Statistics 210A Lecture Notes
Theoretical Statistics

Professor: Will Fithian

Contents

1 Basics of Measure Theory

1.1 Motivation for measure theory

Suppose $X \sim N(0,1)$ has a Normal distribution. We have a probability distribution $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. We can evaluate the expectation of a function by

$$\mathbb{E}[f(X)] = \int f(x)\phi(x)\,dx.$$

If we have a Binomial random variable $Y \sim \text{Binom}(n,p)$, we can write the expectation as

$$\mathbb{E}[f(Y)] = \sum_{y=0}^{n} f(y)q(y).$$

If we have $Z = \max(X, 0)$ and we want to calculate the expectation, we could do

$$\mathbb{E}[f(Z)] = \frac{1}{2}f(0) + \int_0^\infty f(z)\phi(z)\,dz.$$

In general, we could have probability distributions on other sets, such as orthogonal matrices. We want a consistent notation for all of these situations. We want to be able to write something like $\int f(x)\,dP(x)$.

Especially in applied contexts, you may never need the full power of measure theory, but measure theory helps us understand what it means to, for example, condition on an event (and whether we can do so with probability 0 events).

1.2 Measures

Measure theory is a rigorous grounding for probability theory.

Definition 1.1. Let $\mathcal{X}$ be a set. A **measure** μ maps subsets $A \subseteq \mathcal{X}$ to non-negative numbers: $\mu(A) \in [0, \infty]$.

Example 1.1. Let $\mathcal{X}$ be countable (e.g. $\mathcal{X} = \mathbb{Z}$). The **counting measure** is $\#(A) =$ # points in A.

Example 1.2. Let $\mathcal{X} = \mathbb{R}^n$. **Lebesgue measure** is the usual notion of volume: $\lambda(A) = \int \cdots \int_A dx_1 \cdots dx_n$. It is not actually possible to assign a measure to every subset of $\mathbb{R}^n$; it is possible to construct such a non-measurable set using the axiom of choice.

Example 1.3. The standard Gaussian distribution is a measure. If $Z \sim N(0,1)$ and $X = \mathbb{R}$, we can make a measure via $P(A) = \mathbb{P}(Z \in A) = \int_A \phi(x)\,dx$.

In general, the domain of a measure μ is a collection of subsets of $\mathcal{X}$: $\mathcal{F} \subseteq 2^{\mathcal{X}}$. Such a collection is called a **σ-field** and satisfies certain properties.

Example 1.4. If $\mathcal{X}$ is countable, such as with counting measure, then we can take $\mathcal{F} = 2^{\mathcal{X}}$.

Example 1.5. If $\mathcal{X} = \mathbb{R}^n$, we can take $\mathcal{F}$ to be the Borel σ-field, which is the σ-field you get if you want to be able to measure all the open subsets of $\mathbb{R}^n$.

Definition 1.2. Given a **measurable space** $(X, \mathcal{F})$, a **measure** is a map $\mu : \mathcal{F} \to [0, \infty]$ with $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ for disjoint A_i.

Notice that measures can take infinite values, such as $\#(\mathbb{Z}) = \infty$.

Definition 1.3. A **probability measure** is a measure μ with $\mu(\mathcal{X}) = 1$.

1.3 Integration with respect to measures

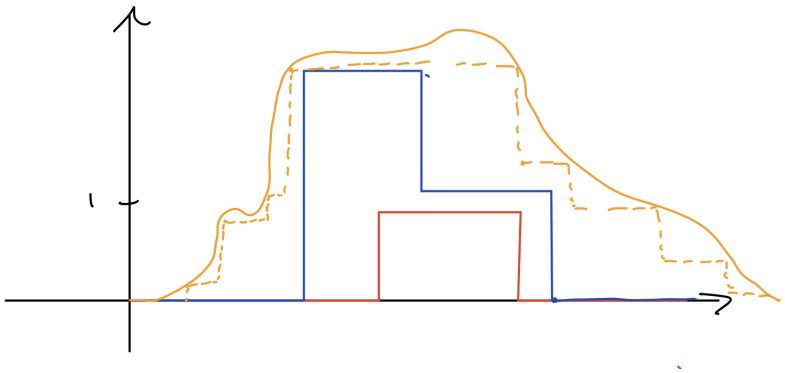
We want to be able to talk about what $\int f(x)\, d\mu(x)$ means for a nice enough function f. Define

$$\int \mathbb{1}_{\{x \in A\}}\, d\mu(x) = \mu(A).$$

Extend this definition to other f by linearity and limits:

$$\int \sum_{j=1}^{n} c_i \mathbb{1}_{\{x \in A_i\}}\, d\mu(x) = \sum_{j=1}^{n} c_i \mu(A_i).$$

With functions like this, we can approximate a wide class of functions:



Example 1.6. With counting measure,

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

Example 1.7. With Lebesgue measure,

$$\int f\, d\lambda = \int \cdots \int f(x)\, dx_1 \cdots dx_n.$$

Example 1.8. With the Gaussian distribution,

$$\int f\, dP = \int f\phi\, dx.$$

With a discrete distribution, we cannot write it as a density with respect to Lebesgue measure. When do we have a density?

1.4 Densities

Definition 1.4. Given $(X, \mathcal{F})$ and two measures P, μ, we say P is **absolutely continuous with respect to** μ (denoted $P \ll \mu$) if $\mu(A) = 0 \implies P(A) = 0$. We also say that μ **dominates** P.

If $P \ll \mu$, then (under mild conditions) we can define the density function $p : \mathcal{X} \to [0.\infty)$ with

$$P(A) = \int_A p(x)\, d\mu(x) = \int \mathbb{1}_{\{x \in A\}} p(x)\, d\mu(x).$$

We can also write

$$\int f(x)\, dP(x) = \int f(x)p(x)\, d\mu(x).$$

The theorem which allows us to do this is the *Radon-Nikodym theorem*. The common notation is $p(x) = \frac{dP}{d\mu}(x)$, where the density is referred to as a **Radon-Nikodym derivative**.

Example 1.9. If P is a probability distribution and μ is Lebesgue measure, we call $p(x)$ a **probability density function (pdf)**.

Example 1.10. If P is a probability distribution and μ is counting measure, we call $p(x)$ a **probability mass function (pmf)**.

1.5 Probability spaces and random variables

How would we talk about an expression like $\mathbb{P}(\frac{X_1 \cdots X_n}{Y} \geq W_{\widehat{\Theta}_{\text{MLE}}})$?

Denote Ω as the **outcome space** and $\omega \in \Omega$ as an **outcome variable**. $A \subseteq \Omega$ is called an **event**, and $\mathbb{P}(A)$ is the *probability of* A. All of these elements together are called a **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 1.5. A **random variable** is a function $X : \Omega \to \mathcal{X}$.

With this definition, we can say things like $X(\omega) = 5$. So Ω contains all the randomness, and if we knew ω, then we would know the value of X.

Definition 1.6. We say that X has **distribution** Q ($X \sim Q$) if

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

The **expectation** is then

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

In practice, we will still calculate expectations as usual.

2 Estimation and Introduction to Exponential Families

2.1 Review of measure theory

Last time, we introduced some ideas from measure theory. Let's review:

A **measure** μ assigns a "weight" to subsets $A \subseteq \mathcal{X}$ (for $A \in \mathcal{F}$).

Example 2.1. The **counting measure** is $\#(A) = \text{card}(A)$.

Example 2.2. Lebesgue measure gives $\lambda(A) = \text{vol}(A)$ (in $\mathbb{R}^n$).

Example 2.3. The **Gaussian distribution** gives $P(A) = \int_A \phi(x)\,dx$.

Measures give rise to integrals:

$$\int f(x)\,d\mu(x) = \begin{cases} \mu(A) & f(x) = \mathbb{1}_{\{x \in A\}} \\ \sum_i c_i \mu(A_i) & f(x) = \sum_i c_i \mathbb{1}_{\{x \in A_i\}} \\ \text{limit} & f(x) \text{ nice enough.} \end{cases}$$

If $P \ll \mu$ (meaning $\mu(A) = 0 \implies P(A) = 0$), there is a **density** $p(x)$ with $p : \mathcal{X} \to [0, \infty)$ such that $\int f\,dP = \int fp\,d\mu$ for all (nice) f.

The **outcome space** Ω containing outcomes ω is equipped with a measure $\mathbb{P}$. Random variables are functions with $X(\omega) \in \mathcal{X}$ (e.g. $\mathcal{X} = \mathbb{R}$). You can think of X "decoding" the randomness ω to tell you what the value in our nicer space $\mathcal{X}$ is. We write $X \sim Q$ if $\mathbb{P}(X \in B) = Q(B)$.

2.2 Estimation

In statistics, there are multiple possible distributions that could have generated the data.

Definition 2.1. A **statistical model** is a family of candidate probability distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for a random variable $X \sim P_\theta$. X is called the **data**, and θ is called the **parameter**.

The data X is observed by the statistical analyst, whereas θ is unobserved by the analyst. For now, θ is fixed and unknown.[1] The goal of estimation is to observe $X \sim P_\theta$ and guess the value of some estimand $g(\theta)$.

Example 2.4. Flip a biased coin n times. The parameter $\theta \in [0, 1]$ is the probability of heads, and $X \sim \text{Binom}(n, \theta)$ is the number of heads after n flips. X has a density $p_\theta(x) = \theta^x (1-x)^{n-x} \binom{n}{x}$ for $x = 0, 1, \ldots, n$ (this is a density with respect to counting measure on $\{0, 1, \ldots, n\}$).

[1] This is a frequentist perspective. With a Bayesian perspective, we may assume that θ follows some distribution.

Definition 2.2. A **statistic** is any function $T(X)$ of X.

In particular, a statistic is not a function of θ. It is something the statistical analyst can calculate.

Definition 2.3. An **estimator** $\delta(X)$ of $g(\theta)$ is a statistic intended to guess $g(\theta)$.

Example 2.5. In our coin flipping example, the natural estimator for θ is $\delta_0(X) = X/n$.

2.3 Loss and risk

How can we tell if an estimator is good?

Definition 2.4. The **loss function** $L(\theta, d)$ measures how badly an estimate is.

Example 2.6. One important loss function is the **squared error loss** $L(\theta, d) = (d - g(\theta))^2$.

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

Definition 2.5. The **risk function** $R(\theta; \delta(\cdot)) = \mathbb{E}_\theta[L(\theta, \delta(X))]$ is the expected loss as a function of θ.

Remark 2.1. The $\mathbb{E}_\theta$ notation refers to the expectation with respect to X, where θ is the true parameter. This is in contrast to other disciplines which use the notation $\mathbb{E}_X$ to denote what variables we are conditioning on in the expectation. We will use the notation $\mathbb{E}[f(X, X') \mid X']$ when we want to only integrate over certain random variables.

Example 2.7. The **mean squared error** is the risk function $\mathrm{MSE}(\theta, \delta_0(\cdot)) = \mathbb{E}_\theta[(\delta(x) - \theta)^2]$.

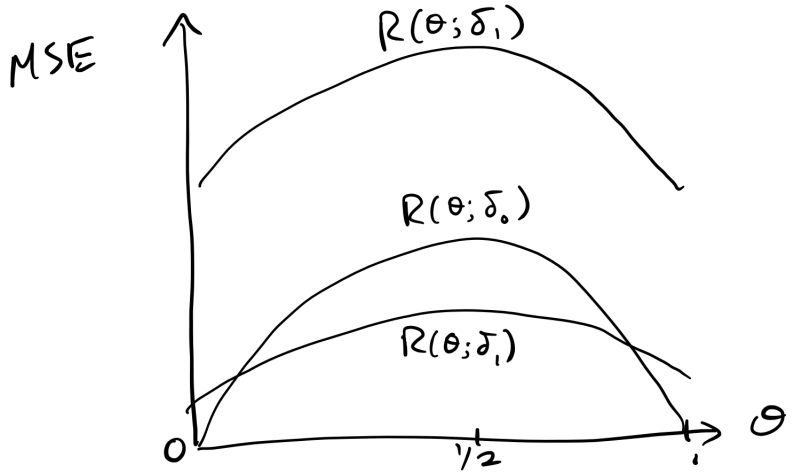
Example 2.8. In our coin flipping example, we have the estimator $\delta_0(X) = X/n$ with $\mathbb{E}_\theta[X/n] = \theta$ (this is an **unbiased estimator**). The loss is

$$\begin{aligned} \mathrm{MSE}(\theta, \delta_0(\cdot)) &= \mathbb{E}_\theta[(\delta(x) - \theta)^2] \\ &= \mathrm{Var}_\theta(X/n) \\ &= \frac{\theta(1-\theta)}{n}. \end{aligned}$$

Here are other choices of estimators. We could take

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

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



There is no estimator which is always the best; if $\theta = 3/4$, then the constant estimator $\delta(X) = 3/4$ would be better than any estimator which has a chance of suggesting anything other than 3/4.

2.4 Comparing estimators

Definition 2.6. An estimator $\delta(X)$ is **inadmissible** if there exists another estimator $\delta^*(X)$ such that

(a) $R(\theta; \delta^*) \leq R(\theta; \delta)$ for all θ,

(b) $R(\theta; \delta^*) < R(\theta; \delta)$ for some θ.

In our previous example, δ_0 rendered δ_1 inadmissible.
Here are some strategies to resolve the ambiguity:

1. Summarize $R(\theta)$ by a scalar:
   (a) **Average-case risk**: Minimize $\int_\Theta R(\theta; \delta)\, d\Lambda(\theta)$. The minimizer δ is called the **Bayes estimator**.
   (b) **Worse-case risk**: Minimize $\sup_{\theta \in \Theta} R(\theta; \delta)$. The minimizer δ is called the **minimax estimator**.
2. Constrain the choice of estimator:
   (a) Only consider **unbiased** $\delta(X)$ ($\mathbb{E}_\theta[\delta(X)] = g(\theta)$).

2.5 Exponential families

Definition 2.7. An s**-parameter exponential family** is a family $\mathcal{P} = \{P_\eta : \eta \in \Xi\}$ with densities $p_\eta(x)$ with respect to a common dominating measure μ on $\mathcal{X}$ of the form

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

where

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

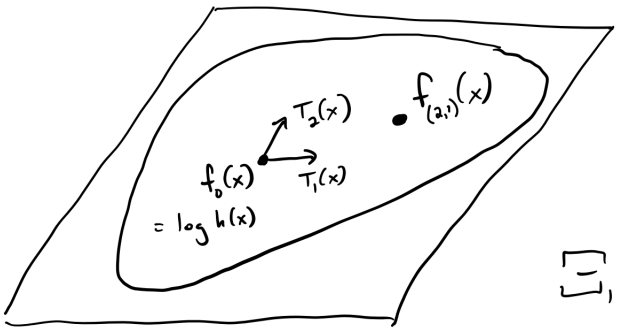
Remark 2.2. $A(\eta)$ is totally determined by h, T, since we always must have $\int_{\mathcal{X}} p_\eta \, d\mu = 1$ for all η. So we can solve

$$A(\eta) = \log \left[\int_\eta e^{\eta^\top T(x)} h(x) \, d\mu(x) \right] \leq \infty.$$

Definition 2.8. p_η is **normalizable** if $A(\eta) < \infty$. The **natural parameter space** is $\Xi_1 = \{\eta : A(\eta) < \infty\}$. We say $\mathcal{P}$ is in **canonical form** if $\Xi = \Xi_1$.

Remark 2.3. $A(\eta)$ is a convex function, so Ξ_1 is a convex set.

In general, you can think of an s-parameter exponential family as describing an s-dimensional hyperplane in the space of log-densities.



3 Exponential Families and Differential Identities

3.1 Examples of exponential families

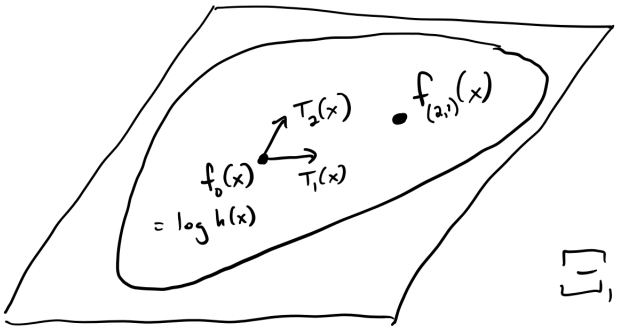
Recall from last time that an s**-parameter exponential family** is a family $\mathcal{P} = \{P_\eta : \eta \in \Xi\}$ with densities

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

with respect to a base measure μ on $\mathcal{X}$. Here,

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

Last time, we mentioned that we can think of an s-parameter exponential family as an s dimensional hyperplane in the space of log densities.



An important thing to note about this picture is that it shows us that the h and T are not unique. Only the span really matters.

Sometimes it is more convenient to use a different parameterization than the natural parameter:

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

Example 3.1. Consider the family of Gaussian distributions, $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Here, $\theta = (\mu, \sigma^2)$. To describe this as an exponential family, we have

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

$$= \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right).$$

So we have

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

In terms of η, we can say

$$p_\eta(x) = \exp\left(\eta^\top \begin{bmatrix}x \\ x^2\end{bmatrix} - A(\eta)\right), \qquad A(\eta) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log(-\pi/\eta_2).$$

Example 3.2. Now suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Then

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

So we have

$$\eta(\theta) = \begin{bmatrix}\mu/\sigma^2 \\ -1/(2\sigma^2)\end{bmatrix}, \qquad T(x) = \begin{bmatrix}\sum_i x_i \\ \sum_i x_i^2\end{bmatrix}, \qquad h(x) = 1, \qquad B(\theta) = nB^{(1)}(\theta).$$

Proposition 3.1. *Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_\eta^{(i)}(x) = e^{\eta^\top T(x) - A(\eta)}h(x)$. Then the distribution of $X = (X_1, \ldots, X_n)$ follows an exponential family with sufficient statistic $\sum_{i=1}^n T(x_i)$ and cumulant generating function $nA(\eta)$.*

Proof.

$$\begin{aligned} X &\sim \prod_{i=1}^n p_\eta^{(i)}(x_i) \\ &= \prod_{i=1}^n e^{\eta^\top T(x_i) - A(\eta)}h(x_i) \\ &= \exp\left(\eta^\top \sum_i T(x_i) - nA(\eta)\right)\prod_{i=1}^n h(x_i). \end{aligned}$$

□

$T(X)$ also follows a closely related exponential family.

Proposition 3.2. *Suppose $X \in \mathcal{X}$ and $T(X) \in \mathcal{T} \subseteq \mathbb{R}^s$ with $h(x) = 1$ and $X \sim p_\eta(x) = e^{\eta^\top T(x) - A(\eta)}$ with respect to μ. For a set $B \subseteq \mathcal{T}$, define $\nu(B) = \mu(T^{-1}(B))$. Then*

$$T(X) \sim q_\eta(t) = e^{\eta^\top t - A(\eta)}$$

with respect to ν.

Example 3.3. In the discrete case, this is

$$\begin{aligned}
\mathbb{P}_\eta(T(X) \in B) &= \sum_{x : T(x) \in B} e^{\eta^\top T(x) - A(\eta)} \mu(\{x\}) \\
&= \sum_{t \in B} \sum_{x : T(x) = t} e^{\eta^\top t - A(\eta)} \mu(\{x\}) \\
&= \sum_{t \in B} e^{\eta^\top t - A(\eta)} \underbrace{\mu(T^{-1}(\{x\}))}_{\nu(\{x\})}.
\end{aligned}$$

So $T \sim e^{\eta^\top t - A(\eta)}$ with respect to ν.

Example 3.4. Let $X \sim \text{Binomial}(n, \theta)$. We can turn this into an exponential family as follows: For $\theta \in (0, 1)$,

$$\begin{aligned}
p_\theta(x) &= \theta^x (1-\theta)^{n-x} \binom{n}{x} \\
&= \left(\frac{\theta}{1-\theta}\right)^x (1-\theta)^n \binom{n}{x} \\
&= \exp\left(x \log \frac{\theta}{1-\theta} + n \log(1-\theta)\right) \binom{n}{x}
\end{aligned}$$

The natural parameter is $\eta(\theta) = \log \frac{\theta}{1-\theta}$.

Example 3.5. Let $X \sim \text{Pois}(\theta)$ with density $p_\lambda(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ with respect to counting measure on $\mathbb{N}$. This is an exponential family

$$p_\lambda(x) = \exp\left((\log \lambda) x - \lambda\right) \frac{1}{x!}$$

with natural parameter $\eta(\lambda) = \log \lambda$.

Most of the families of distributions you can find on, say, Wikipedia, will be exponential families.

3.2 Differential identities for the cumulant generating function

Begin with the equation

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

and then differentiate. Here is a criterion which lets us differentiate under the integral:

Theorem 3.1 (Theorem 2.4 in Keener)**.** *For $f : \mathcal{X} \to \mathbb{R}$, let $\Xi_f = \{\eta \in \mathbb{R}^s : \int |f| e^{\eta^\top T} h\, d\mu < \infty\}$. Then $g(\eta) = \int f e^{\eta^\top T} h\, d\mu$ has continuous partial derivatives of all orders for interior points $\eta \in \Xi_f^0$, and we can find them by differentiating under the integral.*

In particular, letting $f = 1$, we get that $A(\eta)$ has infinitely many partial derivatives in Ξ_1^0. So we can calculate

$$\frac{\partial}{\partial \eta_j} e^{A(\eta)} = \int \frac{\partial}{\partial \eta_j} e^{\eta^\top T(x)} h(x)\, d\mu(x),$$

which gives

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

This shows that

Proposition 3.3.

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

Taking second derivatives, we have

$$\frac{\partial^2 A}{\partial \eta_j \partial \eta_k} e^{A(\eta)} = \int \frac{\partial^2}{\partial \eta_j \partial \eta_k} e^{\eta^\top T(x)} h(x)\, d\mu(x),$$

which gives us

$$\left(\frac{\partial^2 A}{\partial \eta_j} - \frac{\partial A}{\partial \eta_j} \frac{\partial A}{\partial \eta_k} \right) = \int T_j T_k e^{\eta^\top T - A(\eta)} h\, d\mu.$$

So we get

$$\frac{\partial^2 A}{\partial \eta_j \partial \eta_k}(\eta) = \mathbb{E}_\eta[T_j T_k] - \mathbb{E}_\eta[T_j]\, \mathbb{E}_\eta[T_k] = \mathrm{Cov}(T_j, T_k).$$

In total, we get

Proposition 3.4.

$$\nabla^2 A(\eta) = Var_\eta(T(X)).$$

Differentiating repeatedly, we get

$$e^{-A(\eta)} \frac{\partial^{k_1+\cdots+k_s}}{\partial_{\eta 1}^{k_1} \cdots \partial_{\eta_s}^{k_s}} (e^{A(\eta)}) = \mathbb{E}_\eta[T_1^{k_1} \cdots T_s^{k_s}].$$

This is because $M_\eta^T(u) = e^{A(\eta+u)-A(\eta)}$ is the **moment generating function (MGF)** of $T(X)$ when $X \sim p_\eta$:

$$\begin{aligned} M_\eta^{T(X)}(u) &= \mathbb{E}_\eta[e^{u^\top T(X)}] \\ &= \int e^{u^\top T} e^{\eta^\top T - A(\eta)} h \, d\mu \\ &= e^{A(\eta+u)-A(\eta)} \underbrace{\int e^{(\eta+u)^\top T - A(\eta+u)} h \, d\mu}_{=1}. \end{aligned}$$

4 Sufficient Statistics

4.1 Recap: differential identities for exponential families

Last time, we were talking about exponential families $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with densities

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

In natural parameters, we have

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

Last time, we proved some differential identities by starting with the equation

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

and differentiating with respect to η_j. We saw that

$$\nabla A(\eta) = \mathbb{E}_\eta[T(X)], \qquad \nabla^2 A(\eta) = \mathrm{Var}_\eta(T(X)).$$

In general, we have

$$e^{-A(\eta)} \frac{\partial^{k_1 + \cdots + k_s}}{\partial_{\eta 1}^{k_1} \cdots \partial_{\eta s}^{k_s}} (e^{A(\eta)}) = \mathbb{E}_\eta[T_1^{k_1} \cdots T_s^{k_s}].$$

This is saying that $e^{A(\eta+u) - A(\eta)}$ is the **moment generating function** of T:

$$\frac{\partial}{\partial u_j} e^{A(\eta+u) - A(\eta)}|_{u=0} = \left(\frac{\partial}{\partial \eta_j} e^{A(\eta)} \right) \cdot e^{-A(\eta)}.$$

If we take logs, we get that $A(\eta + u) - A(\eta)$ is the **cumulant generating function** of $T(X)$.[2]

Here is another calculation of the MGF for $T(X)$ in an exponential family:

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

[2] We have been calling $A(\eta)$ the CGF, but technically that is only the case where $\eta = 0$.

4.2 Sufficiency

Our motivation is going to be the example of coin flipping.

Example 4.1. Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, so our data is $X \sim \prod_i \theta^{x_i}(1-\theta)^{x_i}$ on $\{0,1\}^n$. Instead of observing the whole sequence, we can observe a summary statistic $T(X) = \sum_i X_i \sim \text{Binom}(n, \theta) = \theta^t(1-\theta)^{n-t}\binom{n}{t}$ on $\{0, 1 \dots, n\}$ which only records the total number of heads. This is a lossy compression of the data $(X_1, \dots, X_n) \mapsto T(X)$. Why can we justify this?

We can think of the information in $(X_1, \dots, X_n)$ as coming in two parts: the first part is $T(X)$, which is the part relevant to estimating θ, and the second part is the ordering, which doesn't depend on θ. The reason that $T(X)$ is the important part for estimating θ is that $T(X)$ is the only part that depends on θ.

Definition 4.1. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model for data X. $T(X)$ is **sufficient** for the model $\mathcal{P}$ if $P_\theta(X \mid T)$ does not depend on θ.

Example 4.2. Continuing our coin flipping example,

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

The interpretation is that we can think of Nature as generating the data in 2 steps:

1. Generate $T(X) \sim P_\theta(T(X))$, dependent on θ.
2. Generate $X \sim P(X \mid T)$, not dependent on θ.

Sufficiency principle: If $T(X)$ is sufficient, then any statistical procedure should depend on the data X only through T.

Why should we believe in this sufficiency principle? Suppose we generate $\widetilde{X} \sim \mathbb{P}(X \mid T)$.

$$\begin{array}{ccccc}
\theta & \xrightarrow{\text{nature}} & T(X) & \longrightarrow & X \\
 & & & \searrow^{\text{us}} & \\
 & & & & \widetilde{X}
\end{array}$$

Then $\widetilde{X} \overset{d}{=} X$, so any estimator gives $\delta(\widetilde{X}) \overset{d}{=} \delta(X)$. So we should always be fine using $T(X)$, since we don't really lose any information by using it. Later, we will see that using sufficient statistics can reduce the loss we incur in estimation.

4.3 Factorization theorem for sufficient statistics

Theorem 4.1 (Fisher-Neyman)**.** *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model with densities $p_\theta(x)$ with respect to a common dominating measure μ. Then T is sufficient for $\mathcal{P}$ if and only if there exist nonnegative functions g_θ, h such that $p_\theta(x) = g_\theta(T(x))h(x)$ for μ-a.e. x.*

Here is a "physics proof." For a careful proof, check Keener.

Proof. ($\Longleftarrow$):

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

($\Longrightarrow$): Take

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

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

for any fixed $\theta_0 \in \Theta$. Then

$$\begin{aligned} g_\theta(T(x))h(x) &= \mathbb{P}(T(X) = T(x))\mathbb{P}_\theta(X = x \mid T(X) = T(x)) \\ &= p_\theta(x). \end{aligned}$$

□

Example 4.3. For exponential families,

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

so T is sufficient for θ.

Example 4.4. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_\theta^{(1)}$ for any model $\mathcal{P}^{(1)} = \{P_\theta^{(1)} : \theta \in \Theta\}$ on $\mathcal{X} \subseteq \mathbb{R}$. $P_\theta^{(1)}$ is invariant to permuting $X = (X_1, \ldots, X_n)$. The **order statistics** $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ are defined by $X_{(k)} =$ the k-th smallest value (counting repeats). For example, if $X = (1, 3, 3, -1)$, then $X_{(1)} = -1, X_{(2)} = 1, X_{(3)} = 3, X_{(4)} = 3$.

If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_\theta^{(1)}$ is any univariate model $\mathcal{P}^{(1)}$, then the order statistics are sufficient. For a more general $\mathcal{X}$, we can say the **empirical distribution**

$$\widehat{P}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(\cdot)$$

is sufficient.

4.4 Minimal sufficiency

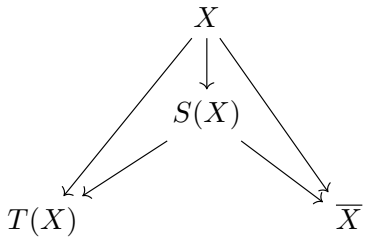
Example 4.5. Consider $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. The following statistics are sufficient:

$$T(X) = \sum_i X_i, \qquad \overline{X} = \frac{1}{n} \sum_i X_i,$$

$$S(X) = (X_{(1)}, \ldots, X_{(n)}), \qquad X = (X_1, \ldots, X_n).$$

It seems like the latter two statistics have more information than $T(X)$ or $\overline{X}$. These are all sufficient statistics (and in fact the data itself is always sufficient), so what should we do with regards to the sufficiency principle? The idea is to find sufficient statistics with the least amount of information, i.e. the ones that cannot recover the others.

Here is a diagram that expresses which statistics have more information than others:



Next time, we will talk about minimal sufficient statistics, which have minimal information while remaining sufficient.

5 Minimal Sufficient and Complete Statistics

5.1 Recap: sufficient statistics

Last time, we talked about sufficient statistics. We said that $T(X)$ is **sufficient** for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if the distribution of $X \mid T(X)$ does not depend on θ. We encountered the **sufficiency principle**, which said that we should only attend to sufficient statistics T in our statistical analysis, rather than the whole data.

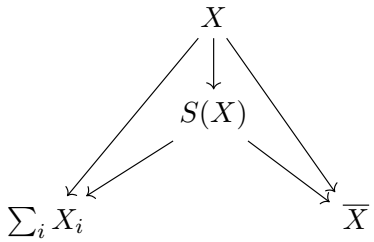
The factorization theorem says that if $\mathcal{P}$ has densities $p_\theta(x)$ with respect to μ, then $T(X)$ is sufficient iff there exist functions g_θ, h such that $p_\theta(x) = g_\theta(T(x))h(x)$. For exponential families, we have

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

Here are a few examples we saw last time:

Example 5.1 (Order statistics)**.** If $X_1, \dots, X_n \in \mathbb{R}$, $X_{(k)}$ is the k-th smallest value (including repeats). Then if $X_1, \dots, X_n \overset{\text{iid}}{\sim} P^{(1)}$ with any model for $P^{(1)}$ on $\mathbb{R}$, then $S(X) = (X_{(1)}, \dots, X_{(n)})$ is sufficient.

Example 5.2. If $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$, then we have the following hierarchy of sufficient statistics:



The higher up statistics in this diagram can be "compressed" more to get the ones at the bottom, which we may think of as minimal sufficient (or the most compressed).

5.2 Minimal sufficient statistics

Proposition 5.1. *If $T(X)$ is sufficient and $T(X) = f(S(X))$, then $S(X)$ is sufficient.*

So statistics with more information than sufficient statistics are also sufficient.

Proof. Using the factorization theorem,

$$\begin{aligned} p_\theta(x) &= g_\theta(T(x))h(x) \\ &= (g_\theta \circ f)(S(x))h(x). \end{aligned}$$

□

Here are the sufficient statistics with the least information.

Definition 5.1. $T(X)$ is **minimal sufficient** if

1) $T(X)$ is sufficient.

2) For any other sufficient statistic $S(X)$, $T(X) = f(S(X))$ for some f (a.s. in $\mathcal{P}$).

5.3 Likelihood functions

We will see that the shape of all likelihood ratios will be minimal sufficient, so any statistic that has the same information will be minimal sufficient.

Definition 5.2. If $\mathcal{P}$ has densities $p_\theta(x)$ with respect to μ the **likelihood function** (resp. **log-likelihood**) is the density (resp. **log-density**), reframed as a *random function* of θ.

$$\text{Lik}(\Theta; X) = p_\theta(X), \qquad \ell(\theta; X) = \log \text{Lik}(\theta; X).$$

If T is sufficient, then

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

Theorem 5.1. *Assume $\mathcal{P}$ has densities p_θ and $T(X)$ is sufficient for $\mathcal{P}$. If*

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

then $T(x)$ is minimal sufficient.

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

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

which is a contradiction. □

5.4 Minimal sufficiency in exponential families

Example 5.3. For an exponential family,

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

Is $T(X)$ minimal? Assume $\text{Lik}(\theta; x) \propto_\theta \text{Lik}(\theta; y)$, We want to show that $T(x) = T(y)$.

$$\begin{aligned}
\text{Lik}(\theta; x) \propto_\theta \text{Lik}(\theta; y) &\iff e^{\eta(\theta)^\top T(x) - B(\theta)} h(x) \propto_\theta e^{\eta(\theta)^\top T(y) - B(\theta)} h(y) \qquad \forall \theta \\
&\iff e^{\eta(\theta)^\top T(x)} = e^{\eta(\theta)^\top T(y)} c(x, y) \qquad \forall \theta \\
&\iff \eta(\theta)^\top T(x) = \eta(\theta)^\top T(y) + a(x, y) \qquad \forall \theta \\
&\iff \eta(\theta)^\top (T(x) - T(y)) = a(x, y) \qquad \forall \theta.
\end{aligned}$$

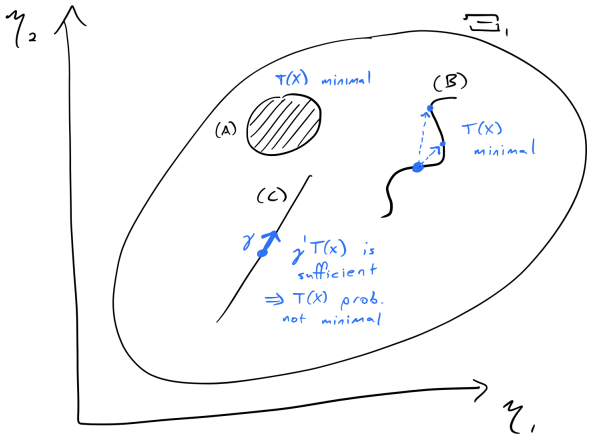
Plug in θ_1 and θ_2 to get 2 different equations and subtract:

$$\begin{aligned}
&\implies (\eta(\theta_1) - \eta(\theta_2))^\top (T(x) - T(y)) = 0 \qquad \forall \theta_1, \theta_2 \\
&\iff T(x) - T(y) \perp \text{span}\{\eta(\theta_1) - \eta(\theta_2) : \theta_1, \theta_2 \in \Theta\}
\end{aligned}$$

If $\text{span}\{\eta(\theta_1) - \eta(\theta_2) : \theta_1, \theta_2 \in \Theta\} = \mathbb{R}^s$, then we will get $T(x) = T(y)$.

Suppose $\eta(\theta) = \begin{bmatrix} \theta \\ 0 \end{bmatrix}$. Then $T_1(x)$ is sufficient. Does this mean that T cannot be minimal sufficient? In a $N(\mu, \sigma^2)$ family with $n = 1$, then $T(X) = \begin{bmatrix} X \\ X^2 \end{bmatrix}$. But if $n = 10$, then $T(x) = \begin{bmatrix} \sum_i X_i \\ \sum_i X_i^2 \end{bmatrix}$ in which case T cannot be recovered from T_1. So in general, it is possible that $T(X)$ may not be sufficient.

Here is a picture of exponential families A, B, and C in the natural parameter space Ξ.



- In exponential family A, the parameter space is locally 2-dimensional, so we get the whole span. Thus, $T(X)$ will be minimal.
- In exponential family B, we still get two vectors that span $\mathbb{R}^2$, so $T(X)$ is still minimal.

- In exponential family C, $\gamma^\top T(x)$ is minimal, where γ lies along the line. But $T(X)$ may not be minimal. If we say $\eta(\theta) = a + \theta\gamma$ with $\theta \in \mathbb{R}$, then $\eta^\top T(x) = a^\top T(x) + \theta\gamma^\top T(x)$.

Example 5.4. If $X \sim N_2(\mu(\theta), I_2) = e^{\mu(\theta)^\top x - B(\theta)} e^{-(1/2)x^\top x}$ with $\theta \in \mathbb{R}$. If $\Theta = \mathbb{R}$, $\mu(\theta) = a + \theta b$ with $a, b \in \mathbb{R}^2$, then

$$p_\theta(x) = e^{\theta(b^\top x) - B(\theta)} e^{-(1/2)(x-2a)^\top x}.$$

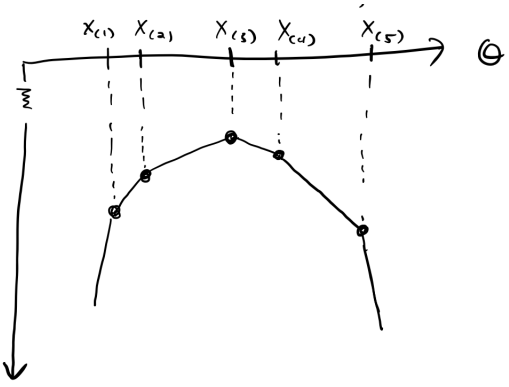
Because $b^\top x$ is sufficient, X is not minimal sufficient.

Example 5.5 (Laplace location family)**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_\theta^{(1)}(x) = \frac{1}{2} e^{-|x-\theta|}$. Then

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

so

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



Here, $(X_{(1)}, \ldots, X_{(n)})$ is minimal sufficient.

In many examples beyond exponential families, there aren't any useful sufficient statistics.

