Theoretical Statistics
STAT 210A

Instructor: Song Mei

Kelvin Lee

UC Berkeley

Contents

*

1 Measure Theory

1.1 Basics

1.1.1 Measures

Definition 1.1.1 (Measure). Given a set $\mathcal{X}$, a **measure** μ maps subsets $A \subseteq \mathcal{X}$ to nonnegative numbers $\mu(A) \in [0, \infty]$.

Example 1.1.2. Let $\mathcal{X}$ be a countable set ($\mathcal{X} = \mathbb{Z}$ for example). Then the **counting measure** ia

$$\mu(A) = \#A = \# \text{ of points in } A.$$

Example 1.1.3. Consider $\mathcal{X} = \mathbb{R}^n$. The **Lebesgue measure** is

$$\lambda(A) = \int \cdots \int_A dx_1 \cdots dx_n = \text{Vol}(A).$$

Example 1.1.4 (Standard Gaussian Distribution).

$$\mathbb{P}(A) = \mathbb{P}(Z \in A) = \int \cdots \int_A \phi(x) dx_1 \cdots dx_n$$

where $Z \sim \mathcal{N}_n(0, I_n)$ and $\phi(x) = \frac{e^{-\frac{1}{2}\sum x_i^2}}{\sqrt{(2\pi)^n}}$.

Because of pathological sets, $\lambda(A)$ is only defined for some subsets $A \subseteq \mathbb{R}^n$. In other words, it is often impossible to assign measures to all subsets A of $\mathcal{X}$. This leads to the idea of a σ-field(σ-algebra).

Definition 1.1.5 (σ-field). A **σ-field** is a collection of sets on which μ is defined, satisfying certain closure properties.

In general, the domain of a measure μ is a collection of subsets $\mathcal{F} \subseteq 2^{\mathcal{X}}$ (power set), and $\mathcal{F}$ must be a σ-field.

Example 1.1.6. Let $\mathcal{X}$ be a countable set. Then $\mathcal{F} = 2^{\mathcal{X}}$. (Counting measure is defined for all subsets).

Example 1.1.7. Let $\mathcal{X} = \mathbb{R}^n$, then $\mathcal{F}$ is the **Borel σ-field** $\mathcal{B}$, the smallest σ-field containing all open rectangles $(a_1, b_1) \times \cdots \times (a_n, b_n)$ where $a_i < b_i \quad \forall i$.

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

Measures let us define integrals that put weight $\mu(A)$ on $A \subseteq \mathcal{X}$.

Define

$$\int \mathbf{1}\{x \in A\} d\mu(x) = \mu(A) \qquad \text{(indicator)}$$

extend to other functions by linearity and limits:

$$\int \left(\sum c_i \mathbf{1}\{x \in A_i\}\right) d\mu(x) = \sum c_i \mu(A_i) \qquad \text{(simple function)}$$

$$\int f(x) d\mu(x) \qquad \text{(approx. by simple functions)}$$

Example 1.1.8.

- *Counting*: $\int f d\# = \sum_{x \in \mathcal{X}} f(x)$.
- *Lebesgue*: $\int f d\lambda = \int \cdots \int f(x) dx_1 \cdots dx_n$.
- *Gaussian*: $\int f dP = \int \cdots \int f(x)\phi(x) dx_1 \cdots dx_n = \mathbb{E}[f(Z)]$.

1.1.2 Measurable Functions

Definition 1.1.9 (Measurable functions)**.** Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and $f : \mathcal{X} \to \mathbb{R}$. Then f is **measurable** if $\forall B \in \mathcal{B}$ (Borel σ-algebra of $\mathbb{R}$)

$$f^{-1}(B) = \{x \in \mathcal{X} : f(x) \in B\} \in \mathcal{A}.$$

1.1.3 Densities

The λ and $\mathbb{P}$ above are closely related and we now want to make this precise.

Given $(\mathcal{X}, \mathcal{F})$, two measures $\mathbb{P}, \mu$, we say that $\mathbb{P}$ is **absolutely continuous** with respect to μ if $\mathbb{P}(A) = 0$ whenever $\mu(A) = 0$.

Notation: $P \ll \mu$ or we say μ *dominates* $\mathbb{P}$.

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

$$\mathbb{P}(A) = \int_A p(x) d\mu(x)$$

$$\int f(x) d\mathbb{P}(x) = \int f(x) p(x) d\mu(x).$$

The density function is also defined as

$$p(x) = \frac{d\mathbb{P}}{d\mu}(x),$$

known as *Radon-Nikodyan derivative.*

Remark. It is useful to turn $\int f d\mathbb{P}$ into $\int f p d\mu$ if we know how to calculate integrals $d\mu$.

If $\mathbb{P}$ is a probability measure, μ is a Lebesgue measure, then $p(x)$ is called **probaility density function** (pdf). If μ is a counting measure, then $p(x)$ is called the **probability mass function** (pmf).

