = \begin{bmatrix} x \\ -x^2 \end{bmatrix}.$$

X is complete sufficient. $T(X)$ is also complete sufficient because it can be computed from X.

Theorem 5.4. *If $T(X)$ is complete sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, then $T(X)$ is minimal sufficient.*

Proof. Assume $S(X)$ is minimal sufficient. Then, $S(X) \overset{\text{a.s.}}{=} f(T(X))$. Note that

$$\mu\big(S(X)\big) = \mathbb{E}_\theta[T(X) \mid S(X)]$$

does not depend on θ. Define $g(t) = t - \mu(f(t))$.

$$\begin{aligned} \mathbb{E}_\theta\big[g\big(T(X)\big)\big] &= \mathbb{E}_\theta[T(X)] - \mathbb{E}_\theta\big[\mu\big(S(X)\big)\big] \\ &= \mathbb{E}_\theta[T(X)] - \mathbb{E}_\theta\big[\mathbb{E}[T(X) \mid S(X)]\big] \end{aligned}$$

$$= 0 \qquad \forall\theta,$$

so $g(T(X)) \overset{\text{a.s.}}{=} 0\ \forall\theta$. Hence, $T(X) \overset{\text{a.s.}}{=} \mu(S(X))$. □

5.2 Ancillarity & Basu's Theorem

Definition 5.5. $V(X)$ is **ancillary** for $\mathcal{P}$ if its distribution does not depend on θ.

Theorem 5.6 (Basu)**.** *If $T(X)$ is complete sufficient, and $V(X)$ is ancillary for $\mathcal{P}$, then*

$$V(X) \perp\!\!\!\perp T(X) \qquad \forall\theta.$$

Proof. We want to show $P_\theta(V \in A, T \in B) = P_\theta(V \in A)P_\theta(T \in B)$. Let $q_A(T) = P_\theta(V \in A \mid T)$.

$$\mathbb{E}_\theta\big[q_A\big(T(X)\big) - p_A\big] = p_A - p_A = 0 \qquad (\forall\theta)$$

since $\mathbb{E}_\theta[P_\theta(V \in A \mid T)] = \mathbb{E}_\theta[\mathbb{E}_\theta[\mathbb{1}_A(V) \mid T]] = P_\theta(V \in A)$, so $q_A(T) \overset{\text{a.s.}}{=} p_A$.

$$\begin{aligned}
P_\theta(V \in A, T \in B) &= \mathbb{E}_\theta[\mathbb{1}_A(V)\,\mathbb{1}_B(T)] \\
&= \mathbb{E}_\theta\big[\mathbb{E}_\theta[\mathbb{1}_A(V)\,\mathbb{1}_B(T) \mid T]\big] \\
&= \mathbb{E}_\theta[\mathbb{1}_B(T) q_A(T)] \\
&= p_A\,\mathbb{E}_\theta[\mathbb{1}_B(T)] \\
&= P_\theta(V \in A)P_\theta(T \in B).
\end{aligned}$$ □

Remark: Ancillarity, completeness, and sufficiency are properties relative to a *family* $\mathcal{P}$. Independence is a property relative to a *distribution* P_θ.

Example 5.7. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Define

$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2.$$

In fact, $\overline{X} \perp\!\!\!\perp S^2$. Let $\mathcal{P}_{\sigma^2} = \{\mathcal{N}(\mu, \sigma^2)^n : \mu \in \mathbb{R}\}$, for $\sigma^2 > 0$ fixed. $\overline{X}$ is complete sufficient for $\mathcal{P}_{\sigma^2}$. S^2 is ancillary. Indeed, define $Y_i = X_i - \mu \sim \mathcal{N}(0, \sigma^2)$. Also, $X_i - \overline{X} = Y_i - \overline{Y}$, so

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$$

has a distribution which does not depend on μ.

5.3 Rao-Blackwell Theorem

5.3.1 Convex Loss Functions

Definition 5.8. f is **convex** if $f(\gamma x + (1-\gamma)y) \le \gamma f(x) + (1-\gamma)f(y)$ for all $\gamma \in (0,1)$ and all $x \ne y$, and f is **strictly convex** if the inequality is replaced with strict inequality.

Theorem 5.9 (Jensen)**.** *If f is convex, then $f(\mathbb{E}[X]) \le \mathbb{E}[f(X)]$. (If f is strictly convex, then the* *inequality is strict unless $X \overset{a.s.}{=} c$ for some c.)*

If $L(\theta, d)$ is convex (in the second argument), it penalizes us for adding extra noise to $\delta(X)$. Let $\tilde{\delta}(X) = \delta(X) + Y$, where Y is mean-zero noise ($Y \perp\!\!\!\perp X$).

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta\big[L\big(\theta, \mathbb{E}_\theta(\tilde{\delta} \mid \delta)\big)\big], \\ R(\theta, \tilde{\delta}) &= \mathbb{E}_\theta\big[\mathbb{E}_\theta\big(L(\theta, \tilde{\delta}) \bigm| \delta\big)\big] \end{aligned}$$

and $L(\theta, \mathbb{E}_\theta(\tilde{\delta} \mid \delta)) \overset{\text{a.s.}}{\leq} \mathbb{E}_\theta(L(\theta, \tilde{\delta}) \mid \delta)$, so $R(\theta, \delta) \leq R(\theta, \tilde{\delta})$.

5.3.2 Rao-Blackwell Theorem

Theorem 5.10 (Rao-Blackwell)**.** *Assume $T(X)$ is sufficient and $\delta(X)$ is an estimator. Let*

$$\bar{\delta}\big(T(X)\big) = \mathbb{E}\big(\delta(X) \bigm| T(X)\big).$$

If $L(\theta, \cdot)$ is convex, then $R(\theta, \bar{\delta}) \leq R(\theta, \delta)$. If $L(\theta, \cdot)$ is strictly convex, then $R(\theta, \bar{\delta}) < R(\theta, \delta)$ unless $\bar{\delta}(T(X)) \overset{a.s.}{=} \delta(X)$.

Proof.

$$\begin{aligned} R(\theta, \bar{\delta}) &= \mathbb{E}_\theta\big[L\big(\theta, \bar{\delta}(X)\big)\big] \\ &= \mathbb{E}_\theta\big[L\big(\theta, \mathbb{E}(\delta \mid T)\big)\big], \\ R(\theta, \delta) &= \mathbb{E}_\theta[L(\theta, \delta)] \\ &= \mathbb{E}_\theta\big[\mathbb{E}_\theta\big(L(\theta, \delta) \bigm| T\big)\big]. \end{aligned}$$

The result follows from $L(\theta, \mathbb{E}_\theta(\delta \mid T)) \leq \mathbb{E}_\theta(L(\theta, \delta) \mid T)$. □

5.4 Bias-Variance Decomposition

$$\begin{aligned} \text{MSE}(\theta, \delta) &= \mathbb{E}_\theta\big[\big(\delta(X) - g(\theta)\big)^2\big] \\ &= \mathbb{E}_\theta\big[\big(\delta(X) - \mathbb{E}_\theta[\delta(X)] + \mathbb{E}_\theta[\delta(X)] - g(\theta)\big)^2\big] \\ &= \mathbb{E}_\theta\big[\big(\delta(X) - \mathbb{E}_\theta[\delta(X)]\big)^2\big] + \mathbb{E}_\theta\big[\big(\mathbb{E}_\theta[\delta(X)] - g(\theta)\big)^2\big] \\ &\quad + 2\,\mathbb{E}_\theta\big[\underbrace{\big(\delta(X) - \mathbb{E}_\theta[\delta(X)]\big)}_{\text{mean zero}}\underbrace{\big(\mathbb{E}_\theta[\delta(X)] - g(\theta)\big)}_{\text{constant}}\big] \\ &= \underbrace{\text{var}_\theta\,\delta(X)}_{\text{variance}} + \underbrace{\big(\mathbb{E}_\theta[\delta(X)] - g(\theta)\big)^2}_{(\text{bias}_\theta\ \delta(X))^2} \end{aligned}$$

Lecture 6

September 12

6.1 UMVU Estimation

6.1.1 Bias-Variance Tradeoff

$$\begin{aligned}\mathrm{MSE}_\theta(\theta,\delta) &= \mathbb{E}_\theta\left[\big(g(\theta)-\delta(X)\big)^2\right] \\ &= \mathrm{var}_\theta\,\delta(X) + \big(\mathbb{E}_\theta[\delta(X)] - g(\theta)\big)^2.\end{aligned}$$

6.1.2 Unbiasedness

$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is the model. $\delta(X)$ is **unbiased** (for $g(\theta)$) if $\mathbb{E}_\theta[\delta(X)] = g(\theta)$, for all $\theta \in \Theta$.

Definition 6.1. $g(\theta)$ is **U-estimable** if there exists any unbiased estimator.

Example 6.2. Let $X \sim \mathrm{Bernoulli}(\theta)$. Then, $\mathbb{E}_\theta[\delta(X)] = \theta\delta(1) + (1-\theta)\delta(0)$. So, θ^2 is not U-estimable. Any function of θ which is U-estimable must be of the form $a\theta + b$.

Definition 6.3. We say that $\delta(X)$ is **uniformly minimum variance unbiased (UMVU)** if $\delta(X)$ is unbiased, and for any unbiased $\tilde{\delta}(X)$, $\mathrm{var}_\theta\,\tilde{\delta}(X) \geq \mathrm{var}_\theta\,\delta(X)$, for all $\theta \in \Theta$.

Theorem 6.4 (Theorem 4.4)**.** *Suppose T is complete sufficient and $g(\theta)$ is U-estimable. Then, there is a unique (up to almost sure equality) UMVU estimator of the form $\delta(T(X))$.*

Proof. Let $\delta_0(X)$ be unbiased and $\delta(T) = \mathbb{E}(\delta_0(X) \mid T)$.

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

so $\delta(T)$ is unbiased. If $\tilde{\delta}(T)$ is unbiased, then $\mathbb{E}_\theta[\delta(T) - \tilde{\delta}(T)] = 0$, for all θ, which implies $\delta(T) \overset{\text{a.s.}}{=} \tilde{\delta}(T)$ by completeness. Suppose $\delta^*(X)$ is unbiased. Then, $\delta(T) \overset{\text{a.s.}}{=} \mathbb{E}(\delta^*(X) \mid T)$, so $\mathrm{var}_\theta\,\delta^* \geq \mathrm{var}_\theta\,\delta$ for all θ (with strict inequality unless $\delta^* \overset{\text{a.s.}}{=} \delta$). □

6.1.3 Interpretation of 6.4

We have two ways to find UMVUE.

1. Find any unbiased $\delta(T)$ (when $T(X)$ is complete sufficient).

2. Find any unbiased $\delta_0(X)$, and then Rao-Blackwellize it.

Remark: Under the hypotheses of 6.4, the same proof works for *any* convex loss.

$\mathcal{P}$ describes a linear transformation from random variables to functions of θ:

$$f(X) \rightsquigarrow \int f(x)\,\mathrm{d}P_{\cdot}(x) = \mathbb{E}_{\cdot}[f(X)].$$

Then, completeness of X is equivalent to saying that this map is one-to-one. For $T(X)$, think of $\mathcal{P}^T$, where P_θ^T is the distribution of $T(X)$.

6.2 Examples

Example 6.5. Take $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$, $\theta \geq 0$, with density

$$p_\theta(x) = \frac{\theta^x \mathrm{e}^{-\theta}}{x!}$$

on $\mathcal{X} = \{0, 1, 2, \dots\}$. The complete sufficient statistic is $T(X) = \sum_{i=1}^n X_i \sim \text{Poisson}(n\theta)$.

$$p_\theta^T(t) = \frac{(n\theta)^t \mathrm{e}^{-n\theta}}{t!}.$$

Estimate $g(\theta) = \theta^2$.

$$\begin{aligned} \delta(T) \text{ unbiased } &\iff \sum_{t=0}^\infty \delta(t) p_\theta^T(t) = \theta^2 \\ &\iff \sum_{t=0}^\infty \delta(t) \frac{n^t}{t!} \theta^t = \theta^2 \mathrm{e}^{n\theta} = \sum_{k=0}^\infty \frac{n^k}{k!} \theta^{k+2}, \qquad \forall \theta > 0 \end{aligned}$$

Match terms: $\delta(0) = \delta(1) = 0$. For $t \geq 2$,

$$\delta(t) \frac{n^t}{t!} = \frac{n^{t-2}}{(t-2)!},$$

so

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

Example 6.6. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} U[0, \theta]$. $T = X_{(n)}$ is complete sufficient. Estimate $g(\theta) = \theta$.

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

Then,

$$\mathbb{E}_\theta[T] = \int_0^\theta t \frac{n}{\theta^n} t^{n-1}\,\mathrm{d}t = \frac{n}{n+1}\theta$$

so $(n+1)T/n$ is UMVU.

The sample mean is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

with

$$\mathbb{E}_\theta[\bar{X}] = \frac{\theta}{2}.$$

So, $2\bar{X}$ is unbiased. Also,

$$\begin{aligned}\mathbb{E}(2\bar{X} \mid T) &= \frac{2}{n}T + \mathbb{E}\Big[\frac{2}{n}\sum_{i=1}^{n-1} \underbrace{Y_i}_{\text{i.i.d. } U[0,T]}\Big] \\ &= \frac{2}{n}T + \frac{2(n-1)}{n}\Big(\frac{T}{2}\Big) \\ &= \frac{n+1}{n}T.\end{aligned}$$

Keener shows that as $n \to \infty$,

$$\begin{aligned}\operatorname{var}_\theta\Big(\frac{n+1}{n}T\Big) &\overset{n\to\infty}{\asymp} n^{-2}, \\ \operatorname{var}_\theta(2\bar{X}) &\overset{n\to\infty}{\asymp} n^{-1},\end{aligned}$$

where $f(n) \asymp g(n)$ means

$$0 < \liminf_{n\to\infty} \frac{f(n)}{g(n)} \le \limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty.$$

In the above example, $(n+1)T/n$ is inadmissible (with respect to MSE). $(n+2)T/(n+1)$ is better. It is well-known that $T \sim \text{Beta}(n, 1)$.

6.3 Log-Likelihood & the Score Function

Let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ be a family of densities with respect to μ, $\Theta \subseteq \mathbb{R}^d$. Assume the densities have a common support: $\{x : p_\theta(x) > 0\}$ is the same for all θ. Define the **log-likelihood function** $\ell(\theta; x) = \log p_\theta(x)$. The **score function** $\nabla_\theta \ell(\theta; x)$ plays a key role. Useful facts (assuming enough regularity): $1 = \int_{\mathcal{X}} e^{\ell(\theta;x)}\, d\mu(x)$, so by differentiating with respect to θ_j,

$$0 = \int \Big(\frac{\partial}{\partial\theta_j}\ell(\theta; x)\Big) e^{\ell(\theta;x)}\, d\mu(x),$$

so $\mathbb{E}_\theta[\nabla_\theta \ell(\theta; x)] = 0$. Differentiating with respect to θ_k,

$$\begin{aligned}0 &= \int \Big(\frac{\partial^2}{\partial\theta_j\partial\theta_k}\ell(\theta; x) + \frac{\partial}{\partial\theta_j}\ell(\theta; x)\frac{\partial}{\partial\theta_k}\ell(\theta; x)\Big) e^{\ell(\theta;x)}\, d\mu(x) \\ &= \mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\ell(\theta; X)\Big] + \underbrace{\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta_j}\ell(\theta; X)\frac{\partial}{\partial\theta_k}\ell(\theta; X)\Big]}_{\operatorname{cov}_\theta((\nabla_\theta\ell(\theta;X))_j,(\nabla_\theta\ell(\theta;X))_k)}\end{aligned}$$

so that

$$\begin{aligned}\operatorname{var}_\theta \nabla_\theta \ell(\theta; X) &= -\mathbb{E}_\theta[\nabla_\theta^2 \ell(\theta; X)] \\ &= J(\theta),\end{aligned}$$

the **Fisher information matrix**.

Lecture 7

September 14

7.1 Log-Likelihood & Score

The **log-likelihood** is $\ell(\theta; x) = \log p_\theta(x)$ (assume $p_\theta(x) > 0$). From $1 = \int \mathrm{e}^{\ell(\theta;x)} \, \mathrm{d}\mu(x)$, we obtain

$$\mathbb{E}_\theta[\underbrace{\nabla \ell(\theta; X)}_{\text{score}}] = 0,$$
$$J(\theta) = \operatorname{var}_\theta \nabla \ell(\theta; X) = -\mathbb{E}[\nabla^2 \ell(\theta; X)].$$

Remark. Recall that $(\ell(\theta; X) - \ell(\theta_0; X))_{\theta \in \Theta}$ is minimal sufficient for fixed θ_0. In a "local neighborhood" of θ_0, we can think of $\nabla \ell(\theta_0; X)$ as "approximately minimal sufficient" or "approximately complete". Consider the "local model" $\mathcal{P}_{\theta_0,\varepsilon} = \{P_{\theta_0+\eta} : \|\eta\| < \varepsilon\}$, then

$$\begin{aligned} p_{\theta_0+\eta}(x) &= \mathrm{e}^{\ell(\theta_0+\eta, x)} \\ &\approx \mathrm{e}^{\eta^\mathsf{T} \nabla \ell(\theta_0; x)} p_{\theta_0}(x). \end{aligned}$$

7.2 Cramèr-Rao Lower Bound

Suppose $\delta(X)$ is unbiased, $\delta(X) : \mathcal{X} \to \mathbb{R}$, for $g(\theta) = \int_{\mathcal{X}} \delta(x) \mathrm{e}^{\ell(\theta;x)} \, \mathrm{d}\mu(x)$.

$$\begin{aligned} \nabla g(\theta) &= \int \delta(x) \nabla \ell(\theta; x) \mathrm{e}^{\ell(\theta;x)} \, \mathrm{d}\mu(x) \\ &= \mathbb{E}_\theta[\delta(X) \nabla \ell(\theta; X)] \\ &= \operatorname{cov}_\theta\big(\delta(X), \nabla \ell(\theta; X)\big). \end{aligned}$$

Suppose $\theta \in \mathbb{R}$. We know $(\operatorname{var}_\theta \delta(X))(\operatorname{var}_\theta \ell'(\theta; x)) \geq \operatorname{cov}_\theta(\delta(X), \ell'(\theta; X))^2$, so

$$\begin{aligned} \operatorname{var}_\theta \delta(X) &\geq \frac{\operatorname{cov}_\theta(\delta(X), \ell'(\theta; X))^2}{J(\theta)} \\ &= \frac{g'(\theta)^2}{J(\theta)}. \end{aligned}$$

When $g(\theta) = a + b\theta$, then

$$\operatorname{var}_\theta \delta(X) \geq \frac{b^2}{J(\theta)}$$

which scales correctly. For the multiparameter case: $\operatorname{var}_\theta \delta(X) \geq (\nabla g(\theta))^\mathsf{T} J(\theta)^{-1} \nabla g(\theta)$.

Example 7.1. Suppose we have i.i.d. samples, $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta^{(1)}(x)$, $\theta \in \Theta$. So,

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

Then,

$$\begin{aligned}
\ell_1(\theta; x) &= \log p_\theta^{(1)}(x), \\
\ell(\theta; x) &= \log p_\theta(x) = \sum_{i=1}^n \ell_1(\theta; x_i), \\
J(\theta) &= \operatorname{var}_\theta \nabla \ell(\theta; X) \\
&= \operatorname{var}_\theta \Big(\sum_{i=1}^n \nabla \ell_1(\theta; X) \Big) \\
&= n \operatorname{var}_\theta \nabla \ell_1(\theta; X) \\
&= n J_1(\theta).
\end{aligned}$$

Thus the lower bound on the variance scales as n^{-1}. In the case of the uniform scale family with density

$$p_\theta(x) = \frac{1}{\theta^n} \mathbb{1}\{x^{(n)} \le \theta\},$$

the log-likelihood $\ell(\theta; x) = -n \log \theta - \infty \, \mathbb{1}\{\theta < x^{(n)}\}$ does not possess sufficient regularity properties to apply the bound.

7.2.1 Efficiency

We say $\delta(X)$ is **efficient** if $\operatorname{var}_\theta \delta(X)$ equals the Cramèr-Rao lower bound (CRLB) (or 70% efficient if $\text{CRLB}/(\operatorname{var}_\theta \delta(X)) = 0.7$). Note that the efficiency is fully determined by the correlation

$$\frac{\text{CRLB}}{\operatorname{var}_\theta \delta(X)} = \operatorname{corr}_\theta \big(\delta(X), \ell'(\theta; X) \big)^2.$$

7.2.2 Exponential Families

We have

$$\begin{aligned}
p_\eta(x) &= \mathrm{e}^{\eta^\mathsf{T} T(x) - A(\eta)} h(x), \\
\ell(\eta; x) &= \eta^\mathsf{T} T(x) - A(\eta) + \log h(x), \\
\nabla \ell(\eta; x) &= T(x) - \mathbb{E}_\eta[T(X)], \\
\operatorname{var}_\eta \nabla \ell(\eta; x) &= \operatorname{var}_\eta T(X) = \nabla^2 A(\eta) = J(\eta).
\end{aligned}$$

7.3 Hammersley-Chapman-Robbins Inequality

$$\begin{aligned}
\frac{p_{\theta+\varepsilon}(x)}{p_\theta(x)} - 1 &= \mathrm{e}^{\ell(\theta+\varepsilon; x) - \ell(\theta; x)} - 1 \\
&\approx \varepsilon^\mathsf{T} \nabla \ell(\theta; x), \\
\mathbb{E}_\theta \Big[\frac{p_{\theta+\varepsilon}(X)}{p_\theta(X)} - 1 \Big] &= \int \Big(\frac{p_{\theta+\varepsilon}}{p_\theta} - 1 \Big) p_\theta \, \mathrm{d}\mu \\
&= 1 - 1 = 0,
\end{aligned}$$

$$\begin{aligned}\operatorname{cov}_\theta\Big(\frac{p_{\theta+\varepsilon}(X)}{p_\theta(X)} - 1, \delta(X)\Big) &= \int \delta\Big(\frac{p_{\theta+\varepsilon}}{p_\theta} - 1\Big) p_\theta \, \mathrm{d}\mu \\ &= \mathbb{E}_{\theta+\varepsilon}[\delta(X)] - \mathbb{E}_\theta[\delta(X)] \\ &= g(\theta+\varepsilon) - g(\theta), \\ \operatorname{var}_\theta \delta(X) &\geq \sup_\varepsilon \frac{(g(\theta+\varepsilon) - g(\theta))^2}{\mathbb{E}_\theta[(p_{\theta+\varepsilon}(X)/p_\theta(X) - 1)^2]}.\end{aligned}$$

Example 7.2 (Curved Exponential Family)**.** For $\theta \in \mathbb{R}$, let $\eta(\theta) \in \mathbb{R}^s$ for $s > 1$.

$$\begin{aligned}p_\theta(x) &= \mathrm{e}^{\eta(\theta)^\mathsf{T} T(x) - B(\theta)} h(x), \\ \nabla \ell(\theta; x) &= \nabla \eta(\theta)^\mathsf{T} T(x) - \nabla B(\theta) \\ &= \nabla \eta(\theta)^\mathsf{T} \big\{T(x) - \nabla A\big(\eta(\theta)\big)\big\} \\ &= \nabla \eta(\theta)^\mathsf{T} (T - \mathbb{E}_\theta[T]).\end{aligned}$$

Example 7.3 (Keener, Example 4.7)**.** Let $X \sim \operatorname{Poisson}(\theta)$ truncated to $\{1, 2, 3, \dots\}$.

$$p_\theta(x) = \frac{\theta^x \mathrm{e}^{-\theta}}{x!(1 - \mathrm{e}^{-\theta})}, \qquad x = 1, 2, 3, \dots$$

Estimate $g(\theta) = \mathrm{e}^{-\theta}$.

$$\mathbb{E}_\theta[\delta(X)] = \sum_{x=1}^\infty \frac{\theta^x \mathrm{e}^{-\theta}}{x!(1 - \mathrm{e}^{-\theta})} \delta(x) = \mathrm{e}^{-\theta},$$

so

$$\begin{aligned}\sum_{x=1}^\infty \frac{\theta^x}{x!} \delta(x) &= 1 - \mathrm{e}^{-\theta} \\ &= 1 - \sum_{k=0}^\infty \frac{(-\theta)^k}{k!} \\ &= \sum_{x=1}^\infty \frac{-(-\theta)^x}{x!}\end{aligned}$$

and therefore:

$$\delta(X) = (-1)^{X+1} = \begin{cases} 1, & X \text{ odd} \\ -1, & X \text{ even} \end{cases}$$

The only unbiased estimator is stupid!