5.5 Complete statistics

Definition 5.3. $T(X)$ is **complete** for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if

$$\mathbb{E}_\theta[f(T)] = 0 \quad \forall\theta \implies f(T) \overset{\text{a.s.}}{=} 0 \quad \forall\theta$$

You should think of this as an upgrading of minimality.

Example 5.6. In the Laplace location family, are there any complete statistics? Let $f(S(X)) = \text{Med}(X) - \overline{X}$. Then $\mathbb{E}_\theta[f(S(x))] = \theta - \theta = 0$, but $\text{Med}(X) \neq \overline{X}$ a.s.

Definition 5.4. Let $\mathcal{P}$ be an exponential family with $p_\theta(x) = e^{\eta(\theta)^\top T(x) - B(\theta)} h(x)$. If $\Xi_0 = \eta(\Theta) = \{\eta(\theta) : \theta \in \Theta\}$ contains an open set, then we say $\mathcal{P}$ is **full-rank**. Otherwise, $\mathcal{P}$ is called **curved**.

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

For a proof, see Lehmann and Romano Theorem 4.3.1.

Going back to our previous examples, in family A, T will be complete, whereas in families B and C, T will probably not be complete.

Theorem 5.3. *If $T(X)$ is complete sufficient for $\mathcal{P}$, then $T(X)$ is minimal.*

Proof. Assume $S(X)$ is minimal sufficient; we will recover T from S. Then $S(X) = f(T(X))$. Define

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

This is a proper statistic (not depending on θ) due to conditioning on the sufficiency of the statistic S. Then let $g(t) = t - m(f(t))$. Then

$$\mathbb{E}_\theta[g(T)] = \mathbb{E}_\theta[T] - \mathbb{E}_\theta[\mathbb{E}_\theta[T \mid S]] = 0 \qquad \forall \theta,$$

so $g(T) \overset{\text{a.s.}}{=} 0$ by completeness. This says that $T \overset{\text{a.s.}}{=} m(S(X))$. □

5.6 Ancillary statistics

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

This is a statistic that we already know without knowing θ.

Theorem 5.4 (Basu)**.** *If $T(X)$ is complete sufficient and $V(X)$ is ancillary, then $T \amalg V$ for all θ.*

Remark 5.1. Completeness is a property of the model, whereas independence is just a property of the distributions.

6 Basu's Theorem, Rao-Blackwell, and Unbiased Estimation

6.1 Recap: Minimal sufficient, complete, and ancillary statistics

Last time we discussed **minimal sufficient** statistics, which are

1) $T(X)$ is sufficient.

2) For any other sufficient statistic $S(X)$, $T(X) = f(S(X))$ for some f (a.s. in $\mathcal{P}$).

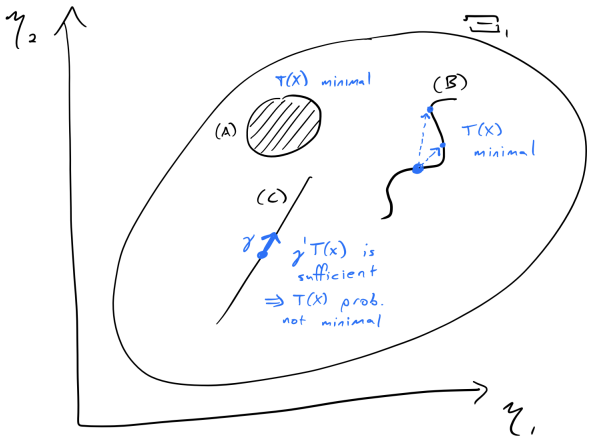
For an s-parameter exponential family with $p_\theta(x) = e^{\eta(\theta)^\top T(x) - B(\theta)} h(x)$, $T(X)$ is minimal if

$$\text{span}\{\eta(\theta_1) - \eta(\theta_2) : \theta_1, \theta_2 \in \Theta\} = \mathbb{R}^s.$$

We also discussed **complete statistics**, which have the property that

$$\mathbb{E}_\theta[f(T(x))] = \theta \quad \forall 0 \implies f(T(x)) \overset{\text{a.s.}}{=} 0.$$

We saw that in an exponential family, $T(X)$ is complete if Ξ contains an open set, but this is not a necessary condition. In the following picture, $T(X)$ will be complete in exponential families A and B but not necessarily in family C.



Family B is usually called a **curved exponential family** since the parameter space is a lower-dimensional space within the natural parameter space.

We saw that completeness is an upgrading of minimality for sufficient statistics:

Theorem 6.1. *Complete sufficient statistics are minimal.*

We also introduced **ancillary statistics** $V(X)$, where the distribution of V doesn't depend on θ.

6.2 Basu's theorem

Theorem 6.2 (Basu)**.** *If $T(X)$ is complete sufficient and $V(X)$ is ancillary for $\mathcal{P}$, then $V(X) \amalg T(X)$ for all $\theta \in \Theta$.*

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

$$\mathbb{P}_\theta(V \in A, T \in B) = \mathbb{P}_\theta(V \in A)\mathbb{P}_\theta(T \in B).$$

This is equivalent to showing

$$\mathbb{P}_\theta(V \in A \mid T \in B) = \mathbb{P}_\theta(V \in A)$$

whenever $\mathbb{P}_\theta(T \in B) > 0$. Let

$$q_A(T(X)) = \mathbb{P}(V \in A \mid T(X)), \qquad p_A = \mathbb{P}(V \in A).$$

Note that q_A, p_A are independent of θ. We have

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

so by completeness of $T(X)$, $q_A(T(X)) \overset{\text{a.s.}}{=} p_A$. □

Remark 6.1. The hypotheses of Basu's theorem apply to a model, whereas the conclusions apply to each distribution. So sometimes, to prove that statistics are independent, we can apply Basu's theorem to submodels of the original model.

Example 6.1. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. We want to show that $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i \amalg S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. Let $\mathcal{Q}_{\sigma^2} = \{N(\mu, \sigma^2)^n : \mu \in \mathbb{R}\}$. In this model, $\overline{X}$ is complete sufficient (which we can verify by writing this as an exponential family). To show that S^2 is ancillary, let $Z_i = X_i - \mu \overset{\text{iid}}{\sim} N(0, \sigma^2)$ (not that these are not statistics, since they suppose the value of μ). Then

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 = \underbrace{\frac{1}{n-1}\sum_{i=1}^n (Z_i - \overline{Z})^2}_{\sim \frac{\sigma^2}{n-1}\chi^2_{n-1}}$$

has distribution not depending on θ. So by Basu's theorem, $\overline{X} \amalg S^2$ for all μ, σ^2. Take note that we split the model into submodels where σ^2 was fixed.

6.3 The Rao-Blackwell Theorem

Why should we use sufficient statistics or complete statistics $T(X)$? The idea is that if θ only depends on $T(X)$, then using anything else would be adding extra randomness, or "noise," that obscures our result. To talk about what effect this has on our loss functions, let's introduce a condition on our loss functions.

Definition 6.1. A function $f(x)$ is **convex** if for any $\gamma \in (0,1)$,

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

The function f is **strictly convex** if the inequality is strict ($<$).

Jensen's inequality says that this extends to general averages, not just the average of two points.

Theorem 6.3 (Jensen's inequality)**.** *If f is convex, then*

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

If f is strictly convex, we get $<$, unless $X \overset{\text{a.s.}}{=} c$.

Here, X could be a random vector.

If the loss $L(\theta; d)$ is convex in d, then we lose by adding extra noise. Jensen's inequality tells us that more the distribution spreads out, the more the average risk increases.

Theorem 6.4 (Rao-Blackwell)**.** *Assume $T(X)$ sufficient and $\delta(X)$ is an estimator for $g(\theta)$. Let $\overline{\delta}(T(X)) = \mathbb{E}[\delta(X) \mid T(X)]$. If $L(\theta; d)$ is convex, then*

$$R(\theta; \overline{\delta}) \leq R(\theta; \delta) \quad \forall \delta.$$

If $L(\theta; d)$ is strictly convex, then the inequality is strict, unless $\overline{\delta} \overset{\text{a.s.}}{=} \delta$.

Proof. The risk is

$$R(\theta; \overline{\delta}) = \mathbb{E}_\theta[L(\theta; \mathbb{E}(\delta \mid T))]$$

By Jensen's inequality (applied to the conditional expectation given T),

$$\begin{aligned} &\leq \mathbb{E}_\theta[\mathbb{E}[L(\theta; \delta) \mid T]] \\ &= \mathbb{E}_\theta[L(\theta; \delta)] \end{aligned}$$

with strict inequality for strict convexity unless $\overline{\delta} \overset{\text{a.s.}}{=} \delta$. □

Remark 6.2. Where did we use sufficiency in the proof? We used it when defining $\overline{\delta}$, where the conditional expectation should not depend on θ.

Turning δ into $\overline{\delta}$ is called **Rao-Blackwellization**.

6.4 Unbiased estimation

Definition 6.2. The **bias** of an estimator $\delta(X)$ for $g(\theta)$ is

$$\text{Bias}_\theta(\delta(X)) = \mathbb{E}_\theta\, \delta(X) - g(\theta).$$

The statistic $\delta(X)$ is **unbiased** for $g(\theta)$ if $\mathbb{E}_\theta[\delta(X)] = g(\theta)$ for all θ.

An unbiased estimator may not always exist.

Definition 6.3. We say $g(\theta)$ is U**-estimable** if there is an estimator $\delta(X)$ that is unbiased for $g(\theta)$.

Definition 6.4. An estimator $\delta(X)$ is **uniform minimum variance unbiased (UMVU)** if for any other unbiased $\widetilde{\delta}$, $\text{Var}_\theta(\delta(X)) \le \text{Var}_\theta(\widetilde{\delta}(X))$.

We could equivalently say $\text{MSE}(\theta;\delta) \le \text{MSE}(\theta;\widetilde{\delta})$.

Theorem 6.5 (Lehmann-Scheffé)**.** *Suppose $T(X)$ is complete sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Then for any U-estimable function $g(\theta)$, there is an a.s. unique UMVU estimator of the form $\delta(T(X))$.*

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

$$\delta(T) = \mathbb{E}[\delta_0 \mid T].$$

This is unbiased because

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

If $\widetilde{\delta}(T)$ is unbiased, then $\mathbb{E}[\delta(T) - \widetilde{\delta}(T)] = 0$ for all θ. So by completeness, $\delta(T) \overset{\text{a.s.}}{=} \widetilde{\delta}(T)$.

Now suppose $\delta^*(X)$ is unbiased. By Rao-Blackwell,

$$\text{MSE}(\theta; \underbrace{\mathbb{E}[\delta^* \mid T]}_{=\delta}) \le \text{MSE}(\theta;\delta^*). \qquad \square$$

Remark 6.3. The picture is the same for any convex loss, not just the mean squared error. For strictly convex loss, the unique UMVU has strictly less loss than any other unbiased estimator.

Remark 6.4. Unbiased estimators are not always the best, but this shows that there is at least a best one.

How do we find an unbiased estimator? Assume T is complete sufficient. We now have two options:

1. Find an unbiased estimator $\delta(T)$ which is a function of T.

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

Example 6.2 (German tank problem[3])**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U[0, \theta]$ with $\theta > 0$.

$$\begin{aligned} p(x) &= \prod_{i=1}^{n} p^{(1)}(x_i) \\ &= \prod_{i=1}^{n} \mathbb{1}_{\{0 \le x_i \le \theta\}} \frac{1}{\theta} \\ &= \frac{1}{\theta^n} \mathbb{1}_{\{0 \le X_{(n)} \le \theta\}}. \end{aligned}$$

Is the maximum complete sufficient?

$$\begin{aligned} \mathbb{P}_\theta(X_{(n)} \le t) &= \left(\frac{t}{\theta} \wedge 1 \right)^n \\ &= \left(\frac{t}{\theta} \right)^n \wedge 1, \end{aligned}$$

so the density is

$$\begin{aligned} p_\theta(t) &= \frac{d}{dt} \mathbb{P}_\theta(X_{(n)} \le t) \\ &= n \frac{t^{n-1}}{\theta^n} \mathbb{1}_{\{t \le \theta\}}. \end{aligned}$$

Suppose that for all $\theta > 0$,

$$0 = \mathbb{E}_\theta[f(T)] = \frac{n}{\theta^n} \int_0^\theta f(t) t^{n-1} \, dt.$$

Then

$$0 = \int_0^\theta f(t) t^{n-1} \, dt,$$

so differentiating with respect to θ tells us that

$$f(\theta) \theta^{n-1} = 0$$

for all $\theta > 0$.

Let's calculate

$$\mathbb{E}_\theta[X_{(n)}] = \frac{n}{\theta^n} \int_0^\theta t \cdot t^{n-1} \, dt$$

[3]Imagine you're hiding in the bushes in World War II, and you count the serial numbers. You observe the largest serial number to try to determine the number of German tanks.

$$= \frac{n}{\theta^n(n+1)}[t^{n+1}]_0^\theta$$
$$= \frac{n}{n+1}\theta.$$

So we can just get an unbiased estimator via

$$\mathbb{E}_\theta\left[\frac{n+1}{n}X_{(n)}\right] = \theta.$$

Another way to get an unbiased estimator is to use $\mathbb{E}_\theta[2X_1] = \theta$. Then you can show that

$$\mathbb{E}[2X_i \mid X_{(n)}] = \frac{n+1}{n}X_{(n)}.$$

7 Computing UMVU Estimators and Lower Bounds for Unbiased Estimation

7.1 Computing UMVU estimators

Last time, we proved **Jensen's inequality** for convex f:

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

The **Rao-Blackwell theorem** told us that if $L(\theta; d)$ is convex in d, $\delta(X)$ is an estimator, and $T(X)$ is sufficient, then $\mathbb{E}[\delta \mid T]$ is better than δ. We also saw that if $T(X)$ is complete sufficient and $g(\theta)$ is U-estimable, there is a unique unbiased estimator of the form $\delta(T)$. It is UMVU (dominates all other unbiased estimators for any convex L). We saw that there were 2 ways to find UMVU estimators:

1. Directly find an unbiased $\delta(T)$.
2. Rao-Blackwellize any unbiased $\delta(X)$.

Example 7.1. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U[0, \theta]$, then $X_{(n)}$ is complete sufficient for estimating θ. We saw that $\frac{n+1}{n} X_{(n)}$ is UMVU. However, Keener shows that among estimators of the form $cX_{(n)}$, $\frac{n+2}{n+1} X_{(n)}$ actually has the best MSE.

Example 7.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Pois}(\theta)$ with $\theta > 0$ and pmf

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

Then $T(X) = \sum_i X_i \sim \text{Pois}(n\theta)$ is complete sufficient with pmf

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

Let's estimate θ^2 with an unbiased estimator. First, we'll use Method 1: $\overline{X}^2$ is not unbiased because $\mathbb{E}[\overline{X}] = \theta$, so $\mathbb{E}[\overline{X}^2] > \theta^2$ by Jensen's inequality. Observe that

$$\begin{aligned} \delta(T) \text{ is unbiased} &\iff \sum_{t=0}^{\infty} \delta(t) p_\theta^T(t) = \theta^2 \quad \forall \theta > 0 \\ &\iff \sum_{t=0}^{\infty} \delta(t) \frac{n^t \theta^t}{t!} = \theta^2 e^{n\theta} \quad \forall \theta > 0. \end{aligned}$$

Write $\theta^2 e^{n\theta} = \sum_{k=0}^{\infty} \frac{n^k \theta^{k+2}}{k!} = \sum_{j=2}^{\infty} \frac{n^{j-2}}{(j-2)!} \theta^j$. So we get $\delta(0) = \delta(1) = 0$, and for $t \geq 2$, $\delta(t) = \frac{n^{t-2}}{(t-2)!} \cdot \frac{t!}{n^t} = \frac{t(t-1)}{n^2}$. We can write this more compactly as

$$\delta(t) = \frac{t(t-1)}{n^2}, \qquad t = 0, 1, \ldots$$

Now we use Method 2, Rao-Blackwellization: We know that $\mathbb{E}_\theta[X_1X_2] = (\mathbb{E}_\theta[X_1])^2 = \theta^2$, so we want to condition X_1X_2 on $T = \sum_i X_i$. Since $X \mid T = t \sim \text{Multinomial}(t, 1/n1_n)$, we can check that $X_1 \mid T = t \sim \text{Binom}(t, 1/n)$ and $X_2 \mid X_1 = x_1, T = t \sim \text{Binom}(t - x_1, 1/(n-1))$. So we can compute

$$\mathbb{E}\left[X_1X_2 \mid \sum_i X_i\right] = \delta(T),$$

as before.

7.2 Differential identities for the score function

Assume that $\mathcal{P}$ has densities p_θ with respect to μ with $\Theta \subseteq \mathbb{R}^d$. Suppose there is a **common support** $\{x : p_\theta(x) > 0\}$ which is the same for all θ. We have the log-likelihood $\ell(\theta; x) = \log p_\theta(x)$.

Definition 7.1. Define the **score function** to be $\nabla\ell(\theta; x)$.

We have

$$p_{\theta+\eta}(x) = e^{\ell(\theta+\eta;x)} \approx p_\theta(x)e^{\eta^\top \nabla\ell(\theta,x)}$$

for small η. So we can think of this as locally looking like an exponential family withe the score function looking like a complete sufficient statistic.

We have differential identities, similar to in an exponential family. Start with

$$1 = \int_{\mathcal{X}} e^{\ell(\theta,x)}\, d\mu(x)$$

Taking $\frac{\partial}{\partial\theta_j}$ on both sides, we get

$$0 = \int_{\mathcal{X}} \frac{\partial}{\partial_j}\ell(\theta; x)e^{\ell(\theta;x)}\, d\mu(x).$$

This gives the identity

$$\mathbb{E}_\theta[\nabla\ell(\theta; X)] = 0.$$

It is important that we are integrating using the same θ that we plug into the score function.

If we differentiate again with respect to θ_k, we get

$$0 = \int_{\mathcal{X}} \left(\frac{\partial^2\ell}{\partial\theta_j\partial\theta_k} + \frac{\partial\ell}{\partial\theta_j}\frac{\partial\ell}{\partial\theta_k}\right) = \mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\ell(\theta; X)\right] + \mathbb{E}\left[\frac{\partial\ell}{\partial\theta_j}(\theta; X)\frac{\partial\ell}{\partial\theta_k}(\theta; X)\right]$$

which gives the identity

$$J(\theta) := \mathbb{E}_\theta[-\nabla^2\ell(\theta; X)] = \text{Var}_\theta(\nabla\ell(\theta; X)).$$

The quantity $J(\theta)$ is called the **Fisher information**.

7.3 The Cramér-Rao lower bound

Let's relate this back to a statistic $\delta(X)$. Suppose

$$g(\theta) = \mathbb{E}_\theta[\delta(X)] = \int_{\mathcal{X}} \delta(x) e^{\ell(\theta;x)}\, d\mu(x).$$

Then

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

If we have only one parameter, so $\theta \in \mathbb{R}$, then Cauchy-Schwarz gives

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

So we get

Theorem 7.1 (Cramér-Rao)**.** *Let $\delta(X)$ be an unbiased estimator for $g(\theta)$. If $\theta \in \mathbb{R}$,*

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

More generally, if $\theta \in \mathbb{R}^d$ and $g(\theta) \in \mathbb{R}$,

$$\operatorname{Var}_\theta(\delta) \geq \nabla g(\theta)^\top J(\theta)^{-1} \nabla g(\theta).$$

Remark 7.1. This technically holds for any estimator δ with $\mathbb{E}_\theta[\delta(X)] = g(\theta)$. We are just interpreting it as $g(\theta)$ coming first and δ being unbiased for $g(\theta)$.

Example 7.3 (iid sample)**.** Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} p_\theta^{(1)}(x)$ with $\theta \in \Theta$, so $X \sim p_\theta(x) = \prod_i p_\theta^{(1)}(x_i)$. Writing $\ell_1(\theta; x_i) = \log p_\theta^{(1)}(x_i)$, we have

$$\ell(\theta; x) = \sum_i \ell_1(\theta, x_i).$$

Then

$$\begin{aligned}J(\theta) &= \operatorname{Var}_\theta(\nabla\ell(\theta;X)) \\ &= n \operatorname{Var}_\theta(\nabla\ell_1(\theta;X_i)) \\ &= nJ_1(\theta),\end{aligned}$$

where $J_1(\theta)$ is the Fisher information in a single observation. So Fisher information scales linearly. This means that the Cramér-Rao lower bound scales like $1/n$.

7.4 The Hammersley-Chapman-Robbins inequality

The Cramér-Rao lower bound requires differentiation under the integral. The Hammersley-Chapman-Robbins inequality gives a more general bound using finite differences. The idea is that

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

for small ε. So in the limit, we will get a similar bound to Cramér-Rao.

Theorem 7.2 (Hammersley-Chapman-Robbins)**.** *Let δ be unbiased for $g(\theta)$, and assume that for some collection of ε, $p_\varepsilon \ll p$. Then*

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

Proof. Observe that

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{p_{\theta+\varepsilon}(x)}{p_\theta(x)} - 1\right] &= \int \left(\frac{p_{\theta+\varepsilon}}{p_\theta} - 1\right) p_\theta \, d\mu \\ &= \int (p_{\theta+\varepsilon} - p_\theta)\, d\mu = 0,\end{aligned}$$

as long as $p_{\theta+\varepsilon} \ll p_\theta$. Furthermore,

$$\begin{aligned}\operatorname{Cov}\left(\delta(X), \frac{p_{\theta+\varepsilon}(X)}{p_\theta(X)} - 1\right) &= \int \delta \left(\frac{p_{\theta+\varepsilon}}{p_\theta} - 1\right) p_\theta \, d\mu \\ &= \int \delta p_{\theta+\varepsilon} \, d\mu - \int \delta p_\theta \, d\mu \\ &= \mathbb{E}_{\theta+\varepsilon}[\delta(X)] - \mathbb{E}_\theta[\delta(X)] \\ &= g(\theta+\varepsilon) - g(\theta).\end{aligned}$$

Using Cauchy-Schwarz, we get

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

So we get

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

This lower bound holds for every ε, so we can take the sup over ε on the right hand side. □

Remark 7.2. If we let $\varepsilon \to 0$, we get the Cramér-Rao lower bound, but taking the sup over ε gives a better bound.

7.5 Efficiency

The Cramér-Rao lower bound is not always achievable.

Definition 7.2. The **efficiency** is

$$\mathrm{eff}_\theta(\delta) = \frac{\mathrm{CRLB}}{\mathrm{Var}_\theta(\delta)} \leq 1.$$

We say that $\delta(X)$ is **efficient** if $\mathrm{eff}_\theta(\delta) = 1$ for all θ.

Note that

$$\mathrm{eff}_\theta(\delta) = \mathrm{Corr}_\theta(\delta(X), \ell'(\theta; X))^2$$

Example 7.4. For exponential families,

$$p_\eta(x) = e^{\eta^\top T(x) - A(\eta)} h(x), \qquad \ell(\eta; x) = \eta^\top T(x) - A(\eta) + \log h(x).$$

So the score is

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

This tells us that the Fisher information is

$$\begin{aligned} \mathrm{Var}_\eta(\nabla \ell(\eta; X)) &= \mathrm{Var}_\eta(T(X)) \\ &= \nabla^2 A(\eta) \\ &= \mathbb{E}_\eta[-\nabla^2 \ell(\eta; X)] \end{aligned}$$

Example 7.5. Consider a curved exponential family with $\theta \in \mathbb{R}$:

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

Then the log-likelihood is

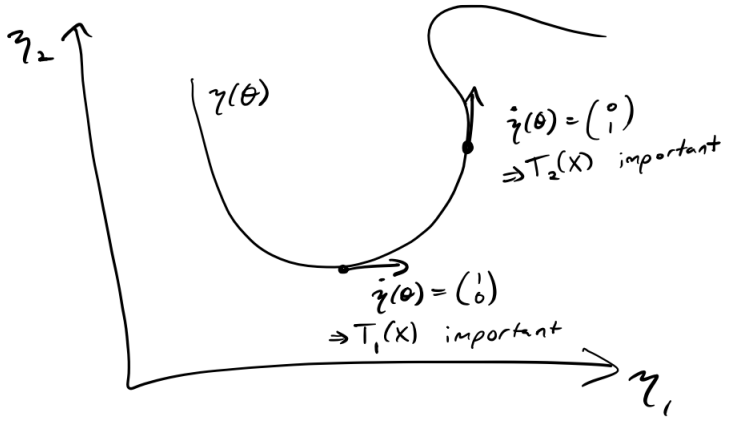
$$\ell(\theta; x) = \eta(\theta)^\top T(x) - B(\theta) - \log h(x),$$

so the chain rule gives the score as

$$\frac{d}{d\theta} \ell(\theta; x) = \dot{\eta}(\theta)^\top T(x) - \dot{B}(\theta)$$

Note that $\frac{d}{d\theta} B(\theta) = \frac{d}{d\theta} A(\eta(\theta)) = \sum_{j=1}^n \dot{\eta}(\theta) \frac{\partial}{\partial \eta_j} A(\eta) = \dot{\eta}(\theta)^\top (\nabla A(\eta))$.

$$= \dot{\eta}(\theta)^\top (T(x) - \mathbb{E}_\eta[T(X)])$$


η_2
$\gamma(\theta)$
$\dot{\gamma}(\theta) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
$\Rightarrow T_2(X)$ important
$\dot{\gamma}(\theta) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$
$\Rightarrow T_1(X)$ important
η_1

8 Bayes Estimation

8.1 Recap: Lower bound for unbiased estimation

Last time, we talked about the **score function**

$$\nabla \ell(\theta; x),$$

where $\ell(\theta; x) = \log p_\theta(x)$ is a log-likelihood. We saw some properties of the score function, like

$$\mathbb{E}_\theta[\nabla \ell(\theta; x)] = 0.$$

The Fisher information was

$$J(\theta) = \operatorname{Var}_\theta(\nabla \ell(\theta; x)) = -\mathbb{E}[\nabla^2 \ell(\theta; x)].$$

If $g(\theta) = \mathbb{E}_\theta[\delta(X)]$ with $g : \mathbb{R}^d \to \mathbb{R}$, then

$$\nabla g(\theta) = \operatorname{Cov}_\theta(\delta(X), \nabla \ell(\theta; X)).$$

Combining this with Cauchy-Schwarz gives the **Cramér-Rao lower bound**

$$\operatorname{Var}_\theta(\delta(X)) \geq \frac{\dot{g}(\theta)^2}{J(\theta)}, \qquad d = 1$$

with multivariate form

$$\operatorname{Var}_\theta(\delta(X)) \geq \nabla g(\theta)^\top J(\theta)^{-1} \nabla g(\theta), \qquad d \geq 1.$$

This gives us a lower bound on how small we can make our risk with unbiased estimation.

Example 8.1. Let $X \sim \text{Binom}(n, \theta)$. Consider two estimators $\delta_0(x) = x/n$ and $\delta_1(X) = \frac{x+3}{n+6}$. The second estimator weights the estimation more towards $1/2$. How can we say that one is better than the other?

To compare these estimators, we previously ruled out all unbiased estimators. However, we can alternatively try to reduce the *average risk*.

8.2 Some problems with unbiased estimation

Unbiased estimation is not always desirable.

Example 8.2. Suppose $X \sim \text{Binom}(50, \theta)$ and $g(\theta) = \mathbb{P}_\theta(X \geq 25)$. The UMVU estimator is

$$\delta(X) = \mathbb{1}_{\{X \geq 25\}},$$

which is somewhat ridiculous because if we saw $X = 25$, we would assume this probability is 1.

Example 8.3. Suppose $X \sim N_d(\theta, I_d)$, where we want to estimate $\|\theta\|_2^2$. The UMVU estimator is $\|X\|_2^2 - d$ because

$$\mathbb{E}[\|X\|_2^2] = \|\theta\|_2^2 + d.$$

This estimator can be < 0, while $\|\theta\|_2^2$ cannot be. So we can always improve on the estimator by instead considering $(\|X\|^2 - d)^+$ instead.

8.3 Bayes estimation from a frequentist viewpoint

We have the model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ for the data X, a loss function $L(\theta; d)$, and the risk $R(\theta; \delta) = \mathbb{E}_\theta[L(\theta; \delta(X))]$.

Definition 8.1. The **Bayes risk** is

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

where $\Theta \sim \Lambda$ and $X \mid \Theta = \theta \sim P_\theta$. This is the average-case risk, integrated with respect to a measure Λ on Ω, called the **prior**.

For now, we assume $\Lambda(\Omega) = 1$. Later, we will allow for $\Lambda(\Omega) = \infty$, which is called an **improper prior**.

Definition 8.2. $\delta(X)$ is a **Bayes estimator** if it minimizes $R_{\text{Bayes}}(\Lambda, \delta)$.

This definition depends on $\mathcal{P}$, Λ, and L. How do we find a Bayes estimator? Fortunately, they are easy to find.

Theorem 8.1. *Suppose $\Theta \sim \Lambda$ and $X \mid \Theta = \theta \sim P_\theta$. Assume that $L(\theta; d) \geq 0$ for all θ, d and that $R_{\text{Bayes}}(\Lambda; \delta_0) < \infty$ for some $\delta_0(X)$. Then*

$$\delta_\Lambda(x) \in \arg\min_d \mathbb{E}[L(\Theta; d) \mid X = x] \text{ for a.e. } x \iff \delta_\Lambda(X) \text{ is Bayes.}$$

So we split up the problem by solving it for any fixed x.

Proof. ($\implies$): Let δ be any other estimator. Then

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

In particular, δ_Λ has finite Bayes risk because we could plug in δ_0 for δ.

($\Longleftarrow$): By contradiction. Let $E_x(d) := \mathbb{E}[L(\Theta; d) \mid X = x]$. Define

$$\delta^*(x) = \begin{cases} \delta_\Lambda(x) & \text{if } \delta_\Lambda(x) \in \arg\min E_x(d) \\ \delta_0(x) & \text{if } E_x(\delta_0(x)) < E_x(\delta_\Lambda(x)) \\ d^*(x) & \text{otherwise,} \end{cases}$$

where $E_x(d^*(x)) < E_x(\delta_\Lambda(x))$. By construction, we have

$$E_x(\delta^*(X)) \leq E_x(\delta_0(X))$$

a.s., so $R_{\text{Bayes}}(\Lambda, \delta^*) < \infty$. We also have

$$E_x(\delta^*(X)) \leq E_x(\delta_\Lambda(X))$$

a.s., with $<$ on a positive measure set. So

$$R_{\text{Bayes}}(\Lambda, \delta^*) \leq R_{\text{Bayes}}(\delta_\Lambda(X)),$$

which is a contradiction. □

8.4 Posterior distributions

Definition 8.3. The conditional distribution of Θ given X is called the **posterior distribution**.

Definition 8.4. When we have densities $\lambda(\theta)$ for a prior and the likelihood $p_\theta(x)$, then the **marginal density** for X is

$$q(x) = \int_\Lambda \lambda(\theta) p_\theta(x)\, d\mu(\theta).$$

The **posterior density** is

$$\lambda(\theta \mid x) = \frac{\lambda(\theta) p_\theta(x)}{q(x)}.$$

In this case, the Bayes estimator is given by

$$\delta_\Lambda = \arg\min_d \int_\Omega L(\theta; d) \lambda(\theta \mid x)\, d\theta.$$

Proposition 8.1. *If $L(\theta; d) = (g(\theta) - d)^2$ is the squared error, then the Bayes estimator is the posterior mean $\mathbb{E}[g(\Theta) \mid X]$ of $g(\Theta)$.*

Proof.

$$\begin{aligned}\mathbb{E}[(g(\Theta)-\delta(X))^2 \mid X] &= \mathbb{E}[(g(\Theta)-\mathbb{E}[g(\Theta)\mid X]+\mathbb{E}[g(\Theta)\mid X]-\delta(X))^2 \mid X] \\ &= \operatorname{Var}(g(\Theta)\mid X)+(\mathbb{E}[g(\Theta)\mid X]-\delta(X))^2,\end{aligned}$$

where the cross term is 0 because $\mathbb{E}[g(\Theta)-\mathbb{E}[g(\Theta)\mid X]\mid X]=0$. This equals $\operatorname{Var}(g(\Theta)\mid X)$ if $\delta(X)\overset{\text{a.s.}}{=}\mathbb{E}[g(\Theta)\mid X]$. □

Let's now consider the **weighted square error** $L(\theta;d)=w(\theta)(g(\theta)-d)^2$. For example, we might take the relative error $L(\theta;d)=(\frac{\theta-d}{\theta})^2$.

Proposition 8.2. *For the weighted square error $L(\theta;d)=w(\theta)(g(\theta)-d)^2$, the Bayes estimator is*

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

Example 8.4 (Beta-Binomial)**.** Suppose $X\mid\Theta=\theta\sim\operatorname{Binom}(n,\theta)=\theta^x(1-\theta)^{n-x}\binom{n}{x}$ with prior $\Theta\sim\operatorname{Beta}(\alpha,\beta)=\theta^{\alpha-1}(1-\theta)^{\beta-1}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Note that in $X\mid\Theta=\theta$, θ is a parameter, whereas in the prior, we are giving a distribution over values of θ. The posterior distribution is

$$\lambda(\theta\mid x)=\frac{\lambda(\theta)p_\theta(x)}{q(x)}$$

Since this will integrate to 1 in θ, we will ignore the quantities not related to θ.

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

So the posterior distribution is a different Beta distribution. Using what we know about the Beta distribution, we have

$$\mathbb{E}[\Theta\mid X]=\frac{X+\alpha}{n+\alpha+\beta}$$

The interpretation is that we have $k=\alpha+\beta$ "pseudo-trials" with α successes. We can write

$$\delta_\Lambda(x)=\frac{x}{n}\cdot\frac{n}{n+\alpha+\beta}+\frac{\alpha}{\alpha+\beta}\cdot\frac{\alpha+\beta}{n+\alpha+\beta}$$

If $n\gg\alpha+\beta$, we can say "the data swamps the prior," whereas for $n\ll\alpha+\beta$, we can say "the prior swamps the data."

Example 8.5 (Normal mean)**.** Suppose $X \mid \Theta = \theta \sim N(\theta, \sigma^2) \propto_\theta e^{-(x-\theta)^2/(2\sigma^2)}$, where σ^2 is known. Take the prior $\Theta \sim N(\mu, \tau^2) \propto_\theta e^{-(\theta-\mu)^2/(2\tau^2)}$. The posterior is

$$\lambda(\theta \mid x) \propto_\theta \exp\left(\theta\left(\frac{x}{\sigma^2} + \frac{\mu}{\tau^2}\right) - \frac{\theta^2}{2}\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)\right).$$

After some algebra,

$$\propto_\theta N\left(\frac{x/\sigma^2 + \mu/\tau^2}{1/\sigma^2 + 1/\tau^2}, \frac{1}{1/\sigma^2 + 1/\tau^2}\right).$$

The posterior mean is

$$\mathbb{E}[\Theta \mid X] = X\frac{1/\sigma^2}{1/\sigma^2 + 1/\tau^2} + \mu\frac{1/\tau^2}{1/\sigma^2 + 1/\tau^2},$$

which is called a **precision-weighted average**.

These examples show that when calculating $\lambda(\theta \mid x)$, we should ignore the parts not depending on θ and try to recognize the resulting shape of the density as a distribution we know already.

9 Priors in Bayesian Estimation

9.1 Recap: Bayesian estimation

Last time, we introduced Bayes estimation, where we want to minimize the **Bayes risk**

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

where $\Theta \sim \Lambda$ and $X \mid \Theta = \theta \sim P_\theta$.

The **Bayes estimator** $\delta_\Lambda(x)$ minimizes

$$\mathbb{E}[L(\Theta; d) \mid X = x]$$

in d. If we have a **prior** density $\lambda(\theta)$ and a likelihood $p_\theta(x)$, then we get the **posterior** density

$$\lambda(\theta \mid x) = \frac{\lambda(\theta) p_\theta(x)}{\int \lambda(\theta) p_\theta(x)\, dx}.$$

Example 9.1 (Beta-Binomial)**.** In this example, $X \mid \theta \sim \text{Binom}(n, \theta) = \theta^x (1-\theta)^{1-x} \binom{n}{x}$ with the prior $\theta \sim \text{Beta}(\alpha, \beta) = \theta^{\alpha-1}(1-\theta)^{\beta-1} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. The posterior distribution is

$$\begin{aligned} \lambda(\theta \mid x) &\propto_\theta \theta^{x+\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \text{Beta}(\alpha + x - 1, \beta + n - x - 1) \end{aligned}$$

It follows that

$$\mathbb{E}[\Theta \mid X] = \frac{X + \alpha}{n + \alpha + \beta}$$

is the Bayes estimator for the squared error loss.

We also had a normal location family with a normal prior which gave us a normal posterior, as well.

9.2 Conjugate priors

Definition 9.1. If the posterior is from the same family as the prior, we say the prior (family) is **conjugate** to the likelihood.

Suppose $X_i \mid \eta \overset{\text{iid}}{\sim} p_\eta(x) = e^{\eta^\top T(x) - A(\eta)} h(x)$ for $i = 1, \dots n$, with $\eta \in \Xi_1 \subseteq \mathbb{R}^s$. For some carrier density $\lambda_0(\eta)$, define the $(s+1)$-parameter exponential family.

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

The sufficient statistic is $\begin{bmatrix} \eta \\ -A(\eta) \end{bmatrix}$ with natural parameter $\begin{bmatrix} k\mu \\ k \end{bmatrix}$. If we take $\lambda_{k\mu,k}$ as our prior, then

$$\begin{aligned}
\lambda(\eta \mid X_1, \ldots, X_n) &\propto_\eta e^{k\mu^\top \eta - kA(\eta)}\, \lambda_0(\eta) \cdot \prod_{i=1}^{n} e^{\eta^\top T(x_i) - A(\eta)} \\
&= \exp\Big((k\mu + n\overline{T}(x))^\top \eta - (k+n)A(\eta) \Big)\, \lambda_0(\eta) \\
&\propto_\eta \lambda_{k\mu + n\overline{T}, k+n}(\eta).
\end{aligned}$$

Here is the interpretation:

1. Suppose we take the prior $\lambda_{k\mu,k}$ and observe X_1. Then the posterior is $\lambda_{k\mu+X_1,k+1}$.
2. Now observe X_2 and update the posterior to get $\lambda_{k\mu+X_1+X_2,k+2}$.
3. ...