1.1.4 Probability Space and Random Variables

Typically, we set up a problem with multiple random variables having various relationships to one another. It is convenient to think of them as functions of an abstract outcome ω.

Definition 1.1.10. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a **probability space**. $\omega \in \Omega$ is called **outcome**. $A \in \mathcal{F}$ is called **event**. $\mathbb{P}(A)$ is called **probability of** A.

Definition 1.1.11. A **random variable** is a function $X : \Omega \to \mathcal{X}$. We say $\mathcal{X}$ has distribution Q, denoted as $X \sim Q$ if $\mathbb{P}(X \in B) = \mathbb{P}(\{\omega : X(\omega) \in B\}) = Q(B)$.

More generally, we could write events involving many random variables $X(\omega), Y(\omega), Z(\omega)$:

$$\mathbb{P}(X \geq Y + Z) = \mathbb{P}(\{\omega : X(\omega) \geq Y(\omega) + Z(\omega)\})$$

Definition 1.1.12. The **expectation** is an integral with respect to $\mathbb{P}$:

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

To do real calculations, we must eventaully boil $\mathbb{P}$ or $\mathbb{E}$ down to concrete integrals/sums/etc. If $\mathbb{P}(A) = 1$, we say that A occurs **almost surely**.

2 Risk and Estimation

2.1 Estimation

Definition 2.1.1 (Statistical Model). A **statistical model** is a family of candidate probability distributions

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

for some random variable $X \sim P_\theta$. $X \in \mathcal{X}$ is called **data** (observed). θ is the **parameter** (unobserved).

The goal of estimation is to observe $X \sim P_\theta$ and guess value of some **estimand** $g(\theta)$.

Example 2.1.2. Suppose we flip a biased coin n times. Let $\theta \in [0, 1]$ be the probability of getting a head and let X be the number of heads after n flips. Then $X \sim \text{Binom}(n, \theta)$ with $p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, which is the density with respect to counting measure on $\mathcal{X} = \{0, \dots, n\}$. A natural estimator would be $\delta_0(X) = \frac{X}{n}$.

Definition 2.1.3 (Statistical Procedure). A **statistical procedure** δ is a mapping from $\mathcal{X}$ to the decision space $\mathcal{D}$:

$$\delta(x) \in \mathcal{D}.$$

In estimation tasks, $\mathcal{D} = \Theta$, δ is called (statistical) estimator. In hypothesis testing tasks, $\mathcal{D} = \{0, 1\}$, δ is called (hypothesis) test.

Question. Is the natural estimator a good estimator? Is there a better one?

Definition 2.1.4 (Statistic). A **statistic** is any function $T(x)$ of data X (not of both X and θ).

Definition 2.1.5 (Estimator). An **estimator** $\delta(X)$ of $g(\theta)$ is a statistic which is intended to guess $g(\theta)$.

2.2 Loss and Risk

Definition 2.2.1 (Loss function). A **loss function** a mapping $L : \Theta \times \mathcal{D} \to \mathbb{R}_{\geq 0}$ such that $L(\theta, \delta)$ measures how bad an estimate is.

Example 2.2.2. One common loss function is the *squared-error loss* $L(\theta, \delta) = (\delta - \theta)^2$.

Typical properties of loss functions:

(i) $L(\theta, \delta) \geq 0 \; \forall \theta, \delta$

(ii) $L(\theta, g(\theta)) = 0 \; \forall \theta$.

Definition 2.2.3 (Risk function)**.** The **risk function** is the expected loss (risk) as a function of θ for an estimator $\delta(\cdot)$.

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

Remark. The subscript θ under $\mathbb{E}$ tells us which parameter value is in effect, NOT what randomness to integrate over.

Example 2.2.4 (Coin flip cont'd)**.** $X \sim \text{Bin}(\theta)$. We have $\delta_0(X) = \frac{X}{n}$. Then $\mathbb{E}_\theta\left[\frac{X}{n}\right] = \theta$ (unbiased). Then

$$R(\theta, \delta) = MSE(\theta; \delta_0) = \text{Var}_\theta\left(\frac{X}{n}\right) = \frac{\theta(1-\theta)}{n}.$$

Other choices:

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

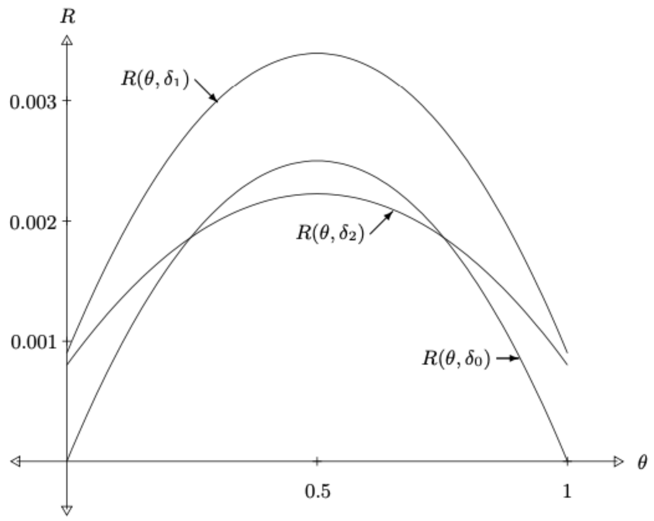

Figure 2.1: Risks for $\delta_0, \delta_1, \delta_2$.