Lecture 8

September 19

8.1 Variance Bounds

Suppose $X \sim \mathcal{N}_2(\mu(\theta), I_2)$, $\mu(\theta) = (\theta, C\sin(\theta/\pi))$, for $\theta \in \mathbb{R}$. Estimate $g(\theta) = \theta$. Then, $\delta(X) = X_1$ is unbiased, $\operatorname{var}_\theta \delta(X) = 1$ for all θ.

CRLB:

$$\begin{aligned}
\ell(\theta; x) &= -\frac{1}{2}\|\mu(\theta) - x\|^2 + \text{constant}, \\
\frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta; x) &= (x - \mu)^\mathsf{T} \nabla\mu(\theta) \\
&= x_1 - \theta + \frac{C}{\pi}\cos\Big(\frac{\theta}{\pi}\Big)\Big(x_2 - C\sin\frac{\theta}{\pi}\Big), \\
J(0) &= 1 + \frac{C^2}{\pi^2}, \\
\operatorname{var}_0 \delta(X) &\ge \frac{1}{1 + C^2/\pi^2}.
\end{aligned}$$

If $C = 0$, then the bound is 1. If $C \to \infty$, then the bound goes to 0.

HCR: For $\theta = 0$, $\varepsilon = 1$,

$$\begin{aligned}
\frac{p_{\theta+\varepsilon}(x)}{p_\theta(x)} &= \mathrm{e}^{x_1 - 1/2}, \\
\operatorname{var}_\theta \delta(X) &\ge \frac{(g(1) - g(0))^2}{\mathbb{E}_\theta[(\mathrm{e}^{X_1 - 1/2} - 1)^2]} \\
&= \frac{1}{\mathrm{e}^1 - 1} \approx 0.58.
\end{aligned}$$

8.2 Bayes Risk, Bayes Estimator

8.2.1 Frequentist Motivation

The model is $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ for the data $X \in \mathcal{X}$. We have a loss $L(\theta, d)$ and thus a risk $R(\theta, \delta)$.

Bayes Risk: Let Λ be a probability measure, i.e., $\Lambda(\Omega) = 1$.

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

$\delta_\Lambda(X)$ is the **Bayes estimator** (for Λ) if it minimizes $R_{\text{Bayes}}(\Lambda, \delta)$.

8.2.2 Bayes Estimator

Theorem 8.1. *Suppose $\Theta \sim \Lambda$ and $X \mid \Theta = \theta \sim P_\theta$. Also, $L(\theta, d) \geq 0$ for all θ, d. If*

$$\mathbb{E}\big[L\big(\Theta, \delta_0(X)\big)\big] < \infty$$

for some δ_0, and $\delta_\Lambda(x)$ minimizes $\mathbb{E}[L(\Theta, d) \mid X = x]$, $\mathcal{P}$-a.e., then δ_Λ is Bayes for Λ.

In this setting,

$$R(\theta, \delta) = \mathbb{E}\big[L\big(\Theta, \delta(X)\big) \bigm| \Theta = \theta\big].$$

Proof of 8.1. Let $\delta(X)$ be another estimator.

$$\begin{aligned} R_{\text{Bayes}}(\Lambda, \delta) &= \mathbb{E}\big[L\big(\Theta, \delta(X)\big)\big] \\ &= \mathbb{E}\Big[\mathbb{E}\big[L\big(\Theta, \delta(X)\big) \bigm| X = x\big]\Big] \\ &\geq \mathbb{E}\Big[\mathbb{E}\big[L\big(\Theta, \delta_\Lambda(X)\big) \bigm| X = x\big]\Big] \\ &= R_{\text{Bayes}}(\Lambda, \delta_\Lambda). \end{aligned}$$

□

Usual Interpretation: Λ is the "prior belief" about θ before seeing data. The posterior (distribution of Θ given X) is the belief after seeing data.

In terms of densities, $\lambda(\theta)$ is the prior density and $p_\theta(x)$ is the likelihood. The posterior density is

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

and $q(x) = \int_\Omega \lambda(\theta) p_\theta(x)\,\mathrm{d}\theta$ is the marginal density of x. δ_Λ minimizes $\int_\Omega L(\theta, d)\lambda(\theta \mid x)\,\mathrm{d}\theta$ for the observed x.

8.2.3 Posterior Mean

If $L(\theta, d) = (g(\theta) - d)^2$, then the Bayes estimator is the posterior mean. We want to minimize

$$\begin{aligned} \int_\Omega \big(g(\theta) - d\big)^2 \lambda(\theta \mid x)\,\mathrm{d}\theta &= \mathbb{E}\big[\big(g(\Theta) - d\big)^2 \bigm| X = x\big] \\ &= \operatorname{var}\big(g(\Theta) \bigm| X = x\big) + (d - \mathbb{E}[g(\Theta) \mid X = x])^2, \end{aligned}$$

so $\delta_\Lambda(x) = \mathbb{E}[g(\Theta) \mid X = x]$. More generally, suppose $L(\theta, d) = w(\theta)(g(\theta) - d)^2$ (for example, we might want to minimize

$$L(\theta, d) = \Big(\frac{\theta - d}{\theta}\Big)^2,$$

the relative error). Then, the Bayes estimator is

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

8.3 Examples

Example 8.2 (Beta-Binomial). Let $X \mid \Theta = \theta \sim \text{Binomial}(n, \theta)$, with likelihood

$$p_\theta(x) = \theta^x (1-\theta)^{n-x} \binom{n}{x}$$

for $x = 0, \dots, n$, and $\Theta \sim \text{Beta}(\alpha, \beta)$, with prior density

$$\lambda(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The posterior density is, for $\theta \in [0,1]$,

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

and so $\Theta \mid X = x \sim \text{Beta}(x+\alpha, n-x+\beta)$. Thus,

$$\begin{aligned} \mathbb{E}(\Theta \mid X) &= \frac{X+\alpha}{n+\alpha+\beta} \\ &= \frac{X}{n} \cdot \frac{n}{n+\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\Big(1 - \frac{n}{n+\alpha+\beta}\Big). \end{aligned}$$

Interpretation: We have $\alpha + \beta$ “pseudo-trials” with α successes.

Example 8.3 (Normal Mean). Let $X \mid \Theta = \theta \sim \mathcal{N}(\theta, \sigma^2/n)$. The likelihood is

$$p_\theta(x) \propto_\theta \mathrm{e}^{-n(x-\theta)^2/(2\sigma^2)}.$$

Also, $\Theta \sim \mathcal{N}(\mu, \tau^2)$ with prior

$$\lambda(\theta) \propto_\theta \mathrm{e}^{-(\theta-\mu)^2/(2\tau^2)}.$$

So,

$$\begin{aligned} \lambda(\theta \mid x) &\propto_\theta \exp\Big\{-\frac{n(x-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu)^2}{2\tau^2}\Big\} \\ &\propto_\theta \exp\Big\{\frac{nx\theta}{\sigma^2} + \frac{\mu\theta}{\tau^2} - \frac{n\theta^2}{2\sigma^2} - \frac{\theta^2}{2\tau^2}\Big\} \\ &= \exp\Big\{\theta\Big(\frac{nx}{\sigma^2} + \frac{\mu}{\tau^2}\Big) - \frac{\theta^2}{2/(n/\sigma^2 + 1/\tau^2)}\Big\} \\ &\propto_\theta \exp\Big\{-\frac{((nx\tau^2 + \mu\sigma^2)/(n\tau^2+\sigma^2) - \theta)^2}{2\sigma^2\tau^2/(\sigma^2 + n\tau^2)}\Big\} \end{aligned}$$

and so

$$\begin{aligned} \Theta \mid X &\sim \mathcal{N}\Big(\frac{nx\tau^2 + \mu\sigma^2}{n\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}\Big), \\ \mathbb{E}(\Theta \mid X) &= \frac{nX\tau^2 + \mu\sigma^2}{n\tau^2 + \sigma^2} = X \cdot \frac{n\tau^2}{\sigma^2+n\tau^2} + \mu \cdot \Big(1 - \frac{n\tau^2}{\sigma^2 + n\tau^2}\Big). \end{aligned}$$

Suppose $\tau^2 = \sigma^2/m$.

$$X \cdot \left(\frac{n\sigma^2/m}{\sigma^2 + n\sigma^2/m}\right) + \mu \cdot \left(1 - \frac{n\sigma^2/m}{\sigma^2 + n\sigma^2/m}\right) = X \cdot \left(\frac{n}{m+n}\right) + \mu \cdot \left(\frac{m}{n+m}\right).$$

Lecture 9

September 21

9.1 Properties of Bayes Estimators

9.1.1 Bayes & Bias

Theorem 9.1. *The posterior mean is biased unless* $\delta_\Lambda(X) \overset{a.s.}{=} g(\Theta)$.

Proof. Assume $\delta_\Lambda(X)$ is unbiased.

$$\begin{aligned} \delta_\Lambda(X) &= \mathbb{E}\big(g(\Theta) \mid X\big) \\ g(\Theta) &= \mathbb{E}\big(\delta_\Lambda(X) \mid \Theta\big) \end{aligned}$$

Condition on X:

$$\begin{aligned} \mathbb{E}\Big[\mathbb{E}\big(\delta_\Lambda(X) g(\Theta) \bigm| X\big)\Big] &= \mathbb{E}\Big[\delta_\Lambda(X)\,\mathbb{E}\big(g(\Theta) \bigm| X\big)\Big] \\ &= \mathbb{E}[\delta_\Lambda(X)^2]. \end{aligned}$$

Condition on Θ:

$$\mathbb{E}\Big[\mathbb{E}\big(\delta_\Lambda(X) g(\Theta) \bigm| \Theta\big)\Big] = \mathbb{E}[g(\Theta)^2].$$

So,

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

□

9.2 Conjugate Priors

If the posterior is from the same family as the prior, we say that the prior is **conjugate**.

Suppose that $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\eta(x) = \mathrm{e}^{\eta^\intercal T(x) - A(\eta)} h(x)$.

Prior: $\lambda_{k,\mu}(\eta) = \mathrm{e}^{k\mu^\intercal \eta - kA(\eta) - B(k,m)} \lambda_0(\eta)$. The sufficient statistic is

$$\begin{bmatrix} \eta \\ A(\eta) \end{bmatrix} \in \mathbb{R}^{s+1}$$

and the natural parameter is $\begin{bmatrix} k\mu \\ k \end{bmatrix}$. Then,

$$\begin{aligned}
\lambda(\eta \mid X_1, \ldots, X_n) &\propto_\eta \left(\prod_{i=1}^n \mathrm{e}^{\eta^\mathsf{T} T(x_i) - A(\eta)}\right) \mathrm{e}^{k\mu^\mathsf{T}\eta - kA(\eta)} \lambda_0(\eta) \\
&= \exp\left\{\left(k\mu + \sum_{i=1}^n T(x_i)\right)^\mathsf{T} \eta - (k+n)A(\eta)\right\} \lambda_0(\eta) \\
&= \lambda_{k+n, \mu \cdot k/(k+n) + \bar{T} \cdot n/(k+n)}(\eta).
\end{aligned}$$

So, starting with prior λ_0 and data

$$\underbrace{\mu, \ldots, \mu}_{k}, \underbrace{T(x_1), \ldots, T(x_n)}_{n}$$

is equivalent to starting with the prior $\lambda_{k,\mu}$ and data $T(x_1), \ldots, T(x_n)$. They both yield the posterior $\lambda_{k+\mu, \mu \cdot k/(k+n) + \bar{T} \cdot n/(k+n)}$.

Likelihood	Prior
$X_i \sim \text{Binomial}(n, \theta)$ $= \theta^x (1-\theta)^{n-x} \binom{n}{x}$	$\Theta \sim \text{Beta}(\alpha, \beta)$ $= \theta^{\alpha-1}(1-\theta)^{\beta-1} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, $k = \alpha + \beta$, $\mu = \frac{\alpha}{\alpha+\beta}$
$X_i \sim \mathcal{N}(\theta, \sigma^2)$ ($\sigma^2 > 0$ known) $= \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(\theta-x)^2/(2\sigma^2)}$	$\Theta \sim \mathcal{N}(\mu, \tau^2)$ $= \frac{1}{\sqrt{2\pi\tau^2}} \mathrm{e}^{-(\theta-\mu)^2/(2\tau^2)}$
$X_i \sim \text{Poisson}(\theta)$ $= \frac{\theta^x \mathrm{e}^{-\theta}}{x!}$	$\Theta \sim \text{Gamma}(k, \sigma)$ $= \frac{1}{\Gamma(k)\sigma^k} \theta^{k-1} \mathrm{e}^{-\theta/\sigma}$

For the gamma prior,

$$\begin{aligned}
\lambda(\theta \mid x) &\propto_\theta \theta^{\sum_{i=1}^n x_i} \mathrm{e}^{-n\theta} \theta^{k-1} \mathrm{e}^{-\theta/\sigma} \\
&= \theta^{k + \sum_{i=1}^n x_i - 1} \mathrm{e}^{-(n+1/\sigma)\theta} \\
&\propto_\theta \text{Gamma}\left(k + \sum_{i=1}^n x_i, \frac{1}{n+1/\sigma}\right).
\end{aligned}$$

Here, $k\sigma =$ "μ" and $1/\sigma =$ "k".

9.3 Where Does the Prior Come From?

1. Previous experience
2. Subjective beliefs
3. Convenience prior

9.3.1 "Objective" Priors

Say $X \sim \mathcal{N}(\theta, 1)$. We could take $\lambda(\theta) = 1$. The problem is that $\Lambda(\mathbb{R}) = \infty$, but this is the limit of $\mathcal{N}(0, \tau^2)$ as $\tau^2 \to \infty$. This is called a "**flat prior**".

A flat prior is not invariant to reparameterization. If $X \sim \text{Binomial}(n, \theta)$, $\theta \sim U[0, 1]$, and we change to the natural parameter

$$\eta = \log \frac{\theta}{1 - \theta}$$

then the flat prior is no longer flat.

The Jeffreys proposed fix is to take $\lambda(\theta) \propto_\theta |J(\theta)|^{1/2}$. For the binomial case, the Jeffreys prior then becomes $\text{Beta}(1/2, 1/2) \propto_\theta \theta^{-1/2}(1 - \theta)^{-1/2}$.

9.3.2 Hierarchichal Priors

In some situations, we want to pool information across multiple "replicates".

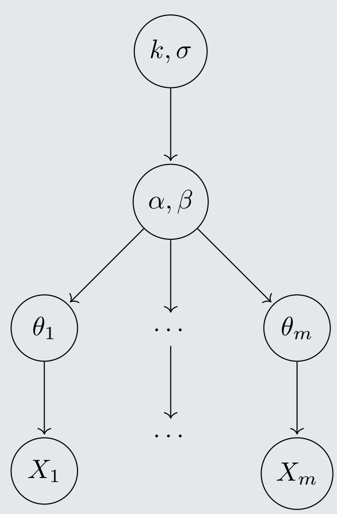
Example 9.2. Predict a batter's batting average after seeing n at-bats. For $i = 1, \ldots, m$, n_i is the number of at-bats, X_i is the number of hits, $X_i \sim \text{Binomial}(n_i, \theta_i)$.

Prior information: We expect performance to "mean-revert".

Bayes: Use a prior $\text{Beta}(\alpha, \beta)$. We want to learn α, β by looking at all players together. So we will use a hierarchichal model $\alpha, \beta \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(k, \sigma)$, $\theta_i \mid \alpha, \beta \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, \beta)$, and

$$X_i \mid \theta_i \stackrel{\text{independent}}{\sim} \text{Binomial}(n_i, \theta_i).$$

This can be represented as a **directed graphical model**.



Lecture 10

September 26

10.1 Normal Means Model

Let $X_i \overset{\text{independent}}{\sim} \mathcal{N}(\mu_i, 1)$ for $i = 1, \ldots, d$. Equivalently, let $X \sim \mathcal{N}_d(\mu, I_d)$ for $\mu \in \mathbb{R}^d$. The natural choice for a prior is the flat prior on μ, which yields the estimator $\delta(X) = X$ for μ.

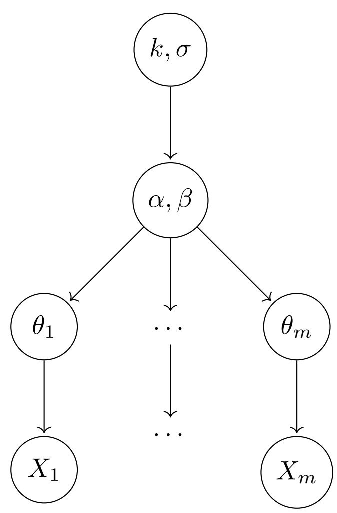
What is the prior for $\rho = \|\mu\|_2 = \sqrt{\sum_{i=1}^d \mu_i^2}$?

$$\begin{aligned}\mathbb{P}(\rho \in [r, r+\varepsilon]) &= \text{vol(shell with radius } r, \text{width } \varepsilon) \\ &\overset{\varepsilon \to 0}{\propto} \rho^{d-1}.\end{aligned}$$

The prior is not agnostic. The estimator then becomes $\mathbb{E}(\rho^2 \mid X) = \|X\|^2 + d$. The UMVU estimator is $\hat{\rho} = \|X\|^2 - d$.

10.2 Hierarchichal Bayes

Directed Graphical Model:



We can factorize the likelihood as

$$p(\alpha, \beta, \theta_1, \ldots, \theta_m, x_1, \ldots, x_m \mid k, \sigma) = p(\alpha, \beta \mid k, \sigma) \prod_{i=1}^m p(\theta_i \mid \alpha, \beta) p(x_i \mid \theta_i).$$

Generically,

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

and the denominator is frequently intractable.

10.3 Markov Chain Monte Carlo (MCMC)/Gibbs Sampler

Definition 10.1. A **(stationary) Markov chain** with transition kernel Q and initial distribution π_0 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 draw a directed graph:

$$X^{(0)} \longrightarrow X^{(1)} \longrightarrow X^{(2)} \longrightarrow X^{(3)} \longrightarrow \cdots$$

so that $\mathbb{P}(X^{(0)} = x^{(0)}, \ldots, X^{(t)} = x^{(t)}) = \pi_0(x^{(0)}) \prod_{i=1}^t Q(x^{(i)} \mid x^{(i-1)})$.

If $\pi(y) = \int_{\mathcal{X}} Q(y \mid x)\pi(x)\,\mathrm{d}x$, we say π is a **stationary distribution** for Q. Under mild conditions, $X^{(t)} \approx \pi$ for "large" t. If $\mathcal{X}$ is finite, then $\pi = \pi Q$, or equivalently, $\pi(Q - I) = 0$, and so convergence says that $\tilde{\pi} Q^t \to \pi$. A sufficient condition for π to be stationary is **detailed balance**: $\pi(x)Q(y \mid x) = \pi(y)Q(x \mid y)\ \forall x, y$.

10.3.1 MCMC

Strategy: Set up a Q for which $\lambda(\theta \mid x)$ is stationary. Start with $\Theta^{(0)} \sim \pi_0$ and run the Markov chain on a computer to $\Theta^{(t)}$. Treat $\Theta^{(t)}$ as a sample from $\lambda(\theta \mid x)$.

Algorithm:

1. Sample $\Theta \sim \pi_0$.
2. For $t = 1, \ldots, B$:
   (a) Sample $\Theta \sim Q(\cdot \mid \Theta)$.
3. Save $\hat{\Theta}^{(1)} \leftarrow \Theta$.
4. For $j = 1, \ldots, m$:
   (a) For $t = 1, \ldots, T$:
      i. Sample $\Theta \sim Q(\cdot \mid \Theta)$.
   (b) Save $\hat{\Theta}^{(m+1)} \leftarrow \Theta$.

10.3.2 Gibbs Sampler

Let $\theta = (\theta_1, \ldots, \theta_d)$ be a parameter vector.

Update Rule. Given $\Theta^{(t-1)}$:

- Sample $\Theta_1^{(t)} \sim \lambda(\theta_1 \mid \Theta_2^{(t-1)}, \ldots, \Theta_d^{(t-1)}, X)$.
- Sample $\Theta_2^{(t)} \sim \lambda(\theta_2 \mid \Theta_1^{(t)}, \Theta_3^{(t-1)}, \ldots, \Theta_d^{(t-1)}, X)$.
- $\vdots$

- Sample $\Theta_d^{(t)} \sim \lambda(\theta_d \mid \Theta_1^{(t)}, \ldots, \Theta_{d-1}^{(t)}, X)$.

The following example exhibits slow mixing.

$$\Theta \sim \frac{1}{2}\mathcal{N}_2(0, I_2) + \frac{1}{2}\mathcal{N}_2(\mu, I_2),$$

where

$$\mu = \begin{bmatrix} 10 \\ 0 \end{bmatrix}.$$

This particular example can be fixed by choosing a different basis.

For a hierarchichal Bayes model:

$$\begin{array}{c} \text{Parents of } \theta_j \\ \big\downarrow \\ \theta_j \\ \big\downarrow \\ \text{Children of } \theta_j \end{array}$$

The parents of θ_j are "fixed hyperparameters" and the children of θ_j are "fixed observed data".

Lecture 11

September 28

11.1 Empirical Bayes

11.1.1 Normal Means Model

Hierarchical Bayes: Let $\tau^2 \sim \lambda(\tau)$ (e.g., $1/\tau^2 \sim \text{Gamma}(k, \sigma)$), $\theta_i \mid \tau^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2)$ for $i = 1, \ldots, n$, and $X_i \mid \tau, \theta_i \overset{\text{independent}}{\sim} \mathcal{N}(\theta_i, 1)$.

Bayesian posterior mean:

$$\begin{aligned}
\delta_i(X) &= \mathbb{E}(\theta_i \mid X) \\
&= \mathbb{E}\big(\mathbb{E}(\theta_i \mid \tau^2, X) \mid X\big) \\
&= \mathbb{E}\Big(\frac{\tau^2}{1+\tau^2} X_i \,\Big|\, X\Big) \\
&= \mathbb{E}\Big(\frac{\tau^2}{1+\tau^2} \,\Big|\, X\Big) X_i.
\end{aligned}$$

Define

$$\zeta = \frac{1}{1+\tau^2}.$$

Since $X_i \mid \tau^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1+\tau^2)$, $X \mid \zeta \sim \mathcal{N}_n(0, \zeta^{-1} I_n)$ with likelihood

$$\Big(\frac{\zeta}{2\pi}\Big)^{n/2} \mathrm{e}^{-\zeta \|X\|^2/2} \propto_\zeta \text{Gamma}\Big(1 + \frac{n}{2}, \frac{2}{\|X\|^2}\Big).$$

For large n, $\|X\|^2$,

$$\begin{aligned}
\mathbb{E}\Big[\text{Gamma}\Big(1 + \frac{n}{2}, \frac{2}{\|X\|^2}\Big)\Big] &= \frac{2+n}{\|X\|^2} \approx \Big(\frac{1}{n}\sum_{i=1}^n X_i\Big)^{-1} \approx \zeta, \\
\operatorname{var} \text{Gamma}\Big(1 + \frac{n}{2}, \frac{2}{\|X\|^2}\Big) &= \Big(1 + \frac{n}{2}\Big) \frac{4}{\|X\|^4} \\
&\approx n^{-1}\Big(\frac{1}{n}\sum_{i=1}^n X_i^2\Big)^{-2} \to 0.
\end{aligned}$$

The likelihood is concentrated near $\approx \zeta$, so for almost any "open-minded" prior,

$$\delta_i(X) \approx \Big(1 - \frac{2+n}{\|X\|^2}\Big) X_i$$

$$\approx (1-\zeta)X_i.$$

This provides motivation for **empirical Bayes**.