If we have a (possibly improper) prior λ_0 and make $k+n$ observations with $\sum_i T(X_i) = k\mu + s$, this is the same as if we had the prior $\lambda_{k\mu,k}$ and observe n observations with $\sum_i T(X_i) = s$.

Example 9.2. Here is a list of some conjugate priors:

Likelihood	Prior
$\text{Binom}(n, \theta)$	$\theta \sim \text{Beta}(\alpha, \beta)$
$N(\theta, \sigma^2)$	$\theta \sim N(\mu, \tau^2)$
$\text{Pois}(\theta)$	$\theta \sim \text{Gamma}(\nu, s)$

People will say that the Beta, for example, is *the* conjugate prior to the Binomial. There can be more than one conjugate prior, which we can get just by changing the carrier distribution.

9.3 Types of priors

Bayesian estimation requires us to have a prior distribution we believe in. In what ways do we do this?

1. **Direct prior or parallel experience:** We can estimate the prior from data. If there is a broad agreement on the prior, corresponding to many observations, the prior may be more meaningful. This gives rise to the following types of Bayesian estimation:
   - Hierarchical Bayes
   - Empirical Bayes
2. **Subjective beliefs:**[4] Here, the prior represents epistemic uncertainty, and the pos-

[4] One may call this the "hardcore" Bayesian perspective.

terior is uncertainty ex post, after observing data and rationally updating.

3. **Convenience prior:** Generally, we have to calculate posteriors. If $\dim(\Omega)$ is large, the posterior is ≈ 0 for most of Ω. This can make it computationally difficult to perform Bayesian estimation, so we might pick a prior which makes the calculation easier, such as a conjugate prior.

4. **"Objective" prior:** We may try to pick a prior which seems to not represent our individual opinion.

   Example 9.3. Suppose $X_i \mid \theta \sim N(\theta, 1)$ for $i = 1, \dots, n$. We could try to use a **flat prior**: $\lambda(\theta) \propto_\theta 1$. This prior is is not a probability distribution, but we can still use it because it gives a valid posterior:

$$\begin{aligned} \lambda(\theta) p_\theta(x) &\propto_\theta e^{\theta \sum_i x_i - n\theta^2/2} \\ &\propto_\theta N(\overline{x}, 1/n). \end{aligned}$$

   The Bayes estimator is $\overline{X}$. The posterior arises naturally as taking taking a limit of priors: $\lim_{\tau^2 \to \infty} N(0, \tau)$.

   The issue with a flat prior is that this is not invariant to reparameterization of the model.

   Example 9.4. Let $X \sim \text{Binom}(n, \Theta)$ with $\Theta \sim U[0, 1]$. Then

$$\mathbb{P}(\Theta \in [0.5, 0.51]) = \mathbb{P}(\Theta \in [0.0001, 0.0101]) = 0.01.$$

   If we let $\eta = \log \frac{\Theta}{1-\Theta}$, then

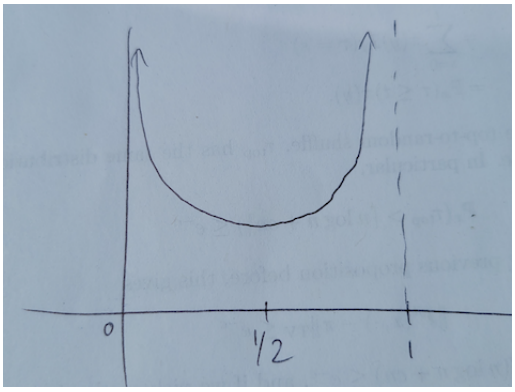
$$\mathbb{P}(\Theta \in [0.5, 0.51]) \approx \mathbb{P}(\eta \in [0, 0.01]),$$

   while

$$\mathbb{P}(\Theta \in [0.0001, 0.0101]) = \mathbb{P}(\eta \in [\log 0.001, \log 0.1]).$$

   Jeffreys proposed using $\lambda(\theta) \propto_\theta |J(\theta)|^{1/2}$. This is called the **Jeffreys prior**, which is invariant under reparameterization. However, the Jeffreys prior can have less of a claim to being agnostic. In the normal case, the Jeffreys prior is the flat prior, but

in the binomial case, the Jeffreys prior looks like this:



Remark 9.1. There has been some controversy about Bayesian vs frequentist statistics. Historically, frequentist statisticians tend to give objections of the form "The object of interest (such as the number of elephants in Africa[5]) is not actually random!" However, if you flip a coin and don't yet look at the result, even though the outcome is certain, there is still epistemic uncertainty about the result.

The Bayesian perspective has the advantage (and disadvantage) of being able to express vague intuitions. Ultimately, making a decision in government may require different statistics from writing a scientific paper. But subjective beliefs and intuitions can often be incorrect.

A practical issue is that it is very difficult to express an opinion of a joint distribution of many random variables.

[5]The elephants in Africa are just standing around, waiting to be counted.

10 Hierarchical Bayes

10.1 Recap: Choosing priors and conjugate priors

We've been talking about Bayesian statistics and estimation. Last time, we talked about 4 ways to choose a prior:

1. Prior or parallel experience
2. Subjective beliefs
3. Convenience prior
4. Objective prior (flat or Jeffreys)

We also gave examples of conjugate priors, where the posterior, $\lambda(\theta \mid x)$, comes from the same family as the prior, $\lambda(\theta)$.

Example 10.1. If $\Theta \sim \text{Beta}(\alpha, \beta)$ and $X \mid \Theta \sim \text{Binom}(n, \Theta)$, then $\Theta \mid X \sim \text{Beta}(\alpha + X, \eta + n - X)$. The Bayes estimator for the mean squared loss is

$$\mathbb{E}[\Theta \mid X] = \frac{\alpha + X}{n + \alpha + \beta}.$$

10.2 Advantages and disadvantages of the Bayes approach

Here are some advantages of the Bayes approach to statistics.

1. **Appealing frequentist properties:** We will show later that Bayes estimators are always admissible. They also minimize average case loss.
2. **Estimator defined straightforwardly:** Compared to something like UMVU estimators, Bayes estimators are much easier to determine. We will see later that it is hard in general to find minimax estimators.
3. **Detailed output:** The posterior distribution gives a lot of information (although there is danger of overestimating the value of our posterior).

Here are some disadvantages.

1. **Difficult to choose prior:** There are many ways to choose a prior, and none of them is always better than the others.
2. **Calculations can be hard:** There is a significant amount of research on how to do the calculations for Bayesian statistics.
3. **Have to have opinions about everything:** If we don't have a parametric model, it may not make sense to come up with a prior.

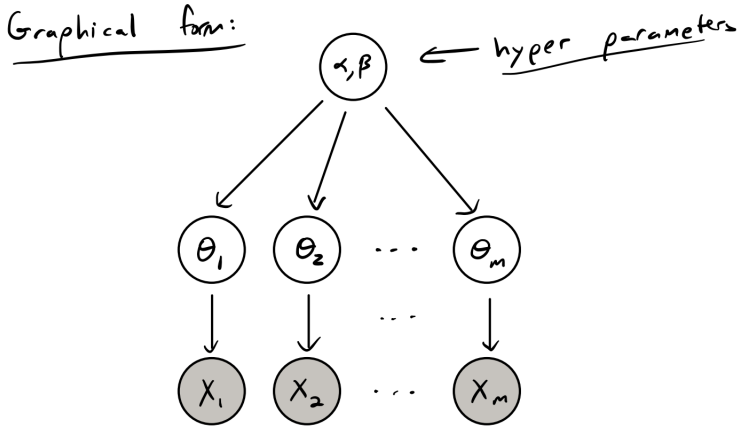
10.3 Hierarchical Bayes and graphical models

What if we want to solve a number of parallel problems at the same time?

Example 10.2. Suppose we want to predict a baseball batter's "true" batting average θ from n at bats. Let X denote the number of hits, with $X \sim \text{Binom}(n, \theta)$. The UMVU estimator is X/n. Most batting averages are between 10% and 30%, so if we observe $X = 4$ hits out of $n = 5$, we want to make sure we are not overestimating the player's batting average. We could use the convenience prior $\text{Beta}(\alpha, \beta)$, which requires us to pick α, β. How should we determine these values? The idea is that we should pool information across players $1, \dots, m$.

Here, $\alpha, \beta \sim \lambda(\alpha, \beta)$ are **hyperparameters**, which govern the distribution of the parameters. Then $\theta \mid \alpha, \beta \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta)$, and $X_i \mid \theta, \alpha, \beta \overset{\text{ind}}{\sim} \text{Binom}(n_i, \theta_i)$.

Let's write this model in a graphical form:



This is called a **directed graphical model**. The graph above is a directed, acyclic graph, and it tells us how the joint density of these $2m+2$ random variables factorizes. If we have a graph (V, E), then the joint density factorizes as

$$p(z_1, \dots, z_m) = \prod_{i=1}^{m} p_i(z_i \mid \text{Pa}(z_i)), \qquad \text{Pa}(z_i) := (z_j : (j \to i) \in E).$$

For our model,

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

10.4 Markov Chain Monte Carlo

This brings us to the idea of **Markov Chain Monte Carlo (MCMC)**: The posterior distribution is

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

where this integral is a high-dimensional integral (which may be difficult to calculate). An extremely successful computational strategy[6] is to set up a Markov chain whose stationary distribution is proportional to the numerator and then run the Markov chain for a long time to get samples from this distribution.

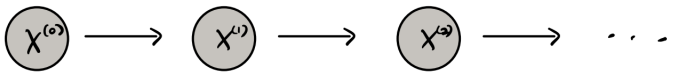
Definition 10.1. A (stationary) **Markov chain** with transition kernel $Q(y \mid x)$ and initial distribution $\pi_0(x)$ is a sequence of random variables $X^{(0)}, X^{(1)}, X^{(2)}, \dots$ such that

$$X^{(0)} \sim \pi_0(x), \qquad X^{(t+1)} \mid X^{(0)}, \dots, X^{(t)} \sim Q(\cdot \mid X^{(t)}).$$

We can think of this as

$$Q(y \mid x) = \mathbb{P}(X^{(t+1)} = y \mid X^{(t)} = x).$$

This is an example of a directed graphical model:



The marginal probability of $X^{(1)}$ is

$$\begin{aligned}\mathbb{P}(X^{(1)} = y) &= \int_{\mathcal{X}} \mathbb{P}(X^{(1)} = y \mid X^{(0)} = x)\pi_0(x)\; d\mu(x) \qquad \text{(for discrete random variables)} \\ &= \int_{\mathcal{X}} Q(y \mid x)\pi_0(x)\,d\mu(x).\end{aligned}$$

Definition 10.2. If

$$\pi(y) = \int_{\mathcal{X}} Q(y \mid x)\pi(x)\,d\mu(x),$$

we say that π is **stationary** for the kernel Q.

A sufficient condition for π to be stationary is **detailed balance**:

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

Proposition 10.1. *Detailed balance implies stationarity.*

[6] This changed the general view of Bayesian statistics in the 90s.

Proof. If we have detailed balance,

$$\int_{\mathcal{X}} \underbrace{Q(y \mid x)\pi(x)}_{=\pi(y)Q(x|y)} \, d\mu(x) = \pi(y) \underbrace{\int_{\mathcal{X}} Q(x \mid y) \, d\mu(x)}_{=1}$$
$$= \pi(y).$$

□

Theorem 10.1. *If a Markov chain with stationary distribution π is*

1. *Irreducible (for any x, y, it is possible to eventually get from x to y),*
2. *Aperiodic (the greatest common divisor of all the possible number of steps for any x to get back to itself is 1),*

then $\text{dist}(X^{(t)}) \xrightarrow{t\to\infty} \pi$*, regardless of the initial distribution.*

10.5 The Gibbs Sampler

Suppose we have a generic parameter vector $\theta = (\theta_1, \ldots, \theta_d)$ and data X. Here is the algorithm:

Initialize $\theta = \theta^{(0)}$

For $t = 1, \ldots, T$,

For $j = 1, \ldots, d$,

Sample $\theta_j \sim \lambda(\theta_j \mid \theta_{\setminus j}, X)$.

Record $\theta^{(t)} = \theta$.

Here are two variations on how we might do the inner loop:

1. Update a random coordinate $J^{(t)} \sim U\{1, \ldots, d\}$.
2. Update all coordinates in a random order.

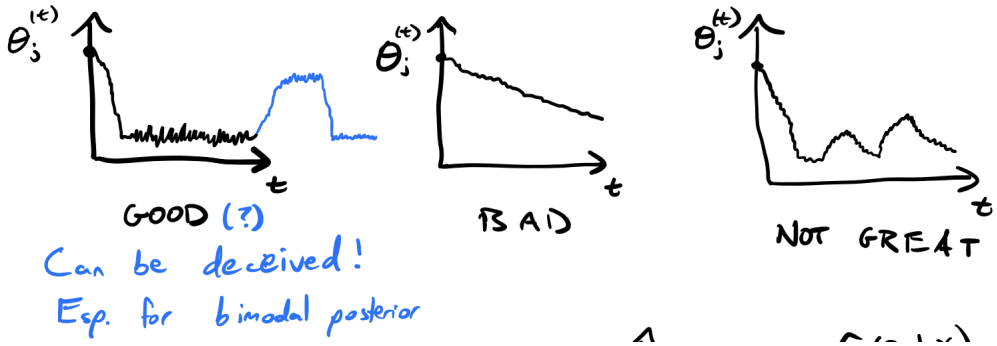
Why is this a good algorithm? If we have a directed acyclic graph, then

$$\lambda(\theta_j \mid \theta_{\setminus j}) \propto_{\theta_j} p(\theta_j \mid \theta_{\text{Pa}(j)}) \prod_{i \in \text{Pa}(j)} p(\theta_i \mid \theta_{\text{Pa}(i)}).$$

In our example, $\theta_j \sim \text{Beta}(\alpha + X_j, n + \alpha + \beta)$ is easy to sample. The α and β will be different every time we sample.

Check that the inner loops satisfies detailed balance, so the posterior distribution of the inner loop is the stationary distribution. This will give us the stationary distribution from the whole algorithm.

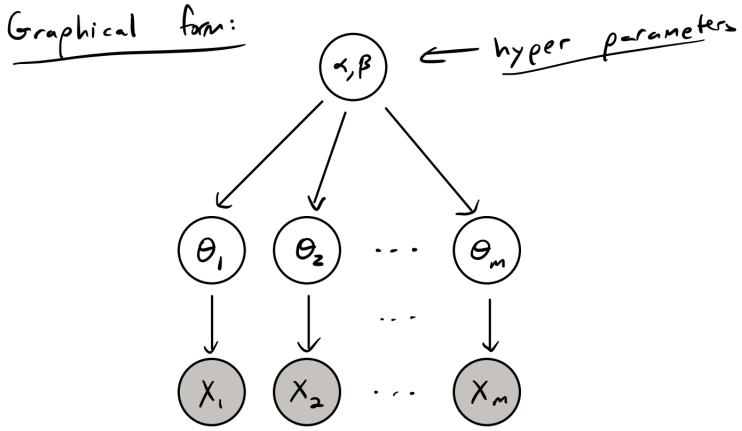
In practice, there can be issues:



11 Hierarchical Bayesian Models and the James-Stein Estimator

11.1 Examples of hierarchical Bayesian models

Last time we talked about hierarchical Bayes models



Example 11.1. In our baseball model last time, we had the **hyperparameters** α, β with $\Theta \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$ and $X_i \mid \Theta_i \sim \text{Binom}(n_i, \Theta_i)$.

This was a directed graphical model with

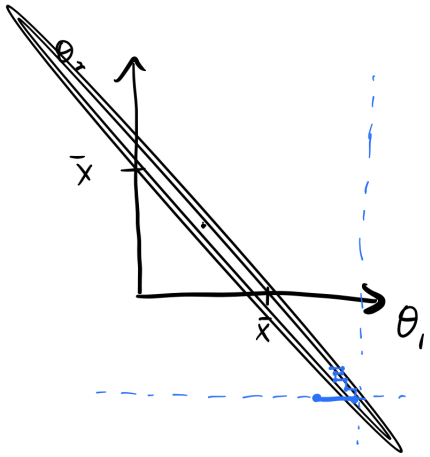
$$p(\gamma, \theta_1, \ldots, \theta_m, x_1, \ldots, x_m) = p(\gamma) \prod_{i=1}^{m} p(\theta_i \mid \gamma) p(x_i \mid \theta_i).$$

We also discussed **Markov chains** with kernels $Q(y \mid x)$; these had a **stationary distribution** π which satisfies $\pi(y) = \int Q(y \mid x)\pi(x)\, dx$. A sufficient (but stronger) condition is **detailed balance**, which requires that $\pi(x)Q(y \mid x) = \pi(y)Q(x \mid y)$ for all x, y.

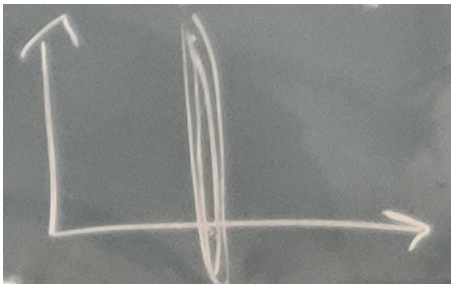
One particularly useful algorithm for sampling in hierarchical models is the **Gibbs sampler**, where we hold all the θ_i fixed except for one at a time and iteratively update our θ_is as we go. Here is an example of where things can go wrong with the Gibbs sampler.

Example 11.2. Let $\Theta_1, \Theta_2 \overset{\text{iid}}{\sim} N(0, 1)$ and $X_i \mid \Theta \overset{\text{iid}}{\sim} N(\Theta_1 + \Theta_2, 1)$ for $i = 1, \ldots, n$. If we

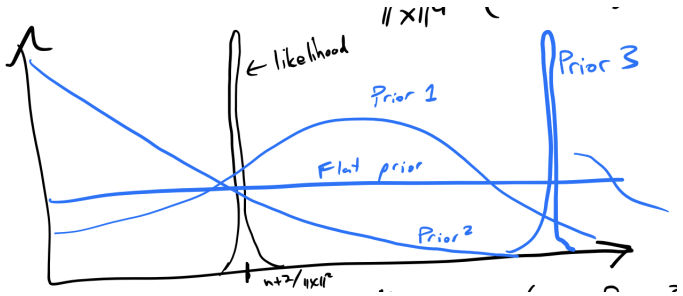
do this, for large n, we will get a very highly correlated posterior distribution:



If we reparameterize the problem with $\beta_1 = \theta_1 + \theta_2$ and $\eta_2 = \theta_1 - \theta_2$, the parameters are much less dependent, so the Gibbs sampler will work better



Another issue would be when we have a bimodal distribution with the two modes having disjoint supports. Then the Gibbs sampler will not be able to jump from 1 of these modes to the other.



This can be a general problem with MCMC.

Example 11.3 (Gaussian hierarchical model)**.** Here is a Gaussian hierarchical model. Let $\tau^2 \sim \lambda(\tau^2)$ (e.g. $1/\tau^2 \sim$ Gamma), $\Theta_i \mid \tau^2 \overset{\text{iid}}{\sim} N(0, \tau^2)$, and $X_i \mid \tau^2, \Theta_i \overset{\text{iid}}{\sim} N(\Theta_i, 1)$ for

$i = 1, \ldots, d$. The posterior mean is

$$\begin{aligned}
\mathbb{E}[\Theta_i \mid X] &= \mathbb{E}[\mathbb{E}[\Theta_i \mid X, \tau^2] \mid X] \\
&= \mathbb{E}\left[\frac{\tau^2}{\tau^2+1} X_i \mid X\right] \\
&= \underbrace{\left(\mathbb{E}\left[\frac{\tau^2}{1+\tau^2} \mid X\right]\right)}_{1-\mathbb{E}[\zeta|X]} X_i,
\end{aligned}$$

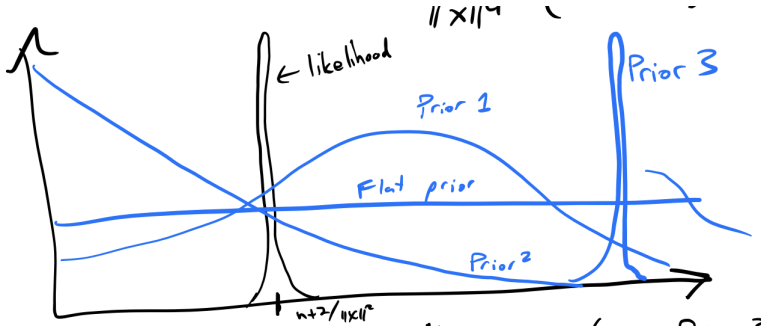
where $\zeta = \frac{1}{1+\tau^2}$. We can think of this as an *optimal shrinkage factor*.

If we marginalize out Θ, we get $X_i \mid \tau^2 \overset{\text{iid}}{\sim} N(0, 1+\tau^2)$. If we think of this as just a problem of estimating τ^2, the sufficient statistic is

$$\begin{aligned}
\frac{\|X\|^2}{d} \mid \tau^2 &\sim \frac{1+\tau^2}{d} \chi^2_d \\
&= (1+\tau^2, 2(1+\tau^2)^2/d),
\end{aligned}$$

where this notation means it is some distribution with mean $1+\tau^2$ and variance $2(1+\tau^2)^2/d$. The likelihood for τ^2 has a sharp peak near $\tau^2 = \frac{\|X\|^2}{d} - 1$ or, equivalently, near $\zeta = \frac{d}{\|X\|^2}$ (for large d).

For any reasonably open-minded prior (not prior 3 in the below figure), $\mathbb{E}[\zeta \mid X] \approx \frac{d}{\|X\|^2}$.



So

$$\mathbb{E}[\Theta_i \mid X] \approx \left(1 - \frac{d}{\|X\|^2}\right) X_i.$$

The moral is that if the prior doesn't matter so much, we can just try to estimate ζ directly from the data. This motivates the idea of **empirical Bayes** models: Write down a hierarchical model and just try to estimate a parameter like ζ using the data. In this way, we don't need to use the Gibbs sampler.

11.2 The James-Stein estimator

Empirical Bayes is a hybrid approach in which we treat the hyperparameters as fixed and treat the paramters as random.

Example 11.4. Think of τ^2 (or of ζ) as a fixed parameter, so we have $X_i \overset{\text{iid}}{\sim} N(0, 1+\tau^2)$ and $\|X\|^2 \sim (1+\tau^2)\chi^2_d$. Then the UMVU estimator for τ^2 is

$$\widehat{\tau}^2 = \frac{\|X\|^2}{d} - 1, \quad \text{which gives} \quad \widehat{\zeta} = \frac{1}{1+\widehat{\tau}^2} = \frac{d}{\|X\|^2}.$$

This is not great because it can be negative. What if we took the UMVUE for ζ? Then we get the James-Stein estimator.

James and Stein proposed that for $d \geq 3$,

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

The interpretation is that $\frac{d-2}{\|X\|^2}$ is the UMVU estimator for ζ:

Proposition 11.1. *If $Y \sim \chi^2_d = \text{Gamma}(d/2, 2)$ with $d \geq 3$, then $\mathbb{E}[1/Y] = \frac{1}{d-2}$.*

Proof.

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

□

Using the proposition,

$$\frac{\|X\|^2}{1+\tau^2} \sim \chi^2_d \implies \zeta^{-1}\, \mathbb{E}\left[\frac{1}{\|X\|^2}\right] = \frac{1}{d-2} \implies \widehat{\zeta} = \frac{d-2}{\|X\|^2}.$$

But the James-Stein estimator is more interesting than just this. Going back to a non-Bayesian model, suppose $X_j \sim N(\theta_j, 1)$ with $\theta \in \mathbb{R}^d$. Then for $d \geq 3$, X is inadmissible as an estimator of θ for the MSE. Say we have n observations:

Proposition 11.2 (James-Stein[7]). *Let* $X_i \overset{\text{iid}}{\sim} N_d(\theta, \sigma^2 I_d)$ *for* $i = 1, \ldots, n$ *with known* $\sigma^2 > 0$. *For*

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

$$\text{MSE}(\theta, \delta_{\text{JS}}) < \text{MSE}(\theta, \overline{X})$$

for all $\theta \in \mathbb{R}^d$.

This says that if we have a bunch of unrelated experiments and we pool the observations together, we can get a better estimator for all of them by combining our observations.

Remark 11.1. We don't need to shrink around 0. For any $\theta_0 \in \mathbb{R}^d$,

$$\delta(X) = \theta_0 + \left(1 - \frac{d-2}{\|X - \theta_0\|^2}\right)(X - \theta_0)$$

renders X itself inadmissible for the mean squared error.

Next time, we will prove this result using Stein's lemma.

[7] This shocking result came out in the 50s, and no one was prepared for it.

12 Analysis of the James-Stein Estimator

12.1 Recap: introduction of the James-Stein estimator

Last time, we discussed the Bayesian model with prior $\Theta_i \overset{\text{iid}}{\sim} N(0, \tau^2)$ and $X_i \mid \Theta \overset{\text{iid}}{\sim} N(\Theta_i, 1)$. This gave $\mathbb{E}[\Theta_i \mid X] = (1-\zeta)X_i$, where $\zeta = \frac{1}{1+\tau^2}$. The **Hierarchical Bayes** approach was to put a prior on ζ, so the posterior mean is

$$\mathbb{E}[\Theta_i \mid X] = (1 - \mathbb{E}[\zeta \mid X])X_i.$$

The **Empirical Bayes** approach was to estimate ζ by an estimator $\widehat{\zeta}(X)$ to get the posterior mean

$$\widehat{E}[\Theta_i \mid X] = (1 - \widehat{\zeta})X_i.$$

This brought us to the **James-Stein estimator**

$$\delta_i^{\text{JS}}(X) = \left(1 - \frac{d-2}{\|X\|^2}\right) X_i.$$

This estimator dominates $\delta(X) = X$, even in the Gaussian sequence model with no Bayesian assumption. In particular,

$$\text{MSE}(\theta; \delta_{\text{JS}}) < \text{MSE}(\theta; X) \qquad \forall \theta \in \mathbb{R}^d.$$

12.2 Linear shrinkage without Bayes assumptions

Suppose $X_i \overset{\text{iid}}{\sim} N(\theta_i, 1)$ with fixed $\theta_1, \ldots, \theta_d \in \mathbb{R}$. Consider the estimator $\delta_\zeta(X) = (1-\zeta)X$ for a fixed parameter ζ. Then

$$\text{MSE}(\theta; \delta_\zeta) = \zeta^2 \|\theta\|^2 + (1-\zeta)^2 d$$

Take the derivative over ζ to optimize:

$$0 = \frac{d}{d\zeta} \text{MSE}(\theta; \delta_\zeta) = 2\zeta\|\theta\|^2 - 2(1-\zeta)d.$$

Solving this gives $\zeta^* = \frac{d}{d+\|\theta\|^2}$. Notice that this is always positive, so the optimal shrinkage is never 0. We can't use this value of ζ because it depends on θ. However, the James-Stein estimator is basically an adaptive ζ.

What if we try to estimate $\|\theta\|$ by using $\|X\|^2$? We have $\frac{1}{d}\|X\|^2 = \frac{1}{d}\sum_{i=1}^d X_i^2$, where each term has mean $\theta_i^2 + 1$ and variance $2 + 4\theta_i^2$. So

$$\frac{1}{d}\|X\|^2 \sim \left(\frac{d + \|\theta\|^2}{d}, \frac{2d + 4\|\theta\|^2}{d^2}\right)$$

This is nice because

$$\frac{\text{standard deviation}}{\text{mean}} = 2\frac{\sqrt{d/2 + \|\theta\|^2}}{d + \|\theta\|^2} \xrightarrow{d\to\infty} 0,$$

so this should exhibit concentration about the mean for large d.

12.3 Stein's lemma

Theorem 12.1 (Stein's lemma, univariate)**.** *Suppose $X \sim N(\theta, \sigma^2)$, and let $h : \mathbb{R} \to \mathbb{R}$ be differentiable with $\mathbb{E}[|\dot{h}(X)|] < \infty$. Then*

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

Proof. Assume without loss of generality that $h(0) = 0$. First assume $\theta = 0$ and $\sigma^2 = 1$ for simplicity. Note that

$$\mathbb{E}[Xh(X)] = \int_0^\infty xh(x)\phi(x)\,dx + \int_{-\infty}^0 xh(x)\phi(x)\,dx.$$

Dealing with these separately,

$$\begin{aligned}\int_0^\infty xh(x)\phi(x)\,dx &= \int_0^\infty x\left[\int_0^x \dot{h}(y)\,dy\right]\phi(x)\,dx \\ &= \int_0^\infty \int_0^\infty \dot{h}(y)\phi(x)\mathbb{1}_{\{y\le x\}}\,dx\,dy \\ &= \int \dot{h}(y)\left[\int_y^\infty x\phi(x)\,dx\right]dy\end{aligned}$$

Using the fact that $\frac{d\phi}{dx} = \frac{d}{dx}\frac{1}{\sqrt{2\pi}}e^{-x^2/2} = -x\frac{1}{\sqrt{2\pi}}e^{-x^2/2} = -x\phi(x)$,

$$= \int_0^\infty \dot{h}(y)\phi(y)\,dy.$$

Similarly,

$$\int_{-\infty}^0 xh(x)\phi(x)\,dx = \int_{-\infty}^0 \dot{h}(y)\phi(y)\,dy.$$

Putting these two together gives

$$\mathbb{E}[Xh(X)] = \int_{-\infty}^\infty xh(x)\phi(x)\,dx = \int_{-\infty}^\infty \dot{h}(y)\phi(y)\,dy = \mathbb{E}[\dot{h}(X)].$$

For a general θ, σ^2, write $X = \theta + \sigma Z$, where $Z \sim N(0,1)$. Then

$$\mathbb{E}[(X-\theta)h(X)] = \sigma\,\mathbb{E}[Zh(\theta+\sigma Z)]$$

Applying the result for $g(Z) = h(\theta + \sigma Z)$ and using the chain rule,

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

□

We want to extend this to the multivariate case. Here is what we replace $\dot{h}$ with:

Definition 12.1. If $h : \mathbb{R}^d \to \mathbb{R}^d$, then the **derivative** is the matrix $Dh \in \mathbb{R}^{d\times d}$ given by

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

Definition 12.2. The **Frobenius norm** of a matrix $A \in \mathbb{R}^{d\times d}$ is

$$\|A\|_F = \left(\sum_{i,j} A_{i,j}^2\right)^{1/2}.$$

Theorem 12.2 (Stein's lemma, multivariate)**.** *Suppose $X \sim N_d(\theta, \sigma^2 I_d)$ with $\theta \in \mathbb{R}^d$, and let $h : \mathbb{R}^d \to \mathbb{R}^d$ be differentiable with $\mathbb{E}[\|Dh\|_F] < \infty$. Then*

$$\mathbb{E}[(X-\theta)^\top h(X)] = \sigma^2\,\mathbb{E}[\mathrm{tr}(Dh(X))] = \sigma^2 \sum_i \mathbb{E}\left[\frac{\partial h_i}{\partial x_i}(X)\right].$$

Proof. The i-th term on the left hand side is

$$\mathbb{E}[(X_i-\theta_i)h_i(X)] = \mathbb{E}[\mathbb{E}[(X_i-\theta_i)h_i(X) \mid X_{\setminus i}]]$$

Conditionally on $X_{\setminus i}$, $X_i \sim N(\theta_i, \sigma^2)$, and $h_i(X)$ is just a function of X_i. So we can apply the univariate lemma.

$$\begin{aligned}
&= \mathbb{E}\left[\sigma^2\,\mathbb{E}\left[\frac{\partial h_i}{\partial x_i}(X) \mid X_{\setminus i}\right]\right] \\
&= \sigma^2\,\mathbb{E}\left[\frac{\partial h_i}{\partial x_i}(X)\right].
\end{aligned}$$

Now sum over i on both sides to get the result. □

Remark 12.1. This differentiability condition can be relaxed somewhat.

12.4 Stein's unbiased risk estimator (SURE)

For our estimator $\delta(X)$, apply Stein's lemma on $h(X) = X - \delta(X)$. Assuming $\sigma^2 > 0$ is known,

$$\begin{aligned}
\mathrm{MSE}(\theta;\delta) &= \mathbb{E}_\theta[\|X-\theta-h(X)\|^2] \\
&= \mathbb{E}_\theta[\|X-\theta\|^2] + \mathbb{E}_\theta[\|h(X)\|^2] - 2\,\mathbb{E}_\theta[(X-\theta)^\top h(X)]
\end{aligned}$$

Since $\frac{1}{\sigma}(X-\theta) \sim \chi^2_d$,

$$= \sigma^2 d + \mathbb{E}_\theta[\|h(X)\|^2] - 2\sigma^2\,\mathbb{E}_\theta[\mathrm{tr}(Dh(X))].$$

So we get the estimator

$$\widehat{R} = \sigma^2 d + \|h(X)\|^2 - 2\sigma^2\,\mathrm{tr}(Dh(X)).$$

Example 12.1. If we take $\delta(X) = X$, then $h(X) = 0$, so $Dh(X) = 0$. In this case, we get

$$\widehat{R} = d\sigma^2 \qquad \forall \theta.$$

Example 12.2. Now look at $\delta_\zeta(X) = (1-\zeta)X$, and let $h(X) = \zeta X$, so $Dh(X) = \zeta I_d$. Then

$$\widehat{R} = \sigma^2 d + \zeta^2 \|X\|^2 - 2\sigma^2 \zeta d.$$

12.5 MSE of the James-Stein estimator

We will take $\sigma^2 = 1$ for simplicity. We have

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

so

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

Then

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

and

$$\frac{\partial h_i}{\partial x_i} = \frac{\partial}{\partial x_i}\frac{d-2}{\|x\|^2} x_i = (d-2)\frac{\|X\|^2 - 2X_i^2}{\|X\|^4}.$$

Summing over i tells us that

$$\mathrm{tr}(Dh(X)) = (d-2)\frac{d\|X\|^2 - 2\|X\|^2}{\|X\|^4} = \frac{(d-2)^2}{\|X\|^2}.$$

So Stein's unbiased risk estimator is

$$\widehat{R} = d + \frac{(d-2)^2}{\|X\|^2} - 2\frac{(d-2)^2}{\|X\|^2} = d - \frac{(d-2)^2}{\|X\|^2}.$$

The risk for the James-Stein estimator is

$$\begin{aligned}
\mathrm{MSE}(\theta; \delta_{\mathrm{JS}}) &= \mathbb{E}[\widehat{R}] \\
&= d - \mathbb{E}\left[\frac{(d-2)^2}{\|X\|^2}\right] \\
&= \mathrm{MSE}(\theta; X) - \mathbb{E}\left[\frac{(d-2)^2}{\|X\|^2}\right].
\end{aligned}$$

This term on the right is the improvement over X.

If $\theta = 0$,

$$\mathrm{MSE}(\theta; \delta_{\mathrm{JS}}) = d - (d-2) = 2.$$

This is a huge improvement for large d! On the other hand, if $\|\theta\| \to \infty$, then

$$\mathrm{MSE}(\theta; \delta_{\mathrm{JS}}) \approx d - \frac{d-2}{\|\theta\|^2} \to d.$$

Remark 12.2. The James-Stein estimator is inadmissible. Here is an estimator that is better:

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

This is also inadmissible because of a "smoothed out" version of this estimator.

Remark 12.3. Here is a more practically useful estimator (when $d \geq 4$) when we have a lot of samples that estimate similar θ_i:

$$\delta^{\mathrm{JS}\,2} = \overline{X} + \left(1 - \frac{d-3}{\|X - \overline{X}\mathbf{1}_d\|}\right)(X - \overline{X}\mathbf{1}_d),$$

where $\overline{X}$ estimates the average value of θ.

Remark 12.4. Should we use the James-Stein estimator in practice?[8] It improves the average risk of the combined problem, but it does not improve the risk of each coordinate individually. So we may not be able to improve our estimation problem by including others' data. If we know more information about each model, it also may not be a good idea to treat them all the same.

[8] Should we go knocking on all the doors of everyone in Berkeley, asking for their samples?

13 Minimax Estimation

13.1 Bayes risk

If we have a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, then we have a few main ideas for choosing an estimator:

1. Constrain the choice of estimator, e.g. unbiased estimation
2. Minimize average-case risk, i.e. Bayes estimation.

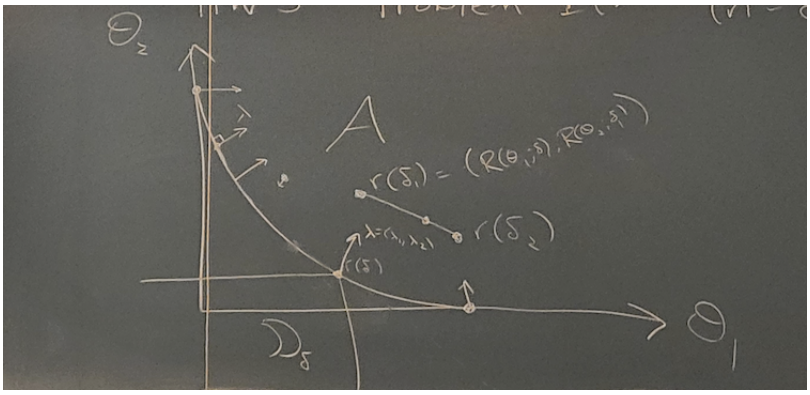
In Bayes estimation, we have a prior Λ with $\Lambda(\Theta) = 1$ (here, Θ is the parameter space). The Bayes estimator (if it exists) minimizes

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

Definition 13.1. The **Bayes risk** for the problem $\Lambda, \mathcal{P}$ is

$$r_\Lambda = \inf_\delta \int R(\theta, \delta)\, d\Lambda(\theta).$$

Example 13.1 (HW 6 Problem 1(c), n=2)**.** In this example, there are only two possible values of θ, θ_1 and θ_2. Then we can plot $r(\delta) = (R(\theta_1; \delta), R(\theta_2; \delta))$.