Empirical Bayes: James-Stein propose (for $n \geq 3$)

$$\delta_i^{\text{JS}}(X) = \Big(1 - \frac{n-2}{\|X\|^2}\Big)X_i.$$

Proposition 11.1. *If $Y \sim \chi_n^2$ for $n \geq 3$, then $\mathbb{E}[Y^{-1}] = (n-2)^{-1}$.*

Proof.

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

since $\Gamma(x+1) = x\Gamma(x)$ (so $\Gamma(n) = (n-1)!$). □

We know that

$$\frac{\|X\|^2}{1+\tau^2} \sim \chi_n^2$$

so that

$$\begin{aligned}
\mathbb{E}\Big[1 - \frac{n-2}{\|X\|^2}\Big] &= 1 - \frac{1}{1+\tau^2} \\
&= 1 - \zeta.
\end{aligned}$$

11.2 Stein's Lemma/SURE

11.2.1 Stein's Lemma

Lemma 11.2 (Stein's Lemma (Univariate))**.** *Suppose $X \sim \mathcal{N}(\theta, \sigma^2)$. Let $h : \mathbb{R} \to \mathbb{R}$ be differentiable and $\mathbb{E}[|h'(X)|] < \infty$. Then, $\mathbb{E}[(X-\theta)h(X)] = \operatorname{cov}(X, h(X)) = \sigma^2\,\mathbb{E}[h'(X)]$.*

Proof. First, assume $\theta = 0$, $\sigma^2 = 1$. Also assume WLOG that $h(0) = 0$.

$$\begin{aligned}
\int_0^\infty xh(x)\phi(x)\,\mathrm{d}x &= \int_0^\infty x\Big[\int_0^x h'(y)\,\mathrm{d}y\Big]\phi(x)\,\mathrm{d}x \\
&= \int_0^\infty\int_0^\infty \mathbb{1}\{y < x\}xh'(y)\phi(x)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^\infty h'(y)\Big[\int_y^\infty x\phi(x)\,\mathrm{d}x\Big]\,\mathrm{d}y.
\end{aligned}$$

It is a nice fact that

$$\frac{\mathrm{d}}{\mathrm{d}x}\phi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2} = -x\phi(x).$$

So,

$$\int_0^\infty xh(x)\phi(x)\,\mathrm{d}x = \int_0^\infty h'(y)\phi(y)\,\mathrm{d}y.$$

A similar argument gives

$$\int_{-\infty}^0 xh(x)\phi(x)\,\mathrm{d}x = \int_{-\infty}^0 h'(x)\phi(x)\,\mathrm{d}x.$$

This gives the result for $\theta = 0$, $\sigma^2 = 1$. For general θ, σ^2, write $X = \theta + \sigma Z$ where $Z \sim \mathcal{N}(0,1)$.

$$\begin{aligned}\mathbb{E}[(X-\theta)h(X)] &= \sigma\,\mathbb{E}[Z\,\underbrace{h(\theta+\sigma Z)}_{g(Z)}] \\ &= \sigma\,\mathbb{E}[g'(Z)] \\ &= \sigma^2\,\mathbb{E}[h'(\theta+\sigma Z)].\end{aligned}$$

□

Definition 11.3. Let $h : \mathbb{R}^d \to \mathbb{R}^d$. Then $Dh \in \mathbb{R}^{d\times d}$ is the matrix with

$$\big(Dh(x)\big)_{i,j} = \frac{\partial h_i}{\partial x_j}(x).$$

Lemma 11.4 (Stein's Lemma (Multivariate))**.** *Let $X \sim \mathcal{N}_d(\theta, \sigma^2 I_d)$, $\theta \in \mathbb{R}^d$, and let $h : \mathbb{R}^d \to \mathbb{R}^d$. If $\mathbb{E}[\|Dh(X)\|_{\mathrm{F}}] = \mathbb{E}[(\sum_{i,j=1}^n Dh(X)_{i,j}^2)^{1/2}] < \infty$, then $\mathbb{E}[(X-\theta)^\mathsf{T} h(X)] = \sigma^2\,\mathbb{E}[\operatorname{tr} Dh(X)]$.*

Proof.

$$\begin{aligned}\mathbb{E}[(X_i-\theta_i)h_i(X)] &= \mathbb{E}\big[\mathbb{E}\big((X_i-\theta_i)h_i(X)\;\big|\;\underbrace{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_d}_{X_{-i}}\big)\big] \\ &= \mathbb{E}\Big[\sigma^2\,\mathbb{E}\Big(\frac{\partial h_i}{\partial x_i}(X)\;\Big|\;X_{-i}\Big)\Big] \\ &= \sigma^2\,\mathbb{E}[Dh(X)_{i,i}].\end{aligned}$$

□

11.2.2 Stein's Unbiased Risk Estimator

We can estimate the MSE for $\delta(X)$ by plugging in $h(X) = X - \delta(X)$. Assume $\sigma^2 = 1$. Then

$$\hat{R} = d + \|h(X)\|^2 - 2\operatorname{tr} Dh(X).$$

So,

$$\begin{aligned}R(\theta,\delta) &= \mathbb{E}_\theta[\|X-\theta-h(X)\|^2] \\ &= \mathbb{E}_\theta[\|X-\theta\|^2] + \mathbb{E}_\theta[\|h(X)\|^2] - 2\,\mathbb{E}_\theta[(X-\theta)^\mathsf{T} h(X)] \\ &= d + \mathbb{E}_\theta[\|h(X)\|^2] - 2\,\mathbb{E}_\theta[\operatorname{tr} Dh(X)] \\ &= \mathbb{E}_\theta[\hat{R}].\end{aligned}$$

This is called **SURE**.

Example 11.5. Let $\delta(X) = X$, where $X \sim \mathcal{N}_d(\theta, I_d)$. Then, $h(X) = 0$, $Dh(X) = 0$, so $\hat{R} = d = R(\theta, \delta)$ for all θ.

Example 11.6. Take $\delta(X) = (1-\zeta)X$, where ζ is fixed. Take $h(X) = \zeta X$. So,

$$Dh(X) = \begin{bmatrix} \zeta & & 0 \\ & \ddots & \\ 0 & & \zeta \end{bmatrix}.$$

Thus,

$$\begin{aligned} \hat{R} &= d + \zeta^2 \|X\|^2 - 2\zeta d \\ &= (1-2\zeta)d + \zeta^2\|X\|^2, \\ R(\theta, \delta) &= (1-2\zeta)d + \zeta^2(\|\theta\|^2 + d) \\ &= (1-\zeta)^2 d + \zeta^2\|\theta\|^2. \end{aligned}$$

11.3 Stein's Paradox

James-Stein Paradox: Under *no* assumptions about $\theta = (\theta_1, \ldots, \theta_n)$, $X_i \overset{\text{independent}}{\sim} \mathcal{N}(\theta_i, 1)$, the "obvious" estimator X is *inadmissible* and dominated by δ^{JS}.

$\delta(X)$ is **location-equivariant** if $\delta(X + a) = \delta(X) + a$. Note that X is UMVU, minimax, and the best location-equivariant estimator.

For *any* value $\theta_0 \in \mathbb{R}^n$, we could shrink toward θ_0 instead.

$$\delta(X) = \Big(1 - \frac{n-2}{\|X - \theta_0\|^2}\Big)X + \frac{n-2}{\|X-\theta_0\|^2}\theta_0.$$

Then, $R_{\text{MSE}}(\theta, \delta^{\text{JS}}) < R_{\text{MSE}}(\theta, X)$ for all $\theta \in \mathbb{R}^n$.

We have:

$$\begin{aligned} \delta^{\text{JS}}(X) &= \Big(1 - \frac{d-2}{\|X\|^2}\Big)X, \\ h(X) &= \frac{d-2}{\|X\|^2}X, \\ \|h(X)\|^2 &= \frac{(d-2)^2}{\|X\|^4}\|X\|^2 = \frac{(d-2)^2}{\|X\|^2}, \\ Dh(X)_{i,i} &= \frac{\partial h_i(X)}{\partial X_i} \\ &= \frac{\partial}{\partial X_i}\frac{(d-2)X_i}{\|X\|^2}. \end{aligned}$$

Lecture 12

October 3

12.1 James-Stein Wrap-Up

12.1.1 SURE

Define:

$$\hat{R} = d + \|h(X)\|^2 - 2\,\operatorname{tr} Dh(X)$$
$$h(X) = X - \delta(X)$$

If $X \sim \mathcal{N}_d(\theta, I_d)$, then $\mathbb{E}_\theta[\hat{R}(X)] = \text{MSE}$. When $\delta(X) = X$, $\text{MSE} = d$. When $\delta(X) = (1-\zeta)X$, then $\text{MSE} = (1-\zeta)^2 d + \zeta^2\|\theta\|^2$.

12.1.2 James-Stein Estimator

$$\begin{aligned}
\delta(X) &= \Big(1 - \frac{d-2}{\|X\|^2}\Big)X \\
h(X) &= \frac{d-2}{\|X\|^2}X \\
\|h(X)\|^2 &= \frac{(d-2)^2}{\|X\|^2} \\
Dh(X)_{i,i} &= \frac{\partial h}{\partial X_i}(X) \\
&= \frac{\partial}{\partial X_i}\frac{(d-2)X_i}{\sum_{j=1}^d X_j^2} \\
&= \frac{\|X\|^2(d-2) - 2(d-2)X_i^2}{\|X\|^4} \\
\operatorname{tr} Dh(X) &= \frac{d-2}{\|X\|^4}(d\|X\|^2 - 2\|X\|^2) \\
&= \frac{(d-2)^2}{\|X\|^2} \\
\hat{R}\big(\delta^{\text{JS}}(X)\big) &= d + \frac{(d-2)^2}{\|X\|^2} - 2\frac{(d-2)^2}{\|X\|^2} \\
&= d - \frac{(d-2)^2}{\|X\|^2} \\
\text{MSE}(\theta, \delta^{\text{JS}}) &= d - \mathbb{E}_\theta\Big[\frac{(d-2)^2}{\|X\|^2}\Big] < d.
\end{aligned}$$

In fact,

$$\delta^{\text{JS}}(X) = \Big(1 - \frac{d-2}{\|X\|^2}\Big)X$$

is inadmissible because $1 - (d-2)/\|X\|^2$ could be negative. We could take

$$\delta^{\text{JS}+}(X) = \Big(1 - \frac{d-2}{\|X\|^2}\Big)_+ X.$$

A more practical estimator might be

$$\delta^{\text{JS},2}(X) = \bar{X} + \Big(1 - \frac{d-3}{\|X - \mathbf{1}\bar{X}\|}\Big)_+ (X - \bar{X}),$$

which also dominates X for $d \geq 4$.

$\text{MSE}((1-\zeta)X) = (1-\zeta)^2 d + \zeta^2\|\theta\|^2$ is never minimized at $\zeta = 0$. The minimum is at $d/(d + \|\theta\|^2)$.

12.2 Hypothesis Testing

Our model is $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. We want to "test":

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \subseteq \Theta \qquad && \text{null hypothesis} \\ H_1 &: \theta \in \Theta_1 \subseteq \Theta \qquad && \text{alternative hypothesis} \end{aligned}$$

Usually, $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \varnothing$. H_0 is the default, and we either "accept" H_0 (fail to reject) or reject H_0 (in favor of H_1).

Example 12.1. $X \sim \mathcal{N}(\theta, 1)$. Test $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$, or $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.

Example 12.2. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_1$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} P_2$. Test $H_0 : P_1 = P_2$ versus $H_1 : P_1 \neq P_2$.

12.2.1 Critical Function/Power Function

Formally describe a test by defining its **critical function** (**test function**).

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

(This is a randomized test.) (In practice, $\phi(\mathcal{X}) = \{0, 1\}$.) For non-randomized tests, the **rejection region** is $R = \{x : \phi(x) = 1\}$ and $\mathcal{X} \setminus R$ is the **acceptance region**.

The **power function** is $\beta(\theta) = \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta(\text{reject } H_0)$, which is the rejection probability if $X \sim P_\theta$.

The **significance level** is $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$. $\alpha = 0.05$ is very common.

Example 12.3. Let $X \sim \mathcal{N}(\theta, 1)$ and we test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. One test is

$$\phi_1(X) = \mathbb{1}\{|X| > z_{\alpha/2}\},$$

where $z_\alpha = \Phi^{-1}(1 - \alpha)$. Other tests are

$$\begin{aligned} \phi_2(X) &= \mathbb{1}\{X > z_\alpha\}, \\ \phi_3(X) &= \mathbb{1}\{X < -z_{\alpha/3} \text{ or } X > z_{2\alpha/3}\}. \end{aligned}$$

How do we compare the power functions?

Lecture 13

October 5

13.1 Review: Testing

Test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$.

The **critical function** is:

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

The **power function** is

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

The **significance level** is $\alpha_\phi = \sup_{\theta \in \Theta_0} \beta_\phi(\theta)$.

Example 13.1. If $X \sim \mathcal{N}(\theta, 1)$ and we test $H_0 : \theta = 0$, $H_1 : \theta \neq 0$, then there are numerous possible power functions and there is not necessarily a best test.

Example 13.2. If $X \sim \mathcal{N}(\theta, 1)$ and we test $H_0 : \theta \leq 0$, $H_1 : \theta > 0$, then there is a single best test: $\phi_2(X) = \mathbb{1}\{X > z_\alpha\}$.

Example 13.3. Let $X \sim \text{Binomial}(n, \theta)$. Test $H_0 : \theta \leq 1/2$ versus $H_1 : \theta > 1/2$. Then,

$$\mathbb{P}_{\theta=1/2}(X \in R) = \frac{1}{2^n} \sum_{x \in R} \binom{n}{x} = \text{multiple of } 2^{-n}.$$

The optimal test will be of the form:

$$\phi(X) = \begin{cases} 0, & X < c \\ \gamma, & X = c \\ 1, & X > c \end{cases}$$

13.2 Neyman-Pearson Lemma

13.2.1 Simple Hypothesis

A **simple hypothesis** is one that fully specifies the sampling distribution. (Θ_0 or Θ_1 is a singleton.) If $\Theta_0 = \{0\}$, $\Theta_1 = \{1\}$, then there exists a unique* best test, which rejects when

$$L(X) = \frac{p_1(X)}{p_0(X)}$$

is large.

$$L(X) = \frac{p_1(X)}{p_0(X)} \in [0, \infty]$$

(undefined if the expression is $0/0$). The test

$$\phi^*(X) = \begin{cases} 0, & L(X) < c \\ \gamma, & L(X) = c \\ 1, & L(X) > c \end{cases}$$

is an optimal level-α test. ϕ^* is called the **likelihood ratio test (LRT)**.

Intuition: The significance level is $\int \phi(x) p_0(x) \, \mathrm{d}\mu(x)$ (buck). The power is $\int \phi(x) p_1(x) \, \mathrm{d}\mu(x)$ (bang).

Proposition 13.4 (Keener 12.1). *Suppose that $c \ge 0$, and ϕ^* maximizes $\mathbb{E}_1[\phi(X)] - c\,\mathbb{E}_0[\phi(X)]$ among all critical functions. If $\mathbb{E}_0[\phi^*(X)] = \alpha$, then ϕ^* maximizes $\mathbb{E}_1[\phi(X)]$ among all level-α critical functions.*

Proof. Suppose $\mathbb{E}_0[\phi(X)] \le \alpha$. Then,

$$\begin{aligned} \mathbb{E}_1[\phi(X)] &\le \mathbb{E}_1[\phi(X)] - c\,\mathbb{E}_0[\phi(X)] + c\alpha \\ &\le \mathbb{E}_1[\phi^*(X)] - c\,\mathbb{E}_0[\phi^*(X)] + c\alpha \\ &= \mathbb{E}_1[\phi^*(X)]. \end{aligned}$$

□

Theorem 13.5 (Neyman-Pearson Lemma). *The LRT with level α is optimal for testing $H_0 : \theta = 0$ versus $H_1 : \theta = 1$.*

Proof. For any test ϕ,

$$\begin{aligned} \mathbb{E}_1[\phi(X)] - c\,\mathbb{E}_0[\phi(X)] &= \int \big(p_1(x) - cp_0(x)\big)\phi(x) \, \mathrm{d}\mu(x) \\ &= \int_{\{p_1 > cp_0\}} |p_1 - cp_0| \phi \, \mathrm{d}\mu - \int_{\{p_1 < cp_0\}} |p_1 - cp_0| \phi \, \mathrm{d}\mu. \end{aligned}$$

Any test maximizing this expression must have $\phi^*(x) = 1$ on $\{p_1(x) > cp_0(x)\}$ and $\phi^*(x) = 0$ on $\{p_1(x) < cp_0(x)\}$. Find c such that

$$\begin{aligned} &\mathbb{P}_0\big(p_1(X) > cp_0(X)\big) \le \alpha, \\ &\mathbb{P}_0\big(p_1(X) < cp_0(X)\big) \le 1 - \alpha. \end{aligned}$$

Take $\gamma \in [0, 1]$ to make the level α. □

Example 13.6. Let $X \sim \mathcal{N}(\theta, 1)$, $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$. Assume $\theta_1 > \theta_0$.

$$\begin{aligned} L(x) = \frac{p_1(x)}{p_0(x)} &= \frac{\mathrm{e}^{-(x-\theta_1)^2/2}}{\mathrm{e}^{-(x-\theta_0)^2/2}} \\ &= \frac{\mathrm{e}^{\theta_1 x - \theta_1^2/2}}{\mathrm{e}^{\theta_0 x - \theta_0^2/2}} \\ &= \mathrm{e}^{(\theta_1-\theta_0)x - (\theta_1^2-\theta_0^2)/2}. \end{aligned}$$

$L(X)$ is strictly monotone in X, so the distribution is continuous.

$$\begin{aligned} \phi^*(X) &= \mathbb{1}\{\mathrm{e}^{(\theta_1-\theta_0)X-(\theta_1^2-\theta_0^2)/2} > c\} && \text{for some } c \\ &= \mathbb{1}\{X > \tilde{c}\} && \text{for } \tilde{c} = \theta_0 + z_\alpha \\ &= \mathbb{1}\{X > \theta_0 + z_\alpha\}. \end{aligned}$$

$\phi^*(X)$ does not depend on θ_1. Thus, ϕ^* is *uniformly most powerful* for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$.

13.3 Uniformly Most Powerful (UMP) Tests

Generally, we say ϕ^* is **level-α UMP** for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ if $\beta_{\phi^*}(\theta) \geq \beta_\phi(\theta)$ for all $\theta \in \Theta_1$, for all ϕ with significance level $\leq \alpha$.

13.3.1 One-Parameter Exponential Families

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_\eta(x) = \mathrm{e}^{\eta T(x) - A(\eta)} h(x)$ $(\eta \in \mathbb{R})$. Test $H_0 : \eta = \eta_0$ versus $H_1 : \eta = \eta_1$ $(\eta_1 > \eta_0)$.

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

ϕ^* rejects when $\sum_{i=1}^n T(X_i)$ is large.

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

There is no dependence on η_1. Therefore, ϕ^* is UMP for $H_0 : \eta = \eta_0$ versus $H_1 : \eta > \eta_0$.

13.3.2 Monotone Likelihood Ratio

Definition 13.7. Let $\mathcal{P} = \{p_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ be a dominated family. Then, $\mathcal{P}$ has **monotone likelihood ratio (MLR)** if there exists a statistic $T(X)$ such that $\theta_1 < \theta_2$ implies $p_{\theta_2}(X)/p_{\theta_1}(X)$ is a non-decreasing function of $T(X)$.

Example 13.8. If $p_\theta(x) = \mathrm{e}^{\eta(\theta)T(x) - B(\theta)} h(x)$, then for $\theta_2 < \theta_1$,

$$\frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} = \mathrm{e}^{(\eta(\theta_1)-\eta(\theta_2))T(x) - (B(\theta_1)-B(\theta_2))}$$

is increasing in $T(x)$ if $\eta(\cdot)$ is increasing.

(We already know that "reject for large T" is UMP for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$.)

Corollary 13.9 (Keener Corollary 12.4)**.** *If p_0, p_1 are not a.s. equal, and ϕ^* is the LRT with level α, then $\mathbb{E}_1[\phi^*(X)] > \alpha$.*

Proof. $\mathbb{E}_1[\phi(X)] = \alpha$ is attainable by $\phi(X) = \alpha$. Therefore, $\mathbb{E}_1[\phi^*(X)] \geq \alpha$. Find $\varepsilon > 0$ and let $B_\varepsilon = \{x : p_1(x) \geq (1+\varepsilon)p_0(x)\}$. Find $\varepsilon > 0$ such that $\mathbb{P}_0(B_\varepsilon) > 0$. Then, $\mathbb{P}_1(B_\varepsilon) > (1+\varepsilon)\mathbb{P}_0(B_\varepsilon)$. If $\mathbb{P}_0(B_\varepsilon) > \alpha$, let:

$$\tilde{\phi}(X) = \begin{cases} 0, & x \notin B_\varepsilon \\ \alpha/\mathbb{P}_0(B_\varepsilon), & x \in B_\varepsilon \end{cases}$$

If $\mathbb{P}_0(B_\varepsilon) \leq \alpha$, let:

$$\tilde{\phi}(X) = \begin{cases} (\alpha - \mathbb{P}_0(B_\varepsilon))/(1 - \mathbb{P}_0(B_\varepsilon)), & x \notin B_\varepsilon \\ 1, & x \in B_\varepsilon \end{cases}$$

One can show that $\mathbb{E}_1[\tilde{\phi}(X)] > \alpha$. □

Lecture 14

October 10

14.1 MLR $\implies$ UMP

Theorem 14.1. *If $\mathcal{P}$ has MLR in $T(X)$, then the test ϕ^* that rejects for large $T(X)$:*

$$\phi^*(X) = \begin{cases} 0, & T(X) < c \\ \gamma, & T(X) = c \\ 1, & T(X) > c \end{cases}$$

1. *is UMP for testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$ among all tests with significance level at most $\alpha = \mathbb{E}_{\theta_0}[\phi^*(X)]$;*
2. *$\beta_{\phi^*}(\theta) = \mathbb{E}_\theta[\phi^*(X)]$ is non-decreasing in θ, and strictly increasing[a] whenever $\beta(\theta) \in (0, 1)$.*
3. *If $\theta_1 < \theta_0$, then ϕ^* minimizes $\mathbb{E}_{\theta_1}[\phi^*(X)]$ among all tests with $\mathbb{E}_{\theta_0}[\phi^*(X)] = \alpha$.*

[a] Provided that the family is identifiable.

Proof. 2. Suppose $\theta_1 < \theta_2$. Then,

$$L(X) = \frac{p_{\theta_2}(X)}{p_{\theta_1}(X)}$$

is non-decreasing in $T(X)$. Therefore, $\phi^*(X)$ is a most powerful LRT for $H_0 : \theta = \theta_1$ versus $H_1 : \theta = \theta_2$, so it is a MP LRT at level $\hat{\alpha}(\theta_1) = \beta_{\phi^*}(\theta_1)$. By 13.9, then $\mathbb{E}_{\theta_2}[\phi^*(X)] \ge \mathbb{E}_{\theta_1}[\phi^*(X)]$ with strict inequality unless both are 0 or 1.

1. Suppose $\theta_1 > \theta_0$ and some other test $\tilde{\phi}$ has level $\le \alpha$. In particular, $\mathbb{E}_{\theta_0}[\tilde{\phi}(X)] \le \alpha$. By the NP Lemma 13.5, $\phi^*(X)$ (the LRT) has power at θ_1 at least $\mathbb{E}_{\theta_1}[\tilde{\phi}(X)]$. By 2, $\phi^*(X)$ has significance level $\le \alpha$, so $\phi^*(X)$ is UMP.

3. Suppose $\theta_1 < \theta_0$, $\mathbb{E}_{\theta_0}[\tilde{\phi}(X)] = \alpha$. If $\tilde{\delta} = \mathbb{E}_{\theta_1}[\tilde{\phi}(X)] < \delta^* = \mathbb{E}_{\theta_1}[\phi^*(X)]$, this contradicts the fact that $\phi^*(X)$ is most powerful for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$.

□

14.2 Two-Sided Tests, UMPU

Setup: $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$. Test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \ne \theta_0$. Assume $T(X) \in \mathbb{R}$ is a summary test statistic, stochastically increasing in θ. $\mathbb{P}_\theta(T(X) \le t)$ is non-increasing in θ ($\mathbb{P}_\theta(T(X) > t)$ is non-decreasing) so bigger θ yields bigger $T(X)$.

Example 14.2. If $X_i \overset{\text{i.i.d.}}{\sim} p_\theta(x) = p_0(x-\theta)$ for $i = 1, \ldots, n$ (a **location family**), $T(X)$ might be the sample mean or sample median.

Example 14.3.

$$X_i \overset{\text{i.i.d.}}{\sim} p_\theta(x) = \frac{1}{\theta} p_1\left(\frac{x}{\theta}\right)$$

for $\theta > 0$, $x \geq 0$, is called a **scale family**.

A **two-tailed test** based on $T(X)$ rejects when $T(X)$ is extreme (big or small).

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

Thus,

$$\begin{aligned} \mathbb{P}_{\theta_0}(\text{reject } H_0) &= \mathbb{P}_{\theta_0}\big(\text{reject because } T(X) \text{ small}\big) + \mathbb{P}_{\theta_0}\big(\text{reject because } T(X) \text{ large}\big) \\ &= \alpha_1 + \alpha_2. \end{aligned}$$

How do we balance α_1 versus α_2? Simplest idea: **equal-tailed test**. $\alpha_1 = \alpha_2 = \alpha/2$.

14.2.1 UMPU Test

We say a test ϕ is **unbiased** if $\mathbb{E}_\theta[\phi(X)] \geq \alpha$ for all $\theta \in \Theta_1$.

14.3 p-Values

Example 14.4. Let $X \sim \mathcal{N}(\theta, 1)$ and test $H_0 : \theta = 0$ versus $H_1 = \theta \neq 0$. The ***p*-value** is

$$p(x) = \mathbb{P}_0(|X| > |x|) = 2\big(1 - \Phi(x)\big),$$

where Φ is the $\mathcal{N}(0,1)$ CDF.

For simplicity, assume that the test statistic has an absolutely continuous distribution so the test is non-randomized for all α.

Setup: Consider a testing problem $\mathcal{P}$, H_0, H_1. Have a test $\phi(X;\alpha)$ for each α. Thus, $\phi(X;\alpha) = \mathbb{1}\{X \in R_\alpha\}$. $\phi(X;\alpha)$ has level exactly α. Assume that the tests are monotone in α: if $\alpha_1 \leq \alpha_2$, then $\phi(X;\alpha_1) \leq \phi(X;\alpha_2)$, or equivalently, $R_{\alpha_1} \subseteq R_{\alpha_2}$.

Definition 14.5. The ***p*-value** is

$$\begin{aligned} p(X) &= \inf\{\alpha : \phi(X;\alpha) = 1\} \\ &= \inf\{\alpha : x \in R_\alpha\}. \end{aligned}$$

Assume $T(X)$ is continuous and $\phi(X;\alpha)$ rejects when $T(X) \geq t_\alpha$. Then,

$$\begin{aligned} p(X) \leq \alpha &\iff \phi(X;\alpha) = 1 \\ &\iff T(X) \geq t_\alpha. \end{aligned}$$

Thus,

$$p(X) = \alpha \iff T(X) = t_\alpha$$

$$\iff \sup_{\theta_0 \in \Theta_0} \mathbb{P}_{T(X^*) \sim P_{\theta_0}}\big(T(X^*) > T(X)\big) = \alpha.$$

For $\theta \in \Theta_0$,

$$\begin{aligned} \mathbb{P}_\theta\big(p(X) \le \alpha\big) &= \mathbb{P}_\theta\big(\phi(X;\alpha) = 1\big) \\ &\le \alpha. \end{aligned}$$

Lecture 15

October 12

15.1 UMPU Tests for Exponential Families

15.1.1 Two-Sided Test (Based on $T(X) \in \mathbb{R}$)

$$\phi(X) = \begin{cases} 0, & T(X) \in (c_1, c_2) \\ 1, & T(X) \in [c_1, c_2]^{\mathsf{c}} \\ \gamma_i, & T(X) = c_i \ (i = 1, 2) \end{cases}$$

Unbiased. $\phi(X)$ is (level-α) **unbiased** if $\mathbb{E}_\theta[\phi(X)] \geq \alpha$ for all $\theta \in \Theta_1$.

Consider a one-parameter exponential family (canonical form) $p_\eta(x) = \mathrm{e}^{\eta T(x) - A(\eta)} h(x)$.
Test $H_0 : \eta = \eta_0$ versus $H_1 : \eta \neq \eta_0$. (15.1)

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\eta} \mathbb{E}_\eta[\phi(X)] &= \int \frac{\mathrm{d}}{\mathrm{d}\eta} \phi(x) \mathrm{e}^{\eta T(x) - A(\eta)} h(x) \, \mathrm{d}\mu(x) \\ &= \int \phi(x) \Big(T(x) - \frac{\mathrm{d}}{\mathrm{d}\eta} A(\eta) \Big) p_\eta(x) \, \mathrm{d}\mu(x) \\ &= \mathbb{E}_\eta\big[\phi(X)\big(T(X) - \mathbb{E}_\eta[T(X)]\big)\big] \\ &= \mathbb{E}_\eta\big[T(X)\big(\phi(X) - \mathbb{E}_\eta[\phi(X)]\big)\big]. \end{aligned}$$

Theorem 15.1 (Keener 12.26)**.** *For the problem* (15.1) *with $\eta_0 \in \Xi^\circ$, there is a two-sided level-α test $\phi^*(X)$ based on $T(X)$ where we choose c_i, γ_i to solve*

$$\mathbb{E}_{\eta_0}[\phi^*(X)] = \alpha, \tag{15.2}$$

$$\mathbb{E}_{\eta_0}\big[T(X)\big(\phi^*(X) - \alpha\big)\big] = 0. \tag{15.3}$$

ϕ^* is UMPU.

Why are (15.2) and (15.3) enough to specify a unique solution for c_i, γ_i? In the continuous case, solving (15.2) makes c_2 an implicit function of c_1. Also in the continuous case, (15.3) is equivalent to

$$\mathbb{E}_{\eta_0}[T(X) \, \mathbb{1}\{T(X) \in R(\phi^*)\}] = \mathbb{E}_{\eta_0}[T(X)] \mathbb{P}_{\eta_0}\big(T(X) \in R(\phi^*)\big),$$

so $\mathbb{E}_{\eta_0}[T(X)] = \mathbb{E}_{\eta_0}[T(X) \mid T(X) \in R(\phi^*)]$.

15.2 Confidence Sets/Intervals

Definition 15.2. Given a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $C(X)$ is a $1-\alpha$ **confidence set for** $g(\theta)$ if $\mathbb{P}_\theta(g(\theta) \in C(X)) \geq 1-\alpha$, for all $\theta \in \Theta$ (a **confidence interval** if $C(X)$ is an interval).

Notes: $C(X)$ is random, not $g(\theta)$. There is a $1-\alpha$ chance that the *procedure* $C(\cdot)$ will produce an interval containing the *fixed* value $g(\theta)$.

Incorrect: "There is a 95% chance that $g(\theta)$ is in the interval $[0.1, 0.2]$ that I just constructed."

15.2.1 Duality of Testing & Interval Estimation

Suppose we have a level-α test $\phi_{\theta_0}(X)$ of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, for each $\theta_0 \in \Theta$. Assume that the tests are non-randomized. Let $C(X) = \{\theta \in \Theta : \phi_\theta(X) < 1\}$ ("all non-rejected θ values"). Then $C(X)$ is a $1-\alpha$ confidence set for θ.

$$\mathbb{P}_\theta\big(\theta \notin C(X)\big) = \mathbb{P}_\theta\big(\phi_\theta(X) = 1\big) \leq \alpha.$$

(For $g(\theta)$, $C(X) = \{g(\theta) : \phi_\theta(x) < 1\}$.)

We say C **inverts** the (family of) tests ϕ_{θ_0}.

Alternatively, suppose we have $C(X)$, a $(1-\alpha)$-level confidence set for θ. Then, $\phi_{\theta_0}(X) = \mathbb{1}\{\theta_0 \notin C(X)\}$ is a level-α test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. $\mathbb{P}_{\theta_0}(\phi_{\theta_0}(X) = 1) = \mathbb{P}_{\theta_0}(C(X) \not\ni \theta_0) \leq \alpha$.

To test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$, we can take

$$\begin{aligned}\phi_{\Theta_0}(X) &= \min_{\theta \in \Theta_0} \phi_\theta(X) \\ &= \mathbb{1}\{\Theta_0 \cap C(X) = \varnothing\}.\end{aligned}$$

For $\theta \in \Theta_0$, $\mathbb{E}_\theta[\phi_{\Theta_0}(X)] \leq \mathbb{E}_\theta[\phi_\theta(X)] \leq \alpha$.

Example 15.3. Let $X \sim \text{Exponential}(\theta)$ with density

$$p_\theta(x) = \frac{1}{\theta} \mathrm{e}^{-x/\theta}$$

for $x, \theta > 0$. Then, $\mathbb{P}_\theta(X \leq x) = 1 - \mathrm{e}^{-x/\theta}$, so if we take the $\alpha/2$-quantile,

$$\frac{\alpha}{2} = 1 - \mathrm{e}^{-x/\theta} \implies x = -\theta \log\Big(1 - \frac{\alpha}{2}\Big).$$

Similarly, the $1 - \alpha/2$ quantile is $x = -\theta \log(\alpha/2)$. Thus, we reject the θ values unless

$$-\theta \log\Big(1 - \frac{\alpha}{2}\Big) \leq X \leq -\theta \log\Big(\frac{\alpha}{2}\Big)$$

which is equivalent to rejecting unless

$$-X^{-1} \log\Big(1 - \frac{\alpha}{2}\Big) \leq \theta^{-1} \leq -X^{-1} \log\Big(\frac{\alpha}{2}\Big).$$

Hence,

$$C(X) = \Big(-\frac{X}{\log(\alpha/2)}, -\frac{X}{\log(1-\alpha/2)}\Big).$$

15.3 Testing with Nuisance Parameters

So far, we have studied one-parameter families.

15.3.1 Nuisance Parameters

The model is $\mathcal{P} = \{P_{\theta,\zeta} : (\theta, \zeta) \in \Omega \subseteq \mathbb{R}^{r+s}\}$. $\theta \in \mathbb{R}^s$ is the parameter of interest and $\zeta \in \mathbb{R}^r$ is the **nuisance parameter**. We test $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$.

Example 15.4. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\nu, \sigma^2)$, where $\mu, \nu \in \mathbb{R}$, $\sigma^2 > 0$, and all parameters are unknown. Test $H_0 : \mu = \nu$ versus $H_1 : \mu \neq \nu$. Then, $\theta = \mu - \nu$ is the parameter of interest and $\zeta = (\mu + \nu, \sigma^2)$ is the nuisance parameter.

Example 15.5. Let $X_i \overset{\text{independent}}{\sim} \text{Poisson}(\lambda_i)$, $\lambda_i > 0$, for $i = 1, 2$. Test $H_0 : \lambda_1 \leq \lambda_2$ versus $H_1 : \lambda_1 > \lambda_2$. The parameter of interest is $\theta = \lambda_1/\lambda_2$ and the nuisance is $\zeta = \lambda_1$ or $\zeta = \lambda_1\lambda_2$. Thus, we equivalently test $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$.

Example 15.6. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$, $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Q$. Test $H_0 : P = Q$ versus $H_1 : P \neq Q$. The nuisance parameter is P, which is infinite-dimensional.

Lecture 16

October 17

16.1 UMPU Testing with Nuisance Parameters

16.1.1 Multiparameter Exponential Families

Model: $p_{\theta,\zeta}(x) = \mathrm{e}^{\theta T(x)+\zeta^\mathsf{T} U(x)-A(\theta,\zeta)}h(x)$, where $\theta \in \mathbb{R}$, $\zeta \in \mathbb{R}^{s-1}$. Test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

Basic idea: eliminate ζ by conditioning on $U(X)$ (condition on sufficient statistics of Θ_0). Under H_0, $\theta = \theta_0$ is known, so $U(X)$ is sufficient under H_0. If we condition on $U(X)$, we get a simple null.

$$\begin{aligned} p_{\theta,\zeta}\big(x \mid U(X) = u\big) &= \frac{\mathrm{e}^{\theta T(x)+\zeta^\mathsf{T} U(x)-A(\theta,\zeta)}h(x)\,\mathbb{1}\{U(x) = u\}}{\int_{\{U(z)=u\}} \mathrm{e}^{\theta T(z)+\zeta^\mathsf{T} U(z)-A(\theta,\zeta)}h(z)\,\mathrm{d}z} \\ &= \mathrm{e}^{\theta T(x)-\hat{A}_u(\theta)}h_u(x). \end{aligned}$$

There is no dependence on ζ, so $T(X)$ is the sole sufficient statistic. We will show that the optimal test rejects when $T(X)$ is extreme given $U(X)$.

Example 16.1. $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ are independent. Test: $H_0 : \lambda_1 = \lambda_2$ versus $H_1 : \lambda_1 \neq \lambda_2$.

$$\begin{aligned} p_\lambda(x) &= \lambda_1^{x_1}\lambda_2^{x_2}\mathrm{e}^{-\lambda_1-\lambda_2}\frac{1}{x_1!x_2!} \\ &= \mathrm{e}^{x_1\log\lambda_1+x_2\log\lambda_2-\lambda_1-\lambda_2}\frac{1}{x_1!x_2!} \\ &\propto_x \mathrm{e}^{(x_1-x_2)(\log\lambda_1-\log\lambda_2)/2+(x_1+x_2)(\log\lambda_1+\log\lambda_2)/2}\frac{1}{x_1!x_2!}. \end{aligned}$$

Thus,

$$\begin{aligned} T(x) &= x_1 - x_2, \\ \theta &= \frac{\log\lambda_1 - \log\lambda_2}{2}, \\ U(x) &= x_1 + x_2, \\ \zeta &= \frac{\log\lambda_1 + \log\lambda_2}{2}. \end{aligned}$$

Now, H_0 is equivalent to $\theta = 0$ and H_1 is equivalent to $\theta \neq 0$. Condition on $U(X) = X_1 + X_2 = u$.

$$\begin{aligned} p_\theta(x \mid x_1 + x_2 = u) &\propto_x e^{(x_1 - x_2)\theta} \frac{u!}{x_1! x_2!} \\ &= e^{(2x_1 - u)\theta} \binom{u}{x_1} \\ &\propto_x e^{x_1 \log(\lambda_1/\lambda_2)} \binom{u}{x_1} \\ &\propto_x \text{Binomial}\Big(u, \frac{\lambda_1}{\lambda_1 + \lambda_2}\Big) \\ &= \text{Binomial}\Big(u, \frac{1}{2}\Big) \end{aligned}$$

under H_0. Reject if $X_1 - X_2$ is extreme given $U(X)$, or equivalently, reject if X_1 is extreme given $X_1 + X_2 = u$. If testing $H_0 : \lambda_1 \leq \lambda_2$, equivalently test $\theta \leq 0$. If testing $H_0 : \lambda_1 \leq 3\lambda_2$, equivalently test $\theta \leq \log(1/3)$.

Theorem 16.2. *Consider testing either $H_0 : \theta = \theta_0$ or $H_0 : \theta \leq \theta_0$ in an exponential family model $\mathcal{P} = \{p_{\theta,\zeta}(x) : (\theta, \zeta) \in \Omega\}$, where $p_{\theta,\zeta}(x) = e^{\theta T(x) + \zeta^\mathsf{T} U(x) - A(\theta,\zeta)} h(x)$. Ω is open, so $\mathcal{P}$ is full-rank. Then, there is a UMPU test of the form: $\phi^*(X) = \psi(T(X), U(X))$ where*

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

for $H_0 : \theta = \theta_0$, or

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

for $H_0 : \theta \leq \theta_0$, where γ is chosen such that

$$\mathbb{E}_{\theta_0}[\phi^*(X) \mid U(X) = u] = \alpha, \qquad \forall u \tag{16.1}$$

$$\mathbb{E}_{\theta_0}\big[T(X)\big(\phi^*(X) - \alpha\big) \bigm| U(X) = u\big] = 0, \qquad \forall u \tag{16.2}$$

(where (16.2) is only for the two-sided version).

Note: There is no dependence on ζ.

Proof Sketch (One-Sided) of 16.2. We need $\beta \leq \alpha$ on Ω_0 (this is the significance level) and we need $\beta \geq \alpha$ (unbiased). Let $\omega = \{(\theta_0, \zeta) : \zeta \in \mathbb{R}^{s-1}\} \cap \Omega$ be the boundary.

Steps:

1. *Any* unbiased test must have $\beta(\theta_0, \zeta) = \alpha$ for all ζ (the power is $\geq \alpha$ on ω, by continuity).
2. Therefore, $\mathbb{E}_{\theta_0}[\phi(X) \mid U = u] = \alpha$ for all u (by completeness).
3. ϕ^* is optimal among tests that condition on u.

Step 1: Recall $\mathbb{E}_{\theta,\zeta}[|\phi(X)|] \leq 1 < \infty$ for all $\theta, \zeta \in \Omega$ so $\mathbb{E}_{\theta,\zeta}[\phi(X)]$ is continuous.

Step 2: Write $\mathcal{Q} = \{q_\zeta(x) = p_{\theta_0,\zeta}(x) : (\theta_0, \zeta) \in \Omega\}$. So, $q_\zeta(x) = \mathrm{e}^{\zeta^\mathsf{T} U(x) - A(\theta_0,\zeta)} \mathrm{e}^{\theta_0 T(x)} h(x)$. $\mathcal{Q}$ is a full-rank one-parameter exponential family with an open parameter space, so $U(X)$ is a complete sufficient statistic for $\mathcal{Q}$. Define $f(u) = \mathbb{E}_{\theta_0}[\phi(X) \mid U(X) = u]$. Then, $\beta(\theta_0, \zeta) = \mathbb{E}_{\theta_0,\zeta}[f(U(X))]$. If $\beta(\theta_0, \zeta) = \alpha$ for all ζ, then $f(U(X)) \overset{\text{a.s.}}{=} \alpha$. Thus, $\phi(X)$ has conditional level α on ω.

Step 3: For $\theta > \theta_0$,

$$\begin{aligned}\mathbb{E}_{\theta,\zeta}[\phi(X)] &= \mathbb{E}_{\theta,\zeta}\big[\mathbb{E}_\theta\big(\phi(X) \bigm| U(X)\big)\big] \\ &\le \mathbb{E}_{\theta,\zeta}\big[\mathbb{E}_\theta\big(\phi^*(X) \bigm| U(X)\big)\big] \\ &= \mathbb{E}_{\theta,\zeta}[\phi^*(X)].\end{aligned}$$

□

Example 16.3. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 > 0$ unknown and test $H_0 : \mu \le 0$.

$$\begin{aligned}p_{\mu,\sigma^2}(x) &= \exp\Big(\frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 - \frac{n\mu^2}{2\sigma^2}\Big)\Big(\frac{1}{2\pi\sigma^2}\Big)^{n/2}, \\ \theta &= \frac{\mu}{\sigma^2}, \\ T(X) &= \sum_{i=1}^n X_i, \\ \zeta &= -\frac{1}{2\sigma^2}, \\ U(X) &= \sum_{i=1}^n X_i^2.\end{aligned}$$

Condition on $U = \|X\|_2^2$. The distribution of X (under $\mu = 0$) is Uniform(sphere of radius $\|X\|_2$).

$$\begin{aligned}p_0(x \mid \|x\|_2^2 = u) &\propto_x \mathrm{e}^{-u/(2\sigma^2)}\, \mathbb{1}\{\|x\|_2^2 = u\} \\ &= \frac{\mathbb{1}\{\|x\|_2^2 = u\}}{\mathrm{vol}(\|x\|_2 S^{n-1})}.\end{aligned}$$

The optimal test rejects when

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$$

is large given $\|X\|_2$, or equivalently, rejects when $\bar{X}/\|X\|_2$ is large given $\|X\|_2$, but this test statistic does not depend on $\|X\|_2$. So, equivalently, the test rejects when

$$\frac{\bar{X}}{\sqrt{S^2/n}} = \frac{\bar{X}}{\sqrt{(\|X\|_2^2 - n\bar{X}^2)/n}}$$

is large, where

$$\begin{aligned}S^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \|X\|_2^2 - n\bar{X}^2.\end{aligned}$$

Rejecting when $T(X)$ is large given $U(X)$ is equivalent to rejecting $f(T(X), U(X))$ is large given $U(X)$ if f is strictly increasing in the first argument. "Reject when $T(X)$ is large/extreme given $U(X)$" $\iff$ "reject when $f(T(X), U(X))$ is large/extreme given $U(X)$" if $f(t, u)$ is strictly increasing in t for each fixed u.

Lecture 17

October 19

17.1 L-Unbiased Decision Rules

δ is **L-unbiased** if $\mathbb{E}_{\theta_0}[L(\theta_0, \delta(X))] \le \mathbb{E}_{\theta_0}[L(\theta, \delta(X))]$, e.g., if $L(\theta, d) = (\theta - d)^2$, then we recover the definition of an unbiased estimator. As another example, we can take $L(\theta, d) = \mathbb{1}\{\theta \notin d\}$.

17.2 Conditioning on Null Sufficient Statistics

We have been discussing exponential families with densities $p_{\theta,\zeta}(x) = \mathrm{e}^{\theta T(x) + \zeta^\mathsf{T} U(x) - A(\theta,\zeta)} h(x)$, where $\theta \in \mathbb{R}$, $\zeta \in \mathbb{R}^{s-1}$, and $H_0 : \theta = \theta_0$.

Example 17.1. Let $X \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$, $Y \sim \mathcal{N}_m(0, \sigma^2 I_m)$. Test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$, where $\sigma^2 > 0$ is unknown. Then,

$$\begin{aligned} p_{\mu,\sigma^2}(x, y) &= \mathrm{e}^{-\|x-\mu\|^2/(2\sigma^2) - \|y\|^2/(2\sigma^2)} \Big(\frac{1}{2\pi\sigma^2}\Big)^{(m+n)/2} \\ &= \mathrm{e}^{(\mu/\sigma^2)^\mathsf{T} x - (\|x\|^2 + \|y\|^2)/(2\sigma^2) - \|\mu\|^2/(2\sigma^2)} \Big(\frac{1}{2\pi\sigma^2}\Big)^{(m+n)/2} \\ &= \mathrm{e}^{\theta^\mathsf{T} T(x) - \zeta U(x,y) - \|\mu\|^2/(2\sigma^2)} \Big(\frac{1}{2\pi\sigma^2}\Big)^{(m+n)/2}. \end{aligned}$$

Here, $\theta \in \mathbb{R}^n$, $\zeta \in \mathbb{R}$.

$$\begin{bmatrix} X \\ Y \end{bmatrix} \mid U \overset{H_0}{\sim} \mathrm{Uniform}(\sqrt{U} S^{n+m-1})$$

$$\frac{1}{\sqrt{U}} \begin{bmatrix} X \\ Y \end{bmatrix} \overset{H_0}{\sim} \mathrm{Uniform}(S^{n+m-1}).$$

Choose some test statistic (a notion of X being "big"). If $R = \|X\|^2$, then reject when $\|X\|^2$ is large given U, or equivalently, reject when $\|X\|/\sqrt{U}$ is large, or equivalently reject for large

$$\frac{\|X\|^2}{\|X\|^2 + \|Y\|^2} = B.$$

Under H_0,

$$\|X\|^2 \sim \sigma^2 \chi_n^2 = \mathrm{Gamma}\Big(\frac{n}{2}, 2\sigma^2\Big),$$

which is independent of

$$\|Y\|^2 \sim \sigma^2 \chi^2_m = \text{Gamma}\Big(\frac{m}{2}, 2\sigma^2\Big)$$

so

$$B \overset{H_0}{\sim} \text{Beta}\Big(\frac{n}{2}, \frac{m}{2}\Big).$$

Then,

$$\mathbb{E}[B] = \frac{n}{m+n}.$$

Equivalently, reject for large

$$\frac{\|X\|^2/n}{\|Y\|^2/m} \sim F_{n,m}.$$

If $V \sim \chi^2_a \perp\!\!\!\perp W \sim \chi^2_b$, then

$$\frac{V/a}{W/b} \sim F_{a,b}.$$

For large n, m, the statistic is ≈ 1.