This is a convex set. The Bayes estimators are the ones on the frontier of this set, the points where the box to the lower left of the point is not in the set. Each vector λ which is normal to this boundary corresponds to a prior.

13.2 Minimax risk, minimax estimators, and least favorable priors

The idea of the minimax risk is to minimize

$$\min_\delta \sup_\theta R(\theta; \delta).$$

Definition 13.2. The minimal achievable sup-risk is called the **minimax risk**,

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

of the problem. An estimator δ^* is **minimax** if it achieves

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

There is a game theoretic interpretation: Imagine we pick our δ first, and then nature tries to maximize the risk (i.e. choosing θ adversarially). The interpretation of Bayes estimation is that nature picks θ (via a prior), and then we try to minimize the risk.

For any proper prior Λ, the Bayes risk is

$$\begin{aligned} r_\Lambda &= \inf_\delta \int R(\theta; \delta)\, d\Lambda(\theta) \\ &\leq \inf_\delta \sup_\theta R(\theta; \delta) \\ &= r^*. \end{aligned}$$

Here is the strategy that nature will pick if it can go first.

Definition 13.3. The **least favorable (LF) prior** is the prior distribution Λ^* that gives the best lower bound:

$$r_{\Lambda^*} = \sup_\Lambda r_\Lambda.$$

We know that

$$\sup_\theta R(\theta; \delta) \geq r^* \geq r_{\Lambda^*} \geq r_\Lambda$$

for any prior Λ. We hope that we can find a prior and an estimator that collapse all these inequalities into equalities.

Theorem 13.1. *If $r_\Lambda = \sup_\theta R(\theta; \delta_\Lambda)$, where δ_Λ is Bayes for Λ, then*

(a) δ_Λ is minimax.

(b) If δ_Λ is the unique Bayes estimator (up to a.s. equality) for Λ, then δ_Λ is the unique minimax estimator.

(c) Λ is the least favorable prior.

Proof.

(a) For any other δ,

$$\begin{aligned}
\sup_\theta R(\theta;\delta) &\geq \int R(\theta;\delta)\, d\Lambda(\theta) \\
&\geq \int R(\theta;\delta_\Lambda)\, d\Lambda(\theta) \qquad (*) \\
&= r_\Lambda \\
&= \sup_\theta R(\theta;\delta_\Lambda).
\end{aligned}$$

(b) Replace $\geq$ with $>$ in the step $(*)$.

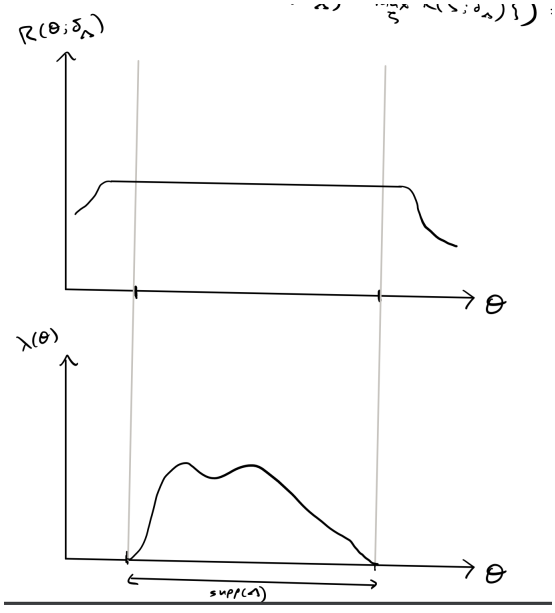
(c) If $\widetilde{\Lambda}$ is any other prior, then

$$\begin{aligned}
r_{\widetilde{\Lambda}} &= \inf_\delta \int R(\theta;\delta)\, d\widetilde{\Lambda} \\
&\leq \int R(\theta;\delta_\Lambda)\, d\widetilde{\Lambda} \\
&\leq \sup_\theta R(\theta;\delta_\Lambda) \\
&= r_\Lambda.
\end{aligned}$$

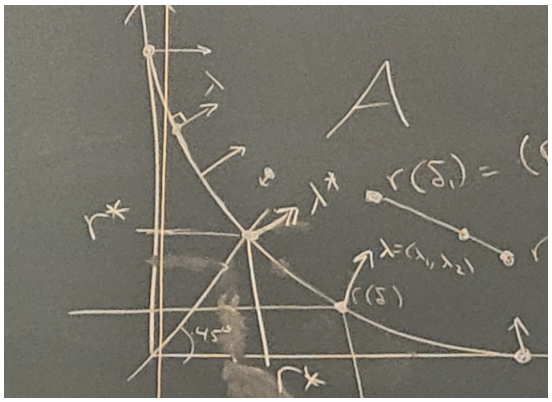
□

Here are sufficient conditions for a minimax estimator:

1. δ is a Bayes estimator whose risk function is constant.
2. δ_Λ is a Bayes estimator with $1 = \Lambda(\{\theta : R(\theta;\delta_\Lambda) = \max_\zeta R(\zeta;\delta_\Lambda)\})$.



In our picture of Bayes estimation, a 45 degree line denotes the points corresponding to estimators with constant risk. The least favorable prior is the corresponding normal vector at the point where this line reaches the boundary of possible risks.



Example 13.2 (Binomial)**.** Suppose $X \sim \text{Binom}(n, \theta)$ with $\theta \in [0, 1]$. We want to estimate θ using the MSE for our risk. Try $\theta \sim \text{Beta}(\alpha, \beta)$, so the Bayes estimator will be

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

Then the Bayes risk is

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

To get a minimax estimator, we want to pick α and β to make this constant in θ. So we set $(\alpha + \beta)^2 = n$ and $2\alpha(\alpha + \beta) = n$ and get $\alpha = \beta = \sqrt{n}/2$. So $\text{Beta}(\sqrt{n}/2, \sqrt{n}/2)$ is the least favorable prior.

This is not such a great estimator, however, since it put a lot of weight around $1/2$. So the pessimistic perspective of minimax estimation can lead us astray for some values of θ.

13.3 Least favorable sequences of priors

Example 13.3. Suppose $X \sim N(\theta, 1)$, and we are estimating θ with the MSE risk. To find the least favorable prior, we would want a flat prior, but this does not give a probability distribution. So we can take, say, $\Lambda_n = N(0, n)$ as a sequence of priors.

Definition 13.4. As sequence $\Lambda_1, \Lambda_2, \ldots$ of priors is **least favorable** if $r_{\Lambda_n} \to \sup_\Lambda r_\Lambda$.

Theorem 13.2. *Suppose $\Lambda_1, \Lambda_2, \ldots$ is any sequence of priors, and suppose δ satisfies*

$$\sup_\theta R(\theta; \delta) = \lim_n r_{\Lambda_n}.$$

Then

(a) δ is minimax.

(b) $\Lambda_1, \Lambda_2, \ldots$ is least favorable.

Proof.

(a) Suppose $\widetilde{\delta}$ is another estimator. Then for all n,

$$\begin{aligned}\sup_\theta R(\theta; \widetilde{\delta}) &\geq \int R(\theta; \widetilde{\delta})\, d\Lambda_n \\ &\geq r_{\Lambda_n}.\end{aligned}$$

Then

$$\sup_\theta R(\theta; \widetilde{\delta}) \geq \lim_n r_{\Lambda_n} = \sup_\theta R(\theta; \delta).$$

(b) If Λ is a prior, then

$$\begin{aligned}r_\Lambda &\leq \int R(\theta; \delta)\, d\Lambda \\ &\leq \sup_\Theta R(\theta; \delta) \\ &= \lim_n r_{\Lambda_n}.\end{aligned}$$

So we get

$$\lim_n r_{\Lambda_n} = \sup_\Lambda r_\Lambda.$$

□

Remark 13.1. If we find the Bayes risk, then we get a lower bound on the minimax risk, and if we provide an estimator, we can get an upper bound on the minimax risk. If these are close, this gives a good estimate of the hardness of a problem.

This is not a very useful measure if your parameter space has some bad corner which you never encounter in practice.

13.4 Bayes estimation example: the Gaussian sequence model

Here is an example of Bayes estimation we did not have time to cover before:

Example 13.4 (Gaussian sequence model)**.** Suppose $X \sim N_d(\theta, I_d)$ for $\theta \in \mathbb{R}^d$. Then the Jeffreys prior on θ is flat. The objective Bayes estimator for Θ is X because the posterior distribution is

$$\lambda(\theta \mid X) \propto_\theta p_\theta(X) \propto_\theta e^{-\|X-\theta\|^2/2} \propto_\theta N_d(X, I_d).$$

What about $\rho^2 = \|\Theta\|^2$? Since $\Theta_i \sim N(X_i, 1)$, $\mathbb{E}[\Theta_i^2 \mid X_i] = 1 + X_i^2$, so

$$\widehat{\rho^2} = \mathbb{E}[\|\Theta\|^2 \mid X] = d + \|X\|^2.$$

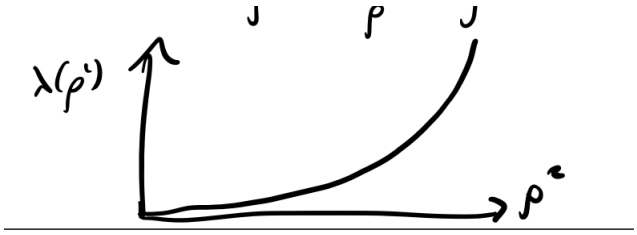
The UMVU estimator is $\widehat{\rho^2}_{\text{UMVU}} = \|X\|^2 - d$ because

$$\mathbb{E}_\theta[\|X\|^2] = d + \|\theta\|^2.$$

Finally, we have the MLE

$$\widehat{\rho^2}_{\text{MLE}} = \|X\|^2.$$

Which one of these estimators is the best? The UMVU estimator is inadmissible because it is negative, but we may not want to rule it out. These all have the same variance, d, and the UMVU estimator has no bias. This serves as a cautionary tale about constructing objective priors. Suppose we took the prior $\Theta \sim N(0, n)$, so $\rho^2 \sim n\chi_d^2$. Then picking an "objective prior" may not produce a good result. In this case, $\lambda(\rho^2) \propto_{\rho^2} (\rho^2)^{(d-1)/2}$.



14 Introduction to Hypothesis Testing

14.1 Null and alternative hypotheses

Suppose we have a model $\mathcal{F} = \{P_\theta : \theta \in \Theta\}$ with data $X \sim P_\theta$, and we want to distinguish between two submodels, the **null hypothesis** $H_0 : \theta \in \Theta_0 \subseteq \Theta$, and the **alternative hypothesis** $H_1 : \theta \in \Theta_1$. If unspecified, $\Theta_1 = \Theta \setminus \Theta_0$.

There is an asymmetry here, where H_0 is considered the "default assumption." We either

1. reject H_0 (conclude $\theta \notin \Theta_0$)
2. fail to reject[9] H_0 (no definite conclusion).

Example 14.1. If $X \sim N(\theta, 1)$, here are common hypothesis tests:

- $H_0 : \theta = 0$ vs $H_1 : \theta > 0$.
- $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$.
- $H_0 : |\theta| \leq \delta$ vs not.

We can also consider nonparametric tests.

Example 14.2. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$ and $Y_1, \dots, T_m \overset{\text{iid}}{\sim}$. We can consider the hypothesis test

$$H_0 : P = Q, \qquad H_1 : P \neq Q.$$

14.2 The power function of a hypothesis test

How can we tell how good our hypothesis test is? We can formally describe a test by its critical function.

Definition 14.1. The **critical function** (or **test function**) of a hypothesis test is

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

The power function tells us how good the test is.

Definition 14.2. The **power function** of a hypothesis test is

$$\beta_\phi(\theta) = \mathbb{E}_\theta[\phi(x)] = \mathbb{P}_\theta(\text{Reject } H_0).$$

[9] We might slip up and say "accept" the null, but really what we are doing is failing to reject the null. Don't say "accept" around non-statisticians.

Definition 14.3. For nonrandomized ϕ, the **rejection region** is

$$R = \{x : \phi(x) = 1\},$$

and the **acceptance region** is

$$A = \mathcal{X} \setminus R.$$

So the power function is $\mathbb{P}_\theta(X \in R)$. We want the power to be large on the alternative hypothesis and small on the null hypothesis. Usually, people refer to the power under the alternative hypothesis, so you want more power for your test.

Definition 14.4. The **significance level** of ϕ is

$$\sup_{\theta \in \Theta_0} \beta_\phi(\theta).$$

We'll say ϕ is a **level-α test** if its significance level is $\leq \alpha$.

The ubiquitous choice is $\alpha = 0.05$.[10]

Example 14.3. Let $X \sim N(\theta, 1)$, where we are testing $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$. Let $z_n = \Phi^{-1}(1 - \alpha)$, where Φ denotes the normal CDF. The usual 2-sided test is

$$\phi_2(X) = \mathbb{1}_{\{|X| > z_{\alpha/2}\}}.$$

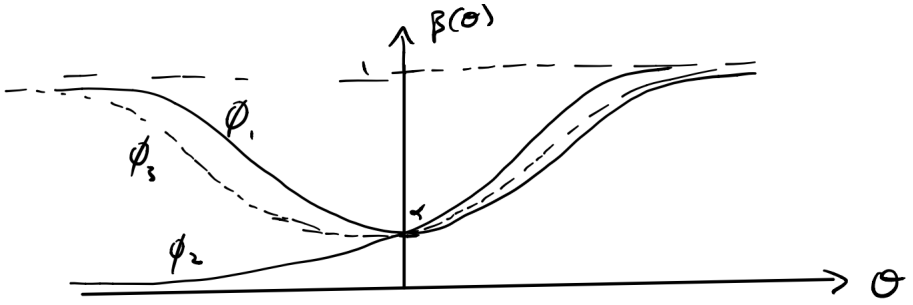
We could also do a 1-sided test

$$\phi_1(X) = \mathbb{1}_{\{X > z_\alpha\}}.$$

Both of these are valid hypothesis tests at level α; the 1-sided test has lower power for $\theta < 0$. We could also try any number of hypothesis tests, such as

$$\phi_3(X) = \mathbb{1}_{\{x < -z_{\alpha/3} \text{ or } X > z_{2\alpha/3}\}}.$$

We can plot the power of these tests against θ:



[10] This is probably ubiquitous because when Fisher came up with the idea of hypothesis testing, he said that he sometimes likes to use the value 0.05. This is probably this most influential offhand remark in the history of science.

Can we tell which hypothesis test is the best? In some situations, there is a best test.

Example 14.4. Let $X \sim (0,1)$ with $H_0 : \theta \leq 0$ vs $H_1 : \theta > 0$. Then the test ϕ_1 is the best possible test (called uniformly most powerful). We will discuss this in detail next time.

So 1-sided tests have a best test. We'll start simple and work our way up to more complicated tests.

Definition 14.5. A **simple hypothesis** is a singleton. A **composite hypothesis** is one that isn't simple.

14.3 Likelihood ratio tests and the Neyman-Pearson lemma

Suppose we test $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$. Without loss og generality, we may assume $\theta_0 = 0$ and $\theta_1 = 1$. Without loss of generality, assume P_0 and P_1 have densities p_0, p_0 (which we may do because P_0 and P_1 are both absolutely continuous with respect to $P_0 + P_1$). The optimal test rejects for large values of $\frac{p_1(x)}{p_0(x)}$.

Definition 14.6. The **likelihood ratio test (LRT)** is of the form

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

where c, γ are chosen so $\mathbb{P}_0(\text{Reject}) = \alpha$.

We will prove that this is the best test, but first, here is some intuition. The power under the alternative hypothesis H_1 is

$$\int_{\mathbb{R}} p_1(x)\, d\mu(x),$$

and the significance level is

$$\int_{\mathbb{R}} p_0(x)\, d\mu(x).$$

We want to maximize the first integral subject to constraint that the second integral equals α. Think of the first integral as the bang, and the second integral as the buck; you want to get the most bang for your buck. If you think about wanting to buy flour from the grocery store with a fixed budget, you'll try to buy bags of flour with the lowest cost per unit until you run out of money. Here, the cost per unit is $\frac{p_1(x)}{p_0(x)}$, and the γ corresponds to the little bit of change you have left over, which you use to buy a fractional bag of flour.

To carry out the proof that the likelihood ratio test is the best test, we would like to use Lagrange multipliers. Since this is over infinitely many parameters, here is a lemma which lets us carry out this optimization.

Proposition 14.1 (12.1 in Keener)**.** *Suppose $c \geq 0$ 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 tests ϕ.

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

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

□

Theorem 14.1 (Neyman-Pearson[11])**.** *The likelihood ratio test wit h signigicance level $= \alpha$ is optimal for testing $H_0 : X \sim P_0$ vs $H_1 : X \sim P_1$ (maximizes $E_1[\phi(X)]$ such that $\mathbb{E}_0[\Phi(X)] \leq \alpha$).*

Proof. We want to maximize the Lagrangian

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

To maximize $\mathcal{L}(\phi; c)$, set

$$\phi(x) = \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_0(x)} > c \\ 0 & \text{if } \frac{p_1(x)}{p_0(x)} < c. \end{cases}$$

Choose the minimum value of c such that

$$\mathbb{P}_0\left(\frac{p_1}{p_0}(X) > c\right) \leq \alpha \leq \mathbb{P}_0\left(\frac{p_1}{p_0}(X) \geq x\right),$$

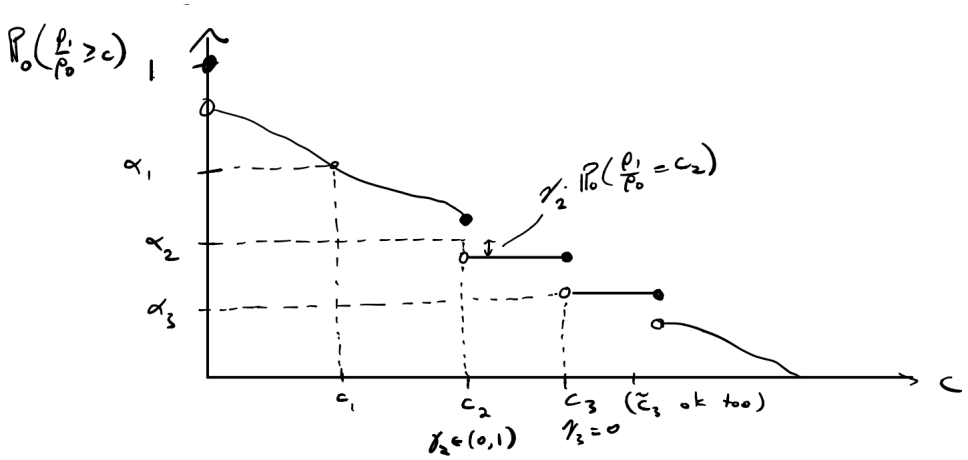
and choose γ to "top up" the significance level to α:

$$\mathbb{P}_0\left(\frac{p_1}{p_0}(X) > c\right) + \gamma\mathbb{P}_0\left(\frac{p_1}{p_0}(X) = c\right) = \alpha.$$

□

[11] This important theorem is often referred to as a lemma.

Here's a picture of how we can pick c_α and γ_α for ϕ^*:



Corollary 14.1 (12.4 in Keener)**.** *If $p_0 \overset{a.s.}{\neq} p_1$ and ϕ is the LRT with level $\alpha \in (0,1)$, then $\mathbb{E}_1[\phi(X)] > \alpha$.*

Proof. We have $\mu(\{p_1 > p_0\}), \mu(\{p_0 > p_1\}) > 0$. We split into a few cases:

$c \geq 1$: We split

$$\mathbb{E}_1[\phi] - \mathbb{E}_0[\phi] = \int_{\{p_1/p_0 > 1\}} |p_1 - p_0| \phi \, d\mu - \int_{\{p_1/p_0 < 1\}} |p_1 - p_0| \phi \, d\mu$$
$$> 0.$$

$c < 1$: This case is similar. □

Example 14.5. Suppose we have a 1-parameter exponential famil $X \sim p_\eta(x) = e^{\eta T(x) - A(\eta)}$. Test the null hypothesis $H_0 : \eta = \eta_0$ vs the alternative $H_1 : \eta = \eta_1 > \eta_0$. The likelihood ratio is

$$\begin{aligned} \frac{p_1(x)}{p_0(x)} &= \frac{e^{\eta_1 T(x) - A(\eta)}}{e^{\eta_0 T(x) - A(\eta)}} \\ &= e^{(\eta_1 - \eta_0)T(x) - (A(\eta_1) - A(\eta_0))} \end{aligned}$$

So the LRT should be to reject when this is large. Since this is a monotone function in $T(x)$, this is the same as saying we reject when $T(x)$ is large. So we can say the test is

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

where we choose c, γ to make

$$\mathbb{P}_{\eta_0}(T(X) > c) + \gamma \mathbb{P}_{\gamma_0}(T(X) = c) = \alpha.$$

Notice that η_1 is nowhere to be found. So this exact test is the best against any alternative η_1, as long as $\eta_1 > \eta_0$. So the best test only depends on the direction of the alternative.

Next time, we will discuss more situations like this, where we have best tests against any alternative in a range of alternatives.

15 One-Sided and Two-Sided Tests

15.1 Recap: Basics of hypothesis testing

Last time, we introduced hypothesis testing, where we have a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and want to distinguish between $H_0 : \theta \in \Theta$ and $H_1 : \theta \in \Theta$ (usually, $\Theta_1 = \Theta \setminus \Theta_0$). The tests were described by a **critical function** $\phi : \mathcal{X} \to [0, 1]$ given by

$$\phi(x) = \begin{cases} 1 & \text{reject} \\ \pi & \text{flip a (biased) coin} \\ 0 & \text{fail to accept.} \end{cases}$$

We defined the **rejection region** $R = \{x : \phi(x) = 1\}$ (ignoring randomization), the **power function** $\beta_\phi(\theta) = \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta(\text{Reject } H_0)$, and the **significance level** $\sup_{\theta \in \Theta_0} \beta_\phi(\theta)$.

Our goal is to obtain the maximum power for $\theta \in \Theta_1$, relative to the constraint that the significance level is at α. There are two types of errors in this setting:

Definition 15.1. A **Type I error** is rejecting the null hypothesis when H_0 is true. A **Type II error** is failing to reject the null hypothesis when H_1 is true.

We introduced the **Likelihood Ratio Test (LRT)** in the case of a simple null $H_0 : \theta = \theta_0$ vs a simple alternative hypothesis $H_1 : \theta = \theta_1$. This test is given by

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

where we choose c, γ such that $\mathbb{E}_0[\phi(X)] = \alpha$. There is a bit of ambiguity because any test of the form (for $c \geq 0$)

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

maximizes $\mathbb{E}_1[\phi(X)] - c\,\mathbb{E}_0[\phi(X)] = \int (p_1 - cp_0)\, d\mu$, as long as we keep the constraint that the significance level is α.

Last time, we had a proposition that said that any test of this form maximizes $\mathbb{E}_1[\phi(X)]$ subject to $\mathbb{E}_\theta[\phi(X)] = \alpha =: \mathbb{E}_1[\phi^*]$. A corollary to this

Example 15.1. If $X \sim p_\eta(x) = e^{\eta T(x) - A(\eta)} h(x)$ is an exponential family with $H_0 : \eta = \eta_0$ and $H_0 : \eta = \eta_1 > \eta_0$, then the LRT gave

$$LR(X) = e^{(\eta_1 - \eta_0)T(X) - (A(\eta_1) - A(\eta_0))},$$

which was monotone in $T(X)$. So we saw that the LRT was dependent only on $T(X)$ and not on the particular value of η_1. So the same exact test is the best for all alternative hypotheses of this form.

15.2 Uniformly most powerful (UMP) tests

Definition 15.2. If $\phi^*(X)$ has significance level α, and for any other level-α test ϕ,

$$\mathbb{E}_\theta[\phi^*(X)] \geq \mathbb{E}_\theta[\phi(X)] \qquad \forall \theta \in \Theta_1,$$

we say that ϕ^* is **uniformly most powerful (UMP)**.

Definition 15.3. A model $\mathcal{P}$ is **identifiable** if $\theta_1 \neq \theta_2 \implies P_{\theta_1} \neq P_{\theta_2}$.

This is just saying that the different values of θ actually mean different things in our model.

Definition 15.4. Assume $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ is identifiable and has densities p_θ for P_θ with respect to μ. We say $\mathcal{P}$ has **monotone likelihood ratios (MLR)** in $T(x)$ if $\frac{p_{\theta_2}}{p_{\theta_1}}(x)$ is a nondecreasing function of $T(x)$ for all $\theta_2 > \theta_1$.

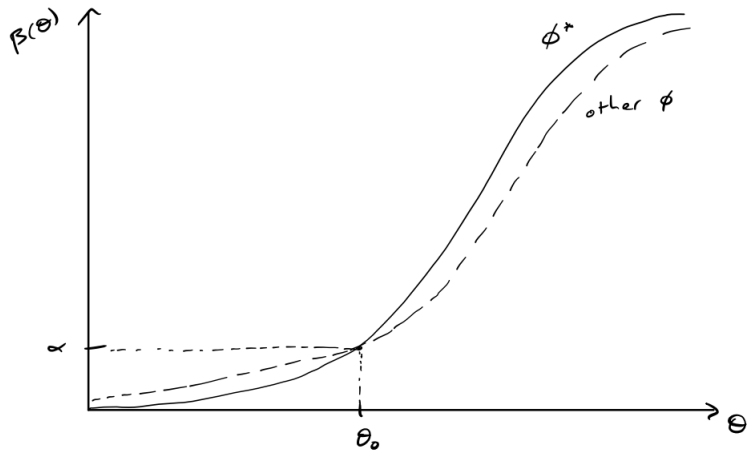
Remark 15.1. This is different from $T(X)$ being **stochastically increasing** in θ, which says that $\mathbb{P}_\theta(T(X) > c)$ is increasing in θ. This condition is enough to construct a valid one-sided test that rejects when T is large, but it will not necessarily be uniformly most powerful.

Theorem 15.1. *Assume $\mathcal{P}$ has MLR in $T(x)$, and test $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$. Let let $\phi^*(x)$ reject for large $T(x)$, where c, γ are chosen so $\mathbb{E}_{\theta_0}[\phi^*(X)] = \alpha$.*

(a) ϕ^* is a UMP level-α test.

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

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



Proof.

(b): if $\theta_1 < \theta_2$, then $\frac{p_{\theta_2}}{p_{\theta_1}}(x)$ is nondecreasing i n$T(X)$. So ϕ^* is a LRT for $H_0 : \theta = \theta_1$ vs $H_1 : \theta = \theta_2$ (at level $\widetilde{\alpha} := \mathbb{E}_{\theta_1}[\phi^*(X)]$). Then the corollary from last time says that $\mathbb{E}_{\theta_2}[\phi^*(X)] > \widetilde{\alpha} = \mathbb{E}_{\theta_1}[\phi^*(X)]$.

(a): If $\theta > \theta_0$, then ϕ^* is the LRT for $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$.

(c): If $\theta_1 < \theta_0$, assume $\mathbb{E}_{\theta_0}[\widetilde{\phi}(X)] = \alpha$. Then both $1 - \phi^*$ and $1 - \widetilde{\phi}$ are level $1 - \alpha$ tests of $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$. But $1 - \phi^*$ is the LRT for this test. Indeed, $\frac{p_{\theta_1}}{p_{\theta_0}}(x) = [\frac{p_{\theta_0}}{p_{\theta_1}}(x)]^{-1}$ is decreasing in T, and $1 - \phi^*$ rejects for small $T(X)$. So ϕ^* maximizes $\mathbb{E}_{\theta_1}[1 - \phi]$ scuh that $\mathbb{E}_{\theta_0}[1 - \phi] \leq 1 - \alpha$. □

15.3 Two-sided tests

What about two-sided alternative hypotheses? Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ with $\theta_0 \in \Theta^0$, where want to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ (this can be generalized to $H_0 : \theta \in [\theta_1, \theta_2]$).

Definition 15.5. $T(X)$ is **stochastically increasing** in θ if $\mathbb{P}_\theta(T(X) \leq t)$ is nonincreasing in θ for all t.

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

Example 15.2. For example, this applies to $X_i \overset{\text{iid}}{\sim} p_0(x - \theta)$ where $T(X)$ is the sample mean or median.

Example 15.3. This also applies to $X_i \overset{\text{iid}}{\sim} \frac{1}{\theta} p_1(x/\theta)$ where $T(X) = \sum_i X_i^2$.

Definition 15.6. The **two-tailed test** rejects when $T(X)$ is extreme in any direction:

$$\phi(x) = \begin{cases} 1 & T(x) < c_1 \text{ or } T(x) > c_2 \\ 0 & T(X) \in (c_1, c_2) \\ \gamma_i & T(x) = c_i, i = 1, 2. \end{cases}$$

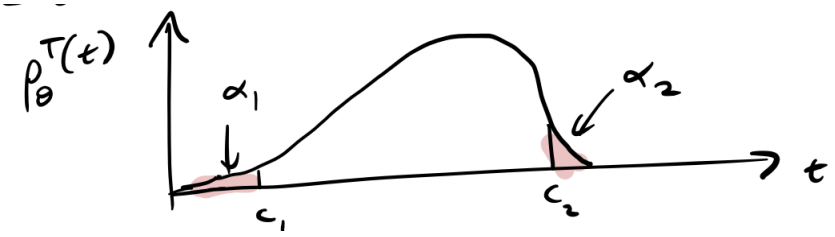
In this setting, we will not usually be able to get a UMP test. We usually have a tradeoff between allocating our type I error to values where θ is large or values where θ is small. Let

$$\alpha = \mathbb{P}_{\theta_0}(T(X) < c_1) + \gamma_1 \mathbb{P}(T(X) = c_1)$$

$$\alpha_2 = \mathbb{P}_{\theta_0}(T(X) > c_2) + \gamma_2 \mathbb{P}(T(X) = c_2).$$

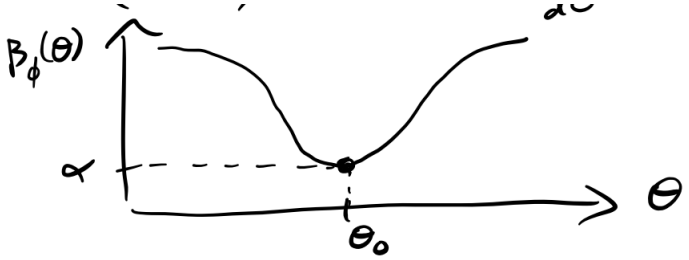
We need $\alpha_1 + \alpha_2 = \alpha$, and we have to balance these considerations. Here are some ideas:

One natural way to do this is to do an **equal-tailed test**, i.e. set $\alpha_1 = \alpha_2 = \alpha/2$.



Definition 15.7. $\phi(x)$ is **unbiased** if

$$\inf_{\theta \in \Theta_1} \mathbb{E}_\theta[\phi(X)] \geq \alpha.$$

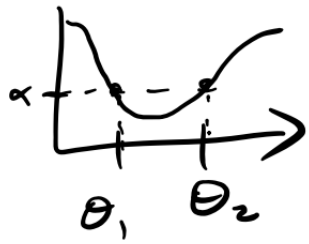


The second idea is to choose an unbiased test.

Theorem 15.2. *Assume $X_i \overset{\text{iid}}{\sim} e^{\theta T(x) - A(\theta)} h(x)$, so the sufficient statistic $\sum_{i=1}^n T(X_i)$. Test $H_0 : \theta \in [\theta_1, \theta_2]$ (with possibly $\theta_1 = \theta_2$) vs the alternative $H_1 : \theta \notin [\theta_1, \theta_2]$. Let $\phi(x)$ be the two-tailed test based on $\sum_{i=1}^n T(X_i)$.*

(a) The unbiased two-tailed test for $\sum_{i=1}^n T(X_i)$ with significance level $= \alpha$ is UMP among all unbiased tests (UMPU).

(b) If $\theta_1 < \theta_2$, the UMPU test solves $\mathbb{E}_{\theta_1}[\phi(X)] = \mathbb{E}_{\theta_2}[\phi(X)] = \alpha$.



(c) If $\theta_1 = \theta_2 = \theta_0$, the UMPU test solves $\mathbb{E}_{\theta_0}[\phi(X)] = \alpha$ and

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

Proof. Proof is in Keener. □

15.4 p-values

Here is an informal definition (if $\phi(x)$ rejects for large $T(x)$): The p-value is

$$\begin{aligned} p(x) &= \text{“}\mathbb{P}_{H_0}(T(X) \geq T(x))\text{.”} \\ &= \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X) \geq T(x)). \end{aligned}$$

Example 15.4. Let $X \sim N(\theta, 1)$, and test $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$. The two-sided test rejects for large $|X|$, and the two-sided p-value is

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

We could instead test $H_0 : |\theta| \leq \delta$ against $H_1 : |\theta| > \delta$. It turns out that we will get

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

Not every test will look like this, so we want a more formal definition.

Definition 15.8. Let $\phi_\alpha(x)$ be a family of tests with $\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi_\alpha(x)] \leq \alpha$ and $\phi_\alpha(x)$ monotone in α. Then the **p-value** is

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

This is the α for which the corresponding test just barely rejects.
For $\theta \in \Theta_0$,

$$\mathbb{P}_\theta(p(X) \leq \alpha) \leq \inf_{\widetilde{\alpha} > \alpha} \mathbb{P}_\theta(\phi_{\widetilde{\alpha}}(X) = 1) \leq \alpha,$$

so the p-value is stochastically larger than $U[0, 1]$ under H_0.

Remark 15.2. The p-value is dependent on not just the data but also the null hypothesis and the hypothesis test we use! This is something many people misunderstand in practice.

16 Confidence Sets and Philosophy of Hypothesis Testing

16.1 Recap: hypothesis tests and p-values

We have been studying hypothesis testing, taking a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and distinguishing between two submodels $H_0 : \theta = \Theta_0$ and $H_1 : \theta = \Theta_1$. The hypothesis test is defined by its **critical function** $\phi(x) \in [0, 1]$.

In a simple null vs simple alternative hypothesis, we saw that it was optimal to reject for large $\frac{p_1}{p_0}(X)$. When we have one real parameter ($\Theta = R$, $\Theta = (0, \infty)$, etc.), this let us analyze 1-sided tests $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$. If $\frac{p_2}{p_1}$ is increasing in $T(x)$, for all $\theta_2 > \theta_1$ (MLR), then the UMP test rejects rejects for large $T(X)$. This is also valid if $T(X)$ is stochastically increasing in θ.

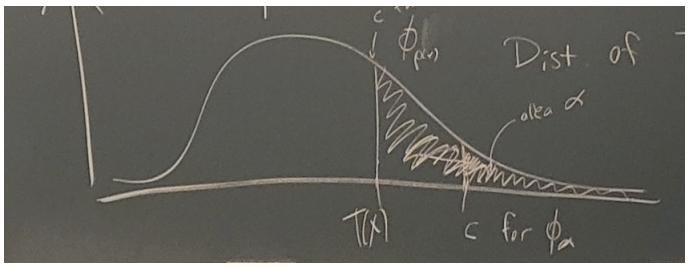
For 2-sided tests, i.e. $H_0 : \theta = \theta_0$ vs $H_1\theta \neq \theta$ (or $H_0 : \theta_1 \leq \theta \leq \theta_2$ vs $H_1 : \theta < \theta_1$ or $\theta > \theta_2$), a 2-sided test rejects for extreme $T(X)$, where $T(x)$ is some test statistic. Here are two ways of making a two tailed test:

- Equal-tailed: Require $\mathbb{P}_{\theta_0}(T(X) > c_2) = \mathbb{P}_{\theta_0}(T(X) < c_1) = \alpha/2$.
- Unbiased: Require $\mathbb{P}_{\theta_0}(T(X) < c_1 \text{ or } > c_2) = \alpha$.

Example 16.1. For an exponential family, the 2-tailed unbiased test is UMPU.

The p**-value** is the level of α for which the test barely rejects:

$$\begin{aligned} p(x) &= \min\{\alpha : \phi_\alpha(x) = 1\} \\ &\overset{\text{often}}{=} \mathbb{P}_{\theta_0}(T(X) \geq T(x)). \end{aligned}$$



The p-value is defined with respect to a family of tests.

For $\theta \in \Theta_0$,

$$\mathbb{P}_\theta(p(X) \leq \alpha) = \mathbb{P}_\theta(\phi_\alpha(X) = 1) \leq \alpha,$$

so $p(X)$ stochastically dominates the uniform distribution on $(0, 1)$.

16.2 Confidence sets

Often, the effect size is a much more relevant question of whether there is an effect or in what direction the effect is.

Definition 16.1. In a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with an estimand $g(\theta)$, $C(X)$ is a $1-\alpha$ **confidence set** for $g(\theta)$ if

$$\mathbb{P}_\theta(C(X) \ni g(\theta)) \geq 1-\alpha \qquad \forall \theta \in \Theta.$$

In other words, the probability that we picked a set containing the estimand is $\geq 1-\alpha$.

Remark 16.1. Note that we have written $C(X) \ni g(\theta)$, rather than the mathematically equivalent $g(\theta) \in C(X)$. This is because $g(\theta)$ is fixed; it is just the bullseye we are shooting for. $C(X)$ is the randomly determined object. People misinterpret this as a statement about $g(\theta)$ conditional on the data, which does not make sense from a frequentist viewpoint.