Example 17.2 (Non-Parametric 2-Sample Testing)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$, $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} Q$. Test $H_0 : P = Q$ versus $H_1 : P \neq Q$. Under H_0, $P = Q$ so (X, Y) is an i.i.d. sample from P of size $n + m$. Let $Z = (X, Y)$, that is:

$$Z_i = \begin{cases} X_i, & i \le n \\ Y_{i-n}, & i > n \end{cases}$$

Then $U(X, Y) = (Z_{(1)}, \ldots, Z_{(n+m)})$. Also,

$$(X, Y) \mid U(X, Y) \overset{H_0}{\sim} \text{Uniform}\{\pi Z : \pi \text{ is a permutation on } (1, \ldots, n+m)\}.$$

Choose any test statistic $T(X, Y)$, e.g., $T(X, Y) = |\bar{X} - \bar{Y}|$, or

$$T(X, Y) = \Big|\frac{1}{n}\sum_{i=1}^{n} \text{rank}(X_i) - \frac{1}{m}\sum_{i=1}^{m} \text{rank}(Y_i)\Big|$$

where $\text{rank}(Z_{(k)}) = k$. Reject when $T(X, Y)$ is (conditionally) large.

17.2.1 "Toy" Linear Model

Let

$$\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} \sim \mathcal{N}_3\Big(\begin{bmatrix} \mu_1 \\ \mu_2 \\ 0 \end{bmatrix}, \sigma^2 I_3\Big).$$

σ^2 is unknown. Test $H_0 : \mu_2 = 0$ versus $H_1 : \mu_2 \neq 0$.

$$p_{\mu_1,\mu_2,\sigma^2}(z) = \mathrm{e}^{-\|z-\mu\|^2/(2\sigma^2)}\Big(\frac{1}{2\pi\sigma^2}\Big)^{3/2}$$

$$\propto_z \exp\Big\{\frac{\mu_2}{\sigma^2} z_2 + \frac{\mu_1}{\sigma^2} z_1 - \frac{1}{2\sigma^2}\|z\|^2\Big\}.$$

Condition on $U = (Z_1, \|Z\|^2)$. Equivalently, condition on

$$(Z_1, \overbrace{Z_2^2 + Z_3^2}^{R^2}).$$

Note that $(Z_2, Z_3) \perp\!\!\!\perp Z_1$. Conditional on U, $Z \overset{H_0}{\sim} \text{Uniform}((Z_1, 0, 0) + RS^1)$. Reject when $|Z_2|$ is large. If $\mu_2 \gg 0$, then $Z_2^2 \gg Z_3^2$. In this case, $Z \approx (Z_1, R, 0)$. If $\mu_2 \ll 0$, then $Z \approx (Z_1, -R, 0)$. Rejecting when $|Z_2|$ is large is equivalent to rejecting when $Z_2^2/Z_3^2 \overset{H_0}{\sim} F_{1,1}$ is large.

Lecture 18

October 24

18.1 Testing in the General Linear Model

18.1.1 Review

Example 18.1. If $X \sim \mathcal{N}_n(\mu, \sigma^2 I_n)$, $Y \sim \mathcal{N}_m(0, \sigma^2 I_m)$ ($X \perp\!\!\!\perp Y$), $H_0 : \mu = 0$, $H_1 : \mu \neq 0$, and σ^2 is unknown, then under H_0, $\|X\|_2^2 \sim \sigma^2 \chi_n^2$, $\|Y\|_2^2 \sim \sigma^2 \chi_m^2$. Also,

$$\frac{\|X\|_2^2}{\|X\|_2^2 + \|Y\|_2^2} \sim \text{Beta}\Big(\frac{n}{2}, \frac{m}{2}\Big),$$
$$\frac{\|X\|_2^2/n}{\|Y\|_2^2/m} \sim F_{n,m}.$$

We can think of

$$\hat{\sigma}^2 = \frac{\|Y\|_2^2}{m}.$$

If σ^2 is known, then we would use

$$\frac{\|X\|_2^2/n}{\sigma^2} \sim \frac{\chi_n^2}{n}.$$

Under H_1,

$$\frac{\|X\|_2^2}{\sigma^2} \sim \text{nc}\chi_n^2\Big(\frac{\|\mu\|_2^2}{\sigma^2}\Big).$$

Example 18.2. For

$$\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ 0 \end{bmatrix}, \sigma^2 I_3\right),$$

$H_0 : \mu_2 = 0$, $H_1 : \mu_2 \neq 0$, then

$$\frac{Z_2^2}{Z_3^2} \sim F_{1,1}.$$

18.1.2 General Linear Model

Basic setup. Observe $Y \sim \mathcal{N}_n(\theta, \sigma^2 I_n)$, where $\sigma^2 > 0$ is possibly unknown. The models/null hypotheses are framed in terms of linear constraints on $\theta \in \mathbb{R}^n$. $\mathcal{P}$ puts $\theta \in \Theta$, for some d-dimensional affine space, and $H_0 : \theta \in \Theta_0 \subseteq \Theta$, where Θ_0 is a d_0-dimensional affine space.

Example 18.3 (One-Sample Testing)**.** $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$, or equivalently we have $Y \sim \mathcal{N}_n(\mu 1_n, \sigma^2 I_n)$, where

$$1_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Let $\theta = \mu 1_n$. Test $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$. Here, $\Theta = 1_n \mathbb{R}$, with $d = 1$, and $\Theta_0 = \{0\}$, with $d_0 = 0$.

Example 18.4 (k-Way ANOVA)**.** Let $Y_{i,j} \sim \mathcal{N}(\mu_j, \sigma^2)$, for $j = 1, \ldots, k$ and $i = 1, \ldots, n_j$. Test $H_0 : \mu_1 = \cdots = \mu_k$. Let $n_+ = \sum_{j=1}^k n_j$.

$$Y = \begin{bmatrix} Y_{1,1} \\ \vdots \\ Y_{1,n_1} \\ Y_{2,1} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{k,n_k} \end{bmatrix} \sim \mathcal{N}_{n_+}(\theta, \sigma^2 I_{n_+})$$

and

$$\theta = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_2 \\ \vdots \\ \mu_k \\ \vdots \\ \mu_k \end{bmatrix} \in \mathbb{R}^{n_+}.$$

Then,

$$\Theta = \operatorname{span}\left\{ \begin{bmatrix} 1_{n_1} \\ 0_{n_2} \\ \vdots \\ 0_{n_k} \end{bmatrix}, \begin{bmatrix} 0_{n_1} \\ 1_{n_2} \\ \vdots \\ 0_{n_k} \end{bmatrix}, \ldots, \begin{bmatrix} 0_{n_1} \\ \vdots \\ 0_{n_{k-1}} \\ 1_{n_k} \end{bmatrix} \right\}$$

with $\dim \Theta = k$. Then, $\Theta_0 = \operatorname{span} 1_{n_+}$ and $\dim \Theta_0 = 1$.

Example 18.5 (Linear Regression)**.** Let $X \in \mathbb{R}^{n\times d}$, $d < n$, and X has full column rank. Then, $Y_i \sim \mathcal{N}(x_i^\mathsf{T}\beta, \sigma^2)$, where x_i^T is the ith row of X, and thus $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$. Test the null hypothesis $H_0 : \beta_{d-s+1} = \beta_{d-s+2} = \cdots = \beta_d = 0$ (s of them). Here, $\Theta = \operatorname{span} X$, $\Theta_0 = \operatorname{span} X_{1:(d-s)}$, and $d_0 = d - s$.

This example subsumes the previous examples. For one-sample testing, take $X = 1_n$, $s = 1$. For k-way ANOVA,

$$X = \left(1_{n_+}, \begin{bmatrix} 1_{n_1} \\ 0_{n_2} \\ \vdots \\ 0_{n_k} \end{bmatrix}, \ldots, \begin{bmatrix} 0_{n_1} \\ \vdots \\ 1_{n_{k-1}} \\ 0_{n_k} \end{bmatrix}\right).$$

Here $s = k - 1$ and

$$\theta = X \begin{bmatrix} \mu_k \\ \mu_1 - \mu_k \\ \vdots \\ \mu_{k-1} - \mu_k \end{bmatrix}.$$

18.1.3 General Strategy

Rotate Y via an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$.

$$Q = \begin{bmatrix} Q_0 & Q_1 & Q_\mathrm{r} \end{bmatrix}$$

where $Q_0 \in \mathbb{R}^{n\times d_0}$ is a basis for Θ_0, $Q_1 \in \mathbb{R}^{n\times(d-d_0)}$ is a basis for $\Theta \cap \Theta_0^\perp$, and $Q_\mathrm{r} \in \mathbb{R}^{n\times(n-d)}$ is a basis for $\Theta^\perp$. Then, let

$$\begin{bmatrix} Z_0 \\ Z_1 \\ Z_\mathrm{r} \end{bmatrix} = Z = Q^\mathsf{T} Y \sim \mathcal{N}\left(\begin{bmatrix} Q_0^\mathsf{T}\theta \\ Q_1^\mathsf{T}\theta \\ Q_\mathrm{r}^\mathsf{T}\theta \end{bmatrix}, \sigma^2 I_n\right) = \mathcal{N}_n\left(\begin{bmatrix} \nu_0 \\ \nu_1 \\ \nu_\mathrm{r} \end{bmatrix}, \sigma^2 I_n\right).$$

So,

$$\begin{aligned} \nu_0 &= Q_0^\mathsf{T}\theta \in \mathbb{R}^{d_0}, \\ \nu_1 &= Q_1^\mathsf{T}\theta \in \mathbb{R}^{d-d_0} = \mathbb{R}^s, \\ \nu_\mathrm{r} &= Q_\mathrm{r}^\mathsf{T}\theta \in \mathbb{R}^{n-d}. \end{aligned}$$

The model puts $\theta \in \Theta$, or equivalently, $\nu_\mathrm{r} = 0$. Then, H_0 puts $\theta \in \Theta_0$, or equivalently, $\nu_1 = \nu_\mathrm{r} = 0$.

	H_0	H_1
ν_0	any $\in \mathbb{R}^{d_0}$	any $\in \mathbb{R}^{d_0}$
ν_1	0_{d-d_0}	$\neq 0_{d-d_0}$
ν_r	0_{n-d}	0_{n-d}

Here, $H_0 : \nu_1 = 0$.

σ^2 *known*: If $s = 1$, then

$$\begin{aligned} \frac{Z_1}{\sigma} &\sim \mathcal{N}\left(\frac{\nu_1}{\sigma}, 1\right) \\ &\overset{H_0}{\sim} \mathcal{N}(0, 1). \end{aligned}$$

This is the Z-test. If $s > 1$,

$$\frac{\|Z_1\|_2^2}{\sigma^2} \overset{H_0}{\sim} \chi_s^2.$$

σ^2 *unknown*: Let

$$\hat{\sigma}^2 = \frac{\|Z_{\mathrm{r}}\|_2^2}{n-d}.$$

For $s = 1$,

$$\frac{Z_1}{\hat{\sigma}} \overset{H_0}{\sim} t_{n-d}.$$

For $s > 1$,

$$\frac{\|Z_1\|_2^2/s}{\hat{\sigma}^2} \overset{H_0}{\sim} F_{s,n-d}.$$

Equivalently,

$$\frac{\|Z_1\|_2^2}{\|Z_1\|_2^2 + \|Z_{\mathrm{r}}\|_2^2} \sim \mathrm{Beta}\Big(\frac{s}{2}, \frac{n-d}{2}\Big).$$

For one-sample testing,

$$Q_0 = \varnothing, \qquad Q_1 = \frac{1}{\sqrt{n}} 1_n, \qquad Q_r = \text{completion to } \mathbb{R}^n.$$

For regression with $s = 1$, $X \in \mathbb{R}^{n\times d}$, and $H_0 : \beta_d = 0$, $d_0 = d - 1$, then $Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$, and

$$Q_0 = \begin{bmatrix} \frac{X_1}{\|X_1\|_2} & \frac{X_{2,\perp}}{\|X_{2,\perp}\|_2} & \cdots & \frac{X_{d-1,\perp}}{\|X_{d-1,\perp}\|_2} \end{bmatrix}$$

where $X_{j,\perp} = \pi^{\perp}_{\mathrm{span}(X_1,\ldots,X_{j-1})} X_j = (I - Q_{0,1:(j-1)} Q_{0,1:(j-1)}^{\mathsf{T}}) X_j$.

$$Q_1 = \begin{bmatrix} \frac{X_{d,\perp}}{\|X_{d,\perp}\|_2} \end{bmatrix}$$

and Q_{r} is the completion to $\mathbb{R}^n$. Then, $\|Z_{\mathrm{r}}\|_2^2 = \|Y\|_2^2 - \|Z_0\|_2^2 - \|Z_1\|_2^2$. Also,

$$Z_1 = \frac{X_{d,\perp}^{\mathsf{T}} Y}{\|X_{d,\perp}\|_2}.$$

We can also write

$$\begin{aligned} \|Z_{\mathrm{r}}\|_2^2 &= \|Y - \overbrace{\pi_{\mathrm{span}\, X} Y}^{\hat{Y}}\|_2^2 \\ &= \mathrm{RSS} \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \end{aligned}$$

Lecture 19

October 26

19.1 Motivation for Large-Sample Theory

Example 19.1. Suppose $X \sim \text{Binomial}(n, \theta)$, and $n = 2000$. We want a CI for θ. We can use

$$\begin{aligned} X &\approx \mathcal{N}\big(n\theta, n\theta(1-\theta)\big) \\ &\approx \mathcal{N}\left(n\theta, n\frac{X}{n}\left(1 - \frac{X}{n}\right)\right), \\ \frac{X - n\theta}{\sqrt{X(1 - X/n)}} &\approx \mathcal{N}(0,1). \end{aligned}$$

Then,

$$\text{CI} = \frac{X}{n} \pm z_{\alpha/2}\sqrt{\frac{(X/n)(1 - X/n)}{n}}.$$

Unless $X/n \approx 0$ or 1, the answer is approximately the same as the exact CI.

Example 19.2. Let $X_i \overset{\text{i.i.d.}}{\sim} p_\theta$ for $i = 1, \ldots, n$ for a "generic" p_θ (under conditions). The MLE gives the approximately optimal estimator. Tests and confidence intervals which are based on the likelihood are approximately optimal.

19.2 Convergence in Probability

Definition 19.3. A sequence of random variables $X_1, X_2, \ldots$ **converges in probability to** X if, for all $\varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$. This is written as $X_n \xrightarrow{\mathbb{P}} X$.

Usually $X = c \in \mathbb{R}$ (constant).

Proposition 19.4 (Chebyshev)**.** *For any random variable X, constant $a > 0$,*

$$\mathbb{P}(|X| > a) \le \frac{\mathbb{E}[X^2]}{a^2}.$$

Proof. Since

$$\mathbb{1}\{|X| > a\} \overset{\text{a.s.}}{\leq} \frac{X^2}{a^2},$$

take expectations. □

Corollary 19.5.

$$\mathbb{P}(|X - \mathbb{E}[X]| > a) \leq \frac{\operatorname{var} X}{a^2}.$$

Corollary 19.6. *If* $\mathbb{E}[X_n] = 0$ *for all* n *and* $\operatorname{var} X_n \to 0$, *then* $X_n \xrightarrow{\mathbb{P}} 0$.

More generally, convergence in probability is defined as $\mathbb{P}(\|X_n - X\| > \varepsilon) \to 0$ for all $\varepsilon > 0$.

Proposition 19.7. *Suppose* $X_1, X_2, \ldots \overset{i.i.d.}{\sim} P$, $\mathbb{E}[X_i] = \mu$, $\operatorname{var} X_i = \sigma^2$. *Then,*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu.$$

Proof. $\mathbb{E}[\overline{X}_n] = \mu$ for all n, and

$$\operatorname{var} \overline{X}_n = \frac{\sigma^2}{n} \to 0.$$ □

Proposition 19.8 (Continuous Mapping Theorem)**.** *If* f *is continuous at* c *and* $X_n \xrightarrow{\mathbb{P}} c$, *then*

$$f(X_n) \xrightarrow{\mathbb{P}} f(c).$$

Proof. Fix $\varepsilon > 0$. There exists $\delta(\varepsilon) > 0$ with $|X_n - c| \leq \delta(\varepsilon) \implies |f(X_n) - f(c)| \leq \varepsilon$. Then, $\mathbb{P}(|f(X_n) - f(c)| > \varepsilon) \leq \mathbb{P}(|X_n - c| > \delta(\varepsilon)) \to 0$. □

Notation: $\xrightarrow{P_\theta}$ means convergence under θ.

Definition 19.9. A sequence of estimators $\delta_n(X^{(n)})$ for $n \geq 1$ is **consistent for** $g(\theta)$ if

$$\delta_n(X^{(n)}) \xrightarrow{P_\theta} g(\theta), \qquad \forall \theta \in \Theta.$$

Recall that $\operatorname{MSE}(\theta, \delta_n) = (\operatorname{bias}_\theta \delta_n(X^{(n)}))^2 + \operatorname{var}_\theta \delta_n(X^{(n)})$. If $\operatorname{bias}_\theta \delta_n(X^{(n)}) \to 0$ and $\operatorname{var}_\theta \delta_n(X^{(n)}) \to 0$, then $\operatorname{MSE}(\theta, \delta_n) \to 0$.

$$\mathbb{P}(|\delta_n(X^{(n)}) - \theta| > \varepsilon) \leq \frac{\operatorname{MSE}(\theta, \delta_n)}{\varepsilon^2} \to 0, \qquad \forall \varepsilon > 0.$$

Let $\delta_n(X^{(n)}) = g(\theta) + B_n k_n$, where $B_n \sim \operatorname{Bernoulli}(\pi_n)$ and $\pi_n \to 0$. If we take

$$k_n = \frac{1}{\pi_n},$$

then $\operatorname{bias}_\theta \delta_n(X^{(n)}) = 1$ for all n. For $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}(|\delta_n(X^{(n)}) - g(\theta)| > \varepsilon) &\leq \mathbb{P}\big(\delta_n(X^{(n)}) \neq g(\theta)\big) \\ &= \pi_n.\end{aligned}$$

19.3 Convergence in Distribution

(a.k.a. **weak convergence**)

Example 19.10.

$$\frac{X}{n} \xrightarrow{P_\theta} \theta$$

for the binomial example. $X \approx \mathcal{N}(n\theta, n\theta(1-\theta))$ is a much more precise and useful statement.

Definition 19.11. A sequence of random variables $X_1, X_2, \ldots$ **converges in distribution** to a RV X with CDF F if $F_n(x) \xrightarrow{n\to\infty} F(x)$ for all x such that F is continuous at x. Notation: $X_n \Rightarrow X$ or $X_n \xrightarrow{\mathsf{d}} X$ or $X_n \xrightarrow{\mathsf{d}} F$ or $X_n \xrightarrow{\mathsf{d}} \mathcal{N}(0,1)$.

Theorem 19.12. *$X_n \Rightarrow X$ iff $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded continuous f.*

This definition generalizes to vectors, matrices, ...

Corollary 19.13. *If g is continuous and $X_n \Rightarrow X$, then $g(X_n) \Rightarrow g(X)$.*

Proof. If f is bounded and continuous, then $f \circ g$ is bounded and continuous, so

$$\mathbb{E}\big[f\big(g(X_n)\big)\big] \to \mathbb{E}\big[f\big(g(X)\big)\big]. \qquad \square$$

Theorem 19.14 (CLT)**.** *If $X_i \sim (\mu, \sigma^2)$ [notation: $\mathbb{E}[X_i] = \mu$, $\operatorname{var} X_i = \sigma^2$] and*

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

then $\sqrt{n}(\overline{X}_n - \mu) \Rightarrow \mathcal{N}(0, \sigma^2)$.

Less formal:

$$\overline{X}_n \approx \mathcal{N}\Big(\mu, \frac{\sigma^2}{n}\Big).$$

Theorem 19.15 (Slutsky)**.** *If $X_n \Rightarrow X$, $Y_n \xrightarrow{\mathbb{P}} c$, then:*

- $X_n + Y_n \Rightarrow X + c$*;*
- $X_n Y_n \Rightarrow cX$*;*
- $X_n / Y_n \Rightarrow X/c$ *if* $c \neq 0$.

Example 19.16. $X_n \sim \text{Binomial}(n, \theta)$. Write $X_n = \sum_{i=1}^n B_i$ where $B_1, B_2, \ldots \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$. Here, $B_i \sim (\theta, \theta(1-\theta))$. The estimator is

$$\hat{\theta} = \frac{X_n}{n}.$$

The LLN 19.7 implies $\hat{\theta} \xrightarrow{\mathbb{P}} \theta$. The CLT 19.14 implies $\sqrt{n}(\hat{\theta} - \theta) \overset{P_\theta}{\Rightarrow} \mathcal{N}(0, \theta(1-\theta))$. If we combine the

LLN 19.7, the CLT 19.14, and Slutsky's Theorem 19.15,

$$\frac{\sqrt{n}(\hat{\theta}-\theta)}{\sqrt{\hat{\theta}(1-\hat{\theta})}} \overset{P_\theta}{\Rightarrow} \mathcal{N}(0,1).$$

Then, our confidence interval is

$$\hat{\theta} \pm \frac{z_{\alpha/2}\sqrt{\hat{\theta}(1-\hat{\theta})}}{\sqrt{n}}.$$

Thus,

$$\begin{aligned}\mathbb{P}_\theta\left(\theta > \hat{\theta} + \frac{z_{\alpha/2}\sqrt{\hat{\theta}(1-\hat{\theta})}}{\sqrt{n}}\right) &= \mathbb{P}_\theta\left(\frac{\sqrt{n}(\theta-\hat{\theta})}{\sqrt{\hat{\theta}(1-\hat{\theta})}} > z_{\alpha/2}\right) \\ &\to 1-\Phi(z_{\alpha/2}) = \frac{\alpha}{2}.\end{aligned}$$

19.3.1 Delta Method

Theorem 19.17 (Delta Method)**.** *If $\sqrt{n}(X_n-\mu) \Rightarrow \mathcal{N}(0,\sigma^2)$, and f is differentiable at μ, then $\sqrt{n}(f(X_n)-f(\mu)) \Rightarrow \mathcal{N}(0,\sigma^2 f'(\mu)^2)$.*

Proof. $f(X_n) = f(\mu) + f'(\mu)(X_n-\mu) + o(X_n-\mu)$, so

$$\begin{aligned}\sqrt{n}\big(f(X_n)-f(\mu)\big) &= \underbrace{f'(\mu)\sqrt{n}(X_n-\mu)}_{\Rightarrow \mathcal{N}(0,\sigma^2 f'(\mu)^2)} + \underbrace{\sqrt{n}o(X_n-\mu)}_{\overset{\mathbb{P}}{\to} 0} \\ &\Rightarrow \mathcal{N}\big(0,\sigma^2 f'(\mu)^2\big).\end{aligned}$$

□

Lecture 20

October 31

20.1 Maximum Likelihood Estimation

For a generic dominated family $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$, the **maximum likelihood estimator (MLE)** is

$$\begin{aligned}\hat{\theta}_{\text{MLE}}(X) &= \arg\max_{\theta \in \Theta} p_\theta(X) \\ &= \arg\max_{\theta \in \Theta} \ell(\theta; X).\end{aligned}$$

Remark 1. The maximizer may not exist, or be unique. It may not be computable.

Remark 2: The MLE of $g(\theta)$ is $g(\hat{\theta}_{\text{MLE}})$.

Example 20.1.

$$\begin{aligned}p_\eta(x) &= \mathrm{e}^{\eta^\mathsf{T} T(x) - A(\eta)} h(x), \\ \ell(\eta; X) &= \eta^\mathsf{T} T(X) - A(\eta) + \log h(X), \\ \nabla \ell(\eta; X) &= T(X) - \nabla A(\eta).\end{aligned}$$

Set $\nabla \ell = 0$ so

$$\begin{aligned}T(X) &= \nabla A(\hat{\eta}) \\ &= \mathbb{E}_{\hat{\eta}}[T(X)].\end{aligned}$$

If there exists $\eta \in \Xi$ with $\mathbb{E}_\eta[T(X)] = T(X)$, then it is the MLE, since

$$\nabla^2 \ell(\eta; X) = -\nabla^2 A(\eta) = -\operatorname{var}_\eta T(X)$$

is negative-definite, unless there exists ν with $\nu^\mathsf{T} T(X) \overset{\mathcal{P}\text{-a.s.}}{=} 0$. If the family is not overparameterized, then we can define the inverse of $\mu(\eta) = \nabla A(\eta)$ as $\psi = \mu^{-1}$, so $\hat{\mu}_{\text{MLE}} = \psi(T)$.

Example 20.2. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\theta)$, $\mathbb{E}_\theta[X_i] = \operatorname{var}_\theta X_i = \theta$. The sufficient statistic is $T(X) = \sum_{i=1}^n X_i$. Then, we take

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \frac{1}{n}\sum_{i=1}^n X_i \\ &\approx \mathcal{N}\Big(\theta, \frac{\operatorname{var}_\theta X_i}{n}\Big)\end{aligned}$$

$$= \mathcal{N}\Big(\theta, \frac{\theta}{n}\Big).$$

since $\mathbb{E}_\theta[T] = n\theta$. More rigorously, $\sqrt{n}(\hat{\theta} - \theta) \Rightarrow \mathcal{N}(0, \theta)$. The natural parameter is $\eta = \log\theta$, so

$$\begin{aligned}\hat{\eta}_{\text{MLE}}(X) &= \log\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big) \\ &\approx \mathcal{N}\Big(\log\theta, \frac{\theta}{n}\Big[\frac{\mathrm{d}}{\mathrm{d}\theta}\log\theta\Big]^2\Big) \\ &= \mathcal{N}\Big(\log\theta, \frac{1}{\theta n}\Big) \\ &= \mathcal{N}\Big(\eta, \frac{\mathrm{e}^{-\eta}}{n}\Big).\end{aligned}$$

Note: For all finite n, $\theta > 0$,

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

Example 20.3 (General One-Parameter Exponential Family)**.** Let $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{e}^{\eta T(x) - A(\eta)} h(x)$, with one parameter $\eta \in \Xi \subseteq \mathbb{R}$. Here, $\mu(\eta) = A'(\eta) = \mathbb{E}_\eta[T(X)]$ is the mean parameter for X_1. Then, $X = (X_1, \ldots, X_n)$ is an exponential family with natural parameter η, sufficient statistic $\sum_{i=1}^n T(X_i)$, and mean parameter $n\mu(\eta)$. Then,

$$\begin{aligned}\mu(\hat{\eta}) &= \frac{1}{n}\sum_{i=1}^n T(X_i), \\ \hat{\eta} &= \psi\Big(\frac{1}{n}\sum_{i=1}^n T(X_i)\Big).\end{aligned}$$

Asymptotically,

$$\begin{aligned}\hat{\eta} &= \frac{1}{n}\sum_{i=1}^n T(X_i) \\ &\approx \mathcal{N}\Big(\mu(\eta), \frac{\operatorname{var}_\eta T(X_i)}{n}\Big) \\ &= \mathcal{N}\Big(\mu(\eta), \frac{A''(\eta)}{n}\Big), \\ \psi(\hat{\eta}) &= \mathcal{N}\Big(\psi\big(\mu(\eta)\big), \frac{A''(\eta)}{n}\psi'\big(\mu(\eta)\big)^2\Big) \\ &= \mathcal{N}\Big(\eta, \frac{1}{A''(\eta) n}\Big), \\ \psi'\big(\mu(\eta)\big) &= \frac{1}{\mu'(\eta)} = \frac{1}{A''(\eta)}\end{aligned}$$

(use the Chain Rule on $\psi(\mu(\eta)) = \eta$). Thus,

$$\sqrt{n}(\hat{\eta} - \eta) \Rightarrow \mathcal{N}\Big(0, \frac{1}{A''(\eta)}\Big).$$

20.2 Asymptotic Relative Efficiency

Previously, we compared estimators via, e.g., MSE, but for any Gaussian estimators, more "concrete" comparisons are possible.

Definition 20.4. Suppose $\hat{\theta}^{(1)}$, $\hat{\theta}^{(2)}$ are asymptotically normal with $\sqrt{n}(\hat{\theta}^{(i)} - \theta) \Rightarrow \mathcal{N}(0, \sigma_i^2)$. The **asymptotic relative efficiency (ARE)** of $\hat{\theta}^{(2)}$ with respect to $\hat{\theta}^{(1)}$ is σ_1^2/σ_2^2.

Example 20.5. If $\sigma_2^2 = 2\sigma_1^2$, then we say $\hat{\theta}^{(2)}$ is 50% as efficient as $\hat{\theta}^{(1)}$.

Interpretation. For large n, if

$$\frac{\sigma_1^2}{\sigma_2^2} = \gamma < 1,$$

then

$$\begin{aligned}\hat{\theta}^{(2)}(X_1, \dots, X_n) &\approx \mathcal{N}\Big(\theta, \frac{\sigma_2^2}{n}\Big) \\ &\overset{\mathsf{d}}{\approx} \hat{\theta}^{(1)}(X_1, \dots, X_{\lfloor \gamma n \rfloor}) \\ &\approx \mathcal{N}\Big(\theta, \frac{\sigma_1^2}{\gamma n}\Big) = \mathcal{N}\Big(\theta, \frac{\sigma_2^2}{n}\Big).\end{aligned}$$

Asymptotically, using $\hat{\theta}^{(2)}$ instead of $\hat{\theta}^{(1)}$ is equivalent to throwing away a $1 - \gamma$ fraction of the data.

Example 20.6 (Sample Median vs. Sample Mean)**.** Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f(x - \theta)$ be symmetric. Keener 8.4 shows that if $\tilde{X}_n$ is the sample median, then

$$\begin{aligned}\sqrt{n}(\tilde{X}_n - \theta) &\Rightarrow \mathcal{N}\Big(0, \frac{1}{4f(0)^2}\Big), \\ \sqrt{n}(\overline{X}_n - \theta) &\Rightarrow \mathcal{N}(0, \operatorname{var} X_1).\end{aligned}$$

Gaussian: Let $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$.

$$\begin{aligned}&\frac{1}{4f(0)^2} = \frac{1}{4(1/(2\pi\sigma^2))} = \frac{\sigma^2\pi}{2}, \\ &\operatorname{var} X_i = \sigma^2,\end{aligned}$$

so the median is $2/\pi \approx 64\%$ as efficient.

Laplace: If

$$X_i \overset{\text{i.i.d.}}{\sim} \frac{1}{2\sigma} e^{-|x|/\sigma}$$

then

$$\begin{aligned}&\frac{1}{4f(0)^2} = \frac{1}{4(1/(4\sigma^2))} = \sigma^2, \\ &\operatorname{var} X_i = 2\sigma^2,\end{aligned}$$

so the mean is $\sigma^2/(2\sigma^2) \approx 50\%$ as efficient.

Lecture 21

November 2

21.1 Asymptotic Distribution of the MLE

Setting: $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta$, "smooth" in θ. Let $\ell_1(\theta; X_i) = \log p_\theta(X_i)$ and

$$\begin{aligned} J_1(\theta) &= \operatorname{var}_\theta \nabla \ell_1(\theta; X_1) \\ &= -\mathbb{E}_\theta[\nabla^2 \ell_1(\theta; X_1)], \\ J(\theta) &= \operatorname{var}_\theta \nabla \ell(\theta; X_1, \dots, X_n) \\ &= n J_1(\theta). \end{aligned}$$

Recall that $\mathbb{E}_\theta[\nabla \ell(\theta; X)] = 0$. We say that an estimator $\hat{\theta}_n$ is **asymptotically efficient** if

$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow \mathcal{N}\big(0, J_1(\theta)^{-1}\big).$$

Today, we will see that under general conditions, $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \Rightarrow \mathcal{N}(0, J_1(\theta)^{-1})$. Also,

$$\sqrt{n}\big(g(\hat{\theta}_{\text{MLE}}) - g(\theta)\big) \Rightarrow \mathcal{N}\big(0, \nabla g(\theta)^\mathsf{T} J_1(\theta)^{-1} \nabla g(\theta)\big)$$

(if g is differentiable).

"Proof" in One Dimension. Let θ_0 denote the true value.

$$\begin{aligned} \frac{1}{\sqrt{n}} \ell'(\theta_0; X) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \ell_1'(\theta_0; X_i) \\ &\overset{P_{\theta_0}}{\Longrightarrow} \mathcal{N}\big(0, J_1(\theta_0)\big) \end{aligned}$$

(by the CLT 19.14). Also,

$$\frac{1}{n} \ell'(\theta_0; X) \xrightarrow{P_{\theta_0}} -J_1(\theta_0)$$

(by the LLN 19.7). Then,

$$\begin{aligned} 0 &= \ell'(\hat{\theta}; X) \\ &= \ell'(\theta_0; X) + (\hat{\theta} - \theta_0)\ell''(\theta_0) + o(|\hat{\theta} - \theta_0|), \\ \sqrt{n}(\hat{\theta} - \theta_0) &\approx \frac{(1/\sqrt{n})\ell'(\theta_0; X)}{-(1/n)\ell''(\theta_0; X)}. \end{aligned}$$

Since the numerator $\xRightarrow{P_{\theta_0}} \mathcal{N}(0, J_1(\theta_0))$ and the denominator $\xrightarrow{P_{\theta_0}} J_1(\theta_0)$, then

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \mathcal{N}\big(0, J_1(\theta_0)^{-1}\big). \qquad \square$$

Remark 1. We need the MLE to be consistent.

Remark 2: We need the second derivative to have finite expectation near θ_0.

21.2 Asymptotic Distribution of the MLE, Take 2

Theorem 21.1 (Keener Theorem 9.14). *Setup: $X_1, \dots, X_n \overset{i.i.d.}{\sim} p_\theta$ are from a dominated family*

$$\mathcal{P} = \{p_\theta \mid \theta \in \Theta \subseteq \mathbb{R}\}.$$

1. *Twice-differentiable log-likelihood: For all $\theta \in \Theta$, for all $x \in \mathcal{X}$, $p_\theta(x) > 0$ and $\ell(\theta; x)$ has two continuous derivatives.*
2. *Fisher information: $\mathbb{E}_\theta[\ell'(\theta; X)] = 0$ and $\operatorname{var}_\theta \ell'(\theta; X) = -\mathbb{E}_\theta[\ell''(\theta; X)] \in (0, \infty)$.*
3. *"Tame" second derivative (locally): For all $\theta \in \Theta^\circ$, there exists $\varepsilon > 0$ such that*

$$\mathbb{E}_\theta\Big[\sup_{\tilde{\theta} \in [\theta-\varepsilon, \theta+\varepsilon]} |\ell_1''(\tilde{\theta}; X)|\Big] < \infty.$$

4. *The MLE is consistent.*

Then, for all $\theta \in \Theta^\circ$, $\sqrt{n}(\hat{\theta} - \theta) \Rightarrow \mathcal{N}(0, J_1(\theta)^{-1})$.

Lemma 21.2. *Suppose $X_n \Rightarrow X$ and $\mathbb{P}(B_n) \to 1$ as $n \to \infty$. Then, for arbitrary random variables Z_n, $n \geq 1$, $Y_n = X_n \mathbb{1}_{B_n} + Z_n \mathbb{1}_{B_n^c} \Rightarrow X$.*

Proof. Fix $\varepsilon > 0$. $\mathbb{P}(|Z_n \mathbb{1}_{B_n^c}| > \varepsilon) \leq \mathbb{P}(B_n^c) \to 0$. Also, $\mathbb{P}(|\mathbb{1}_{B_n} - 1| > \varepsilon) \leq \mathbb{P}(B_n^c) \to 0$. Apply Slutsky's Theorem 19.15. $\square$

Proof of 21.1. Fix $\theta_0 \in \Theta^\circ$, choose $\varepsilon > 0$ for which

(a) $[\theta_0 - \varepsilon, \theta_0 + \varepsilon] \subseteq \Theta^\circ$ and

(b) $\mathbb{E}[\sup_{\tilde{\theta} \in [\theta_0 - \varepsilon, \theta_0 + \varepsilon]} |\ell''(\tilde{\theta}, X)|] < \infty$ by 3.

Let $B_n = \{|\hat{\theta}_n - \theta_0| < \varepsilon\}$. Then, $\mathbb{P}(B_n) \to 1$ by 4. On B_n, we have

$$0 = \ell'(\hat{\theta}_n; X) = \ell'(\theta_0; X) + (\hat{\theta}_n - \theta_0)\ell''(\tilde{\theta}_n; X)$$

for some $\tilde{\theta}_n$ between $[\theta_0, \hat{\theta}_n]$. Hence,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{(1/\sqrt{n})\ell'(\theta_0; X)}{-(1/n)\ell''(\tilde{\theta}_n; X)}$$

and the numerator $\xRightarrow{P_{\theta_0}} \mathcal{N}(0, J_1(\theta_0))$. We want the denominator $\xrightarrow{P_{\theta_0}} J_1(\theta_0)$. If $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$, then

$\tilde{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$ also. This implies

$$\frac{1}{n}\ell''(\tilde{\theta}_n; X) \xrightarrow{P_{\theta_0}} \mathbb{E}_{\theta_0}[\ell_1''(\theta_0; X_1)]$$

(for reasons we will defer until next time). The behavior on B_n^{c} does not affect the limit. □

21.2.1 Dimension $d > 1$

$$\begin{aligned}
\frac{1}{\sqrt{n}}\nabla\ell(\theta_0; X) &\overset{P_{\theta_0}}{\Longrightarrow} \mathcal{N}_d\big(0, J_1(\theta_0)\big), \\
-\frac{1}{n}\nabla^2\ell(\theta_0; X) &\xrightarrow{P_{\theta_0}} J_1(\theta_0), \\
0 &= \nabla\ell(\theta_0; X) + \nabla^2\ell(\theta_0; X)(\hat{\theta}_n - \theta_0) + o(\|\hat{\theta}_n - \theta_0\|), \\
\sqrt{n}(\hat{\theta}_n - \theta_0) &\approx \underbrace{\Big(-\frac{1}{n}\nabla^2\ell(\theta_0; X)\Big)}_{\xrightarrow{P_{\theta_0}} J_1(\theta_0)}{}^{-1} \underbrace{\frac{1}{\sqrt{n}}\nabla\ell(\theta_0; X)}_{\overset{P_{\theta_0}}{\Longrightarrow}\mathcal{N}(0, J_1(\theta_0))}.
\end{aligned}$$

Example 21.3 (Gaussian)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_0, 1)$. Then,

$$\begin{aligned}
\ell(\theta; X) &= \log\Big\{\Big(\frac{1}{\sqrt{2\pi}}\Big)^n \mathrm{e}^{-\sum_{i=1}^n (X_i - \theta)^2/2}\Big\} \\
&= n\overline{X}_n\theta - \frac{n\theta^2}{2} - \frac{n}{2}\log(2\pi) - \frac{\|X\|_2^2}{2}, \\
\ell'(\theta; X) &= n(\overline{X}_n - \theta) \sim \mathcal{N}(0, n), \\
\ell''(\theta; X) &= -n = -nJ_1(\theta), \\
\sqrt{n}(\underbrace{\hat{\theta}_n}_{\overline{X}_n} - \theta_0) &= \frac{(1/\sqrt{n})\ell'(\theta_0; X)}{-(1/n)\ell''(\theta_0; X)} \\
&\sim \mathcal{N}(0, 1)
\end{aligned}$$

since the numerator is $\sim \mathcal{N}(0, 1)$ and the denominator is 1.

Lecture 22

November 7

22.1 Consistency of MLE

Last time, we needed

$$-\frac{1}{n}\ell''(\tilde{\theta}_n; X) \xrightarrow{P_{\theta_0}} \mathbb{E}_{\theta_0}\Big[-\frac{1}{n}\ell''(\theta_0; X)\Big] = J(\theta_0).$$

We had $\tilde{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$ and

$$-\frac{1}{n}\ell''(\theta_0; X) \xrightarrow{P_{\theta_0}} J(\theta_0).$$

Setup: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$ for $\theta_0 \in \Theta$. Note that $\ell_n(\theta; X) = \sum_{i=1}^n \log p_\theta(X)$ and $\hat{\theta}_n = \arg\max_{\theta\in\Theta} \ell_n(\theta; X)$. We want $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$.

Recall the **Kullback-Leibler divergence**

$$D_{\mathrm{KL}}(\theta_0 \parallel \theta) = \mathbb{E}_{\theta_0}\Big[\log\frac{p_{\theta_0}(X_1)}{p_\theta(X_1)}\Big].$$

Then,

$$\begin{aligned}
-D_{\mathrm{KL}}(\theta_0 \parallel \theta) &\le \log \mathbb{E}_{\theta_0}\Big[\frac{p_\theta(X_1)}{p_{\theta_0}(X_1)}\Big] \\
&\le \log \int_{\{x: p_{\theta_0}(x) > 0\}} \frac{p_\theta}{p_{\theta_0}} p_{\theta_0}\,\mathrm{d}\mu \\
&\le \log 1 = 0.
\end{aligned}$$

Also, $-D_{\mathrm{KL}}(\theta_0 \parallel \theta) < 0$ unless $p_\theta = p_{\theta_0}$ (unless $P_\theta = P_{\theta_0}$). If $\mathcal{P}$ is identifiable (all P_θ are distinct), then $D_{\mathrm{KL}}(\theta_0 \parallel \theta) > 0$ if $\theta \ne \theta_0$. Write

$$\begin{aligned}
W_n(\theta) &= \frac{1}{n}\big(\ell_n(\theta; X) - \ell_n(\theta_0; X)\big) \\
&= \frac{1}{n}\sum_{i=1}^n \ell_1(\theta; X_i) - \frac{1}{n}\sum_{i=1}^n \ell_1(\theta_0; X_i), \\
\mathbb{E}_{\theta_0}[W_n(\theta)] &= -D_{\mathrm{KL}}(\theta_0 \parallel \theta).
\end{aligned}$$

Game Plan

1. We want $\sup_{\theta\in\Theta}|W_n(\theta) - \mathbb{E}_{\theta_0}[W_n(\theta)]| \xrightarrow{P_{\theta_0}} 0$.
2. Prove consistency for compact Θ.
3. Generalize to non-compact Θ.

22.2 Uniform Convergence of Random Functions (Stochastic Processes)

For a compact set K, let $C(K) = \{f : K \to \mathbb{R} : f \text{ continuous}\}$. For $f \in C(K)$, let $\|f\|_\infty = \sup_{t\in K}|f(t)|$. We say $f_n \to f$ in $\|\cdot\|_\infty$ if $\|f_n - f\|_\infty \to 0$ (**uniform convergence**).

Lemma 22.1 (Lemma 9.1 (Keener)). *Let $W \in C(K)$ be random with $\mathbb{E}[\|W\|_\infty] < \infty$, then $\mathbb{E}[W(t)]$ is continuous in t and $\sup_{t\in K} \mathbb{E}[\sup_{s:\|s-t\|<\varepsilon}|W(s) - W(t)|] \to 0$ as $\varepsilon \downarrow 0$.*

Theorem 22.2 (Weak Law). *Let $W_1, W_2, \ldots$ be in $C(K)$, where K is compact. Let $\mu(t) = \mathbb{E}[W(t)]$. Assume $\mathbb{E}[\|W\|_\infty] < \infty$. Let*

$$\overline{W}_n = \frac{1}{n}\sum_{i=1}^n W_i.$$

Then, $\|\overline{W}_n - \mu\|_\infty \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$.

Theorem 22.3 (Theorem 9.4 (Keener)). *Let G_n, $n \geq 1$, be random functions in $C(K)$, K is compact, and g be a fixed function in $C(X)$ with $\|G_n - g\|_\infty \xrightarrow{\mathbb{P}} 0$.*

1. *If $t_n \xrightarrow{\mathbb{P}} t^* \in K$, where t^* is fixed, then $G_n(t_n) \xrightarrow{\mathbb{P}} g(t^*)$.*
2. *If g is maximized at a unique value t^* and t_n maximizes G_n, then $t_n \xrightarrow{\mathbb{P}} t^*$.*
3. *If $K \subseteq \mathbb{R}$ and $g(t) = 0$ has a unique solution t^*, and t_n solves $G_n(t_n) = 0$, then $t_n \xrightarrow{\mathbb{P}} t^*$.*

Proof. 1.

$$\begin{aligned}|G_n(t_n) - g(t^*)| &\leq |G_n(t_n) - g(t_n)| + |g(t_n) - g(t^*)| \\ &\leq \underbrace{\|G_n - g\|_\infty}_{\xrightarrow{\mathbb{P}}0} + \underbrace{|g(t_n) - g(t^*)|}_{\xrightarrow{\mathbb{P}}0}.\end{aligned}$$

(This completes the proof from last time.)

2. Fix $\varepsilon > 0$ and let $K_\varepsilon = K \setminus B_\varepsilon(t^*)$. K_ε is compact. Let

$$M = g(t^*) = \sup_{t\in K} g(t),$$
$$M_\varepsilon = \sup_{t\in K_\varepsilon} g(t) < M.$$

Define $\delta = M - M_\varepsilon > 0$. If

$$\|G_n - g\|_\infty < \frac{\delta}{2},$$

then

$$\sup_{t\in K} G_n(t) \geq G_n(t^*) > M - \frac{\delta}{2},$$
$$\sup_{t\in K_\varepsilon} G_n(t) < M_\varepsilon + \frac{\delta}{2} = M - \frac{\delta}{2}.$$

So,

$$\mathbb{P}\big(t_n \in B_\varepsilon(t^*)\big) \geq \mathbb{P}(\|G_n - g\|_\infty \leq \delta)$$
$$\to 1.$$

3. The proof is similar to 2.

□

Theorem 22.4. *If Θ is compact, $\mathbb{E}_{\theta_0}[\|W_1\|_\infty] < \infty$ and $\log p_\theta(x)$ is continuous in θ for a.e. x, and $P_\theta \neq P_{\theta_0}$ for all $\theta \neq \theta_0$, then $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$ if $\hat{\theta}_n \in \arg\max_{\theta\in\Theta} \ell(\theta; X)$.*

Proof. Let

$$W_i(\theta) = \log \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}.$$

The W_i are i.i.d. in $C(\Theta)$. The mean is

$$\mu(\theta) = \mathbb{E}_{\theta_0}[W_i(\theta)]$$
$$= -D_{\mathrm{KL}}(\theta_0 \,\|\, \theta).$$

Since $\mu(\theta_0) = 0$ and $\mu(\theta) < 0$ for all $\theta \neq \theta_0$, then μ has a unique maximizer θ_0. $\hat{\theta}_n$ maximizes

$$\overline{W}_n = \frac{1}{n}\sum_{i=1}^n W_i.$$

So, $\|\overline{W}_n - \mu\|_\infty \xrightarrow{\mathbb{P}} 0$ by the Weak Law 22.2. Apply 22.3, 2. □

As an example of why uniform convergence is important, consider $K = [0,1]$, $g(t) = t$ (maximized at $t = 1$), and

$$G_n(t) = g(t) + \mathbb{1}\Big\{|t - U_n| < \frac{1}{n}\Big\}$$

where $U_n \sim \text{Uniform}[0,1]$. Then,

$$t^* = 1,$$
$$t_n = \Big(U_n + \frac{1}{n}\Big) \wedge 1.$$

Here, $\mathbb{P}(|t_n - t^*| < \varepsilon) \to \varepsilon$. However,

$$\mathbb{P}\big(G_n(t) \neq g(t)\big) \leq \frac{2}{n}.$$

Theorem 22.5. *Suppose $\Theta = \mathbb{R}^d$, $p_\theta(x)$ is continuous in θ for a.e. x, $P_{\theta_1} \neq P_{\theta_2}$ for all $\theta_1 \neq \theta_2$, and for all x, $p_\theta(x) \to 0$ as $\theta \to \infty$. If*

- $\mathbb{E}_{\theta_0}[\|\mathbb{1}_K W_1\|_\infty] < \infty$ *for all compact K,*
- $\mathbb{E}_{\theta_0}[\sup_{\|\theta\|>a} W_1(\theta)] < \infty$ *for some $a > 0$,*

then $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$.