This should not be called a "confidence" set because confidence is a Bayesian notion. This should really be called an "interval estimate" instead.

16.3 Duality of confidence sets and testing

How do we make confidence sets? Suppose for every value a, we have a level-α test $\phi(x;\alpha)$ for $H_0 : g(\theta) = a$ vs $H_1 : g(\theta) \neq a$. Let

$$\begin{aligned} C(X) &= \{a : \phi(X;a) < 1\} \\ &= \{\text{all non-rejected values}\}. \end{aligned}$$

Then for every θ,

$$\mathbb{P}_\theta(C(X) \not\ni g(\theta)) = \mathbb{P}_\theta(\phi(X;a) = 1) \leq \alpha.$$

Note that the two appearances of θ on the left hand side need to be the same θ.

Remark 16.2. Why don't we need a correction for multiple testing, if we are making uncountably many tests? There is only one true null, so we only have 1 chance to make a type I error.

The above procedure is called **inverting a test** to get a confidence set. We can go the other way: We could rekect $H_0 : \theta \in \Theta_0$ if $C(X) \cap \Theta_0 = \varnothing$. For $\theta \in \Theta_0$,

$$\mathbb{P}_\theta(\text{test rejects}) = \mathbb{P}_\theta(\theta \notin C(X)) \leq \alpha.$$

Example 16.2. A **confidence interval** is a confidence set $C(X)$ which is an interval $[C_1(X), C_2(X)]$. This is usually obtained by inverting a two-sided test.

Example 16.3. An **upper confidence bound** is $C_2(X)$, where $C(X) = (-\infty, C_2(X)]$, and a **lower confidence bound** is $C_1(X)$, where $C(X) = [X_1(X), \infty)$. These are usually obtained by inverting a one-sided test.

Definition 16.2. A upper/lower confidence bound is called **uniformly most accurate (UMA)** if it inverts a UMP test. A confidence interval is called **UMAU** if it inversets a UMPU test.

Example 16.4. Suppose we observe $X \sim \text{Exp}(\theta) = \frac{1}{\theta} e^{-x/\theta}$ with $\theta > 0$. The CDF is $\mathbb{P}_\theta(X \le x) = 1 - e^{-x/\theta}$.

- To get a lower confidence bound for θ, invert the one-sided test for $H_0 : \theta \le \theta_0$. Solve

  $$\alpha = \mathbb{P}_{\theta_0}(X > c(\theta_0)) = e^{-c(\theta_0)/\theta_0}$$

  to get

  $$c(\theta_0) = \theta_0(-\log \alpha) > 0.$$

  Now

  $$\begin{aligned} \phi(x; \theta_0) = 0 &\iff X \le c(\theta_0) \\ &\iff \theta_0 \ge \frac{X}{-\log \alpha}. \end{aligned}$$

  So the confidence region is $C(X) = [\frac{X}{-\log \alpha}, \infty)$.

- For an upper confidence bound, a similar argument gives $C(X) = (-\infty, \frac{X}{-\log(1-\alpha)}]$.

- For a confidence interval derived from inverting an equal-tailed test, the equal-tailed test is

  $$\phi^{2T}\alpha(X; \theta_0) = \phi^{\ge\theta_0}_{\alpha/2}(X; \theta_0) + \phi^{\le\theta_0}_{\alpha/2}(X; \theta_0),$$

  where these tests test $H_0 : \theta = \theta_0$, $H_0 : \theta \ge \theta_0$, and $H_0 : \theta \le \theta_0$, respectively. Then the confidence interval is

  $$\begin{aligned} C(X) &= \big[\tfrac{X}{-\log(\alpha/2)}, \infty\big) \cap \big(-\infty, \tfrac{X}{-\log(1-\alpha/2)}\big] \\ &= \big[\tfrac{X}{-\log(\alpha/2)}, \tfrac{X}{-\log(1-\alpha/2)}\big]. \end{aligned}$$

16.4 Philosophy: misinterpreting hypothesis tests and objections to hypothesis testing

Here are some ways people misinterpret hypothesis tests:

1. If $p < 0.05$, then "there is an effect."

2. If $p > 0.05$, then "there is no effect."

The hypothesis test does not eliminate uncertainty; it just describes or quantifies the uncertainty.

3. If $p = 10^{-6}$, then " the effect is huge."
4. If $p = 10^{-6}$, then "the data are significant," and everything about our model is incorrect.
5. The effect confidence interval for men is $[0.2, 3.2]$ and for women is $[-0.2, 3.8]$, therefore "there is an effect for men and not for women."

Hypothesis tests ask specific questions about specific data sets under specific modeling assumptions using a specific testing method. Top tier medical journals, for example, let people publish claims by reporting p-values without saying what their model was or how they tested the data. But even if we do hypothesis testing right, here are some more objections:

1. Why should we ever test $H_0 : \theta = 0$? Maybe exact zero effects don't exist! Here are some responses:
   (a) One answer is that we could test something else, for example $H_0 : |\theta| \leq \delta$, where δ is some minimum effect size we care about. However, in a $N(\theta, \sigma^2)$ model, the power of this δ test $= \alpha + O((\delta/\sigma)^2)$
   (b) Usually, directional claims are justified.
   (c) In a 2-sample problem, we can test $H_0 : P = Q$ vs $H_1 : P \neq Q$, so this is harder to answer in non-parametric problems.
2. People only like frequentist results like p-values and confidence intervals because they mistake them for Bayesian results.
3. p-values ignore $\mathbb{P}(\text{Data} \mid H_0)$ and only look at $\mathbb{P}(\text{Data} \mid H_1)$. The data could be more likely under the null than under the alternative.
4. Maybe we should use something else instead of hypothesis testing, since scientists often misuse hypothesis tests.

17 Nuisance Parameters, Tests for Multiparameter Exponential Families, and Permutation Tests

17.1 Nuisance parameters

We have been looking at tests with one real parameter $\theta \in \Theta \subseteq \mathbb{R}$. In a one sided test, $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$, we reject for large $T(X)$. This is valid if $T(X)$ is stochasticly increasing in θ and UMP if we the density has MLR in $T(X)$.

For two-sided tests, $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, we reject for extreme values of $T(X)$. We get valid **directional inference** if $T(X)$ is tochastically increasing, and this is UMPU if we have an exponential family and calibrate $c_1, c_2(x_1, x_2)$. So

$$\frac{d\text{Power}}{d\theta} = 0$$

at θ_0.

What about tests with multiple parameters?

Now, our model is $\mathcal{P} = \{P_{\theta,\lambda} : (\theta, \lambda) \in \Omega\}$, and we want to test $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. We call θ the **parameter of interest** and λ the **nuisance paramater**. λ can affect our hypothesis test, even if we are only interested in estimating θ.

Example 17.1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\nu, \sigma^2)$, where μ, ν, σ^2 are unknown. We want to test $H_0 : \mu = \nu$ vs $H_i : \mu - \nu$. Here, we only care about $\theta = \mu - \nu$, so $\lambda = (\mu + \nu, \sigma)$ or $\lambda = (\mu, \sigma)$.

Example 17.2. Let $X_0 \sim \text{Binom}(n_0, \pi_0)$ and $X_1 \sim \text{Binom}(n_1, \pi_1)$ with $X_0 \amalg X_1$. Here, n_0, n_1 are known (not nuisance parameters). We want to test $H_0 : \pi_1 \leq \pi_0$ vs $H_1 : \pi_1 > \pi_0$. A nice choice of θ is the log **odds ratio** $\theta = \log \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$, which we can use to write the null hypothesis as $\theta \leq 0$. In this case, $\lambda = \pi_0$.

17.2 Dealing with nuisance parameters in hypothesis tests for multiparameter exponential families

Suppose we have an exponential family $X \sim p_{\theta,\lambda}(x) = e^{\theta^\top T(x) + \lambda^\top U(x) - A(\theta,\lambda)} h(x)$ with $\theta \in \mathbb{R}^s$ and $\lambda \in \mathbb{R}^r$ both unknown. The distribution of $X \mid U(X)$ only depends on θ. This blocks the dependence on λ. Proceed in steps:

1. Make a sufficiency reduction to T, U:

$$(T(X), U(X)) \sim q_{\theta,\lambda}(t, u) e^{\theta^\top t + \lambda^\top u - A(\theta,\lambda)} g(t, u)$$

2. Condition on U to get

$$q_\theta(t \mid u) = \frac{q_{\theta,\lambda}(t, u)}{\int q_{\theta,\lambda}(z, u)\, dz} = e^{\theta^\top t - B_u(\theta)} g(t, u)$$

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

If $H_0 : \theta \leq \theta_0$, then

$$\phi(x) = \mathbb{1}_{\{T(X) > c_\alpha(u(x))\}},$$

where

$$\mathbb{E}_{\theta,\lambda}[\phi(X) \mid U(X)] \leq \alpha, \qquad \forall \theta \in \Theta_0.$$

Conditional control of the Type I error rate is *stronger* than marginal control of the Type I error rate.

Remark 17.1. We may not want conditional control of the Type I error if we don't need to get rid of a nuisance parameter because requiring this may give a less powerful test.

Theorem 17.1. *Assume $\mathcal{P}$ is a full-rank exponential family with densities*

$$p_{\theta,\lambda}(x) = e^{\theta^\top T(x) + \lambda^\top U(x) - A(\theta,\lambda)} h(x),$$

where $\theta \in \mathbb{R}$, $\lambda \in \mathbb{R}^r$, $(\theta, \lambda) \in \Omega$ is open, and θ_0 is possible.

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

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

where $c(u)$ and $\gamma(u)$ are chosen so that

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

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

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

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

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

$$\mathbb{E}_{\theta_0}[T(X)(\phi^*(X) - \alpha) \mid U(X) = u] = 0.$$

We will prove this theorem next time.

Example 17.3. Suppose $X_i \overset{\text{ind}}{\sim} \text{Pois}(\mu_i)$ $i = 1, 2$, where we want to test $H_0 : \mu_1 \le \mu_2$ vs $H_1 : \mu_1 > \mu_2$.[12] If we let $\eta_i = \log \mu_i$, then

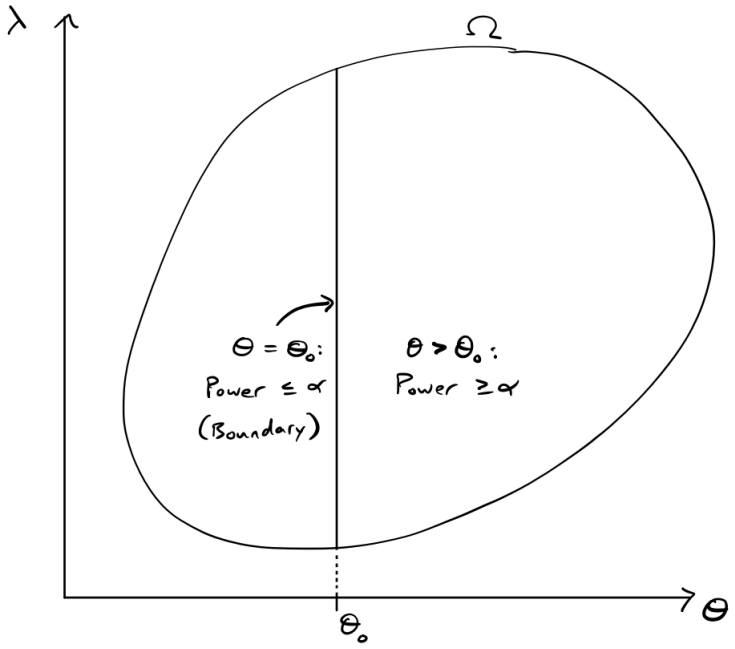
$$p_\mu(x) = \prod_{i=1,2} \frac{\mu_i^{x_i} e^{-\mu_i}}{x_i!} = e^{x_1\eta_1 + x_2\eta_2 - (e^{\eta_1} e^{\eta_2})} \frac{1}{x_1! x_2!} \quad = e^{x_1(\mu_1 - \mu_2) + (x_1+x_2)\eta_2 - (\cdots)} \frac{1}{x_1! x_2!}.$$

The null hypothesis is $H_0 : \mu_1 \le \mu_2 \iff \eta_1 \le \eta_2 \iff \eta_1 - \eta_2 \le 0$. So our test is to reject when X_1 is *conditionally* large given $X_1 + X_2$.

$$\begin{aligned}
\mathbb{P}_\theta(X_1 = x_1 \mid X_0 + X_1 = u) &= \frac{e^{x_1\theta + u\lambda - A(\theta,\lambda)} \frac{1}{x_1!}(u - x_1)!}{\sum_{z=0}^{u}(\cdots)} \\
&\propto_\theta e^{x_1\theta} \frac{1}{x_1!(u - x_1)!} \\
&\propto \text{Binom}(u, \tfrac{e^\theta}{1+e^\theta}) \\
&= \text{Binom}(X_0 + X_1, \tfrac{\mu_1}{\mu_1+\mu_2}).
\end{aligned}$$

Here is a sketch of the proof of the theorem:

Proof. We can think of the power function as a function on the set Ω:



[12] For example, Professor Fithian's wife has gotten pooped on by a bird 4 times in her life, and Professor Fithian has only gotten pooped on once. We can test if Professor Fithian's wife is more unlucky than average. This is the real reason to learn statistics.

1. We must have power $= \alpha$ on the boundary $\theta = \theta_0$.
2. On the boundary, U is complete sufficient. Then $\mathbb{E}[_{\theta_0,\lambda}[\phi(X)] = \alpha$ for all λ, so $\mathbb{E}_{\theta_0,\lambda}[\phi(X) \mid U(X)] \overset{a.s.}{=} \alpha$.
3. ϕ^* must then be the best among conditional tests.

The idea is the same for the two-sided tests, where we have constant power on the null hypothesis. □

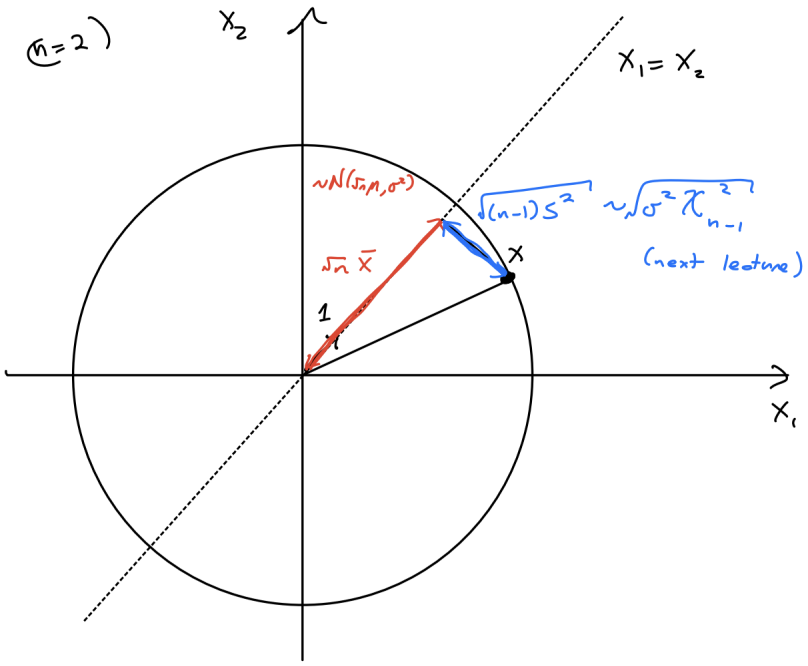
Example 17.4. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where μ, ν, σ^2 are unknown. Test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. We have

$$\rho_{\mu,\sigma^2}(x) = e^{\frac{\mu}{\sigma^2}\sum_i x_i - \frac{1}{2\sigma^2}\sum_i x_i^2 - \frac{n\mu}{\sigma^2}} \left(\frac{1}{2\pi\sigma^2}\right)^{n/2}.$$

Condition on $\sum_i X_i^2 = \|X\|^2 = U(X)$. Then under H_0, $X \mid \|X\|^2 = u \overset{H_0}{\sim} \text{Unif}(\sqrt{n}\mathbb{S}^{n-1})$ is uniform on the sphere. This is equivalent to $\frac{X}{\|X\|} \overset{H_0}{\sim} \text{Unif}(\mathbb{S}^{n-1})$, where $\frac{X}{\|X\|} \amalg \|X\|^2$. The UMPU test rejects when $\sum_i X_i$ is extreme given $\|X\|^2$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, so

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 \\
&= \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - 2\overline{X}\sum_{i=1}^n X_i + n\overline{X}^2\right) \\
&= \frac{1}{n-1}(\|X\|^2 - n\overline{X}^2) \\
&= \frac{1}{n-1}\left(\|X\|^2 - \left(\frac{1}{\sqrt{n}}\mathbf{1}_n^\top X\right)^2\right).
\end{aligned}$$

This means $(n-1)S^2 = \|\operatorname{Proj}_{\overline{X}_0}^{\perp} X\|$. Here is the picture when $n = 2$:



Reject for extreme $\frac{\sqrt{n}\overline{X}}{\sqrt{\|X\|^2 - n\overline{X}^2}}$ (note that this is monotone in $\overline{X}$). This is

$$\frac{\sqrt{n}\overline{X}/\|X\|}{\sqrt{1 - n\overline{X}^2/\|X\|^2}},$$

which is a function of $X/\|X\|$. So this is independent of $U = \|X\|^2$. This statistic is a scaled version of the **T-statistic**:

$$\frac{1}{\sqrt{n-1}} \frac{\sqrt{n}\overline{X}}{\sqrt{S^2}},$$

where $\frac{\sqrt{n}}{\overline{X}}\sqrt{S^2} \overset{H_0}{\sim} t_{n-1}$.

17.3 Permutation tests

Example 17.5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} Q$ be independent, where we want to test $H_0 : P = Q$ vs $H_1 : P \neq Q$. Under H_0, $X_1, \ldots, X_n, Y_1, \ldots, X_n \overset{\text{iid}}{\sim} P$. So condition on the complete sufficient statistic for the null hypothesis (which is not complete for the alternative!): Define $(Z_1, \ldots, Z_{n+m}) = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$. Under H_0, the

order statistics $U(Z) = (Z_{(1)}, \ldots, Z_{(n+m)})$ is complete sufficient. So let $S_{n+m} = \{\pi : \text{permutation on } n+m \text{ elements}\}$. Then

$$Z = (X, Y) \mid U \overset{H_0}{\sim} \text{Unif}(\{\pi U : \pi \in S_{n+m}\}).$$

For *any* test statistic Y, if $P = Q$, then

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

In practice, we can do a Monte Carlo version of this; sample $\pi_1, \ldots, \pi_B \overset{\text{iid}}{\sim} \text{Unif}(S_{n+m})$. Then $Z, \pi_1 Z, \pi_2 Z, \ldots, \pi_B Z \overset{\text{iid}}{\sim} \text{Unif}(S_{n+m} U)$. Let the p-value be

$$p = \frac{1}{1+B} \sum_{b=1}^{B} \mathbb{1}_{\{T(Z) \leq T(\pi_b Z)\}}$$
$$\overset{H_0}{\sim} \text{Unif}(\{\tfrac{1}{1+B}, \ldots, \tfrac{B+1}{1+B}\}).$$

So if we take a test statistic and apply it to all permutations of the data, if the original data looks special, then we should reject.

18 Hypothesis Tests for Gaussian Models

18.1 Recap: hypothesis testing with nuisance parameters

Last time, we discussed hypothesis testing with nuisance parameters. If we have an exponential family $X \sim e^{\theta T(x)+\lambda^\top U(x)-A(\theta,\lambda)}h(x)$ with the one-sided test $H_0 : \theta \le \theta_0$ vs $H_1 : \theta > \theta_0$, then the UMPU test rejects for conditionally large $T \mid U$. If we have the two-sided test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \ne \theta_0$, the UMPU test rejects for conditionally extreme $T \mid U$. Here, $P_\theta(T \mid U)$ depends only on θ (U is sufficient after fixing θ).

We saw the nonparametric test where if $X_i \overset{\text{iid}}{\sim} P$, $Y_i \overset{\text{iid}}{\sim} Q$, then we can test $H_0 : P = Q$ vs $H_1 : P \ne Q$ by conditioning on the pooled order statistics. There are various choices of test statistics to use for permutation tests with various properties.

18.2 Distributions related to Gaussians

Example 18.1 (χ^2 distribution)**.** If $Z_1, \dots, Z_d \overset{\text{iid}}{\sim} N(0,1)$, then

$$V = \sum_{i=1}^{d} Z_i^2 \sim \chi^2_d = \text{Gamma}(d/2, 2)$$

with

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

Note that the standard deviation grows slower than the mean, so as $d \to \infty$,

$$\frac{V}{d} \xrightarrow{p} 1.$$

That is,

$$\mathbb{P}(|\tfrac{V}{d} - 1| \ge \varepsilon) \to 0$$

for all $\varepsilon > 0$. This is what we would expect from the law of large numbers. The central limit theorem tells us that $V \approx N(d, 2d)$ because $\sqrt{d}(\frac{V}{d} - 1) \xrightarrow{d} N(0, 2)$.

Example 18.2 (t-distribution)**.** If $Z \sim N(0,1)$ and $V \sim \chi^2_d$ with $Z \amalg V$, then

$$\frac{Z}{\sqrt{V}/d} \sim t_d,$$

the **Student's T-distribution**, where $t_d \approx N(0,1)$ for large d.

Example 18.3 (F-distribution)**.** If $V_1 \sim \chi^2_{d_1}$ and $V_2 \sim X^2_{d_2}$ with $V_1 \amalg V_2$, then

$$\frac{V_1/d_1}{V_2/d_2} \sim F_{d_1,d_2},$$

the F-**distribution**, which has 2 degrees of freedom. $F_{d_1,d_2} \approx \chi^2_{d_1}$ if $d_2 \to \infty$. If $t \sim t_d$, then

$$T^2 \sim \frac{Z^2}{V/d} \sim F_{1,d}.$$

Example 18.4. If $Z \sim N_d(\mu, \Sigma)$ with $A \in \mathbb{R}^{k \times d}$ and $b \in \mathbb{R}^k$, then

$$AZ + b \sim N(A\mu + b, A\Sigma A^\top).$$

18.3 Analysis of the one-sample t-test

We saw earlier that if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with both $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ unknown and we test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$, then the UMPU test says to reject for extreme $\overline{X}$ given $\|X\|^2$. We could also say to reject for large $|\overline{X}|/\|X\|$; this gets rid of the conditioning given $\|X\|^2$, since under the null, $|\overline{X}|/\|X\| \perp\!\!\!\perp \|X\|^2$. Equivalently, we can reject for large values of

$$\frac{n\overline{X}}{\|X\|^2 - n\overline{X}^2} = \frac{n\overline{X}^2/\|X\|^2}{1 - \overline{X}^2/\|X\|^2}.$$

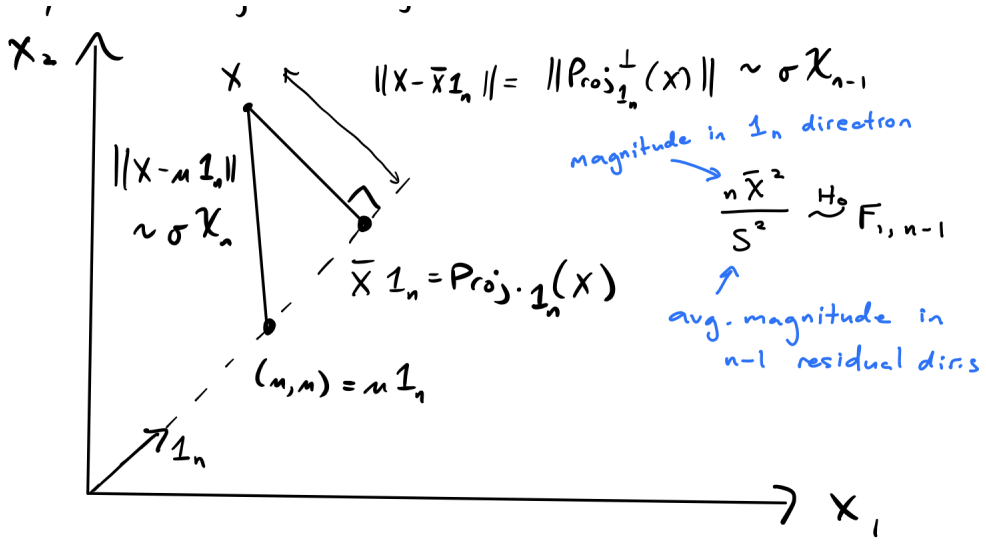
If we didn't want to square this, we could equivalently reject for extreme

$$\frac{\sqrt{n}\overline{X}}{\sqrt{S^2}}, \overset{H_0}{\sim} t_{n-1},$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. This has a T-distirbution because

$$S^2 = \frac{1}{n-1}\|\operatorname{Proj}^\perp_{\mathbf{1}_n} X\|^2 \sim \sigma^2 \chi_{n-1}.$$

Here is a picture for $n = 2$:



What's happening geometrically is that

$$\frac{n\overline{X}}{S^2} \sim F_{1,n-1}$$

is a ratio of squared magnitudes in different directions. If $\mu = 0$, then no direction should be special. Let's make this more precise with linear algebra.

Here is a change of basis: Let

$$Q = \begin{bmatrix} | & & | \\ q_1 & \cdots & q_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & \\ q_1 & Q_r \\ | & \end{bmatrix}$$

with $q_1 = \frac{1}{\sqrt{n}}\mathbf{1}_n$ and $q_2, \ldots, q_n$ completing this to an orthonormal basis. Then $Q^\top Q = QQ^\top = I_n$, and $X \sim N_n(\mu\mathbf{1}_n, \sigma^2 I_n)$. Let

$$Z = Q^\top X = \begin{bmatrix} q_1^\top X \\ Q_r^\top X \end{bmatrix} = \begin{bmatrix} \sqrt{n}\overline{X} \\ Q_r^\top X \end{bmatrix}.$$

Then

$$Q^\top X \sim N_n\left(\begin{bmatrix} \sqrt{n}\mu \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \sigma^2 I_n\right),$$

so

$$\|Q_r^\top X\|^2 = \|Q^\top X\|^2 - n\overline{X}^2 = \|X\|^2 - n\overline{X}^2 = (n-1)S^2.$$

This tells us that

$$Q_r^\top X \sim N_{n-1}(0, \sigma^2 I_{n-1}).$$

Here, we have

$$(n-1)S^2 \sim \sigma^2 \chi^2_{n-1}, \qquad \sqrt{n}\overline{X} \sim N(\sqrt{n}\mu, \sigma^2).$$

18.4 Canonical linear model

Assume

$$Z = \begin{bmatrix} Z_0 \\ Z_1 \\ Z_r \end{bmatrix} \sim N_n\left(\begin{bmatrix} \mu_0 \\ \mu_1 \\ \mu_r \end{bmatrix}, \sigma^2 I_n\right)$$

where Z_i has dimension d_i with $n - d_0 - d_1 = d_r$. Here, $\mu_0 \in \mathbb{R}^{d_0}$, $\mu_1 \in \mathbb{R}^{d_1}$, $\mu_r \in \mathbb{R}^{d_r}$. We want to test $H_0 : \mu_1 = 0$ vs $H_1 : \mu_1 \neq 0$.

This is an exponential family, with

$$p(z) \propto e^{\frac{\mu_1^\top}{\sigma^2} Z_1 + \frac{\mu_0^\top}{\sigma^2} Z_0 - \frac{1}{2\sigma^2}\|z\|^2} h(z).$$

We want to condition on (i.e. ignore) the nuisance parameter Z_0.

- If σ^2 is known and $d_1 = 1$, then the UMPU test rejects for large/small/extreme values of
$$\frac{Z_1}{\sigma} \overset{H_0}{\sim} N(0,1), \qquad (Z\text{-test}).$$
- For known σ^2 and $d_1 \geq 1$, reject for large values of
$$\frac{|Z_1||^2}{\sigma^2} \overset{H_0}{\sim} \chi^2_{d_1} \qquad (\chi^2\text{-test}[13]).$$
- For $d_1 = 1$ and unknown σ^2, condition on Z_0, $\|Z\|^2 = \|Z_0\|^2 + Z_1^2 + \|Z_r\|^2$. We reject for conditionally large/small/extreme Z_1, which is the same as conditioning on extreme values of $\frac{Z_1}{\sqrt{Z_1^2+\|Z_r\|^2}}$. This is equivalent to rejecting for extreme values of
$$\frac{Z_1}{\sqrt{\|Z_1\|^2}/d_1} \overset{H_0}{\sim} t_{d_1} \qquad (t\text{-test}).$$
- If $d_1 \geq 1$ and σ^2 is unknown, condition on $Z_0, \|Z_1\|^2 + \|Z_r\|^2$. Reject for large
$$\frac{\|Z_1\|^2/d_1}{\|Z_r\|^2/d_r} \overset{H_0}{\sim} F_{d_1,d_r} \qquad (F\text{-test}).$$

Remark 18.1. This is related to the Beta distribution.

$$\|X_1\|^2 \sim \sigma^2 \,\text{Gamma}(d_1/2, 2\sigma^2)$$

and

$$\|Z_r\|^2 \sim \sigma^2 \chi^2_{d_r} = \text{Gamma}(d_r/2, 2\sigma^2).$$

so

$$\frac{\|Z_1\|^2}{\|Z_1\|^2 + \|Z_r\|^2} \sim \text{Beta}(d_1/2, d_r/2).$$

More generally, if $U \sim \text{Beta}(d_1/2, d_r/2)$, then

$$\frac{U/d_1}{(1-U)/d_r} \sim F_{d_1,,d_2}.$$

The normalized residual vector is

$$\frac{1}{d_r}\|Z_r\|^2 \sim \frac{\sigma^2}{d_r}\chi^2_{d_r} \approx \sigma^2.$$

[13] There are a number of hypothesis tests refered to as a "χ^2-test."

If we write

$$\widehat{\sigma}^2 = \frac{\|Z_r\|^2}{d_r},$$

then the t-statistic is

$$\frac{Z_1}{\widehat{\sigma}},$$

and the F-statistic is

$$\frac{1}{d_1}\frac{\|Z_1\|^2}{\widehat{\sigma}^2}.$$

This is a solution to a problem presented in a nice form. Next time, we will talk about how to use a change of basis to solve more general Gaussian model problems.

19 General Linear Model for Gaussian Hypothesis Tests

19.1 Recap: Canonical linear model for Gaussian hypothesis tests

Last time, we the χ^2 distribution: if $Z_1, \dots, Z_d \overset{\text{iid}}{\sim} N(0,1)$, then $V = \|Z\|^2 \sim \chi^2_d$. We also had the t distribution, where if $Z \sim N(0,1) \amalg V \sim \chi^2_d$, then $Z/\sqrt{V/d} \sim t_d$. We also had the F-distribution, where if $V_1 \sim \chi^2_{d_1} \amalg V_2 \sim \chi^2_{d_2}$, then $\frac{V_1/d_1}{V_2/d_2} \sim F_{d_1,d_2}$.

In our canonical linear model, we had

$$Z = \begin{bmatrix} Z_0 \\ Z_1 \\ Z_r \end{bmatrix} \sim N_n \left(\begin{bmatrix} \mu_0 \\ \mu_1 \\ \mu_r \end{bmatrix}, \sigma^2 I_n \right)$$

where Z_i has dimension d_i with $n - d_0 - d_1 = d_r$. Here, $\mu_0 \in \mathbb{R}^{d_0}$, $\mu_1 \in \mathbb{R}^{d_1}$, $\mu_r \in \mathbb{R}^{d_r}$. We are interesting in testing $H_0 : \mu_1 = 0$ vs $H_1 : \mu_1 \neq 0$. We ended up with 4 cases last time:

	σ^2 known	σ^2 unknown
$d_1 = 1$	$Z_1/\sigma \overset{H_0}{\sim} N(0,1)$	$Z_1/\widehat{\sigma} \overset{H_0}{\sim} t_{n-d}$
$d_1 \geq 1$	$\|Z_1\|^2/\sigma^2 \overset{H_0}{\sim} \chi^2_{d_1}$	$\frac{\|Z_!\|^2/d_1}{\|Z_r\|^2/(n-d)} = \frac{\|Z_1\|^2/d_1}{\widehat{\sigma}^2} \overset{H_0}{\sim} F_{d_1,n-d}$

where $\widehat{\sigma}^2 = \frac{\|Z_r\|^2}{d_r}$.

19.2 General linear model for testing Gaussian means

In the general linear model for testing Gaussian means, we have $Y \sim N_n(\theta, \sigma^2 I_n)$ with $\sigma^2 > 0$. We want to test $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta \setminus \Theta_0$, where $\Theta_0 \subseteq \Theta \subseteq \mathbb{R}^n$ are subspaces. Denote $dd_0 = \dim(\Theta_0)$ and $d = \dim(\Theta) = d_0 + d_1$.

Let

$$Q = \begin{bmatrix} Q_0 & Q_1 & Q_r \end{bmatrix},$$

where Q_0 is an orthonormal basis for Θ_0, Q_1 is an orthonormal basis for $\Theta \cap \theta_0^\perp$, and Q_r is an orthonormal basis for $\mathbb{R}^n \cap \Theta^\perp$. Then

$$Z = Q^\top Y \sim N \left(\begin{bmatrix} Q_0^\top \theta \\ Q_1^\top \theta \end{bmatrix}, \sigma^2 I_n \right).$$

In this basis, we are testing $H_0 : Q_1^\top \theta = 0$ vs $H_1 : Q_1^\top \theta \neq 0$.

19.3 Linear regression

Let $x_i \in \mathbb{R}^d$ be fixed, and let $Y_i = x_i^\top \beta + \varepsilon_i$, where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Then $Y \sim N(X\beta, \sigma^2 I_n)$, where

$$X = \begin{bmatrix} - & x_1 & - \\ & \vdots & \\ - & x_d & - \end{bmatrix} = \begin{bmatrix} | & & | \\ X_1 & \cdots & X_n \\ | & & | \end{bmatrix} \in \mathbb{R}^{n \times d}$$

Assume that X has full column rank. Our model is to estimate $\theta = \mathbb{E}[Y] = X\beta$, where $\theta \in \text{span}(X_1, \ldots, X_d)$. Our null hypothesis is $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{d_1} = 0$, where $1 \le d_1 \le d$. This is the same as $\theta \in \text{span}(X_{d_1+1}, \ldots, X_d)$ (or $\{0\}$ if $d_1 = d$). In this model, we have $Q_0 = \text{Proj}_{\text{span}(x_{d_1+1},\ldots,x_d)}$ and $Q_1 = \text{Proj}_{\text{span}(x_1,\ldots,x_d)\cap\Theta_0^\perp}$.

We have

$$\begin{aligned}\|Z_r\| &= \|Y - \text{Proj}_\Theta Y\|^2 \\ &= \|Y - X\widehat{\beta}_{\text{OLS}}\|^2 \\ &= \sum_{i=1}^n (Y_i - x_i^\top \beta)^2,\end{aligned}$$

the residual sum of squares (RSS). Here,

$$\widehat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top Y = \underset{\beta\in\mathbb{R}^d}{\arg\min} \|Y - X\beta\|^2 = \underset{\theta\in\Theta}{\arg\min} \|Y - \theta\|^2.$$

Note that

$$\|Z_1\|^2 + \|Z_r\|^2 = \|Y - \text{Proj}_{\Theta_0}(Y)\| = \text{RSS}_0\,.$$

The F-statistic is

$$\frac{\|Z_1\|^2/(d-d_0)}{\|Z_r\|^2/(n-d)} = \frac{(\text{RSS}_0 - \text{RSS})/(d-d_0)}{\text{RSS}\,/(n-d)} \overset{\text{H}_0}{\sim} F_{d-d_0,n-d}.$$

If $d = 1$, let $X_0 = \begin{bmatrix} X_2 & \cdots & X_d \end{bmatrix} \in \mathbb{R}^{d_0\times n}$. Then let

$$\begin{aligned}X_{1\perp} &= X_1 - \text{Proj}_{\Theta_0}(X_1) \\ &= X_1 - X_0 \underbrace{(X_0^\top X_0)^{-1} X_0^\top X_1}_{\gamma} \\ &= X_1 - X_0\gamma\end{aligned}$$

To make X_1 special, write $\theta = X\beta = X_{1\perp}\beta_1 + X_0(\beta_{-1} + \gamma\beta_1) = X_{1\perp}\beta_1 + X_0\delta$. Then

$$\begin{aligned}\begin{bmatrix}\widehat{\beta}_1 \\ \widehat{\delta}\end{bmatrix} &= ((X_{1\perp}X_0)^\top (X_{1\perp}X_0))^{-1}(X_{1\perp}X_0)^\top Y \\ &= \begin{bmatrix}(X_{1\perp}^\top X_{1\perp})^{-1} & 0 \\ 0 & (X_0^\top X_0)^{-1}\end{bmatrix}\begin{bmatrix}X_{1\perp}^\top Y \\ X_0^\perp Y\end{bmatrix},\end{aligned}$$

so

$$\widehat{\beta}_1 = \frac{X_{1\perp}^\top Y}{\|X_{1\perp}\|^2} = \frac{Z_1}{\|X_{1\perp}\|}.$$

Here $Q = [q_1] = \frac{X_{1\perp}}{\|X_{1\perp}\|}$, so $Z_1 = q_1^\top Y = \frac{X_{1\perp}^\top Y}{\|X_{1\perp}\|}$.