Lecture 23

November 9

23.1 Finish MLE Consistency

Compact Θ: If Θ is compact, $\mathbb{E}_\theta[\|W\|_\infty] < \infty$, $p_\theta(x)$ is continuous in θ for a.e. x, and $P_\theta \neq P_{\theta_0}$ for all $\theta \neq \theta_0$, then $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$.

Theorem: Consistency of MLE. If $\Theta = \mathbb{R}^d$, $\hat{\theta}_n \in \arg\max_{\theta\in\Theta} \ell_n(\theta; X)$, $p_\theta(x)$ is continuous in θ for a.e. x and $p_\theta(x) \to 0$ as $\|\theta\| \to \infty$, $\mathbb{E}_{\theta_0}[\|\mathbb{1}_K W_1\|_\infty] < \infty$ for all $K \subseteq \mathbb{R}^d$ compact, where

$$W_i(\theta) = \ell_1(\theta; X_i) - \ell_1(\theta_0; X_i),$$
$$\overline{W}_n(\theta) = \frac{1}{n}\sum_{i=1}^n W_i(\theta),$$

and $\mathbb{E}_{\theta_0}[\sup_{\|\theta\|>a} W_1(\theta)] < \infty$ for some $a > 0$, then $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$.

Proof of 22.5. $p_\theta \to 0$ as $\|\theta\| \to \infty$, so $\sup_{\|\theta\|>b} W_1(\theta) \to -\infty$ as $b \to \infty$. By Dominated Convergence, $\mathbb{E}_{\theta_0}[\sup_{\|\theta\|>b} W_1(\theta)] \to -\infty$. Choose b for which $\mathbb{E}_{\theta_0}[\sup_{\|\theta\|>b} W_1(\theta)] < -\delta$ for some $\delta > 0$. Note that $\mathbb{E}_{\theta_0}[W_1(\theta_0)] = 0$ so $\|\theta_0\| \leq b$. Define

$$\tilde{\theta}_n = \arg\max_{\|\theta\|\leq b} \overline{W}_n(\theta) \xrightarrow{P_{\theta_0}} \theta_0$$

(since $K_b = \{\|\theta\| \leq b\}$ is compact). Then,

$$\sup_{\|\theta\|>b} \overline{W}_n(\theta) \leq \frac{1}{n}\sum_{i=1}^n \sup_{\|\theta\|>b} W_i(\theta) \xrightarrow{P_{\theta_0}} -\delta < 0.$$

So,

$$\begin{aligned}\mathbb{P}_{\theta_0}(\hat{\theta}_n \neq \tilde{\theta}_n) &\leq \mathbb{P}\Big(\sup_{\|\theta\|>b} \overline{W}_n(\theta) \geq \overline{W}_n(\theta_0)\Big) \\ &\leq \mathbb{P}\Big(\sup_{\|\theta\|>b} \overline{W}_n > -\frac{\delta}{2}\Big) + \mathbb{P}\Big(\overline{W}_n(\theta_0) \leq -\frac{\delta}{2}\Big) \\ &\to 0.\end{aligned}$$

$\square$

Example 23.1. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x) = p_0(x - \theta)$ for $\theta \in \mathbb{R}$. Assume:

- p_θ is continuous and bounded ($\sup_{x \in \mathbb{R}} p_0(x) = R < \infty$),
- $p_0(x) \to 0$ as $x \to \pm\infty$,
- $\int |\log p_0(x)| p_0(x) \, \mathrm{d}x < \infty$.

Then,

$$\begin{aligned}
\mathbb{E}_{\theta_0}\Big[\sup_{\theta \in \mathbb{R}} W_1(\theta)\Big] &= \mathbb{E}_{\theta_0}\Big[\sup_{\theta \in \mathbb{R}} \log \frac{p_0(X - \theta)}{p_0(X - \theta_0)}\Big] \\
&= \log R - \mathbb{E}_{\theta_0}[\log p_0(X - \theta_0)] \\
&= \log R - \mathbb{E}_0[\log p_0(X)] \\
&= \log R - \int_{\mathbb{R}} \big(\log p_0(x)\big) p_0(x) \, \mathrm{d}x \\
&< \infty.
\end{aligned}$$

23.2 Likelihood-Based Tests

23.2.1 Multidimensional MLE Distribution

Setup: $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$, where $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ is unknown.

- p_θ is "smooth" in θ (e.g., twice continuously differentiable).
- $\hat{\theta}_{\text{MLE}} \xrightarrow{P_{\theta_0}} \theta_0$.
- $\theta_0 \in \Theta^\circ$.

Expanding around θ_0,

$$\begin{aligned}
0 &= \nabla \ell(\hat{\theta}_n; X) \\
&\approx \nabla \ell(\theta_0; X) + \nabla^2 \ell(\theta_0; X)(\hat{\theta}_n - \theta_0), \\
\sqrt{n}(\hat{\theta}_n - \theta_0) &\approx \underbrace{\Big(-\frac{1}{n} \nabla^2 \ell(\theta_0; X)\Big)}_{\xrightarrow{P_{\theta_0}} J_1(\theta_0)}{}^{-1} \underbrace{\Big(\frac{1}{\sqrt{n}} \nabla \ell(\theta_0; X)\Big)}_{\overset{P_{\theta_0}}{\Longrightarrow} \mathcal{N}(0, J_1(\theta_0))} \\
&\overset{P_{\theta_0}}{\Longrightarrow} \mathcal{N}\big(0, J_1(\theta_0)^{-1}\big).
\end{aligned}$$

23.2.2 Wald-Type Confidence Regions/Tests

If

$$\frac{1}{n} \hat{J}_n \xrightarrow{P_{\theta_0}} J_1(\theta_0) \succ 0,$$

then $\hat{J}_n^{1/2}(\hat{\theta}_n - \theta_0) \overset{P_{\theta_0}}{\Longrightarrow} \mathcal{N}(0, I)$. So, $\|\hat{J}_n^{1/2}(\hat{\theta}_n - \theta_0)\|_2^2 \overset{P_{\theta_0}}{\Longrightarrow} \chi_d^2$. We can reject $H_0 : \theta = \theta_0$ if

$$\|\hat{J}_n^{1/2}(\hat{\theta}_n - \theta_0)\|_2^2 > \chi_d^2(\alpha).$$

We can also construct a confidence region:

$$\|\hat{J}_n^{1/2}(\hat{\theta}_n - \theta)\|_2^2 \le c \iff \hat{J}_n^{1/2}(\hat{\theta}_n - \theta) \in \sqrt{c} B_1(0)$$

$$\iff \theta \in \hat{\theta}_n + \sqrt{c}\hat{J}_n^{-1/2} B_1(0).$$

Popular choices:

$$\hat{J}_n = nJ_1(\hat{\theta}_n) = n \operatorname{var}_\theta \nabla \ell(\theta; X)\Big|_{\hat{\theta}_n},$$
$$\hat{J}_n = -\nabla^2 \ell(\hat{\theta}_n; X).$$

The second estimator is usually preferred, since it takes into account how informative the actual dataset is.

Conditionality Principle: Flip a coin; with probability 1/2, $X \sim \mathcal{N}(\mu, 1)$ ($Z = 1$), and with probability 1/2, $X \sim \mathcal{N}(\mu, 9)$ ($Z = 2$). Test $H_0 : \mu = 0$. A natural idea would be: if $Z = 1$, reject if $|X| > z_{\alpha/2}$, and if $Z = 2$, reject if $|X| > 3z_{\alpha/2}$. This is not the same as the Neyman-Pearson test. The Conditionality Principle says that we should condition on whatever information is available.

Example 23.2 (Logistic Regression)**.** Suppose

$$\mathbb{P}(Y_i = 1 \mid X_i = x) = \frac{e^{\beta^\mathsf{T} x}}{1 + e^{\beta^\mathsf{T} x}}$$

for $x \in \mathbb{R}^d$.

1. Solve numerically for $\hat{\beta} = \arg\max_{\beta \in \mathbb{R}^d} \ell(\beta; X, Y)$.
2. Find $\hat{J}^{-1} = (-\nabla^2 \ell(\hat{\beta}; X, Y))^{-1}$.

Since $\hat{\beta} \approx \beta + \mathcal{N}(0, \hat{J}^{-1})$, the confidence region for β is $\hat{\beta} + \sqrt{c}\hat{J}^{-1/2} B_1(0)$. Also, $\hat{\beta}_j \approx \beta_j + \mathcal{N}(0, (\hat{J}^{-1})_{j,j})$, so the interval is $\beta_j \in \hat{\beta}_j \pm \sqrt{(\hat{J}^{-1})_{j,j}} z_{\alpha/2}$. Note that $\sqrt{c}$ scales as $\sqrt{d}$.

If $S \subseteq [d]$, write $\hat{J}^{-1} = \hat{\Sigma}$, and then $\beta_S \in \sqrt{\chi^2_{|S|}(\alpha)} (\hat{\Sigma}_{S,S})^{1/2} + \hat{\beta}$, and the constant in front now scales as $\sqrt{|S|}$.

Lecture 24

November 14

24.1 Score Test/Region

24.1.1 Wald

If

$$\hat{J}_1 \xrightarrow{P_{\theta_0}} J_1(\theta_0) \succ 0$$

then $\sqrt{n}\hat{J}_1^{1/2}(\hat{\theta}_n - \theta_0) \overset{P_{\theta_0}}{\Longrightarrow} \mathcal{N}_d(0, I_d)$.

Test: $\|\sqrt{n}\hat{J}_1^{1/2}(\hat{\theta}_n - \theta_0)\|_2^2 \overset{P_{\theta_0}}{\Longrightarrow} \chi_d^2$, so we use the region $\theta_0 \in \hat{\theta}_n + \sqrt{\chi_d^2(\alpha)}\hat{J}_1^{-1/2}n^{-1/2}B_1(0)$.

Some choices for $\hat{J}_1$ are $J_1(\hat{\theta}_n)$ and $-n^{-1}\nabla^2\ell_n(\hat{\theta}_n; X_1, \ldots, X_n)$.

Example 24.1. Let $X \sim \text{Binomial}(n, \theta)$, so

$$\hat{\theta}_n = \frac{1}{n}$$

and

$$J_n(\theta) = \frac{n}{\theta(1-\theta)}.$$

So,

$$\hat{J}_1 = \left[\left(\frac{1}{n}\right)\left(1 - \frac{1}{n}\right)\right]^{-1} = J_1(\hat{\theta}_n).$$

For $\alpha = 0.05$, the interval becomes

$$\hat{\theta} \pm 1.96\widehat{\text{SE}}(\hat{\theta}_n)$$

where

$$\widehat{\text{SE}}(\hat{\theta}_n) = \frac{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}{\sqrt{n}}.$$

Thus, the interval is approximately

$$\hat{\theta} \pm 1.96\frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}} = \frac{1}{n} \pm 1.96\frac{1}{n}$$

which falls outside of the parameter space.

24.1.2 Score Test

$$\frac{1}{\sqrt{n}}\nabla \ell_n(\theta_0; X_1, \ldots, X_n) \xrightarrow{P_{\theta_0}} \mathcal{N}_d\big(0, J_1(\theta_0)\big).$$

Reject $H_0 : \theta = \theta_0$ if

$$\left\| \frac{1}{\sqrt{n}} J_1(\theta_0)^{-1/2} \nabla \ell_n(\theta_0; X_1, \ldots, X_n) \right\|_2^2 > \chi_d^2(\alpha).$$

Example 24.2 (Exponential Family)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_\eta(x) = \mathrm{e}^{\eta^\mathsf{T} T(x) - A(\eta)} h(x)$. Then,

$$\nabla \ell(\eta; X) = \sum_{i=1}^n \big(T(X_i) - \mathbb{E}_\eta[T(X_1)]\big),$$
$$J_1(\eta) = \operatorname{var}_\eta T(X_1) = \nabla^2 A(\eta).$$

Reject if $(\sum_{i=1}^n (T(X_i) - \mu(\eta)))^\mathsf{T} (n \nabla^2 A(\eta))^{-1} (\sum_{i=1}^n (T(X_i) - \mu(\eta))) > \chi_d^2(\alpha)$.

Example 24.3 (Pearson's χ^2 Test)**.** Let

$$\begin{aligned}(N_1, \ldots, N_d) &\sim \text{Multinomial}\big(n, (\pi_1, \ldots, \pi_d)\big) \\ &= \pi_1^{N_1} \cdots \pi_d^{N_d} \frac{n!}{N_1! \cdots N_d!} \mathbb{1}\Big\{\sum_{i=1}^d N_i = n\Big\}.\end{aligned}$$

Test $H_0 : \pi = \pi^{(0)}$ (note the constraint $\sum_{j=1}^d \pi_j = 1$). The test statistic is

$$\sum_{i=1}^d \frac{(N_i - n\pi_i^{(0)})^2}{n\pi_i^{(0)}} \xrightarrow{P_{\pi^{(0)}}} \chi_{d-1}^2.$$

This is a score test.

24.2 Generalized Likelihood Ratio Test/Region

Expand ℓ around $\hat{\theta}$.

$$\begin{aligned}&\ell_n(\theta_0; X_1, \ldots, X_n) - \ell_n(\hat{\theta}_n; X_1, \ldots, X_n) \\ &\quad \approx \cancel{\nabla \ell_n(\hat{\theta}_n; X_1, \ldots, X_n)(\theta_0 - \hat{\theta}_n)} + \frac{1}{2}(\theta_0 - \hat{\theta}_n)^\mathsf{T} \nabla^2 \ell_n(\hat{\theta}_n; X_1, \ldots, X_n)(\theta_0 - \hat{\theta}_n),\end{aligned}$$

so

$$\begin{aligned}2\big(\ell_n(\hat{\theta}_n; X_1, \ldots, X_n) - \ell_n(\theta_0; X_1, \ldots, X_n)\big) &\approx \underbrace{\big(\sqrt{n}(\hat{\theta}_n - \theta_0)\big)^\mathsf{T}}_{\xrightarrow{P_{\theta_0}} \mathcal{N}(0, J_1(\theta_0)^{-1})} \underbrace{\Big(-\frac{1}{n}\nabla^2 \ell_n(\hat{\theta}_n; X_1, \ldots, X_n)\Big)}_{\xrightarrow{P_{\theta_0}} J_1(\theta_0)} \big(\sqrt{n}(\hat{\theta}_n - \theta_0)\big) \\ &\xrightarrow{P_{\theta_0}} \chi_d^2.\end{aligned}$$

24.2.1 Generalized LRT with Nuisance Parameters

Test $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta \setminus \Theta_0$. The GLRT statistic is $G_n^2 = 2(\ell_n(\hat{\theta}_n; X_1, \ldots, X_n) - \ell_n(\hat{\theta}_0; X_1, \ldots, X_n))$ where $\hat{\theta}_0 \in \arg\max_{\theta \in \Theta_0} \ell_n(\theta; X_1, \ldots, X_n)$. If Θ_0 is a d_0-dimensional manifold in Θ, and $\theta_0 \in (\operatorname{relint} \Theta_0) \cap \Theta^\circ$ and $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$, with additional regularity conditions, then $G_n^2 \overset{P_{\theta_0}}{\Longrightarrow} \chi^2_{d-d_0}$. Asymptotically near Θ_0, we have

$$\ell_n(\theta; X_1, \ldots, X_n) \approx \ell_n(\hat{\theta}_n; X_1, \ldots, X_n) + \frac{1}{2} \| J_n(\theta_0)^{-1/2} (\theta - \hat{\theta}_n) \|_2^2.$$

Assume that we have parameterized the problem so $J_1(\theta_0) = \text{id}$. Then,

$$\hat{\theta}_0 \approx \operatorname*{arg\,min}_{\theta \in \Theta_0} \| \theta - \hat{\theta}_n \|_2^2 = \text{projection}_{\Theta_0}(\hat{\theta}_n).$$

So, the GLRT $\approx \| \hat{\theta}_n - \text{projection}_{\Theta_0}(\hat{\theta}_n) \|_2^2 \approx \chi^2_{d-d_0}$.

Lecture 25

November 16

25.1 Plug-In Estimators, Bootstrap

Example 25.1. We observe $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$, where the $X_i \in \mathbb{R}$. We want to estimate the median $\theta(P)$. The "obvious estimator" (for n odd) is $\hat{\theta}_n = X_{((n+1)/2)} = \theta(\hat{P}_n)$. This is a "**plug-in estimator**". Here, $\hat{P}_n$ is the empirical distribution $n^{-1}\sum_{i=1}^n \delta_{X_i}$ (δ_x is the point mass at x).

Questions: What is $\operatorname{var}_P \hat{\theta}_n$? $\operatorname{bias}_P \hat{\theta}_n$?

We know that as $n \to \infty$,

$$\sqrt{n}\big(\theta(\hat{P}_n) - \theta(P)\big) \overset{P}{\Rightarrow} \mathcal{N}\Big(0, \frac{1}{4p(\theta(P))^2}\Big)$$

(assuming $p(\theta(P)) > 0$, where p is the density for P).

- We do not know if P has a density, or if $p(\theta(P)) > 0$.
- This answer could be "very" asymptotic.

We want to estimate $\sigma^2(P) = \operatorname{var}_P \hat{\theta}_n$. A natural estimator is

$$\begin{aligned}\hat{\sigma}_n^2 &= \sigma^2(\hat{P}_n) \\ &= \operatorname{var}_{\hat{P}_n} \hat{\theta}(X^*).\end{aligned}$$

We are *"integrating" over* possible samples $X_1^*, \ldots, X_n^* \overset{\text{i.i.d.}}{\sim} \hat{P}_n$. For fixed A, $\hat{P}_n(A) \xrightarrow{\text{a.s.}} P(A)$.

1. For $b = 1, \ldots, B$ $(= 200)$:
   (a) Sample $X_1^{*,b}, \ldots, X_n^{*,b}$ from the original data set with replacement.
   (b) Compute $\hat{\theta}^{*,b} = \hat{\theta}(X_1^{*,b}, \ldots, X_n^{*,b})$.

Then,

$$\begin{aligned}\overline{\theta}^* &= \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{*,b}, \\ \hat{\sigma}_n^2 &= \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*,b} - \overline{\theta}^*)^2 \\ &\xrightarrow{B\to\infty} \operatorname{var}_{\hat{P}_n} \hat{\theta}(X^*).\end{aligned}$$

Similarly, $\hat{\beta}_n = \overline{\theta}^* - \hat{\theta}_n$ is the bootstrap estimate of the bias.

25.1.1 Bias Correction

$$\begin{aligned}\widehat{\text{bias}}\,\hat{\theta}_n &= \text{bias}_{\hat{P}_n}\,\hat{\theta}(X^*)\\ &= \mathbb{E}_{\hat{P}_n}[\hat{\theta}(X^*) - \theta(\hat{P}_n)].\end{aligned}$$

Thus, we can use

$$\tilde{\theta}_n = \hat{\theta}_n - \text{bias}_{\hat{P}_n}\,\hat{\theta}(X^*)$$

as a substitute for the ideal estimator

$$\theta_n = \hat{\theta}_n - \text{bias}_P\,\hat{\theta}_n.$$

25.1.2 Bootstrapping for the Maximum

Let $M(X) = X_{(n)}$. Assume P is continuous so there are no ties. Let $X_i^* \overset{\text{i.i.d.}}{\sim} \hat{P}_n$. Then,

$$\begin{aligned}\mathbb{P}\big(M(X^*) = M(X)\big) = 1 - \mathbb{P}(X_i^* \neq X_{(n)})^n &= 1 - \Big(1 - \frac{1}{n}\Big)^n\\ &\approx 1 - \mathrm{e}^{-1}\\ &\approx 63\%.\end{aligned}$$

25.2 Bootstrap Confidence Intervals

Bootstrap CIs start with a **root** $R_n(X, \theta(P)) \in \mathbb{R}$. Assume the root has law

$$L_n(r; P) = \mathbb{P}_P\big\{R_n\big(X, \theta(P)\big) \le r\big\}.$$

If P is known, we can use L_n to get a confidence region for θ: $C_\alpha(X, P) = \{\theta : L_n(R_n(X, \theta(P)); P) \le 1 - \alpha\}$. Then,

$$\begin{aligned}\mathbb{P}_P\big(\theta(P) \in C_\alpha(X; P)\big) &= \mathbb{P}_P\big\{L_n\big(R_n(X, \theta); P\big) \le 1 - \alpha\big\}\\ &\le 1 - \alpha\end{aligned}$$

(with equality if R_n is continuous).

Example 25.2. If $R_n = |\hat{\theta}_n - \theta(P)|$, then $C_\alpha(X; P) = \hat{\theta}_n \pm L_n^{-1}(1 - \alpha; P)$.

Example 25.3. If

$$R_n = \frac{|\hat{\theta}_n - \theta(P)|}{\hat{\sigma}_n},$$

then $C_\alpha(X; P) = \hat{\theta}_n \pm \hat{\sigma}_n L_n^{-1}(1 - \alpha; P)$.

Example 25.4. If $R_n = \|\hat{\theta}_n - \theta(P)\|_\infty$, then

$$C_\alpha(X; P) = [(\hat{\theta}_n)_1 \pm L_n^{-1}(1 - \alpha; P)] \times \cdots \times [(\hat{\theta}_n)_d \pm L_n^{-1}(1 - \alpha; P)].$$

Problem: We do not know P.

Solution: Use $\hat{P}_n$.

In 25.2, use $C_\alpha(X; \hat{P}_n) = \hat{\theta}_n \pm L_n^{-1}(1 - \alpha; \hat{P}_n)$. We need $L_n^{-1}(1 - \alpha; \hat{P}_n) \to L_n^{-1}(1 - \alpha; P)$. Usually, we see something like $L_n(r; \hat{P}_n) \to \Phi(r)$.

1. For $b = 1, \ldots, B$:

   (a) Sample $X_1^{*,1}, \ldots, X_n^{*,b} \overset{\text{i.i.d.}}{\sim} \hat{P}_n$.

   (b) $\hat{\theta}^{*,b} = \hat{\theta}(X^{*,b})$.

   (c) $R^{*,b} = R_n(X^{*,b}, \theta(\hat{P}_n))$.