The variance of the OLS estimator is

$$\mathrm{Var}(\widehat{\beta}_1) = \mathrm{Var}\left(\frac{Z_1}{\|X_{1\perp}\|}\right) = \frac{\sigma^2}{\|X_{1\perp}\|^2}.$$

So the standard error of $\widehat{\beta}_1$ is

$$\mathrm{s.\,e.}(\widehat{\beta}_1) = \frac{\sigma}{\|X_{1\perp}\|}.$$

The t-statistic is

$$\frac{q_1^\top Y}{\sqrt{\mathrm{RSS}/(n-d)}} = \frac{\widehat{\beta}_1}{\widehat{\sigma}/\|X_{1\perp}\|} = \frac{\widehat{\beta}_1}{\widehat{\mathrm{s.\,e.}}(\widehat{\beta}_1)}.$$

19.4 One way ANOVA (fixed effect)

ANOVA is short for "analysis of variates."

Our model has $Y_{k,i} \overset{\text{ind}}{\sim} \mu_k + \varepsilon_{k,i}$, where $\varepsilon_{k,i} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ with $k = 1, \ldots, m$ and $i = 1, \ldots, n$. We want to test $H_0 : \mu_1 = \cdots = \mu_m = \mu$ for any $\mu \in \mathbb{R}$. Then the null has dimension $d_0 = 1$, and the whole model has dimension $d = m$. The residual dimension is $d_r = m(n-1)$.

If we concatenate everything into 1 long vector,

$$Q_0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \qquad Q_1 = \text{basis for orthogonal complement of } \mathbf{1}_{mn}.$$

Denote

$$\overline{Y}_k = \frac{1}{n}\sum_i Y_{k,i}. S_k^2 = \frac{1}{n-1}\sum_i (Y_{k,i} - \overline{Y}_k)^2,$$

$$\overline{Y} = \frac{1}{mn}\sum_k \sum_i Y_{k,i}, \qquad S_0^2 = \frac{1}{mn-1}\sum_k \sum_i (Y_{k,i} - \overline{Y})^2.$$

Then

$$\mathrm{RSS} = \sum_k \sum_i (Y_{k,i} - \overline{Y}_k)^2 = (n-1)\sum_k S_k^2 = \|Y\|^2 - n\sum_k \overline{Y}_k^2,$$

$$\mathrm{RSS}_0 \sum_k \sum_i (Y_{k,i} - \overline{Y})^2 = (mn-1)S_0^2 = \|Y\|^2 - mn\overline{Y}^2.$$

The F-statistic is

$$\frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(m-1)}{\mathrm{RSS}/(m(n-1))} = \frac{\frac{n}{m-1}(\sum_k \overline{Y}_k^2 - m\overline{Y}^2)}{\frac{1}{m(n-1)}\sum_k \sum_i (Y_{k,i} - \overline{Y}_i)^2} = \frac{\text{between variance}}{\text{within variance}}.$$

An equivalent test statistic would be

$$\frac{\mathrm{RSS}_0 - \mathrm{RSS}}{\mathrm{RSS}_0} = \frac{\|Z_1\|^2}{\|Z_1\|^2 + \|Z_r\|^2}.$$

This is asking "by what percentage does the residual sum of squares goes down?"" or "what fraction of the variance is explained by adding these extra variables to the model?" We reject when the residual variance goes down by a larger than expected percentage.

20 Convergence, Consistency, and Limit Theorems

20.1 A note about linear regression

Last time, we discussed linear regression, where we have $x_i \in \mathbb{R}^d$ and $y_i = x_i^\top \beta + \varepsilon_i$, where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Then we can write the density as

$$\begin{aligned} p(y) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - x_i^\top \beta)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}y^\top y + \frac{1}{\sigma^2}(X\beta)^\top y - \frac{\beta^\top X^\top X\beta}{2\sigma}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|y\|^2 + \frac{\beta^\top}{\sigma^2}(X^\top y) - A(\beta)\right). \end{aligned}$$

Then $X^\top y, \|y\|^2$ are sufficient iff $(X^\top X)^{-1}X^\top y$ and $\|y\|^2$ are sufficient. This is equivalent to the OLS estimator $\widehat{\beta}$ and RSS $= \|y\|^2 - \|X\widehat{\beta}\|^2$ being sufficient. So we can make a suffiicencty reduction to $\widehat{\beta}$, $\widehat{\sigma}^2$. Here, one can show that $\widehat{\beta} = (X^\top X)^{-1}X^\top y \sim N_d(\beta, \sigma^2(X^\top X)^{-1})$with $\widehat{\beta}\amalg\widehat{\sigma}^2$. Note that this is d-dimensional, rather than n-dimensional, so we have a dimensionality reduction.

20.2 Convergence and consistency

Let $X_1, X_2, \cdots \in \mathbb{R}^d$ be random variables.

Definition 20.1. X_n **converges in probability** to c, written $X_n \xrightarrow{p} c$, if

$$\mathbb{P}(\|X_n - c\| > \varepsilon) \to 0 \qquad \forall \varepsilon > 0.$$

This says that X_n becomes roughly constant.

Definition 20.2. X_n **converges in distribution** to X, written $X_n \xrightarrow{D} X$ or $X_n \implies X$, if $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, continuous functions f.

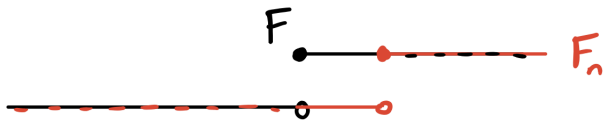
This says that when n is large, the distribution of X_n looks a lot like the distribution of X.

Theorem 20.1. *If $X_1, X_2, \cdots \in \mathbb{R}$, let the CDFs be $F_n(x) = \mathbb{P}(X_n \le x)$ and $F(x) = \mathbb{P}(X \le x)$. Then $X_n \implies X$ iff $F_n(x) \to F(x)$ for all x such that F is continuous at x.*

This is a weaker version of pointwise convergence, and convergence in distribution is sometimes called **weak convergence**. Here is why we want to only consider continuity points:

Example 20.1. Let $X_n \sim \delta_{1/n}$ and $X \sim \delta_0$. We want our definition to say $X_n \implies X$. The CDFs are

$$F_x(x) = \mathbb{1}_{\{1/n \le x\}}, \qquad F(x) = \mathbb{1}_{\{0 \le x\}}.$$



This example suggests that convergence in probability and in distribution are related.

Proposition 20.1. *$X_n \xrightarrow{p} c$ if and only if $X_n \implies \delta_c$.*

The kind of convergence we care most about in statistics is consistency:

Definition 20.3. If $\mathcal{P}_n = \{P_{\theta,n} : \theta \in \Theta\}$ with $X_n \sim P_{n,\theta}$, then we say that $\delta_n(X_n)$ is **consistent** for $g(\theta)$ if $\delta_n(X_n) \xrightarrow{p} g(\theta)$ for all θ, i.e.

$$\mathbb{P}_\theta(\|\delta_n(X_n) - g(\theta)\| > \varepsilon) \to 0.$$

20.3 Limit theorems

20.3.1 The law of large numbers and the central limit theorem

Theorem 20.2 ((Weak) law of large numbers)**.** *Let $X_1, X_2, \dots$ be iid random vectors, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. If $\mathbb{E}[\|X_i\|] < \infty$ and $\mathbb{E}[X_i] = \mu$, then $\overline{X}_n \xrightarrow{p} \mu$.*

Remark 20.1. You may have seen a stronger version of this theorem, in which we can prove that $\overline{X}_n \to \mu$ **almost surely**. In statistics, we are interested in convergence in probability, so this will suffice for our purposes.

Theorem 20.3 (Central limit theorem)**.** *Let $X_1, X_2, \dots \in \mathbb{R}^d$ be iid random vectors, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Assume that $\mathbb{E}[X_i = \mu$ and $\mathrm{Var}(X_i) = \Sigma < \infty$. Then*

$$\sqrt{n}(\overline{X}_n - \mu) \implies N_d(0, \Sigma).$$

20.3.2 The continuous mapping theorem

Here are three tools for how we propagate convergence to other kinds of random variables:

Theorem 20.4 (Continuous mapping)**.** *Let $X_1, X_2, \dots$ be random variables, and let g be a continuous function. If $X_n \implies X$, then $g(X_n) \implies g(X)$. In particular, if $X_n \xrightarrow{p} c$, then $g(X_n) \xrightarrow{p} g(c)$.*

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

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

20.3.3 Slutsky's theorem

Theorem 20.5 (Slutsky)**.** *Assume* $X_n \implies X$ *and* $Y_n \xrightarrow{p} c$*. Then*

$$X_n + Y_n \implies X + c, \qquad X_n Y_n \implies X \cdot c, \qquad \frac{X_n}{Y_n} \implies \frac{X}{c}$$

(where we assume $c \neq 0$ *for the last one).*

Proof. Here is a sketch: The first step is to show that $(X_n, Y_n) \implies (X, c)$. Then apply the continuous mapping theorem. □

20.3.4 The delta method

Last is the delta method, which informally says that if $X_n \approx (\mu, \sigma^2)$ with σ^2 small and f is differentiable, then $f(X_n) \approx N(f(\mu), \sigma^2 \dot{f}(\mu)^2)$.

Theorem 20.6 (Delta method)**.** *If* $\sqrt{n}(X_n - \mu) \implies N(0, \sigma^2)$ *and* $f(x)$ *is continuously differentiable at* μ*, then*

$$\sqrt{n}(f(X_n) - f(\mu)) \implies N(0, \sigma^2 \dot{f}(\mu)^2).$$

Proof. Here is the idea: Write $f(X_n) = f(\mu) + \dot{f}(\zeta_n)(X_n - \mu)$, where ζ_n is between μ and X_n (by the mean value theorem). $X_n \xrightarrow{p} \mu$ because $X_n - \mu \xrightarrow{p} 0$. Then $\zeta_n \xrightarrow{p} 0$, as well, because

$$\mathbb{P}(\|\zeta_n - \mu\| > \varepsilon) \leq \mathbb{P}(\|X_n - \mu\| > \varepsilon) \to 0.$$

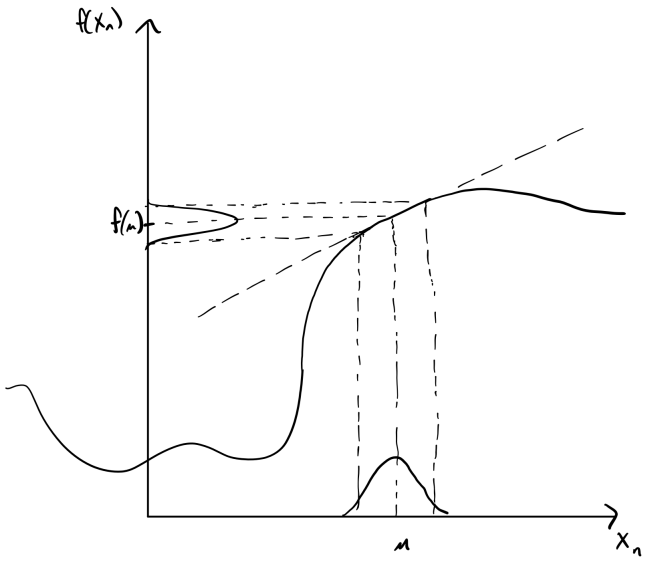
So, by the continuous mapping theorem applied to $\dot{f}$,

$$\sqrt{n}(f(X_n) - f(\mu)) = \underbrace{\dot{f}(\zeta_n)}_{\xrightarrow{p} \dot{f}(\mu)} \underbrace{\sqrt{n}(X_n - \mu)}_{\implies N(0, \sigma^2)},$$

So by Slutsky's theorem, $\sqrt{n}(f(X_n) - f(\mu)) \implies N(0, \sigma^2 \dot{f}(\mu)^2)$. □

Remark 20.2. We don't need to have $\sqrt{n}$ in the front. The theorem is still true if we replace $\sqrt{n}$ with a_n, as long as $a_n \to \infty$. Where in the proof did we use that $\sqrt{n} \to \infty$? This was necessary for the fact that $X_n \xrightarrow{p} \mu$.

Here is a picture of the delta method:



There is also a multivariate version:

Theorem 20.7 (Delta method, multivariate)**.** *If* $\sqrt{n}(X_n - \mu) \implies N(0, \Sigma)$ *and* $f : \mathbb{R}^d \to \mathbb{R}$ *is continuously differentiable at* μ*, then*

$$\sqrt{n}(f(X_n) - f(\mu)) \implies N(0, \nabla^\top \Sigma \nabla f).$$

The proof is the same as the univariate case.

Example 20.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (\mu, \sigma^2)$, and let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} (\nu, \tau^2)$ be independent of the X_i. Suppose we estimate $(\mu + \nu)^2$ with $(\overline{X} + \overline{Y})^2$. We can say a few things:

1. By the law of large numbers, $\overline{X} \xrightarrow{p} \mu$ and $\overline{Y} \xrightarrow{p} \nu$ as $n \to \infty$. The function $f(x, y) = (x + y)^2$ is continuous, so $f(\overline{X}, \overline{Y}) \xrightarrow{p} f(\mu, \nu)$. In other words,

   $$(\overline{X} + \overline{Y})^2 \xrightarrow{p} (\mu + \nu)^2,$$

   so $(\overline{X} + \overline{Y})^2$ is consistent for $(\mu + \nu)^2$.

2. The central limit theorem says that $\sqrt{n}(\overline{X} - n) \implies N(0, \sigma^2)$ and $\sqrt{n}(\overline{Y} - n) \implies N(0, \tau^2)$. Here,

   $$\frac{\partial f}{\partial x}(x, y) = \frac{\partial f}{\partial y}(x, y) = 2(x + y).$$

So the delta method tells us that

$$f(\overline{X}, \overline{Y}) \approx N\left(f(\mu,\nu), \frac{1}{n}\nabla f(\mu,\nu)^\top \begin{bmatrix}\sigma^2 & 0\\ 0 & \tau^2\end{bmatrix} \nabla f\right)$$
$$= N\left((\mu+\nu)^2, 4(\mu+\nu)^2(\sigma^2+\tau^2)/n\right).$$

More rigorously,

$$\sqrt{n}((\overline{X}+\overline{Y})^2 - (\mu+\nu)^2) \implies N(0, 4(\mu+\nu)^2(\sigma^2+\tau^2)).$$

3. What if $(\mu+\nu)^2 = 0$? Then

$$\sqrt{n}((\overline{X}-\overline{Y})^2 - (\mu+\nu)^2) \xrightarrow{p} 0.$$

We also know

$$\sqrt{n\overline{X}} + \sqrt{n\overline{Y}} \implies N(0, \sigma^2+\tau^2),$$

so if we square this sum,

$$n(\overline{X}+\overline{Y})^2 \implies (\sigma^2+\tau^2)\chi_1^2.$$

If we keep getting things converging to 0, we can keep blowing up the error to find what the distribution of the error rate is in this way.

21 Maximum Likelihood Estimation and Asymptotic Efficiency

21.1 Recap: Convergence in probability and distribution

Last time we introduced notions of convergence. We had

- Convergence in probability:

$$X_n \xrightarrow{p} c \quad \text{if} \quad \mathbb{P}(\|X_n - c\| > \varepsilon) \to 0 \quad \forall \varepsilon > 0.$$

- Convergence in distribution (sometimes called **weak convergence**[14]):

$$X_n \implies X \quad \text{if} \quad \mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)] \quad \forall \text{ bounded, continuous } f.$$

If $X_n, X \in \mathbb{R}$, then $X_n \implies X$ if and only if $F_n(x) \to F(x)$ for all x where F is continuous at x, where $F_n(x) = \mathbb{P}(X_n \leq x)$ and $F(x) = \mathbb{P}(X \leq x)$.

We had a few theorems will allow us to extend convergence to more random variables:

Theorem 21.1 (Continuous mapping)**.** *If f is continuous,*

$$X_n \to pX \implies f(X_n) \xrightarrow{p} f(X), \qquad X_n \to DX \implies f(X_n) \xrightarrow{D} f(X).$$

Theorem 21.2 (Slutsky)**.** *If $X_n \implies X$ and $Y_n \implies c$, then*

$$X_n + Y_n \implies X + c, \qquad X_n \cdot Y_n \implies cX, \qquad X_n / Y_n \implies X/c \quad (c \neq 0),$$

Theorem 21.3 (Delta method)**.** *Suppose $g(n)(X_n - \mu) \implies N_d(0, \Sigma)$, where $g(n) \to \infty$. Then for $f : \mathbb{R}^d \to \mathbb{R}^k$, where*

$$Df(x) = \begin{bmatrix} - & \nabla f_1(x)^\top & - \\ & \vdots & \\ - & \nabla f_k(x)^\top & - \end{bmatrix}$$

exists and is continuous at μ, then $g(n)(f(X_n) - f(\mu)) \implies N_k(0, Df(\mu)\Sigma Df(\mu)^\top)$.

[14] The real reason this is called weak convergence is that it corresponds to convergence of the distribution measures in a weak topology on $BC(\mathbb{R}^n)^*$, the dual space of the bounded continuous functions on $\mathbb{R}^n$.

21.2 Maximum likelihood estimators

Definition 21.1. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a dominated family with densities p_θ for P_θ with respect to μ. The **maximum likelihood estimator (MLE)** is

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

If we are worried about whether this exists, i.e. if the maximum is achieved, we can just take some ε tolerance instead. For now, we won't worry about that.

Remark 21.1. This is invariant to parametrization. If we have two different parameterizations θ and $\eta(\theta)$, then $\widehat{\eta}_{\mathrm{MLE}} = \eta(\widehat{\theta}_{\mathrm{MLE}})$. This is not the case for, for example, the UMVU estimator.

Example 21.1. Let $p_\eta(x) = e^{\eta^\top T(x) - A(\eta)} h(x)$. The log likelihood is

$$\ell(\eta; x) = \eta^\top T(x) - A(\eta) + \log h(x),$$

so

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

Note that $\nabla\ell$ is concave. If we set it equal to 0, we get something like a method of moments estimator.

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

Let $\mu = \psi(\eta) = \nabla A$. Then $\widehat{\eta} = \psi^{-1}(T(X))$.

Example 21.2. Let $X_i \overset{\text{iid}}{\sim} e^{\eta T(x) - A(\eta)} h(x)$ with $\eta \in \Xi \subseteq \mathbb{R}$. Then

$$\widehat{\eta} = \psi^{-1}(\overline{T}), \qquad \overline{T} = \frac{1}{n}\sum_{i=1}^n T(X_i).$$

Assume $\eta \in \Xi^o$ and $\dot{\psi}(\eta) = \ddot{A}(\eta) > 0$. Then ψ^{-1} is continuous, so

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

By the law of large numbers, $\overline{T} \xrightarrow{p_\eta} \mu$. Here, we write p_η to emphasize that this convergence depends on η. So the continuous mapping theorem gives consistency: $\psi^{-1}(\overline{T}) \xrightarrow{p_\eta} \eta$.

The central limit theorem gives

$$\sqrt{n}(\overline{T} - \mu) \implies N(0, \operatorname{Var}_\eta T(X_1)) = N(0, \ddot{A}(\eta)),$$

where the Fisher information is $J_1^\eta(\mu) = \ddot{A}(\mu)^{-1}$. The delta method gives

$$\sqrt{n}(\widehat{\eta} - \eta) = \sqrt{(}\psi^{-1}(\overline{T}) - \eta) \implies N(0, \tfrac{1}{\ddot{A}(\eta)^2}\ddot{A}(\eta)) = N(0, \ddot{A}(\eta)^{-1}).$$

Recall that $J_1^\eta(\mu) = \operatorname{Var}_\eta(T(X_1)) = \ddot{A}(\eta)$. Asymptotically, $\widehat{\eta}$ is unbiased and achieves the Cramér-Rao lower bound because $J_n^\eta(\eta) = n\ddot{A}(\eta)$.

What do we mean by asymptotically unbiased?

Example 21.3. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Pois}(\theta) = \frac{\theta^x e^{-\theta}}{x!}$, and let $\eta = \log\theta$. Then $\widehat{\eta} = \log\overline{X}$. The central limit theorem says $(\overline{X} - \theta) \implies N(0, \theta)$, so the delta method tells us that

$$\sqrt{n}(\widehat{\eta} - \eta) = \sqrt{n}(\log\overline{X} - \log\theta) \implies N(0, \tfrac{1}{\theta^2}\theta) = N(0, \theta^{-1}).$$

What if $\overline{X} = 0$? In fact, $\widehat{\eta}$ has bias $-\infty$ and infinite variance, so the bias does not converge to 0. What we mean by asymptotically unbiased is that the scaled limiting distribution has no bias.

If you are a glass half-empty person, you might say that we can never use $\widehat{\eta}$, since it will always have infinite mean squared error. But if you are a glass half-full person, you might say that

$$\mathbb{P}_\eta(\overline{X} = 0) = \mathbb{P}_\theta(X_1 = 0)^n = e^{-n\theta},$$

which is an exponentially decaying probability of anything bad happening.

Proposition 21.1. *If $X_n \implies X$, Z_n is arbitrary, and B_n is an event such that $\mathbb{P}(B_n) \to 0$, then*

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

Proof. Observe that $Z_n \xrightarrow{p} 0$:

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

So $Z_n \mathbb{1}_{B_n} \xrightarrow{p} 0$. Since $\mathbb{1}_{B_n^c} \xrightarrow{p} 1$, use Slutsky's theorem to get the result. □

21.3 Asymptotic efficiency

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} p_\theta(x)$ with $\theta \in \mathbb{R}^d$. Assume that p_θ is "smooth" in θ, e.g. it has 2 continuous integrable derivatives. Let

$$\ell_1(\theta; X_i) = \log p_\theta(X_i), \qquad \ell_n(\theta, X) = \sum_{i=1}^n \ell_1(\theta; X_i).$$

Recall the Fisher information for a single observation is

$$H_1(\theta) = \mathrm{Var}_\theta(\nabla \ell_1(\theta; X_i)).$$

The likelihood ratio, which captures everything about the data, looks like

$$\frac{\mathrm{Lik}(\theta+\delta; X)}{\mathrm{Lik}(\theta; X)} = \log(\ell_n(\theta+\delta) - \ell_n(\theta)) \approx \log(\delta^\top \nabla \ell_n(\theta)).$$

The Fisher information for n observations is

$$J_n(\theta) = \mathrm{Var}_\theta(\nabla \ell_n(\theta; X)) = nJ_1(\theta).$$

Recall that $\mathbb{E}[\nabla \ell_1(\theta)] = 0$.

Definition 21.2. An estimator $\widehat{\theta}_n$ is **asymptotically efficient** if

$$\sqrt{n}(\widehat{\theta}_n - \theta) \overset{P_\theta}{\implies} N(0, J_1(\theta)^{-1}).$$

Really this is a sequence of estimators converging, but this is usually understood from context. For continuously differentiable $g(\theta) : \mathbb{R}^d \to \mathbb{R}$,

$$\sqrt{n}(g(\widehat{\theta}_n) - g(\theta)) \implies N(0, \nabla g(\theta)^\top J_1(\theta)^{-1} \nabla g(\theta)).$$

You may recognize this as the Cramér-Rao lower bound.

Let θ_0 be the true value, and let θ be a generic value of the parameter. We will maximize $\ell_n(\theta; X)$ by setting $\nabla \ell_n(\widehat{\theta}_{\mathrm{MLE}}) = 0$. We know $\nabla \ell_1(\theta_0; X_i) \overset{\mathrm{iid}}{\sim} (0, J_1(\theta_0))$, so by the central limit theorem,

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0; X) = \sqrt{n} \frac{1}{n} \sum_{i=1}^n \nabla \ell_1(\theta_0, X_i) \implies N(0, J_1(\theta_0)).$$

Now calculate the second derivative: Using the law of large numbers,

$$\frac{1}{n} \nabla^2 \ell_n(\theta_0, X) \xrightarrow{p} \mathbb{E}_{\theta_0}[\nabla^2 \ell(\theta_0; X_i)] = -J_1(\theta).$$

Here is an informal proof of why the MLE should be asymptotically efficient.

Proof. Assume

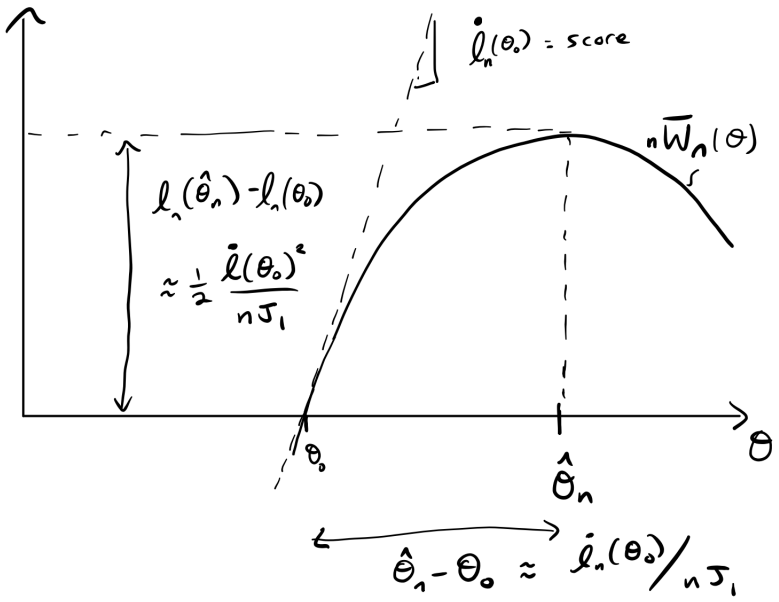
$$0 = \nabla \ell_n(\widehat{\theta}_n; X) = \nabla \ell_n(\theta_0) + \nabla^2 \ell_n(\theta_0)(\widehat{\theta}_n - \theta),$$

using a Taylor expansion. Then

$$\sqrt{(}\widehat{\theta}_n - \theta_0) \approx \underbrace{\left(-\frac{1}{n} \nabla^2 \ell_n(\theta_0)\right)^{-1}}_{\xrightarrow{p} J_1(\theta_0)^{-1}} \underbrace{\left(\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0)\right)}_{\implies N(0, J_1(\theta_0))} \implies N(0, J_1(\theta_0)^{-1}),$$

which gives asymptotic efficiency. □

What's missing from this proof? To do our Taylor expansion, we need to first show that $\widehat{\theta}_n$ is close to θ_0; that is, we want to show consistency: $\widehat{\theta}_n \xrightarrow{p} \theta_0$.



22 Asymptotic Consistency of the Maximum Likelihood Estimator

22.1 Recap: Maximum likelihood estimation

Last time, we introduced maximum likelihood estimation. If our model is $\mathcal{P}$ with densities $p_\theta(x)$ with respect to μ and if our sample is $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_{\theta_0}$, then the **maximum likelihood estimator (MLE)** is

$$\begin{aligned}
\widehat{\theta}_n &= \arg\max_{\theta \in \Theta} p_\theta(X) \\
&= \arg\max_{\theta \in \Theta} \ell_n(\theta; X) \\
&= \arg\max_{\theta \in \Theta} \sum_{i=1}^n \underbrace{\ell_1(\theta; X_i) - \ell_1(\theta_0, X_i)}_{W_i(\theta)} \\
&= \arg\max_{\theta \in \Theta} \overline{W}_n(\theta),
\end{aligned}$$

where

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

We are interested in how $\overline{W}_n$ converges to its expectation.

Last time, we made a quadratic expansion of the log-likelihood (or a linear expansion of the score),

$$0 = \nabla \ell_n(\widehat{\theta}_n) = \nabla \ell_n(\theta_0) + \nabla^2 \ell(\widetilde{\theta}_n)(\widehat{\theta}_n - \theta_0),$$

where $\widetilde{\theta}_n$ is some value given by the mean value theorem. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \left(-\frac{1}{n} \nabla^2 \ell(\widehat{\theta}_n) \right)^{-1} \underbrace{\left(\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0) \right)}_{\Longrightarrow N_d(0, J_1(\theta_0))}.$$

We want to say that the first term converges in probability to $J_1(\theta_0)^{-1}$. We need a few ingredients:

- $\widehat{\theta}_n \xrightarrow{p} \theta_0$.
- $J_1(\theta_0) \succ 0$.
- We need to deal with a random function at the random value $\widehat{\theta}_n$.

22.2 Pointwise convergence of likelihood ratio averages

We can say $\overline{W}_n(\theta)$ is a sample mean of iid $W_1(\theta), \ldots, W_n(\theta)$. Recall the **KL-Divergence**

$$D_{\mathrm{KL}}^{(1)}(\theta_0 \mid\mid \theta) = \mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(X_1)}{p_\theta(X_1)}\right].$$

Then by Jensen's inequality,

$$\begin{aligned} -D_{\mathrm{KL}}^{(1)}(\theta_0 \mid\mid \theta) &\leq \log \mathbb{E}_{\theta_0}\left[\frac{p_{\theta_0}(X_1)}{p_\theta(X_1)}\right] \\ &\leq \log 1 \\ &= 0. \end{aligned}$$

Since log is strictly concave, this is a strict inequality unless $p_{\theta_0} = p_\theta$.

Now let's calculate the expectation of the Ws:

$$\begin{aligned} \mathbb{E}_{\theta_0}[\overline{W}_n(\theta)] &= \mathbb{E}_{\theta_0}[W_i(\theta)] \\ &= \mathbb{E}_{\theta_0}[\ell_1(\theta; X_1) - \ell_1(\theta_0; X_i)] \\ &= -D_{\mathrm{KL}}(\theta_0 \mid\mid \theta) \\ &< 0, \end{aligned}$$

unless $p_{\theta_0} = p_0$. Then

$$\overline{W}_n(\theta) \xrightarrow{p} -D_{\mathrm{KL}}(\theta_0 \mid\mid \theta) < 0$$

unless $p_{\theta_0} = p_0$. We need a way to make this convergence uniform.

22.3 Uniform convergence of random functions

Definition 22.1. For a compact K, let $C(K)$ be the set of all continuous functions $f : K \to \mathbb{R}$.

Definition 22.2. For any $f \in C(K)$, the L^∞ **norm is**

$$\|f\|_\infty = \sup_{t \in K} |f(t)|.$$

Definition 22.3. We say that $f_n \to f$ in this norm (f_n **converges uniformly** to f) if $\|f_n - f\|_\infty \to 0$.

Theorem 22.1 (Law of large numbers for random functions)**.** *Assume K is comapct, and $W_1, W_2, \cdots \in C(K)$ are iid with $\mathbb{E}[\|W_i\|_\infty] < \infty$. Let $\mu(t) = \mathbb{E}[W_i(t)]$. Then $\mu(t) \in C(K)$, and*

$$\left\| \frac{1}{n}\sum_{i=1}^n W_i - \mu \right\|_\infty \xrightarrow{p} 0.$$

*That is, $\frac{1}{n}\sum_{i=1}^n W_i \to \mu$ **uniformly in probability**.*

We won't prove this.

Theorem 22.2 (9.4 in Keener)**.** *Let $G_1, G_2, \ldots$ be random functions in $C(K)$ with K compact. Assume that $\|G_n - g\|_\infty \xrightarrow{p} 0$ for some fixed $g \in C(K)$. Then*

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

Proof.

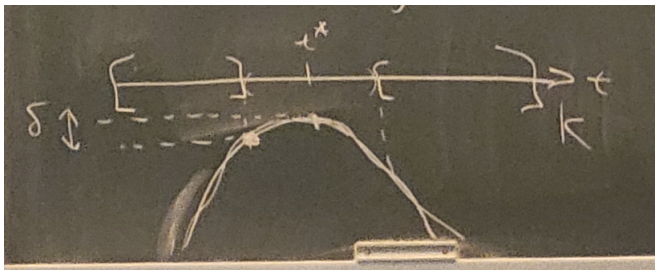
1.

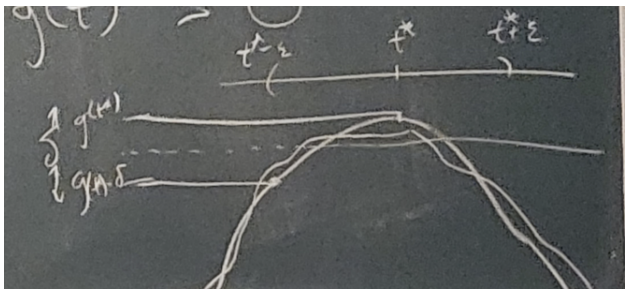
$$\begin{aligned} |G_n(t_n) - g(t^*)| &\le |G_n(t_n) - g(t-n)| + |g(t_n) - g(t^*)| \\ &\le \underbrace{\|G_n - g\|_\infty}_{\xrightarrow{p} 0} + \underbrace{|g(t_n) - g(t^*)|}_{\xrightarrow{p} 0}, \end{aligned}$$

   where the second term converges to 0 in probability by the continuous mapping theorem. So $G_n(t_n) \xrightarrow{p} g(t^*)$.

2. Fix $\varepsilon > 0$, and let $B_\varepsilon(t^*) = \{t : \|t - t^*\| < \varepsilon\}$. Let $K_\varepsilon = K \setminus B_\varepsilon(t^*)$; this intersection is also compact. Let

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





If $t_n \in K_\varepsilon$, then

$$G_n(t_n) \leq \max_{t \in K_\varepsilon} g(t) + \|G_n - g\|_\infty. = g(t^*) - \delta + \|G_n - g\|_\infty.$$

We also know that

$$G_n(t_n) \geq G_n(t^*) \geq g(t^*) - \|G_n - g\|_\infty.$$

Subtracting these inequalities gives

$$2\|G_n - g\|_\infty \geq \delta.$$

The probability of this is going to 0 by assumption, so $\mathbb{P}(t_n \in K_\varepsilon) \to 0$.

3. The proof of this is analogous to the proof of the second statement. □

What if we don't need the exact maximizer or if there is no exact maximizer? We can modify part 2 of the theorem:

Theorem 22.3. *Let $G_1, G_2, \ldots$ be random functions in $C(K)$ with K compact. Assume that $\|G_n - g\|_\infty \xrightarrow{p} 0$ for some fixed $g \in C(K)$. Then if g is maximized at a unique value $t^* \in K$ and $G_n(t_n) = \max_t G_n(t) - \alpha_n$ with $\alpha_n \to 0$, then $t_n \xrightarrow{p} t^*$.*

Proof. We can repeat the same argument, except this time we get

$$F_n(t_n) \geq G_n(t^*) - \alpha_n \geq g(t^*) - \|G_n - g\|_\infty - \alpha_n.$$

This gives

$$2\|G_n - g\|_\infty \geq \delta - \alpha_n,$$

and the proof still works. □

22.4 Consistency results for the MLE

Theorem 22.4 (Consistency of the MLE for compact Θ)**.** *Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_{\theta_0}$, where $\mathcal{P}$ has continuous densities p_θ for $\theta \in \Theta$. Assume that*

- *Θ is compact,*
- *$\mathbb{E}_{\theta_0}[\|W_i\|_\infty] = \mathbb{E}_{\theta_0}[\|\ell_1(\theta; X_i) - \ell_1(\theta_0; X_i)\|_\infty] < \infty$,*
- *The model $\mathcal{P}$ is identifiable.*

Then $\widehat{\theta}_n \xrightarrow{p} \theta_0$ if $\widehat{\theta}_n \in \arg\max \ell_n(\theta; X)$.

So it doesn't matter which value we pick for the MLE; we still get consistency.

Proof. Since the densities are continuous, $W_i \in C(\Theta)$. They are iid with mean $\mu(\theta) = -D_{\mathrm{KL}}(\theta_0 \,||\, \theta)$, where $\mu(\theta_0) = 0$ and $\mu(\theta) < 0$ for all $\theta \neq \theta_0$. So θ_0 uniquely maximizes μ. By definition, $\widehat{\theta}_n$ maximizes $\overline{W}_n$, so $\|\overline{W}_n - \mu\|_\infty \xrightarrow{p} 0$ by the law of large numbers. Now apply the previous theorem. □

Here is a way (but not the only way) to restrict our attention to a compact set.

Theorem 22.5 (Keener 9.11 with slightly stronger assumptions)**.** *Let* $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_{\theta_0}$, *where the model* $\mathcal{P}$ *has continuous densities* p_θ *for* $\theta \in \Theta \subseteq \mathbb{R}^d$. *Assume*

- *The model is identifiable.*
- *For all compact* $K \subseteq \mathbb{R}^d$, $\mathbb{E}[\sup_{\theta \in K} |W_i(\theta|] < \infty$.
- *There exists an* $r > 0$ *such that*

$$\mathbb{E}\left[\sup_{\|\theta - \theta_0\| > r} W_i(\theta)\right] < 0.$$

Then $\widehat{\theta}_n \xrightarrow{p} \theta_0$ *if* $\widehat{\theta}_n \in \arg\max \ell_n(\theta; X)$.

Proof. Let $A = \{\theta : \|\theta - \theta_0\| > r\}$, and let $\alpha = \mathbb{E}[\sup_{\theta \in A} W_i(\theta)] < 0$. Then

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

So

$$\mathbb{P}(\widehat{\theta}_n \in A) \leq \mathbb{P}(\overline{W}_n(\theta_0) \leq \sup_{\theta \in A} \overline{W}_n(\theta_)) \xrightarrow{0},$$

as $\alpha \xrightarrow{p} 0$ implies $\sup_{\theta \in A} \overline{W}_n(\theta) \xrightarrow{p} 0$. Now let

$$\widehat{\theta}_n^A = \widehat{\theta}_n \mathbb{1}_{\{\widehat{\theta}_n \in A^c\}} + \theta_0 \mathbb{1}\{\widehat{\theta}_n \in A\} \xrightarrow{p} \theta_0.$$

Then $\widehat{\theta}_n \xrightarrow{p} \theta_0$. □

23 Asymptotic Consistency of the MLE and Likelihood-Based Hypothesis Tests

23.1 Recap: Uniform convergence of random functions

Last time, we were interested in uniform convergence of the random functions given by the sample mean of $W_i(\theta; X_i) = \ell_1(\theta; X_i) - \ell_1(\theta_0; X_i)$. The nice thing about these is that

$$\mathbb{E}[W_i(\theta)] = D_{\mathrm{KL}}(\theta \mid\mid \theta_0),$$

which is ≤ 0, with equality iff $P_\theta = P_{\theta_0}$. We saw that $\widetilde{\theta}_n \xrightarrow{p} \theta_0$ if the W_i are continuous and $\|\overline{W}_n - \mathbb{E}[\overline{W}_n]\|_\infty \xrightarrow{p} 0$ on compact Θ (otherwise, we need an extra argument).

We also proved the helpful fact

Proposition 23.1. *If $\|G_n - g\|_\infty \xrightarrow{p} 0$, $t_n \xrightarrow{p} t$, and G_n, g are continuous with compact domain, then*

$$G_n(t_n) \xrightarrow{p} g(t).$$

23.2 Asymptotic distribution of the MLE

Theorem 23.1. *Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_{\theta_0}$, where $\theta_0 \in \Theta^o \subseteq \mathbb{R}^d$. Assume that*

- *$\widehat{\theta}_n \xrightarrow{p} \theta_0$, where $\widehat{\theta}_n \in \operatorname{argmax}_{\theta\in\Theta} \ell_n(\theta; X)$*
- *In some neighborhood $B_\varepsilon(\theta_0) = \{\theta : \|\theta - \theta_0\| \leq \varepsilon\} \subseteq \Theta^o$,*
  - *(i) $\ell_1(\theta; X)$ has 2 continuous derivatives on $B_\varepsilon(\theta_0)$ for all x.*
  - *(ii) $\mathbb{E}_{\theta_0}[\sup_{\theta\in B_\varepsilon} \|\nabla^2 \ell_1(\theta; X_i)\|] < \infty$.*
  - *(iii) Fisher information condition:*

$$\mathbb{E}_{\theta_0}[\nabla \ell_1(\theta_0; X_i)] = 0, \qquad \operatorname{Var}_{\theta_0}(\nabla \ell_1(\theta)) = -\mathbb{E}_\theta[\nabla^2 \ell_1(\theta_0)] \succ 0.$$

Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \implies N_d(0, J_1(\theta_0)^{-1}),$$

i.e. the MLE is asymptotically efficent.

The conditions in this theorem can be relaxed somewhat.

Proof. Let A_n be the event $\{\|\widehat{\theta}_n - \theta_0\| \geq \varepsilon\}$. Then $\mathbb{P}_{\theta_0}(A_n) \to 0$ by assumption. All we care about is what happening on A_n^c. On A_n^c, $\widehat{\theta}_n \in B_\varepsilon(\theta_0)$, and

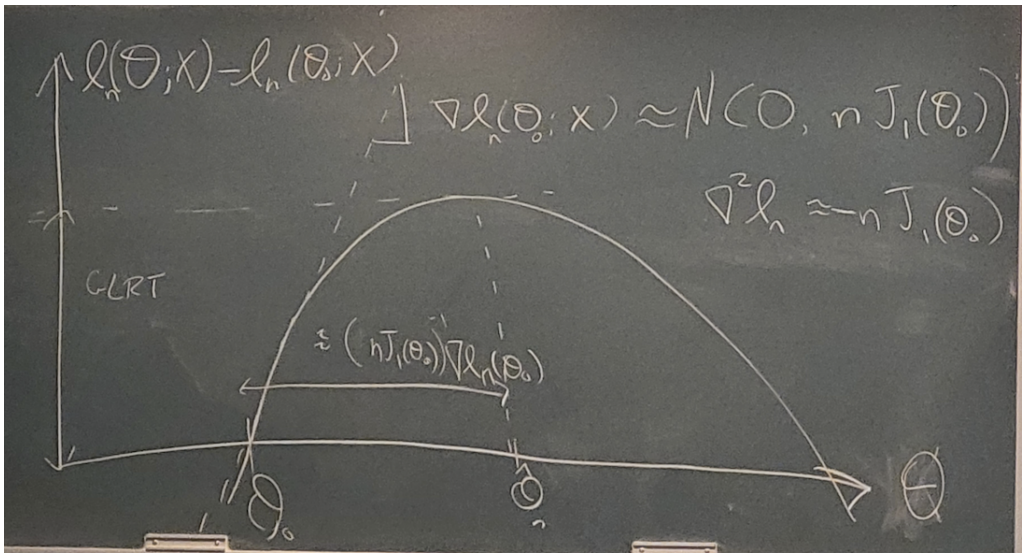
$$\begin{aligned} 0 &= \nabla \ell_n(\widehat{\theta}_n; X) \\ &= \nabla \ell_n(\theta_0; X) + \nabla^2 \ell_n(\widetilde{\theta}_0; X)(\widehat{\theta}_n - \theta_0) \end{aligned}$$

for some $\widetilde{\theta}_n$. Now

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \underbrace{\left(\frac{1}{n}\nabla^2 \ell_n(\widetilde{\theta}_n)\right)^{-1}}_{\xrightarrow{p} J_1(\theta_0)^{-1}} \underbrace{\frac{1}{\sqrt{n}}\nabla \ell_n(\theta_0)}_{\Longrightarrow N_d(0, J_1(\theta))},$$

$$\Longrightarrow N_d(0, J_1(\theta_0)^{-1}). \qquad \square$$

The proof basically says that the second derivative of the likelihood is approximately non-random and equals the Fisher information.



If the fisher information is very large, the second derivative of the likelihood function is huge at θ_0. This makes the likelihood more strongly peaked, so the MLE won't be so far from θ_0.

23.3 Likelihood-based hypothesis tests

We can develop likelihood-based tests based on measuring different aspects of the above MLE picture. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_\theta(x)$, where $p_\theta(x)$ is "smooth" in θ. Assume that

$$\mathbb{E}_\theta[\nabla \ell_1(\theta; X_i)] = 0, \qquad \text{Var}_\theta(\nabla \ell_1(\theta; X_i)) = -\mathbb{E}_\theta[\nabla^2 \ell_1(\theta; X_i)] = J_1(\theta) \succ 0,$$

and $\widehat{\theta}_{\text{MLE}} \xrightarrow{p} \theta_0$. Then if $\theta = \theta_0$,

$$\frac{1}{\sqrt{n}}\nabla \ell_n(\theta_0) \Longrightarrow N_d(0, J_1(\theta_0),$$

$$-\frac{1}{n}\nabla^2 \ell_n(\theta_0) \xrightarrow{p} J_1(\theta_0),$$

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \Longrightarrow N_d(0, J_1(\theta)^{-1}).$$

23.3.1 Wald-type confidence regions

Assume we have an estimator $\widehat{J}_n \succ 0$ such that $\frac{1}{n}\widehat{J}_n \xrightarrow{p} J_1(\theta_0) \succ 0$. Then

$$(J_1(\theta_0))^{1/2}\sqrt{n}(\widehat{\theta}_n - \theta_0) \implies N_d(0, I_d),$$

and by Slutsky's theorem,

$$\widehat{J}_n^{1/2}(\widehat{\theta}_n - \theta_0) \implies N_d(0, I_d).$$

To get a test statistic, we can do the simplest (but not always the best) thing and take the 2-norm:

$$\|\widehat{J}_n^{1/2}(\widehat{\theta}_n - \theta_0)\|^2 \implies \chi^2_d.$$

Here,

$$\mathbb{P}(\|\widehat{J}_n^{1/2}(\widehat{\theta}_n - \theta_0)\|^2 > \chi^2_d(\alpha)) \to \alpha,$$

where $\chi^2_d(\alpha)$ is the upper-α quantile.

To test $H_0 : \theta = \theta_0$, we reject if $\|\widehat{J}_n^{1/2}(\widehat{\theta}_n - \theta_0)\|_2^2 > \chi^2_d(\alpha)$. Equivalently, we can say we reject θ_0 iff $\widehat{J}_n^{1/2}(\widehat{\theta}_n - \theta_0) \notin B_{\chi^2_d(\alpha)}(0)$. So we can reject θ_0 if and only if $\theta_0 \notin \widehat{\theta} + \widehat{J}_n^{1/2} B_{\chi^2_d(\alpha)}(0)$. This gives a *confidence ellipsoid*.

Here are some options for $\widehat{J}_n$:

1.

$$\begin{aligned}\widehat{J}_n &= nJ_1(\widehat{\theta}_n) \\ &= n \operatorname{Var}_\theta(\nabla \ell_n(\theta; X))|_{\theta=\widehat{\theta}_n} \\ &= n \operatorname{Var}_{\widehat{\theta}_n}(\nabla \ell_n(\widehat{\theta}_n; X))\end{aligned}$$

2. Observed Fisher information:

$$\widehat{J}_n = -\nabla^2 \ell_n(\widehat{\theta}_n; X)$$

The observed Fisher information is generally preferred and is used in practice. We can get a *Wald interval* for θ_j by

$$\theta_n \approx N_d(\theta_0, J_n(\theta_0)^{-1}),$$

which tells us that

$$\widehat{\theta}_{n,j} \approx N(\theta_{0,j}, (J_n(\theta_0)^{-1})_{j,j}).$$

So the **univariate Wald interval** for θ_j is

$$\begin{aligned}C_j &= \widehat{\theta}_{n,j} \pm \widehat{\text{s.e.}}(\widehat{\theta}_{n,j}) z_{\alpha/2} \\ &= \widehat{\theta}_{n,j} \pm \sqrt{(\widehat{J}_n^{-1})_{j,j}} z_{\alpha/2}\end{aligned}$$

23.3.2 The score test

Here is a test which only assumes normality of the Fisher information. Test $J_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. Then

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0; X) \overset{H_0}{\implies} N_d(0, J_1(\theta_0)),$$

and the **score statistic** looks like

$$J_n(\theta_0)^{-1/2} \nabla \ell_n(\theta_0; X) \overset{H_0}{\implies} N_d(0, I_d).$$

So we reject H_0 if $\|J_n(\theta_0)^{-1/2} \nabla \ell_n(\theta_0; X)\|^2 > \chi^2_d(\alpha)$.

If $d = 1$, this looks like

$$\frac{\dot{\ell}_n(\theta_0)}{\sqrt{J_n(\theta_0)}} \implies N(0, 1).$$

This is actually invariant of parameterization. For simplicity of notation, assume $d = 1$ for now. Let $\theta = g(\zeta)$ with $\dot{\zeta}) > 0$ be a reparameterization, and denote $q_\zeta(x) = p_{g(\zeta)}(x)$. Then the score is

$$\begin{aligned} \dot{\ell}^{(\zeta)}(\zeta, x) &= \frac{d}{d\zeta} \log p_{g(\zeta)}(x) \\ &= \dot{\ell}(g(\zeta)) \dot{g}(\zeta) \end{aligned}$$

by the chain rule. The Fisher information is

$$J^{(\zeta)}(\zeta) = J^{(\theta)}(g(\zeta) \dot{g}(\zeta)^2.$$

So the score statistic is unchanged by the parameterization.

Example 23.1. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} e^{\eta^\top T(x) - A(\eta)} h(x)$ be an s-parameter exponential family. THen

$$\nabla \ell_n(\eta) = \left(\sum_{i=1}^n T(X_i) \right) - n\mu(\zeta), \qquad \text{where } \mu(\eta) = \mathbb{E}_\eta[T(X_i)].$$

Then

$$\left\| J_n(\eta_0)^{-1/2} \left(\sum_i T(X_i) - n\mu(\eta) \right) \right\|_2^2 \implies \chi^2_d$$

gives us our test. In particular, if $d = 1$, we get

$$\frac{\sum_i T(X_i) - n\mu(\eta_0)}{\sqrt{n \operatorname{Var}_{\eta_0}(T(X_1))}} \overset{H_0}{\implies} N(0, 1),$$

so this is a Z-test.

The test statistic for the score test is

$$\|(J_1(\theta_0))^{-1/2}\tfrac{1}{\sqrt{n}}\nabla\ell_n(\theta_0)\|^2,$$

while the test statistic for the Wald test is

$$\|\widehat{J}_1^{1/2}\sqrt{n}(\widehat{\theta}_n-\theta_0)\|^2,$$

where $\sqrt{n}(\widehat{\theta}_n-\theta_0)\approx J_1(\theta_0^{-1})\frac{1}{n}\nabla\ell_n(\theta_0)$. So these are asymptotically the same test.

24 Generalized Likelihood Ratio Tests, Asymptotic Relative Efficiency, and Pearson's χ^2 Test

24.1 Recap: Likelihood-ratio based hypothesis tests

We have been assuming a parametric model $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_{\theta_0}(x)$, where $\theta_0 \in \Theta^o \subseteq \mathbb{R}^d$. $p_\theta(x)$ sufficiently regular in θ. We have the MLE

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \ell_n(\theta; X),$$

which we assume converges in probability to θ_0. The central limit theorem tells us that

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0; X) \implies n_d(0, J_1(\theta_0)),$$

where we can think of $\nabla \ell_n$ as a complete sufficient statistic for all the likelihood ratios. We had the Taylor expansion

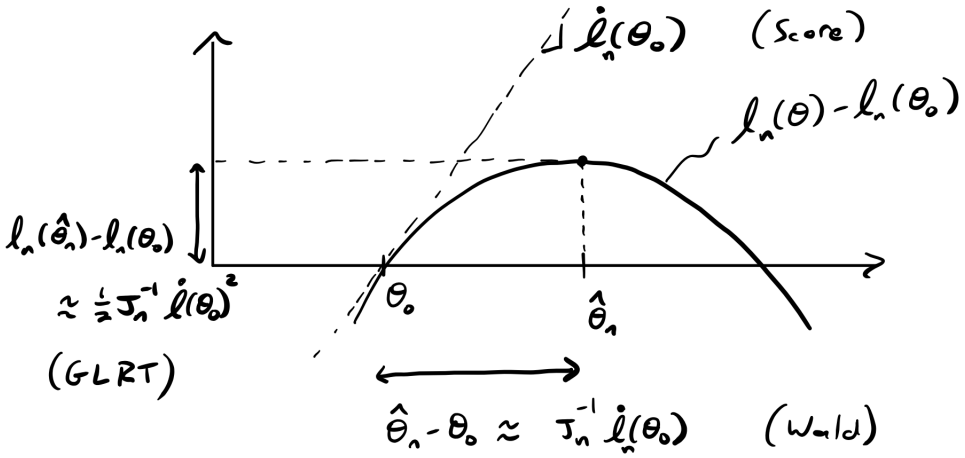
$$0 = \nabla \ell_n(\widehat{\theta}_n) = \nabla \ell_n(\theta_0) + \nabla^2 \ell_n(\widetilde{\theta}_n)(\widehat{\theta}_n - \theta_0),$$

which told us that

$$\begin{aligned} \sqrt{n}(\widehat{\theta}_n - \theta_0) &= \underbrace{\left(-\frac{1}{n} \nabla^2 \ell_n(\widetilde{\theta}_n) \right)^{-1}}_{\xrightarrow{p} J_1(\theta_0)^{-1}} \underbrace{\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0)}_{\implies N(0, J_1(\theta_0))} \\ &\implies N_d(0, J_1(\theta_0)^{-1}). \end{aligned}$$

We have following picture of the second order Taylor approximation of the log-likelihood

$$\ell_n(\theta) - \ell_n(\theta_0) \approx \dot{\ell}_n(\theta_0)(\theta - \theta_0) - \frac{1}{2} J_n(\theta_0)(\theta - \theta_0)^2.$$



Different parts of this picture give us different likelihood-based test statistics for hypothesis testing.

For large n,

$$2(\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)) \approx \|J_n^{1/2}(\widehat{\theta}_n - \theta_0)\|^2,$$

which gives us the Wald test. Looking at

$$2(\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)) \approx \|J_n(\theta_0)^{-1/2}\nabla\ell_n(\theta_0)\|^2,$$

gives us the score test, and

$$2(\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)) \approx \|J_n(\theta)^{1/2}(\widehat{\theta}_n - \theta_0)\|^2,$$

gives us the generalized likelihood ratio test. This is looking at the vertical distance in the above picture.

24.2 Generalized likelihood ratio tests

24.2.1 GLRT with a simple null

Suppose we want to test $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. We have

$$\begin{aligned}
\ell_n(\theta_0) - \ell_n(\widehat{\theta}_n) &= \cancelto{0}{\nabla\ell_n(\widehat{\theta}_n)} + \frac{1}{2}(\theta_0 - \widehat{\theta}_n)^{-1}\nabla^2\ell_n(\widetilde{\theta}_n)(\theta_0 - \widehat{\theta}_n) \\
&= -\frac{1}{2}\| \underbrace{(-\tfrac{1}{n}\nabla^2\ell_n(\widetilde{\theta}_n))^{1/2}}_{\xrightarrow{p} J_1(\theta_0)^{1/2}} \underbrace{\sqrt{n}(\theta_0 - \widehat{\theta}_n)}_{\Longrightarrow N_d(0, J_1(\theta_0)^{-1})} \|^2 \\
&\Longrightarrow -\frac{1}{2}\chi^2_d.
\end{aligned}$$

This means that

$$2(\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0)) \implies \chi^2_d.$$

We should reject θ_0 if and only if

$$\ell_n(\theta_0) \leq \ell_n(\widehat{\theta}_n) - \frac{1}{2}\chi^2_d(\alpha).$$

This has some of the advantages of the Wald test, such as invariance under parameterization, but without requiring the confidence set to always be an ellipsoid.

24.2.2 GLRT with a composite null or with nuisance parameters

Theorem 24.1. *Suppose we are testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \notin \Theta_0$. Assume that*

- *$\Theta \subseteq \mathbb{R}^d$, where $\Theta_0 \subseteq \Theta$ is a d_0-dimensional manifold contained in Θ^o.*

- θ_0 *is in the relative interior of* Θ_0.
- $\widehat{\theta}_n \xrightarrow{p} \theta_0$ *with smooth likelihood.*
- $J_1(\theta) \succ 0$.

Then

$$2(\ell_n(\widehat{\theta}_n) - \ell_n(\widehat{\theta}_0)) \implies \chi^2_{d-d_0},$$

where $\widehat{\theta}_0 = \arg\max_{\theta \in \Theta_0} \ell_n(\theta; X)$.

Here is an informal derivation.

Proof. Assume without loss of generality that $\theta_0 = 0$ and $J_1(0) = I_d$. Then $\widehat{\theta}_n \approx N_d(0, \frac{1}{n} I_d)$, and locally ($\theta \approx 0$), $\nabla^2 \ell_n(\theta) \approx -nI_d$. Then

$$\ell_n(\theta) - \ell_n(\widehat{\theta}_n) \approx \frac{n}{2} \|\theta - \widehat{\theta}_n\|^2.$$

Then

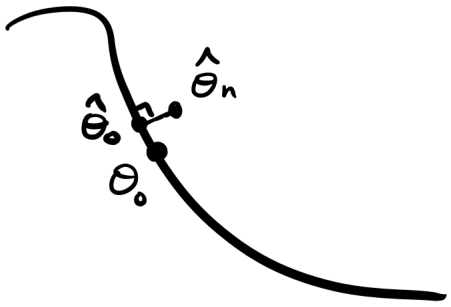
$$\widehat{\theta}_0 \approx \arg\min_{\theta \in \Theta_0} \|\theta - \widehat{\theta}_n\|^2 = \mathrm{Proj}_{\Theta_0}(\widehat{\theta}_n).$$

This means that -1 times the test statistic looks like

$$\begin{aligned} 2(\ell_n(\widehat{\theta}_0) - \ell(\widehat{\theta}_n)) &\approx -n\|\widehat{\theta}_n - \mathrm{Proj}_{\Theta_0}(\widehat{\theta}_n)\|^2 \\ &= -\|\mathrm{Proj}^{\perp}_{\Theta_0}(\underbrace{\sqrt{n}\widehat{\theta}_n}_{\approx N(0, I_d)})\|^2 \\ &\implies -\chi^2_{d-d_0}. \end{aligned}$$

□

Here is a picture when $d = 2$ and $d_0 = 1$.



The segment looks like χ^2_1.

24.3 Asymptotic relative efficiency

Suppose $\widehat{\theta}_n^{(i)}$ with $i = 1, 2$ are two estimators with $d = 1$ and

$$\sqrt{n}(\widehat{\theta}_n^{(i)} - \theta_0) \implies N(0, \sigma_i^2.$$

Definition 24.1. The **asymptotic relative efficiency (ARE)** of $\widehat{\theta}^{(2)}$ with respect to $\widehat{\theta}^{(1)}$ is σ_1^2/σ_2^2.

This has a nice interpretation of telling us that using an inefficient estimator is really like throwing away a fraction of our data set. Suppose $\sigma_1^2/\sigma_2^2 = \gamma \in (0, 1)$. Then

$$\begin{aligned}\widehat{\theta}^{(1)}_{\lfloor \gamma n \rfloor}(X_1, \ldots, X_{\lfloor \gamma n \rfloor}) &\approx N(\theta_0, \sigma_2^2/n) \\ &\overset{D}{\approx} \widehat{\theta}_n^{(2)}(X_1, \ldots, X_n).\end{aligned}$$

24.4 Pearson's χ^2 test for goodness of fit

Let $N = (N_1, \ldots, N_d) \sim \text{Multinom}(n, \pi)$, where $\pi = (\pi_1, \ldots, \pi_d)$ with $\sum_j \pi_j = 1$ and all $\pi_j > 0$. The multinomial density is

$$p_\theta(N) = \frac{n! \pi_1^{N_1} \cdots \pi_d^{N_d}}{N_1! \cdots N_d!} \mathbb{1}_{\{\sum_j N_j = n\}}.$$

We can parameterize this as a $d-1$-parameter exponential family by

$$\pi_j = \begin{cases} \frac{1}{1+\sum_{k>1} e^{\eta_k}} & j = 1, \\ \frac{e^{\eta_j}}{1+\sum_{k>1} e^{\eta_k}} & j > 1 \end{cases}$$

so that

$$\eta_j = \log(\pi_j + \pi_1).$$

We can calculate the score

$$\nabla \ell_n(\eta, N) = (N_2, \ldots, N_d) - (n\pi_2, \ldots, n\pi_d).$$

The variance of the score is

$$\begin{aligned}\text{Var}_\eta(\nabla \ell_n(\eta; N)) &= \begin{bmatrix} n\pi_2(1-\pi_2) & \cdots & -\pi_i \pi_j & \cdots \\ & \ddots & & \\ & & & n\pi_d(1-\pi_d) \end{bmatrix} \\ &= n(\text{diag}(\pi_{2-d}) - \pi_{2-d}\pi_{2-d}^\top)\end{aligned}$$

If we use the formula

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u},$$

we get

$$J_n(\eta)^{-1} = \frac{1}{2}(\text{diag}(\pi_{2,\ldots,d}^{-1}) + \pi_1^{-1}\mathbf{1}_{d_1}\mathbf{1}_{d_2}^\top).$$

After some algebra, it follows that the score test fo $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$ is

$$\begin{aligned}
\nabla\ell_n(\eta_0)J_n^{-1}(\eta_0)\nabla\ell_n(\eta_0) &= (N_{2,\ldots,d} - n\pi_{2,\ldots,d})^\top(\tfrac{1}{n}(\text{diag}(\pi_{2,\ldots,d}^{-1}) + \pi_0\mathbf{1}))(N_{2,\ldots,d} - n\pi_0) \\
&= \sum_{j>1}\frac{(N_j - n\pi_j)^2}{n\pi_j} - \frac{1}{n\pi_n}\mathbf{1}^\top(N_{2,\ldots,d} + n\pi_{2,\ldots,d})^2 \\
&= \sum_j \frac{(N_j - n\pi_j)^2}{n\pi_j}.
\end{aligned}$$

This is the test statistic for Pearson's χ^2 test.

25 Introduction to Bootstrap

25.1 Recap: Comparison of bootstrap to other kinds of inference

So far we have done:

- Exact, finite-sample inference
  - Requires special structure
  - No reliance on asymptotic approximation
  - Parametric or non-parametric (e.g. permutation tests)
- Parametric, asymptotic inference
  - Simple ideas, leading to asymptotically optimal results.
  - Only relies on regularity conditions

Today, we will study asymptotic, nonparametric inference.

25.2 Functionals and plug-in estimators

Suppose we have a nonparametric iid sample $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$. We want inference on some "parameter" $\theta(P) \subseteq \mathbb{R}^d$. More precisely, we want a functional $\theta(P)$.

Example 25.1. If the sample space $\mathcal{X} \subseteq \mathbb{R}$, we could look at

$$\theta(P) = \text{median}(P).$$

Example 25.2. If the sample space $\mathcal{X} \subseteq \mathbb{R}^d$, we could look at

$$\theta(P) = \lambda_{\max}(\text{Var}_P(X_i)).$$

Example 25.3. If we are doing linear regression, we could look at

$$\theta(P) = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_P[(Y_i - \theta^\top X_i)^2],$$

where $(X_i, Y_i) \overset{\text{iid}}{\sim} P$.

Example 25.4. More generally, we can set

$$\begin{aligned}\theta(P) &= \arg\min_{\theta \in \Theta} D_{\text{KL}}(P \mid\mid P_\theta) \\ &= \arg\max_{\theta} \mathbb{E}_P[\ell_1(\theta; X_i)].\end{aligned}$$

Note that in these cases, we may have many distribution with the same value $\theta(P)$.

Definition 25.1. The **empirical distribution** of $X_1, \dots, X_n$ is the random measure

$$\widehat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \qquad \widehat{P}_n(A) = \frac{\#\{i : X_i \in A\}}{n}.$$

Definition 25.2. The **plug-in estimator** of $\theta(P)$ is $\widehat{\theta}_n = \theta(\widehat{P}_n)$.

Example 25.5. If $\theta(P)$ is the median, $\widehat{\theta}_n$ is the sample median.

Example 25.6. If $\theta(P) = \lambda_{\max}(\mathrm{Var}_P(X_i))$, then the plug-in estimator is $\lambda_{\max}(\text{samplevariance})$.

Example 25.7. For linear regression, the plug-in estimator is the OLS etimator.

Example 25.8. For the minimizer of the KL-divergence, the plug-in estimator is the MLE for $\{P_\theta : \theta \in \Theta\}$.

25.3 Convergence of plug-in estimators

Does using the plug-in estimator work? It depends. Whether $\widehat{P}_n \xrightarrow{p} P$ depends on what distance we use. We have pointwise convergence, $\widehat{P}_n(A) \xrightarrow{p} P(A)$ for all A by the weak law of large numbers. We can consider convergence in the *total variation distance* (i.e. uniform convergence of these functions):

$$\sup_A |\widehat{P}_n(A) - P(A)| \xrightarrow{p} 0?$$

This is true if the sample space $\mathcal{X}$ is finite. However, if $\mathcal{X} = \mathbb{R}$ and P is continuous, this is not true because if $A^* = \{x_1, \dots, x_n\}$, then $P(A^*) =)$ but $\widehat{P}_n(A^*) = 1$.

If $X \subseteq \mathbb{R}$, we can look at convergence of the CDFs:

$$\sup_x |\widehat{P}_n((-\infty, x]) - P((-\infty, x])| \xrightarrow{p} 0 \qquad \forall x \in \mathbb{R}.$$

We want the functional $\theta(\cdot)$ to be continuous with respect to some topology in which $\widehat{P}_n \xrightarrow{p} P$, so $\theta(\widehat{P}_n) \xrightarrow{p} \theta(P)$ by the continuous mapping theorem.

Here is a counterexample to keep in mind, so you don't think that bootstrap always works:

Example 25.9. Let

$$\theta(P) = \begin{cases} 1 & \mathbb{P}_{X_1, X_2 \overset{\text{iid}}{\sim} P}(X_1 = X_2) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

If P is continuous, then $\theta(P) = 0$. But $\theta(\widehat{P}_n) = 1$ for all n.

25.4 Bootstrap standard errors

Suppose $\widehat{\theta}_n$ is any estimator of $\theta(P)$. We want to know the standard error $\text{s.e.}(\widehat{\theta}_n) = \sqrt{\text{Var}_P(\widehat{\theta}(X_1, \ldots, X_n))}$.

The only thing here we don't know is P, so we will plug in $\widehat{P}_n$:

$$\widehat{\text{s.e.}}(\widehat{\theta}_n) = \sqrt{\text{Var}_{\widehat{P}_n}(\theta_n^*)}.$$

Here, the star notation is just to make sure we know that $\widehat{\theta}_n^*$ is a random variable drawn from $\widehat{P}_n$. We can write

$$\text{Var}_{\widehat{P}_n}(\widehat{\theta}_n^*) = \text{Var}_{X_i^* \overset{\text{sim}}{\sim} \widehat{P}_n}(\widehat{\theta}_n(X_1^*, \ldots, X_n^*)).$$

Often, bootstrap is defined algorithmically.

How do we calculate this? We will use Monte Carlo with $\widehat{P}_n$ instead of P: For $b = 1, \ldots, B$, Sample $X_1^{*b}, \ldots, X_n^{*b} \overset{\text{iid}}{\sim} \widehat{P}_n$ (resampling n values with replacement from $X_1, \ldots, X_n$), and let $\widehat{\theta}^{*b} = \widehat{\theta}(X_1^{*b}, \ldots, X_n^{*b})$. Then let $\overline{\theta}^* = \frac{1}{B}\sum_{b=1}^B \widehat{\theta}^{*b}$, so the standard error is

$$\widehat{\text{s.e.}}(\widehat{\theta}_n) = \sqrt{\frac{1}{B}\sum_{b=1}^B (\widehat{\theta}^{*b} - \overline{\theta}^*)^2}.$$

25.5 Bootstrap Bias Estimation/Correrction

Let $\widehat{\theta}_n$ be some estimator. What is its bias?

$$\text{Bias}_P(\widehat{\theta}_n) = \mathbb{E}_P[\widehat{\theta}_n - \theta(P)].$$

The idea is to plug in $\widehat{P}_n$ for P:

$$\widehat{\text{Bias}}(\widehat{\theta}_n) = \text{Bias}_{\widehat{P}_n}(\widehat{\theta}_n^*) = \mathbb{E}_{\widehat{P}_n}[\widehat{\theta}_n^* - \theta(\widehat{P}_n)].$$

We can calculate this using Monte Carlo: Sample $X_1^{*b}, \ldots, X_n^{*b} \overset{\text{iid}}{\sim} \widehat{P}_n$, and calculate the estimator $\widehat{\theta}^{*b} = \widehat{\theta}(X^{*b})$. Then we have the average $\overline{\theta}^* = \frac{1}{B}\sum_{b=1}^B \widehat{\theta}^{*b}$, so we can calculate

$$\widehat{\text{Bias}}(\widehat{\theta}_n) = \overline{\theta}^* - \theta(\widehat{P}_n).$$

Remark 25.1. The advantage of thinking of this as a plug-in estimator instead of just defining it algorithmically is that it becomes more conceptually clear why we subtract $\theta(\widehat{P}_n)$ instead of $\theta(P)$.

We can then define the **Bias-corrected estimator**

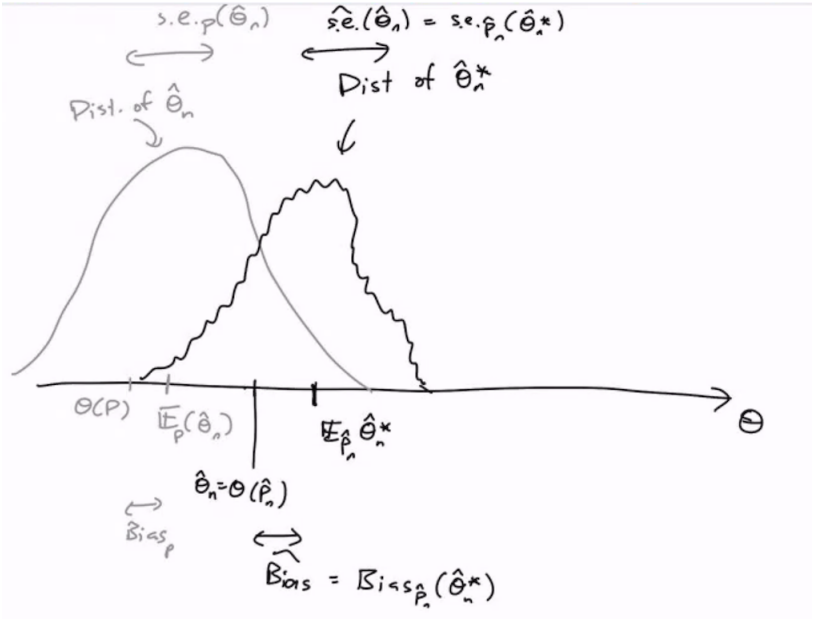
$$\widehat{\theta}_n^{\text{BC}} = \widehat{\theta}_n - \widehat{\text{Bias}}(\widehat{\theta}_n)$$

If $\theta(\widehat{P}_n) = \widehat{\theta}_n$,

$$= 2\widehat{\theta}_n - \overline{\theta}^*.$$

Remark 25.2. If we know the actual bias, it's always better to subtract it because we reduce the bias while keeping the variance the same. However, it is not always better to subtract out the estimated bias because the estimate could be wrong. In particular, the estimate of the bias might be noisy, so it might introduce some variance. Typically, $\widehat{\theta}_n^{\text{BC}}$ has a lower bias but a higher variance than $\widehat{\theta}_n$.

Here is a picture. The things that we can't see are in gray, and what we can see is in black.



Here is a table of analogies between the "real world" and "bootstrap world."

	"Real world"	"Bootstrap world"
Samplingdistribution	P	$\widehat{P}_n$
Parameter	$\theta(P)$	$\theta(\widehat{P}_n)$ (maybe $\widehat{\theta}_n$)
Dataset	$X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$	$X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \widehat{P}_n$
Estimator	$\widehat{\theta}_n(X_1, \ldots, X_n)$	$\widehat{\theta}_n^*(X_1^*, \ldots, X_n^*)$
Standard error of estimator	$\sqrt{\text{Var}_P(\widehat{\theta}_n)}$	$\sqrt{\text{Var}_{\widehat{P}_n}(\widehat{\theta}_n^*)}$
Bias of estimator	$\text{Bias}_P(\widehat{\theta}_n)$	$\text{Bias}_{\widehat{P}_n}(\widehat{\theta}_n^*)$

26 Bootstrap Confidence Intervals and Double Bootstrap

26.1 Recap: Bootstrap methods

Bootstrap is an asymptotic nonparametric method, where we use the empirical distribution as an asymptotic approximation to the true distribution. Anything we want to do with the true distribution, we substitute in the empirical distribution and call it a day.

If we have a nonparametric model $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$ with "parameter" $\theta(P)$ (not necessarily 1 to 1), then we discussed the notion of a plug-in estimator $\widehat{\theta}_n(X) = \theta(\widehat{P}_n)$, where $\widehat{P}_n$ is an estimator of P. A typical choice is the **empirical distribution** $\widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. (Note: There ar eother choices, estpeically for non-i.i.d. sampling models, e.g. time series.)

Remark 26.1. Bootstrap is often conflated with permutation tests. They are both nonparametric and involve resampling from the data, but they have very different underlying statistical logic. The permutation test is an exact, finite sample method; if you take enough permutations, you will get the exact conditional distribution of the test statistic under the null hypothesis. On the other hand, bootstrap is an approximation which only becomes accurate asymptotically.

We have seen two bootstrap algorithms so far:

- If $\widehat{\theta}_n(X)$ is any estimator we want, its standard error is

$$\text{s. e.}(\widehat{\theta}_n(X)) = \sqrt{\text{Var}_{X_i \overset{\text{iid}}{\sim} P}(\widehat{\theta}_n(X))},$$

  and the **bootstrap standard error** is

$$\widehat{\text{s. e.}}(\widehat{\theta}_n(X)) = \sqrt{\text{Var}_{X_i^* \overset{\text{iid}}{\sim} \widehat{P}_n}(\widehat{\theta}_n(X^*))}.$$

- If $\widehat{\theta}_n(X)$ is any estimator we want, its bias is

$$\text{Bias}(\widehat{\theta}_n) = \mathbb{E}_{X_i \overset{\text{iid}}{\sim} P}[\widehat{\theta}_n(X)] - \theta(P),$$

  and the **bootstrap bias estimator** is

$$\widehat{\text{Bias}}(\widehat{\theta}_n) = \mathbb{E}_{X_i^* \overset{\text{iid}}{\sim} \widehat{P}_n}[\widehat{\theta}_n(X^*)] - \theta(\widehat{P}_n).$$

  We also have the **bias corrected bootstrap estimator**

$$\widehat{\theta}_n^{\text{BC}} = \widehat{\theta}_n - \widehat{\text{Bias}}.$$

26.2 Bootstrap confidence intervals

Suppose we want a confidence interval for $\theta(P)$. Instead of inverting a hypothesis test, we can define a random variable $R_n(X, P) = \widehat{\theta}_n(X) - \theta(P)$ for any estimator $\widehat{\theta}_n$; if we know the distribution of R_n, we can construct the confidence interval using a point estimate for R_n.

Define the CDF

$$G_{n,P}(r) = \mathbb{P}_P(\widehat{\theta}_n(X) - \theta(P) \leq r).$$

The lower $\alpha/2$ quantile is

$$r_1 = G_{n,P}^{-1}(\alpha/2),$$

and the upper $\alpha/2$ quantile is

$$r_1 = G_{n,P}^{-1}(1 - \alpha/2).$$

Then

$$\begin{aligned} 1 - \alpha &= \mathbb{P}_P(r_1 \leq \widehat{\theta}_n - \theta \leq r_2) \\ &= \mathbb{P}_P(\theta \in [\widehat{\theta}_n - r_2, \widehat{\theta}_n - r_1]) \end{aligned}$$

The interval we get only depends on $G_{n,P}$.