(For example, $R^{*,b} = |\hat{\theta}(X^{*,b}) - \theta(\hat{P}_n)|$.) Let r be the $1 - \alpha$ quantile of $\{R^{*,1}, \ldots, R^{*,B}\}$. Then,

$$C_\alpha(X) = \{\theta : R_n(X, \theta) \leq r\}.$$

Lecture 26

November 21

Lecturer: Xiao Li

26.1 Global Testing

Setup: $X \sim \mathcal{N}_d(\theta, I_d)$, where $\theta \in \mathbb{R}^d$. Test $H_0 : \theta = 0$ vs. $\theta \neq 0$. Write $X = \theta + \varepsilon$, for $\varepsilon \sim \mathcal{N}_d(0, I_d)$.

Applications:

1. detection of chemical weapons
2. detection of KBOs in the Kuiper Belt

Suppose that we observe $X_1, \ldots, X_d$.

Test Statistic 1: $\max_{i=1,\ldots,d} |X_i|$ (max test).

Test Statistic 2: $\sum_{i=1}^d X_i^2$ (χ^2 test).

26.1.1 Power of the Max Test

Lemma 26.1.

$$\frac{1}{2}\Big(1 - \frac{1}{z^2}\Big)\frac{\phi(z)}{z} \leq 1 - \Phi(z) \leq \frac{\phi(z)}{z}$$

where Φ is the CDF of $\mathcal{N}(0,1)$ and ϕ is the density of $\mathcal{N}(0,1)$.

Lemma 26.2.

$$\frac{\max_{i=1,\ldots,d} |X_i|}{\sqrt{2\log d}} \xrightarrow{\mathbb{P}} 1 \qquad \text{as } d \to \infty.$$

Proof.

$$\begin{aligned}\mathbb{P}\Big(\max_{i=1,\ldots,d} |X_i| \leq x\Big) &= \Phi(x)^d \\ &= \big[1 - \big(1 - \Phi(x)\big)\big]^d\end{aligned}$$

$$\to \begin{cases} 1, & \frac{x}{\sqrt{2\log d}} < 1 \\ 0, & \frac{x}{\sqrt{2\log d}} > 1 \end{cases}$$

□

In comparison, if $X_1, \ldots, X_n$ are Cauchy, then

$$\frac{\max_{i=1,\ldots,n} X_i}{n} \xrightarrow{\mathsf{d}} f,$$

where f is the density

$$f(x) = \exp\Big(-\frac{1}{x}\Big)\,\mathbb{1}\{x > 0\}.$$

Consider the regime where $\theta_1 = \cdots = \theta_k = \mu > 0$, $\theta_{k+1} = \cdots = \theta_d = 0$, and $k(d) = d^\beta$ for some $\beta \in (0,1)$.

Theorem 26.3. *Suppose $\mu(d) = \sqrt{2r\log d}$, $r > 0$.*

(a) If $r > (1-\sqrt{\beta})^2$, then the power of the max test $\to 1$.

(b) If $r < (1-\sqrt{\beta})^2$, then the power $\to \alpha$.

Proof. $\max_{i=1,\ldots,d}|X_i| = \max\{\max_{i=1,\ldots,k}|X_i|, \max_{i=k+1,\ldots,d}|X_i|\}$. Also,

$$\begin{aligned} \frac{\max_{i=1,\ldots,k}|X_i|}{\sqrt{2\log d}} &\geq \frac{1}{\sqrt{2\log d}}\Big(\sqrt{2r\log d} + \sqrt{2\log k}\frac{\max_{i=1,\ldots,k}\varepsilon_i}{\sqrt{2\log k}}\Big) \\ &\xrightarrow{\mathbb{P}} \sqrt{r} + \sqrt{\beta} \begin{cases} > 1, & \text{if } r > (1-\sqrt{\beta})^2 \\ < 1, & \text{if } r < (1-\sqrt{\beta})^2 \end{cases} \end{aligned}$$

since

$$\frac{\max_{i=1,\ldots,k}\varepsilon_i}{\sqrt{2\log k}} \xrightarrow{\mathbb{P}} 1$$

and $k = d^\beta$. So, the power of the max test is

$$\begin{aligned} \mathbb{P}\Big(\max_{i=1,\ldots,d}|X_i| \geq \sqrt{2\log d}\big(1+o(1)\big)\Big) &\geq \mathbb{P}\Big(\max_{i=1,\ldots,k}|X_i| \geq \sqrt{2\log d}\big(1+o(1)\big)\Big) \\ &\to 1 \end{aligned}$$

if $r > (1-\sqrt{\beta})^2$. Otherwise,

$$\begin{aligned} &\mathbb{P}\Big(\max_{i=1,\ldots,d}|X_i| > \sqrt{2\log d}\big(1+o(1)\big)\Big) \\ &\quad\leq \underbrace{\mathbb{P}\Big(\max_{i=1,\ldots,k}|X_i| > \sqrt{2\log d}\big(1+o(1)\big)\Big)}_{\to 0} + \mathbb{P}\Big(\max_{i=k+1,\ldots,d}|X_i| > \sqrt{2\log d}\big(1+o(1)\big)\Big) \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}\Big(\max_{i=k+1,\ldots,d}|X_i| > \sqrt{2\log d}\big(1+o(1)\big)\Big) &= \mathbb{P}\Big(\max_{i=k+1,\ldots,d}|\varepsilon_i| > \sqrt{2\log d}\big(1+o(1)\big)\Big) \\ &\leq \mathbb{P}\Big(\max_{i=1,\ldots,d}|\varepsilon_i| > \sqrt{2\log d}\big(1+o(1)\big)\Big) \end{aligned}$$

$$\to \alpha$$

if $r < (1-\sqrt{\beta})^2$. □

26.1.2 Power of the χ^2 Test

Under $H_0 : \theta = 0$, $\mathbb{E}[X_i^2] = 1$ and $\operatorname{var} X_i^2 = 2$. By the CLT 19.14,

$$\frac{1}{\sqrt{d}}\Big(\sum_{i=1}^d X_i^2 - d\Big) \Rightarrow \mathcal{N}(0,2).$$

and the cutoff is $\chi_d^2(\alpha) = d + \sqrt{2d}z_{1-\alpha} + o(\sqrt{d})$.

Under H_1, $\theta \neq 0$, so $\mathbb{E}[X_i^2] = 1 + \theta_i^2$ and $\operatorname{var} X_i^2 = 4\theta_i^2 + 2$. By the CLT 19.14,

$$\sum_{i=1}^d X_i^2 \approx \mathcal{N}(d + \|\theta\|_2^2, 4\|\theta\|_2^2 + 2d)$$

or equivalently,

$$\frac{1}{\sqrt{d}}\Big(\sum_{i=1}^d X_i^2 - d\Big) \approx \mathcal{N}\Big(\frac{\|\theta\|_2^2}{\sqrt{d}}, 2 + 4\frac{\|\theta\|_2^2}{d}\Big).$$

If $\|\theta\|_2^2/\sqrt{2d} \gg 1$, the power is very high. If $\|\theta\|_2^2/\sqrt{2d} \ll 1$, the power is $\approx \alpha$. Here,

$$\frac{\|\theta\|_2^2}{\sqrt{2d}} = \frac{k\mu^2}{\sqrt{2d}}$$

since $\theta_1 = \theta_2 = \cdots = \theta_k = \mu$.

26.1.3 Comparison of the Tests

β	χ^2 test needs	max test needs
1/2	$\mu > 3$	$\mu > 0.29\sqrt{2\log d}$
1/4	$\mu > 3d^{1/8}$	$\mu > 0.5\sqrt{2\log d}$
3/4	$\mu > 3d^{-1/8}$	$\mu > 0.13\sqrt{2\log d}$

If $\beta \in (1/4, 1/2)$, there is another optimal test (Donoho and Jin, 2004).

Lecture 27

November 28

27.1 Multiple Testing

Setup: $X \sim P \in \mathcal{P}$. Test $H_{0,i}$, $i = 1, \ldots, n$. Return an accept/reject decision for each i.

Example 27.1. Let $X_i \overset{\text{independent}}{\sim} \mathcal{N}(\mu_i, 1)$ for $i = 1, \ldots, n$. Test $H_{0,i} : \mu_i = 0$ vs. $H_{1,i} : \mu_i \neq 0$.

Last time, we considered testing $H_0 = \bigcap_{i=1}^n H_{0,i} : \mu = 0$.

Example 27.2. Let $p_i \in [0, 1]$, $i = 1, \ldots, n$. Test $H_{0,i} : p_i \sim \text{Uniform}[0, 1]$ vs. $H_{1,i} : p_i$ is not larger than Uniform$[0, 1]$.

GWAS: There is a 2×2 table for each of n SNPs.

	diseased	controls
wild-type		
mutant		

Basic problem: Observe X, return a set $\mathcal{S}(X) \subseteq \{1, \ldots, n\}$ of rejections.

Variants of the decision problem:

1. Look at the best (largest) X_i, test whether it is actually the best μ_i.
2. Look at the best X_i, return a confidence interval for only the mean corresponding to $X_{(1)}$.
3. Return a CI for every μ_i.
4. Return a CI for μ_i through μ_j.
5. Return intervals for $\mu_{\mathcal{S}(X)}$.

"Bad" things happen if we do not correct for multiplicity.

Example 27.3. Suppose that in the independent Gaussian example, $\mu_i = 0$, for all i, and test all $H_{0,i}$ at level α. $\mathbb{E}[\#\text{rejections}] = \alpha n$, so $\mathbb{P}(\text{at least 1 false rejection}) \xrightarrow{n \to \infty} 1$.

27.2 Familywise Error Rate (FWER)

Classic Proposal (Pre-1995): Control the **FWER (familywise error rate)**, i.e.,

$$\text{FWER} = \mathbb{P}(\text{make at least 1 type I error}).$$

In multiple testing,

$$\begin{aligned}\text{FWER} &= \sup_{P\in\mathcal{P}} \mathbb{P}_P(\text{any true } H_{0,i} \text{ is rejected}) \\ &= \sup_{P\in\mathcal{P}} \mathbb{P}_P\big(\mathcal{H}_0(P)\cap\mathcal{S}(X)\neq\varnothing\big),\end{aligned}$$

where $\mathcal{H}_0(P) = \{i : H_{0,i} \text{ is true}\}$ and $\mathcal{S}(X) = \{i : H_{0,i} \text{ is rejected}\}$.

27.2.1 Bonferroni Correction

Reject $H_{0,i}$ iff

$$p_i \le \frac{\alpha}{n}.$$

Then,

$$\begin{aligned}\mathbb{P}(\text{any false rejections}) &= \mathbb{P}\Big(\bigcup_{i\in\mathcal{H}_0}\{H_{0,i} \text{ rejected}\}\Big) \\ &\le \sum_{i\in\mathcal{H}_0}\mathbb{P}(H_{0,i} \text{ rejected}) \\ &\le |\mathcal{H}_0|\frac{\alpha}{n} \\ &\le \alpha.\end{aligned}$$

If $p_1,\ldots,p_n$ are known to be independent, we can do a bit better.

Šidák's Correction: Reject $H_{0,i}$ if $p_i \le \tilde{\alpha}_n$, where

$$\begin{aligned}\tilde{\alpha}_n &= 1-(1-\alpha)^{1/n} \\ &\approx \frac{\alpha}{n} \qquad \text{for large } n.\end{aligned}$$

Now, FWER $= \alpha$ if all $H_{0,i}$ are true and the p-values are independent and uniform.

27.2.2 Correlated Test Statistics

Example 27.4 (Pairwise Comparisons)**.** Let $X_i \overset{\text{independent}}{\sim} \mathcal{N}(\mu_i, 1)$. Write $X_i = \mu_i + \varepsilon_i$, where

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

Test $H_{0,i,j} : \mu_i = \mu_j$, for $i, j = 1,\ldots,n$. There are a total of $\binom{n}{2} \approx n^2/2$ hypotheses. Since

$$\frac{X_i - X_j}{\sqrt{2}} \overset{H_{0,i,j}}{\sim} \mathcal{N}(0,1),$$

we may reject all $H_{0,i,j}$ with $|X_i - X_j| > \sqrt{2}z_{\alpha/(2\binom{n}{2})}$.

More powerful: reject $H_{0,i,j}$ if $|X_i - X_j| > r_\alpha$, where $\mathbb{P}(\max_{i,j=1,\ldots,n}|\epsilon_i - \varepsilon_j| > r_\alpha) = \alpha$. Then,

$$\begin{aligned}\mathbb{P}(\text{any false rejection}) &= \mathbb{P}(|X_i - X_j| > r_\alpha \text{ for any } i,j \text{ with } \mu_i = \mu_j) \\ &\le \mathbb{P}(|\varepsilon_i - \varepsilon_j| > r_\alpha \text{ for any } i,j) = \alpha.\end{aligned}$$

This is **Tukey's Honestly Significant Difference (HSD) Procedure**. HSD is not much better than Bonferroni's correction if n is large.

$$\max_{i=1,\ldots,n}|\varepsilon_i - \varepsilon_j| = \max_{i=1,\ldots,n}\varepsilon_i + \max_{i=1,\ldots,n}(-\varepsilon_i)$$

$$= 2\sqrt{2\log n}\big(1 + o_{\mathrm{p}}(1)\big).$$

So, $r_\alpha \approx 2\sqrt{2\log n}$. In comparison, $\sqrt{2}z_{\alpha/(2\binom{n}{2})} \approx \sqrt{2}\sqrt{2\log\binom{n}{2}} \approx 2\sqrt{2\log n}$. For $n = 6$, the difference is like 4.0 vs. 4.1. The difference is more important if σ^2 is estimated instead of known.

Example 27.5 (Scheffé's S-Method)**.** Test, for all linear combinations, $H_{0,\nu} : \mu^\mathsf{T}\nu = 0$, for all $\nu \in S^{n-1}$. Reject $H_{0,\nu}$ when $|X^\mathsf{T}\nu| > \chi_n(1-\alpha)$. Why?

$$\begin{aligned}
\mathbb{P}(\text{any false rejections}) &= \mathbb{P}\big(|X^\mathsf{T}\nu| > \chi_n(1-\alpha), \text{any } \nu \text{ with } \nu^\mathsf{T}\mu = 0\big) \\
&\le \mathbb{P}\Big(\max_{\|\nu\|_2 = 1} |\varepsilon^\mathsf{T}\nu| > \chi_n(1-\alpha)\Big) \\
&= \mathbb{P}\big(\|\varepsilon\|_2 > \chi_n(1-\alpha)\big) = \alpha.
\end{aligned}$$

Here, $\chi_n(1-\alpha) \approx \sqrt{n}$, which is a significant loss.

More generally, let $X_i \overset{\text{independent}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$ for $i = 1, \dots, n$. Test $H_{0,\nu} : \nu^\mathsf{T}\mu = 0$, for $\nu \in \Xi \subseteq \mathbb{R}^n$. Suppose that we have an independent estimator $\hat{\sigma}^2 \sim \sigma^2\chi_d^2$ ($\perp\!\!\!\perp \varepsilon$). Reject $H_{0,\nu}$ if

$$\frac{|X^\mathsf{T}\nu|}{\hat{\sigma}\|\nu\|_2} \ge c_\alpha,$$

where

$$\mathbb{P}\Big(\sup_{\nu\in\Xi} \frac{|\varepsilon^\mathsf{T}\nu|}{\hat{\sigma}\|\nu\|_2} > c_\alpha\Big) = \alpha.$$

27.3 Simultaneous CIs & Deduced Inference

Closely related: $X \sim P \in \mathcal{P}$. There are many parameters of interest, $\theta_1(P), \dots, \theta_n(P)$. Construct $C_1, \dots, C_n$, and $\mathrm{FWER} = \sup_{P\in\mathcal{P}} \mathbb{P}_P(\theta_i \notin C_i \text{ for any } i)$.

Example 27.6 (Gaussian, Unknown Variance)**.** Suppose $\theta_i = \mu_i$, $\theta_{i,j} = \mu_i - \mu_j$, or $\theta_\nu = \mu^\mathsf{T}\nu$ for $\nu \in \Xi$. Return $C_\nu = X^\mathsf{T}\nu \pm \hat{\sigma}\|\nu\|_2 c_\alpha$. Interpret these confidence intervals as giving a confidence region for $\mu \in \mathbb{R}^n$, defined as $\{\mu : \mu \text{ is covered by all } C_i\}$. Then, $R(X) = \{\mu : \mu^\mathsf{T}\nu \in C_\nu,\ \forall \nu \in \Xi\}$, so

$$\begin{aligned}
\mathbb{P}_\mu\big(\mu \in R(X)\big) &= \mathbb{P}_\mu(C_\nu \ni \mu^\mathsf{T}\nu,\ \forall \nu \in \Xi) \\
&= 1 - \alpha.
\end{aligned}$$

27.3.1 Deduced Intervals

We want an interval for $\mu^\mathsf{T}\nu^*$ for $\nu^* \notin \Xi$.

$$C_{\nu^*}(X) = \Big[\inf_{\mu\in R(X)} \mu^\mathsf{T}\nu^*, \sup_{\mu\in R(X)} \mu^\mathsf{T}\nu^*\Big].$$

Then,

$$\begin{aligned}
\mathbb{P}_\mu\big(\mu^\mathsf{T}\nu^* \in C_{\nu^*}(X)\big) &\ge \mathbb{P}_\mu\big(\mu \in R(X)\big) \\
&= 1 - \alpha.
\end{aligned}$$

Lecture 28

November 30

28.1 False Discovery Rate

28.1.1 Motivation for FDR Control

Suppose we test 1000 hypotheses at level 0.05. We get 53 rejections. Under FWER control, we instead test at level 0.05/1000. If instead we test 1000000 hypotheses, then FWER control tests at level 0.05/1000000, which is unappealing.

With FDR control, we may get 530 rejections, of which 40 are false discoveries.

28.2 Benjamini-Hochberg Procedure (1995)

Recall:

$$\mathcal{H}_0 = \{i : H_{0,i} \text{ is true}\},$$
$$\mathcal{S}(X) = \{i : H_{0,i} \text{ is rejected}\}.$$

Define $R(X) = |\mathcal{S}(X)|$, the number of rejections, and $V(X) = |\mathcal{S}(X) \cap \mathcal{H}_0|$, the number of false discoveries. Define

$$\begin{aligned} \text{FDP} &= \begin{cases} \dfrac{V}{R}, & R \geq 1 \\ 0, & V = R = 0 \end{cases} \\ &= \frac{V}{R \vee 1}, \end{aligned}$$

the "**false discovery proportion**". Then, $\text{FDR} = \mathbb{E}[\text{FDP}]$.

Benjamini-Hochberg Procedure: We have p-values $p_1, \ldots, p_n$.

1. Order the p-values. $p_{(1)} \leq \cdots \leq p_{(n)}$.
2. Find
$$\hat{R} = \max\Big\{r : p_{(r)} \leq \frac{\alpha r}{n}\Big\}.$$
3. Reject $H_{(1)}, \ldots, H_{(\hat{R})}$.

28.2.1 BH as "Empirical Bayes" Interpretation

What does

$$p_{(r)} \leq \frac{\alpha r}{n}$$

have to do with FDR? Consider rejecting all H_i with $p_i \leq t$, where t is fixed in $[0,1]$. Define

$$\mathcal{S}_t(X) = \{i : p_i \leq t\}.$$

Then, we can define $R_t = |\mathcal{S}_t|$, $V_t = |\mathcal{S}_t \cap \mathcal{H}_0|$, FDP_t, etc. What is FDR_t? We can estimate it from data. We want to maximize the number of rejections, or equivalently maximize t, subject to

$$\frac{V_t}{R_t \vee 1} \leq \alpha.$$

Problem: We cannot observe V_t.

Solution:

$$\begin{aligned}
\mathbb{E}[V_t] &= \mathbb{E}\Big[\sum_{i \in \mathcal{H}_0} \mathbb{1}\{p_i \leq t\}\Big] \\
&= \sum_{i \in \mathcal{H}_0} \mathbb{P}(p_i \leq t) \\
&= t|\mathcal{H}_0| \leq tn.
\end{aligned}$$

So,

$$\widehat{\text{FDP}}_t = \frac{nt}{R_t \vee 1}$$

is a conservative estimator of FDP_t.

BH Procedure (equivalent):

1. Find $\hat{t} = \max\{t : \widehat{\text{FDP}}_t \leq \alpha\}$.
2. Reject H_i if $p_i \leq \hat{t}$.

$$\widehat{\text{FDP}}_t = \frac{np_{(r)}}{r} \leq \alpha \iff p_{(r)} \leq \frac{\alpha r}{n}.$$

It is not clear that $\widehat{\text{FDP}}_{\hat{t}} \geq \text{FDP}_{\hat{t}}$.

28.2.2 BH Proof

Elegant proof due to Storey, Taylor, and Siegmund: $\text{FDR} = \mathbb{E}[\text{FDP}_{\hat{t}}]$. We can write

$$\text{FDP}_t = \frac{V_t}{R_t \vee 1} = \underbrace{\frac{nt}{R_t \vee 1}}_{\widehat{\text{FDP}}_t} \underbrace{\frac{V_t}{nt}}_{M_t}.$$

Assume that $p_1, \ldots, p_n$ are independent. For $i \in \mathcal{H}_0$, assume $p_i \sim \text{Uniform}[0,1]$. Define

$$\mathcal{F}_t = \sigma\big((p_i)_{i \notin \mathcal{H}_0}, (p_i \vee t)_{i \in \mathcal{H}_0}\big).$$

If $s \leq t$, then $\mathcal{F}_t \subseteq \mathcal{F}_s$. So, $(\mathcal{F}_t)_{t=1}^{0}$ is a filtration.

Proposition 28.1. *(a)* $(M_t)_{t=1}^{\alpha/n}$ *is a MG with respect to* $(\mathcal{F}_t)_{t=1}^{\alpha/n}$.

(b) $\hat{t}$ *is a stopping time.*

Then,

$$\begin{aligned}\mathbb{E}[\mathrm{FDP}_{\hat{t}}] &= \mathbb{E}[\widehat{\mathrm{FDP}}_{\hat{t}} \cdot M_{\hat{t}}] \\ &= \alpha\, \mathbb{E}[M_{\hat{t}}] \\ &= \alpha\, \mathbb{E}[M_1] = \alpha \frac{|\mathcal{H}_0|}{n}\end{aligned}$$

because

$$\begin{aligned}M_1 &= \frac{V_1}{n} \\ &= \frac{|\mathcal{H}_0|}{n}.\end{aligned}$$

Proof of 28.1. $\mathcal{F}_t = \sigma((p_i)_{i \notin \mathcal{H}_0}, (p_i \vee t)_{i \in \mathcal{H}_0})$. For $s \le t$,

$$\begin{aligned}\mathbb{E}(M_s \mid \mathcal{F}_t) &= \frac{1}{ns}\, \mathbb{E}\big(V_s \bigm| (p_i \vee t)_{i \in \mathcal{H}_0}\big) \\ &= \frac{1}{ns} \sum_{i \in \mathcal{H}_0} \mathbb{P}(p_i \le s \mid p_i \vee t).\end{aligned}$$

Now,

$$\mathbb{P}(p_i \le s \mid p_i \vee t) = \begin{cases} 0, & p_i > t \quad [p_i \vee t > t] \\ \dfrac{s}{t}, & p_i < t \quad [p_i \vee t = t] \end{cases}$$

so

$$\begin{aligned}\mathbb{E}(M_s \mid \mathcal{F}_t) &= \frac{1}{ns} \sum_{i \in \mathcal{H}_0} \frac{s}{t}\, \mathbb{1}\{p_i \le t\} \\ &= \frac{1}{nt} V_t = M_t.\end{aligned}$$

Can we evaluate $\{\hat{t} \ge t\}$ based on $\mathcal{F}_t$? $\{\hat{t} \ge t\} = \{\widehat{\mathrm{FDP}}_s \le \alpha \text{ for some } s > t\}$. □

The martingale proof is fragile, but the problems can be repaired:

- FDR $\le \alpha$ if the nulls are $\ge$ Uniform$[0, 1]$ and "positively dependent".
- FDR $\le \alpha \log n$, so we could use the BH Procedure with level $\alpha/(\log n)$.