If we don't know P, then we can use $\widehat{P}_n$ instead:

$$G_{n,\widehat{P}_n}(r) = \mathbb{P}_{X^* \overset{\text{iid}}{\sim} \widehat{P}_n}(\widehat{\theta}(X^*) - \theta(\widehat{P}_n) \leq r).$$

This depends only on the sample X. Using this CDF in the above calculation gives us the **bootstrap confidence interval**

$$C_{n,\alpha}(X) = [\widehat{\theta}_n(X) - \widehat{r}_2, \widehat{\theta}_n(X) - \widehat{r}_1],$$

where

$$\widehat{r}_1 = G_{n,\widehat{P}_n}^{-1}(\alpha/2), \qquad \widehat{r}_2 = G_{n,\widehat{P}_n}^{-1}(1 - \alpha/2).$$

Here is the procedure in practice:

1. For $b = 1, \ldots, B$, let $X_1^{*b}, \ldots, X_n^{*b} \overset{\text{iid}}{\sim} \widehat{P}_n$.
2. For $b = 1, \ldots, B$, let $R_n^{*b} = \widehat{\theta}_n(X^{*b}) - \theta(\widehat{P}_n)$.
3. Return $\widehat{G}_n(r) = \frac{1}{B} \sum_{k=1}^{B} \mathbb{1}_{\{R_n^{*b} \leq r\}}$
4. Invert this to recover $\widehat{r}_1$ and $\widehat{r}_2$.

This is not the only way to make a bootstrap confidence interval. Other examples of estimators we could use bootstrap with for confidence intervals are

- The **studentized root**
$$R_n(X, P) = \frac{\widehat{\theta}_n(X) - \theta(P)}{\widehat{\sigma}(X)}.$$
- The **relative error**
$$R_n(X, P) = \frac{\widehat{\theta}_n(X)}{\theta(P)}.$$

With the studentized root,

$$C_{n,\alpha} = [\widehat{\theta}_n - r_2\widehat{\sigma}, \widehat{\theta}_n - r_1\widehat{\theta}],$$

where we can estimate r_1, r_2 using a the plug-in estimator R_n.

Remark 26.2. Our first version of the bootstrap confidence interval works best when $G_{n,P}$ is not so sensitive to varying P.

26.3 Double bootstrap

Bootstrap is an approximation. Is it a good approximation? Suppose we have, for example, a bootstrap confidence interval

$$C_{n,\alpha} = [\widehat{\theta}_n(X) - \widehat{r}_2(X)\widehat{\sigma}(X), \widehat{\theta}_n(X) - \widehat{r}_1(X)\widehat{\sigma}(X)].$$

What is the probability

$$\mathbb{P}_{X_i \overset{\text{iid}}{\sim} P}(\widehat{\theta}_n(P) \in C_{n,\alpha}(X))?$$

We can use bootstrap to estimate this:

$$\mathbb{P}_{X_i^* \overset{\text{iid}}{\sim} \widehat{P}_n}(\widehat{\theta}_n(\widehat{P}_n) \in C_{n,\alpha}(X^*))?$$

Suppose we estimate that $C_{n,0.1}$ has $\approx 87\%$ coverage, but $C_{n,0.08}$ has $\approx 90\%$ coverage. Then we want the latter confidence interval. In particular, we are using bootstrap to calibrate the confidence level of the confidence interval.

Remark 26.3. We could do this bootstrap "tuning" of α using any confidence interval, not just one that was originally obtained through bootstrap.

Here is how we can implement this α "tuning" in practice:

1. For $a = 1, \ldots, A$, let $X_1^{*a}, \ldots, X_n^{*a} \overset{\text{iid}}{\sim} \widehat{P}_n$.
2. Calculate $C_{n,\alpha'}(X^{*a})$ for α' in some grid (try $\alpha' = 10\%, 9\%, 8\%$, etc.) using whatever method you are using to obtain a confidence interval (bootstrap or not).

   We can specify this in particular for the double bootstrap:

(a) Let $\widehat{P}_n^{*a} = \frac{1}{n}\sum_{i=1}^n \delta_{X_i^*}$.

(b) For $b = 1, \ldots, B$,

i. Let $X_1^{**a,b}, \ldots, X_n^{**a,b} \overset{\text{iid}}{\sim} \widehat{P}_n^{*a}$.

ii. Let

$$R_n^{**a,b} = \frac{\widehat{\theta}_n(X^{**a,b}) - \theta(\widehat{P}_n^{*a})}{\widehat{\sigma}(X^{**a,b})}.$$

(c) Let $G_n^{*a} = \text{ecdf}(R_n^{**a,1}, \ldots, R_n^{**a,B})$.

(d) For α' in the grid, let

$$C_{n,\alpha'}(X^{*a}) = [\widehat{\theta}_n^* - \widehat{\sigma}^{*a}\widehat{r}_2(G_n^{*a}), \widehat{\theta}_n^* - \widehat{\sigma}^{*a}\widehat{r}_1(G_n^{*a})].$$

3. For α' in this grid, let

$$\widehat{\text{Coverage}}(\alpha') = \frac{1}{A}\sum_{a=1}^{A} \mathbb{1}_{\{C_{n,\alpha'}(X^{*a}) \ni \theta(\widehat{P}_n)\}}.$$

4. Take $\widehat{\alpha} = \max\{\alpha' : \widehat{\text{Coverage}}(\alpha') \geq 1 - \alpha\}$, and return $C_{n,\widehat{\alpha}}(X)$.

Remark 26.4. This seems like circular logic, where this method will suffer from the same issues as the original bootstrap confidence interval. The heuristic idea is that the double bootstrap confidence interval may be less sensitive to changes in P than the original confidence interval.

27 Introduction to Multiple Hypothesis Testing

27.1 Correcting p-values to account for multiple hypotheses

Suppose $X \sim P_\theta \in \mathcal{P}$. We have hypotheses $H_{0,i} : \theta_{\in\theta_{0,i}}$ for $i = 1, \dots, m$. We will let

$$\mathcal{R}(X) = \{i : H_{0,i} \text{ is rejected}\}, \qquad \mathcal{H}_0(\theta) = \{i : \theta \in \Theta_{0,i}\}$$

and denote $R(X) = |\mathcal{R}|$ and $m_0 = |\mathcal{H}_0|$. The central issue is that the more hypotheses we test, the more likely we are to reject a hypothesis by chance.

Example 27.1. Let $X_i \overset{\text{iid}}{\sim} N(\theta_i, 1)$ with $H_{0,i} : \theta_i = 0$. Reject $H_{0,i}$ if $|X_i| > z_{\alpha/2}$. Then

$$\mathbb{P}_0(\text{any } H_{0,i} \text{ rejected}) = 1 - (1-\alpha)^m \xrightarrow{m\to\infty} 1.$$

Definition 27.1. The **familywise error rate (FWER)** is

$$\mathbb{P}_\theta(\text{any false rejections}) = \mathbb{P}_\theta(\mathcal{R} \cap \mathcal{H}_0 \neq \varnothing).$$

The classical view of multiple testing to say that we want

$$\sup_\theta \text{FWER}_\theta \leq \alpha.$$

Remark 27.1. Why should we make a correction for the FWER? If you conduct 10 experiments and submit your analysis to a journal, they'll require you to make a familywise error correction. But if you submit the 1 experiment each to 10 journals, then no one will hassle you.

Sometimes, when we are testing many hypotheses, we care individually about each one. But sometimes, such as if we are testing hypotheses for a large number of genes, where we should expect most of our null hypotheses to be true, we might be okay with some percentage of our hypotheses being falsely rejected.

One way we can account for multiple hypotheses is to alter our p-values. Denote the p-values by $p_1(X), p_2(X), \dots, p_m(X)$. Here are some procedures for altering the p-values:

Example 27.2 (Šidák's correction)**.** Assume $p_i \geq U[0,1]$ for $i \in \mathcal{H}_0$. If the p_i are independent and we reject if $p_i \leq \widetilde{\alpha}_m$,

$$\begin{aligned} 1 - \alpha &= \mathbb{P}\theta(\text{no false rejection}) \\ &= \mathbb{P}(p_i \geq \widetilde{\alpha}_m \, \forall i \in \mathcal{H}_0) \\ &\geq (1 - \widetilde{\alpha}_m)^{m_0} \\ &\geq (1 - \widetilde{\alpha}_m)^m. \end{aligned}$$

If we solve this, we get

$$\widetilde{\alpha}_m = 1 - (1-\alpha)^{1/m}.$$

If α is small, this is close to α/m.

Here is what we can do if we don't necessarily have independence.

Example 27.3 (Bonferroni correction)**.** Bonferroni rejects if $p_i \leq \alpha/m$. Then

$$\begin{aligned}\mathbb{P}_\theta(\text{any false rejection}) &= \mathbb{P}_\theta\left(\bigcup_{i\in\mathcal{H}_0}\{p_i \leq \alpha/m\}\right)\\ &\leq \sum_{i\in\mathcal{H}_0}\mathbb{P}_\theta(p_i \leq \alpha/m)\\ &\leq m_0 \cdot \frac{\alpha}{m}\\ &= \alpha\frac{m_0}{m}.\end{aligned}$$

The Bonferroni correction is still conservative. Here is a strictly better procedure:

Example 27.4 (Holm's procedure)**.**

Step 0: Order the p-values from small to large:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

Let $H_{(i)}$ denote the hypothesis corresponding to $p_{(i)}$.

Step 1: If $p_{(i)} \leq \alpha/m$, reject $H_{(i)}$ and continue. Otherwise, stop and accept all null hypotheses.

$\vdots$

Step k: If $p_{(k)} \leq \frac{\alpha}{m-k+1}$, reject $H_{(k)}$ and continuous. Otherwise, stop and accept all hypotheses.

$\vdots$

Step m: If $p_{(m)} \leq \alpha$, reject $H_{(m)}$.

We can analyze this procedure by

$$R^{\text{Holm}} = \max\left\{r : p_{(i)} \leq \frac{\alpha}{m+i-1}\ \forall i \leq r\right\}.$$

We reject $H_{(1)}, \ldots, H_{(R^{\text{Holm}})}$.

Proposition 27.1. *Holm's procedure controls* FWER $\leq \alpha$.

Proof. Let $p_0^* = \min\{p_i : i \in \mathcal{H}_0\}$. Then

$$\mathbb{P}(p_0^* \le \alpha/m_0) \le \alpha$$

by the union bound. We claim that if $p_0^* > \alpha/m_0$, there are no false rejections. Let $k = \#\{i : p_i \le p_0^*\} \le m - m_0 + 1$. Then

$$p_{(k)} = p_0^* > \frac{\alpha}{m_0} \ge \frac{\alpha}{m-k+1}.$$

Then Holm makes $< k$ rejections. □

27.2 The closure principle

Holm's procedure is an instance of the more general **closure principle**, which is used in a lot of modern developments in multiple testing methodology.

For $S \subseteq [m]$, let H_s be the hypothesis where all H_i are true for $i \in S$: $\theta \in \bigcap_{i \in S} \Theta_{0,i}$. Assume we have a level-α test for each subset. For example, we could reject H_s if $\min_{i \in S} p_i \le \alpha/|S|$.

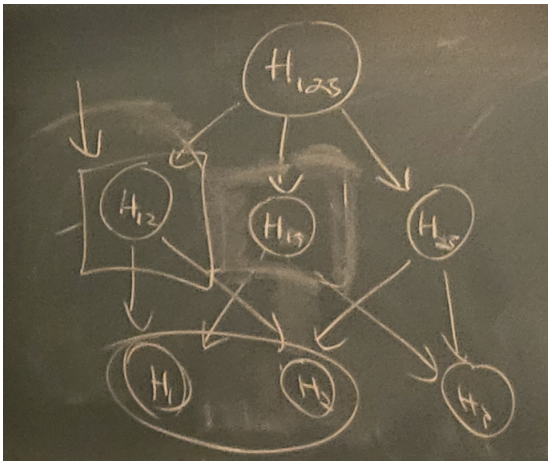
Step 1: Provisionally reject H_S if the corresponding marginal test ϕ_S rejects.

Step 2: Reject H_i if H_s is rejected for every $S \ni i$.

We can analyze the closure principle as follows:

$$\begin{aligned}\mathbb{P}(\text{any false rejections}) &\le \mathbb{P}(H_{\mathcal{H}_0} \text{ is rejected in Step 1}) \\ &\le \alpha.\end{aligned}$$

Here is a picture:



If H_1 and H_2 are the null hypotheses that are true, then they are protected as long as we don't reject the hypothesis $H_{1,2}$.

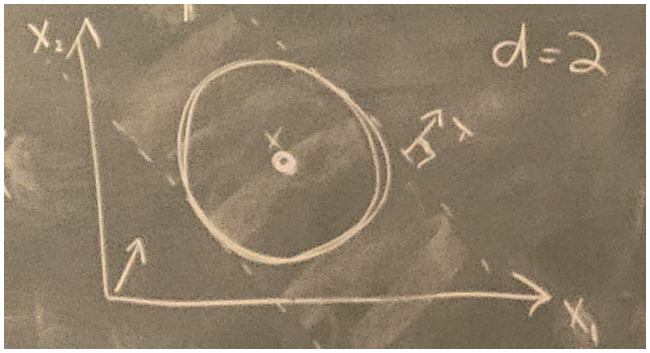
Remark 27.2. This might seem computationally inefficient, but as in our description of Holm's procedure, there can be computationally tractable ways to implementing this.

27.3 Testing with dependence

Example 27.5 (Scheffe's S-method)**.** Let $X \sim N_d(\theta, I_d)$ with $\theta \in \mathbb{R}^d$, and test $H_\lambda : \theta^\top \lambda = 0$ for $\lambda \in \mathbb{S}^{d-1}$ (this is uncountably infinitely many hypotheses). Reject H_λ if $(X^\top \lambda)^2 \geq \chi^2_d(\alpha) \approx 3 + 3\sqrt{d}$ if $\alpha = 0.05$. This controls the FWER because

$$\begin{aligned}\sup_{\lambda:\theta^\top\lambda=0} (X^\top\lambda)^2 &\leq \sup_{\lambda\in\mathbb{S}^{d-1}} ((X-\theta)^\top\lambda)^2 \\ &= \|X-\theta\|^2 \\ &\sim \chi^2_d.\end{aligned}$$

Scheffe's method is a deduction from a spherical confidence region for θ: We can use $\|X - \theta\|^2 \sim \chi^2_d$ to get a confidence region for X. We get a confidence interval for θ^λ via $X^\top\lambda \pm \sqrt{\chi^2_d(\alpha)} \approx X^\top\lambda \pm \sqrt{d}$.



28 Simultaneous Confidence Bounds for Multiple Hypothesis Testing

28.1 Recap: Multiple testing

Last time, we began discussing multiple hypothesis testing, where $X \sim P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with hypotheses $H_1, \dots, H_m$ ($H_i : \theta \in \Theta_{0,i}$). The setup includes individual p-values $p_1(X), \dots, p_m(X)$, rejection set $\mathcal{R}(X)$, and true null set $\mathcal{H}_0$.

The classical approach was to control the **familywise error rate (FWER)**,

$$\mathbb{P}_\theta(\text{any false rejections}).$$

The **Bonferroni correction**, a popular procedure, says to reject H_i if $p_i \le \alpha/m$ and works under arbitrary dependence. We also learned about a direct improvement, the **closure principle**, with an **intersection null** for $S \subseteq \{1, \dots, m\}$. Here, $H_S : H_i$ true for all $i \in S$, which is equivalent to $\theta \in \bigcap_{i \in S} \Theta_{0,i}$. Then we use some **local test** $\phi_S(X)$ which is valid for H_S. The closed testing procedure rejects H_i if $\phi_S(X) = 1$ for all $S \ni i$.

28.2 Simultaneous upper confidence bounds via closed testing

Definition 28.1. Suppose $g_1(\theta), \dots, g_m(\theta)$ are estimands. Then $C_1(X), \dots, C_m(X)$ are **simultaneous confidence bounds** if

$$\mathbb{P}_\theta(\text{any } g_i(\theta) \notin C_i(X)) \le \alpha.$$

We can use the closed testing procedure to get an upper confidence bound on the number of null indices in S, $|\mathcal{H}_0 \cap S|$.

Example 28.1. Suppose we are looking at an experiment for the brain, and each voxel i, a tiny region of the brain, corresponds to a null hypothesis H_i (about how the voxel behaves in testing vs control). If we look at a region S of the brain, the scientist gives the subset S, the software will give back $U_S(X)$.

Proposition 28.1. *If we take*

$$U_S(X) = \max_{\phi_{S_0}(x)=0} |S \cap S_0|,$$

we get simultaneous confidence bounds.

Proof.

$$\mathbb{P}_\theta(\text{any } U_S(X) \le |S \cap \mathcal{H}_0(\theta)|) \le \mathbb{P}_\theta(\phi_{\mathcal{H}_0}(X) = 1)$$

because the first event is a subset of the other. Indeed, if $\phi_{\mathcal{H}_0}(X) = 0$, then $\mathcal{H}_0$ is going to be one of the S_0 sets we take the max over. In this case,

$$U_S(X) = \max_{\phi_{S_0}(x)=0} |S \cap S_0| \ge |S \cap \mathcal{H}_0(\theta)|. \qquad \square$$

We can get simultaneous confidence bounds for the proportion of null indices by looking at $U_S(X)/|S|$. Goeman, Solari, and other coauthors have developed this procedure in a series of papers.

28.3 Simultaneous confidence intervals for the Gaussian sequence model

Example 28.2 (Gaussian sequence model)**.** Suppose have $X \sim N_d(\theta, I_d)$ with $\theta \in \mathbb{R}^d$ and we want simultaneous confidence intervals for $\theta_1, \dots, \theta_d$. Let c_α be the upper α quantile of $\max_{i=1,\dots,d} |X_i - \theta_i|$. Then if we take $C_i(X) = (X_i - c_\alpha, X_i + c_\alpha)$, these are simultaneous confidence intervals for θ_i. Why? If any $\theta_i \notin C_i(X)$, then $|X_i - \theta_i| > c_\alpha$; in particular, $\max_{i=1,\dots,d} |X_i - \theta_i| > c_\alpha$. In this case, we can show that $c_\alpha = z_{\widetilde{\alpha}_d/2}$, where $\widetilde{\alpha} - d$ is the Šidák correction.

What if we want to make pairwise comparisons? We can deduce a confidence interval for $\theta_i - \theta_j$ from the intervals for θ_i, θ_j.

$$|(\theta_i - \theta_j) - (X_i - X_j)| \le |X_i - \theta_i| + |X_j - \theta_j|,$$

so we could construct a confidence interval with $2c_\alpha$. But this is not very good. Instead, let c'_α be the upper-α quantile of $\max_{i,j} |(X_i - X_j) - (\theta_i - \theta_j)| = \max_{i,j} |Z_i - Z_j|$, where $Z = X - \theta$; this is something we can directly simulate. Then, let

$$C_{i,j}(X) = (X_i - X_j - c'_\alpha, X_i - X_j + c'_\alpha).$$

This is called **Tukey's Honestly Significant Difference procedure (HSD)**.

More generally, we may want simultaneous confidence intervals $c_\lambda(X)$ for $\lambda^\top \theta$, there $\lambda \in \mathbb{S}^{d-1}$. Let

$$\begin{aligned} c'_\alpha &= \text{upper-}\alpha \text{ quantile of } \sup_{\lambda \in \mathbb{S}^{d-1}} |\lambda^\top (X - \theta)| \\ &= \text{upper-}\alpha \text{ quantile of } \|X - \theta\|_2 \\ &\sim \chi_d(\alpha). \end{aligned}$$

28.4 Simultaneous confidence intervals in linear regression

Example 28.3 (Linear regression)**.** Suppose we have $\frac{\widehat{\beta} - \beta}{\sigma} \sim N_d(0, (X^\top X)^{-1})$ with $\beta \in \mathbb{R}^d$ and $\frac{\widehat{\sigma}^2}{\sigma^2} \sim \frac{1}{n-d}\chi^2_{n-d}$, and suppose we want simultaneous confidence intervals for $\beta_1, \dots, \beta_d$. Let c_α be the upper-α quantile of $\max_{j=1,\dots,d} |\widehat{\beta}_j - \beta_j|/\widehat{\sigma}$. We can directly simulate

$$\frac{\widehat{\eta} - \beta}{\widehat{\sigma}} = \frac{N_d(0, (X^\top X)^{-1})}{\sqrt{\frac{1}{n-d}\chi^2_{n-d}}}.$$

This has what is known as a **multivariate t distribution**. If we want simultaneous confidence intervals for β_i, then we can use

$$C_j = (\widehat{\beta} - j - \widehat{\sigma} c_\alpha, \widehat{\beta} - j + \widehat{\sigma} c_\alpha).$$

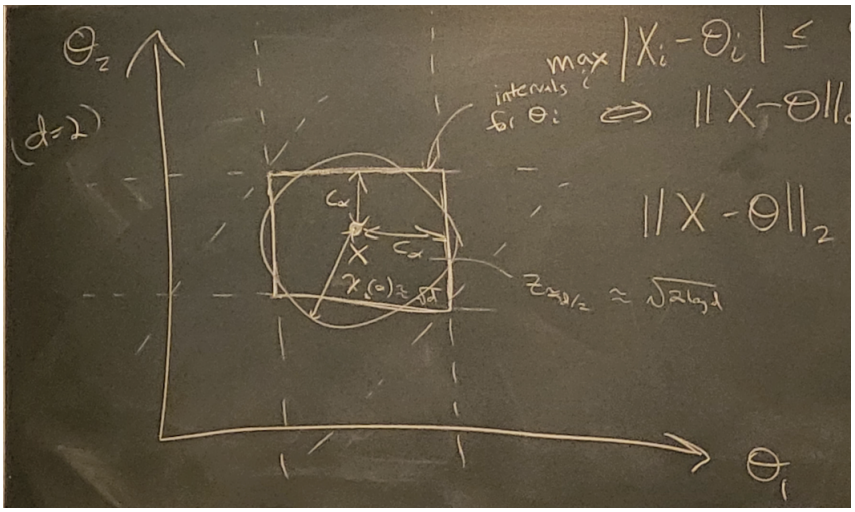
If any $\beta_j \notin C_j$, then, as before,

$$\max_{i=1,\dots,d} |\widehat{\beta}_i - \beta_i| > c_\alpha \widehat{\sigma}.$$

We can do the same procedure with Tukey's HSD, where we let $c'_\alpha = \max_{i,j} |Z_i - Z_j|$ with $Z = (\widehat{\beta} - \beta)/\widehat{\sigma}$ and use the intervals

$$C_{i,j}(X) = (X_i - X_j - c'_\alpha, X_i - X_j + c_\alpha).$$

Observe that $\max_i |X_i - \theta_i| \le c_\alpha \iff \|X_i - \theta\|_\alpha \le C_\alpha$. Alternatively, we could try to control $\|X - \theta\|_2 \le \chi_d(\alpha)$.



Our method involves constructing this rectangle and projecting it onto each of the axes. The naive method of estimating $\theta_i - \theta_j$ from before is projecting in the direction of $\theta_i - \theta_j$; so the projection we use may make a difference.

Example 28.4. Consider testing the global null $H_0 : \theta = 0$. The max test rejects if $\max_| X_i| > c_\alpha \approx \sqrt{2 \log d}$, and the χ^2 test rejects if $\|X\|_2^2 \ge \chi_d^2(\alpha) \approx d + 3\sqrt{d}$. If θ is 1-sparse (only $\theta_1 \ne 0$), then the max test needs $|\theta_1| > \sqrt{2 \log d}$, whereas the χ^2 test needs $|\theta_1 = \|\theta\|_2 \approx d^{1/4}$. If θ is dense, the χ^2 test is vastly more powerful, but if θ is sparse, then the max test is vastly more powerful.

Next time, we will discuss controlling what is known as the false discovery rate.

29 Multiple Testing via Control of the False Discovery Rate

29.1 False discovery rate

In our multiple testing setup, we have data $X \sim P_\theta$, hypotheses $H_i : \theta \in \Theta_{0,i}$ for $i = 1, \dots, m$, and p-values $p_1, \dots, p_m$. We also denote the rejection set as $\mathcal{R}(X) \subseteq \{1, \dots, m\}$and the true null set as $\mathcal{H}_0 \subseteq \{1, \dots, m\}$. We have been trying to control the **familywise error rate (FWER)**,

$$\mathbb{P}_\theta(|\mathcal{H}_0 \cap \mathcal{R}| \geq 1) \leq \alpha.$$

However, if we are making several hundred rejections, it might be okay if we only have a few false alarms.

Definition 29.1. Benjamini and Hochberg (1995)[15] defined the **false discovery proportion (FDP)**

$$\text{FDP} = \frac{V}{R \vee 1}, \qquad V = |\mathcal{H}_0 \cap \mathcal{R}|, \qquad R = |\mathcal{R}|.$$

This is the probability is that a randomly selected rejection is a false one, which we want to control. The maximum in the denominator is just so if $R = 0$, we don't divide by 0.

Definition 29.2. Benjamini and Hochberg also define the **false discovery rate (FDR)**

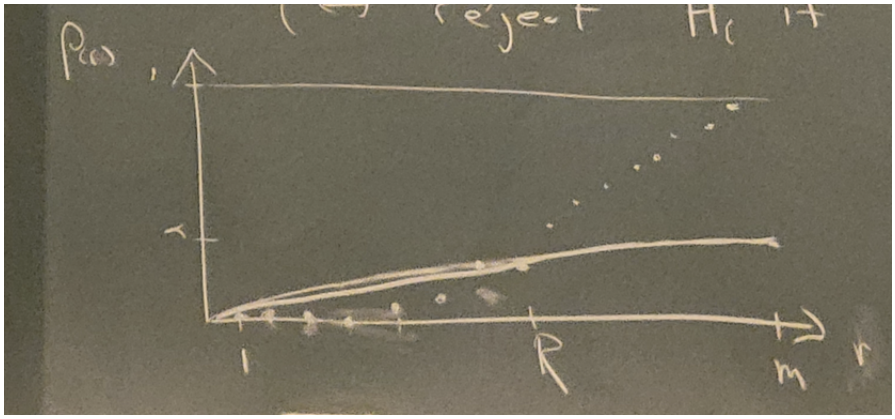
$$\text{FDR} = \mathbb{E}_\theta[\text{FDP}].$$

Benjamini and Hochberg didn't just introduce the FDR; they introduced a way to control it.

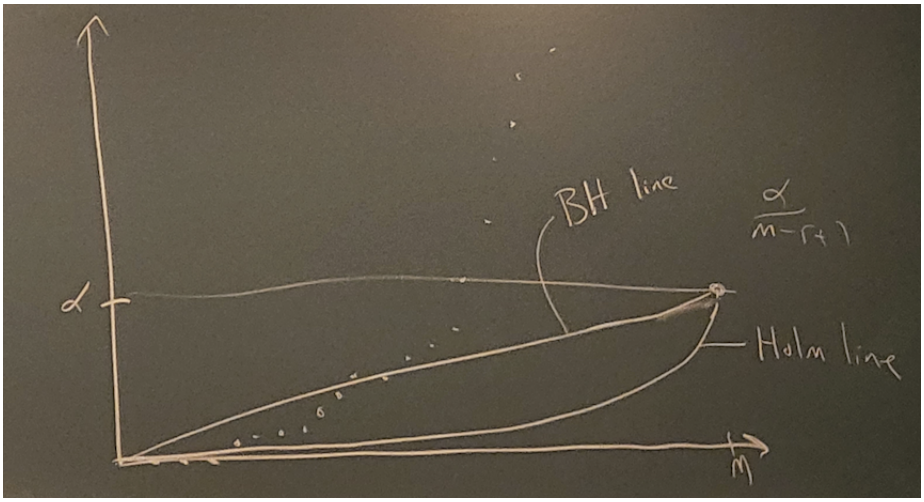
[15] They proposed this in 1988, but this radical idea of accepting some false discoveries took 7 years for any journal to accept. Professor Fithian has heard that this is the most cited paper in the entire field of statistics.

29.2 The Benjamini-Hochberg procedure

Let the p-values have order statistics $p_{(1)} \leq \cdots \leq p_{(n)}$. Then let $R^{\text{BH}} = \max\{r : p_{(r)} \leq \frac{\alpha r}{m}\}$, so the R^{BH} rejection set is $H_{(1)}, \ldots, H_{(R^{\text{BH}})}$. That is, we reject H_i if $p_i \leq \frac{\alpha R^{\text{BH}}}{m}$.



In this procedure, we reject all the hypotheses with p-values up until the last point which is below the line; even if a point is above the line, we reject it as long as there is a further point which is below the line. We can compare this to Holm's procedure, which has a lower line, since we are comparing $p_{(k)}$ to $\frac{\alpha}{m-k+1}$:



If $m = 10000$, then for Holm's procedure to make $R = 100$ rejections, $p_{(R)} \leq \frac{\alpha}{9901}$. But for BH to make 100 rejections, we need $p_{(R)} \leq \frac{\alpha 100}{10000} = \frac{\alpha}{100}$.

Remark 29.1. One issue with controlling the FDR instead of the FWER is that you can cheat. Suppose you have 5000 hypotheses you care about, but you can't make any

rejections. Then you can throw in 10000 clearly false hypotheses and be able to make a lot more rejections.

To understand this procedure, first consider rejecting H_i iff $p_i \le t$ for some fixed t. What is the false discovery proportion? Suppose $t = 5/m$. Then we expect about 5 rejections of null hypotheses. If we get 100 rejections, then we can say with more confidence that we must have had some correct rejections.

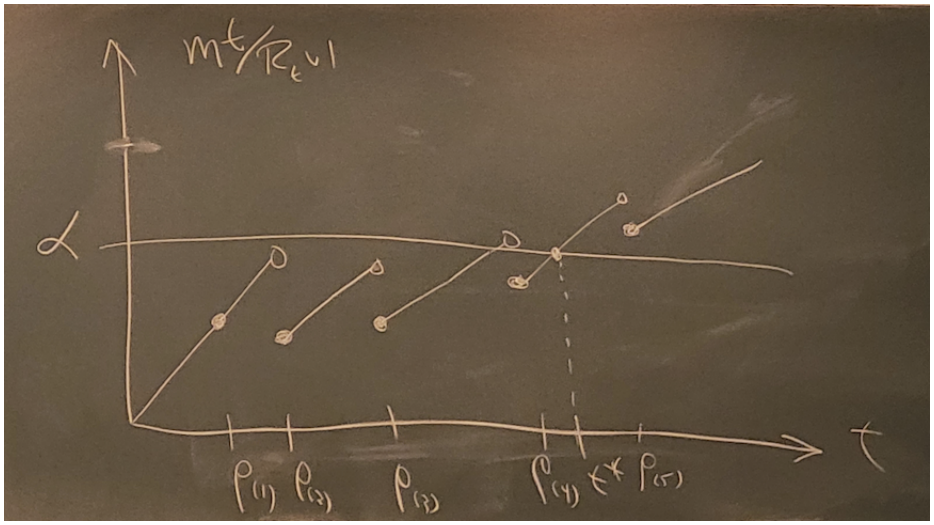
An equivalent formulation of the Benjamini-Hochberg procedure is to define

$$\widehat{\mathrm{FDP}}_t = \frac{mt}{R_t \vee 1}, \qquad R_t = \#\{i : p_i \le t\}.$$

Then we can let

$$t^*(X) = \max\{t : \widehat{\mathrm{FDP}}_t \le \alpha\}$$

and reject H_i if $p_i \le t^*$.



This is equivalent because the rejection set only depends on the order statistics of the p-values and does not actually need the information of t^*; we reject $H_{(1)}, \dots, H_{(R)}$, where

$$\begin{aligned} R &= \max\{r : \widehat{\mathrm{FDP}}_{p_{(r)}} \le \alpha\} \\ &= \max\{r : \tfrac{mp_{(r)}}{r} \le \alpha\} \\ &= \max\{r : p_{(r)} \le \tfrac{\alpha r}{m}\}. \end{aligned}$$

29.3 Finite sample control of FDR using the Benjamini-Hochberg procedure

This makes sense on controlling the FDR from an asymptotic perspective (if we let the number of samples and rejections both go to infinity), but there are many interesting

multiple testing problems where we only reject, say, 10 hypotheses. Asymptotic control is philosophically unsatisfactory here, but fortunately, we do have finite sample control with the Benjamini-Hochberg procedure.

Theorem 29.1. *The Benjamini-Hochberg procedure controls* $\mathrm{FDR} \le \alpha$.

Here is a celebrated proof due to Stoiey, Taylor, and Siegmund (2002) based on optional stopping of a martingale. Since we are looking at the last time the line crosses the α threshold, we need to index time backwards, starting from $t = 1$. This proof assumes that the p-values p_i are independent and that $p_i \sim U[0,1]$ for $i \in \mathcal{H}_0$.

Proof. Then define $V_t = \#\{i \in \mathcal{H}_0 : p_i \le t\} \le R_t$. Then we estimate

$$\mathrm{FDP}_t = \frac{V_t}{R \vee 1}$$

by

$$\widehat{\mathrm{FDP}}_t = \frac{mt}{R_t \vee 1}.$$

This gives

$$\mathrm{FDP}_t = \widehat{\mathrm{FDP}}_t \cdot \underbrace{\frac{V_t}{mt}}_{:=Q_t}.$$

This quotient Q_t is what we will apply the optional stopping argument to. This gives

$$\begin{aligned}\mathrm{FDR} &= \mathbb{E}[\mathrm{FDP}_{t^*}] \\ &= \mathbb{E}[\widehat{\mathrm{FDP}}_{t^*} \cdot Q_{t^*}] \\ &= \alpha\, \mathbb{E}[Q_{t^*}]\end{aligned}$$

Using the optional stopping theorem,

$$\begin{aligned}&= \alpha\, \mathbb{E}[Q_1] \\ &= \alpha \frac{m_0}{m}.\end{aligned}$$

It now remains to show that Q_t is a martingale and t^* is a stopping time with respect to the filtration $\mathcal{F}_t = \sigma(p_i \vee t, i = 1, \dots, m)$; we could alternatively use $\mathcal{F}_t = \sigma(V_s : s \ge t)$. Conditional on $\mathcal{F}_t$, we know V_s and $\widehat{\mathrm{FDP}}_s$ for all $s \ge t$. As a result, t^* is a stopping time ($\mathbb{1}_{\{t^* \ge s\}}$ is $\mathcal{F}_t$-measurable). To check that this is a martingale, we have for $s < t$ that

$$\mathbb{E}[V_s \mid V_t = v] = v\frac{s}{t}.$$

(More precisely, we have that $\mathbb{E}[V_s \mid \mathcal{F}_t] = V_t \frac{s}{t}$.) So

$$\mathbb{E}\left[\frac{V_s}{ms} \mid \frac{V_t}{mt} = q\right] = \frac{1}{ms} \cdot (qmt) \cdot \frac{s}{t} = q.$$

(More precisely, we have $\mathbb{E}[Q_s \mid \mathcal{F}_t] = Q_t$.) □

Here is another proof:

Proof. Define $B_i = \mathbb{1}_{\{H_i \text{ rejected}\}}$. The we can decompose

$$\frac{V}{R \vee 1} = \sum_{i \in \mathcal{H}_0} \frac{V_i}{R \vee 1}.$$

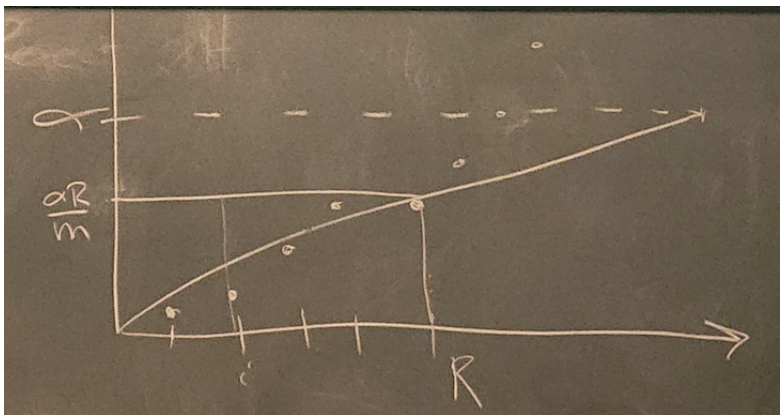
By the linearity of expectation, we can say that

$$\text{FDR} = \sum_{i \in \mathcal{H}_0} \underbrace{\mathbb{E}\left[\frac{\mathbb{1}_{\{i \text{ rejected}\}}}{R \vee 1}\right]}_{\text{want to show } \leq \alpha/m}.$$

Assume that $p_1, \ldots, p_m$ are independent. Then condition on p_{-i}. We will be in good shape if we can show that

$$\mathbb{E}\left[\frac{\mathbb{1}_{\{i \text{ rejected}\}}}{R \vee 1} \mid p_{-i}\right] \leq \frac{\alpha}{m}.$$

Rewrite the indicator as $\mathbb{1}_{\{p_i \leq \alpha R/m\}}$. We would like to pull out R, but R is not a deterministic function of p_{-i}. The key observation (which is generalizable) is that if p_i were already being rejected and we send it to 0, then it is still rejected:



Define $R^{(i)} = R(p_{-i}, 0)$. We claim that on the event $\{p_i \leq \frac{\alpha R}{m}\}$, $R^{(i)} = R$. So we can look at

$$\mathbb{E}\left[\frac{\mathbb{1}_{\{p_i \leq \frac{\alpha R}{m}\}}}{R \vee 1} \mid p_{-i}\right] = \mathbb{E}\left[\frac{\mathbb{1}_{\{p_i \leq \frac{\alpha R^{(i)}}{m}\}}}{R^{(i)}} \mid p_{-i}\right]$$

$$
\begin{aligned}
&= \frac{1}{R^{(i)}} \mathbb{P}\left(p_i \leq \frac{\alpha R^{(i)}}{m} \mid p_{-i} \right) \\
&= \frac{1}{R^{(i)}} \frac{\alpha R^{(i)}}{m} \\
&= \frac{\alpha}{m}.
\end{aligned}
$$

□

Professor Fithian and a collaborator were able to generalize this proof to non-independent p_i by conditioning on something other than p_{-i}.