δ_1 is bad but δ_0, δ_2 are ambiguous. We need some notion to determine which estimator to pick.

Definition 2.2.5 (Inadmissible)**.** An estimator δ is **inadmissible** (bad) if $\exists \delta^*$ with

(i) $R(\theta; \delta^*) \leq R(\theta; \delta) \; \forall \theta \in \Theta$, and

(ii) $R(\theta; \delta^*) < R(\theta; \delta)$ for some $\theta \in \Theta$

From the previous example, we see that δ_1 is inadmissible.

Back to the issue regarding the ambiguity of the comparison between two estimators. Here are some strategies to resolve that ambiguity:

1. Summarize $R(\theta)$ by a scalar (collapse the risk function)

   (i) Average-case risk: minimize the **Bayes risk function**

   $$R_B(\Lambda;\delta) \coloneqq \int_\Theta R(\theta;\delta)d\Lambda(\theta)$$

   with some measure Λ. The resulting δ is called the **Bayes estimator**, and Λ is the **prior**.

   (ii) Worst-case risk: minimize the **maximum risk function**

   $$R_M(\Theta;\delta) \coloneqq \sup_{\theta\in\Theta} R(\theta,\delta).$$

   over $\delta : \mathcal{X} \to \mathbb{R}$. This is a **minimax estimator**, which is closely related to Bayes.

   Remark. We do not consider the best-case risk because the constant estimator would always ignore the data, which makes it a bad estimator.

2. Constrain the choice of estimator.

   (i) (Classical choice) Only consider unbiased δ. $\mathbb{E}_\theta[\delta(X)] = g(\theta)\,\forall\theta\in\Theta$.

2.2.1 Unbiasedness

Definition 2.2.6 (Unbiased estimator)**.** An **unbiased estimator** is a $\delta : \mathcal{X} \to \Theta$ such that

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

Remark. There sometimes exists an unbiased estimator that is uniformly optimal among all unbiased estimators.

3 Exponential Families

3.1 s-parameter Exponential Family

Definition 3.1.1 (s-parameter exponential family)**.** An **s-parameter exponential family** is a family of probability densities $\mathcal{P} = \{p_\eta : \eta \in \Xi\}$ with respect to a common measure μ on $\mathcal{X}$ of the form

$$p_\eta(x) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x), \quad x \in \mathcal{X}$$

where

- $T : \mathcal{X} \to \mathbb{R}^s$ is a **sufficient statistic**
- $h : \mathcal{X} \to \mathbb{R}$ is a **carrier/base density**
- $\eta \in \Xi \subseteq \mathbb{R}^s$ is a **natural parameter**
- $A : \mathbb{R}^s \to \mathbb{R}$ is a **cumulant generating function** (CGF)

Note that the CGF $A(\eta)$ is totally determined by h, T since we must have $\int_{\mathcal{X}} p_\eta d\mu = 1 \, \forall \eta$. Hence,

$$A(\eta) = \log \int_X \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x) d\mu(x).$$

p_η is only normalizable iff $A(\eta) < \infty$.

Definition 3.1.2 (Natural parameter space)**.** The **natural parameter space** is the set of all allowable (normalizable) η:

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

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

Remark. Note that Ξ_1 is determined by T, h, η. We could take $\Xi \subset \Xi_1$ if we wanted. $A(\eta)$ is convex $\implies$ Ξ_1 is convex (from homework).

Interpretation of Exponential Families:

- Start with a base density p_0.
- Apply an **exponential tilt**:
  1. multiply by $e^{\eta^\top T}$

2. renormalize (if possible)

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

Sometimes it is more convenient to use a different parameterization:

$$p_\theta(x) = \exp\left\{\eta(\theta)^\top T(x) - B(\theta)\right\} h(x), \quad \text{where } B(\theta) = A(\eta(\theta)).$$

Example 3.1.3 (Gaussian Family)**.** Consider $X \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}, \sigma^2 > 0$. $\theta = (\mu, \sigma^2)$.

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{ \underbrace{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2}_{\eta(\theta)^\top T(x)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)}_{B(\theta)} \right\} \cdot \underbrace{1}_{h(x)} \end{aligned}$$

This is a two-parameter exponential family with $\eta(\theta) = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}$ and $T(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$, $h(x) = 1$, and $B(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$.

Remark. $h(x)$ can also be $1/\sqrt{2\pi}$ if we did not include the factor $1/(\sqrt{2\pi}\sigma)$ into the exponentiation. In that case, $B(\theta) = \mu^2/(2\sigma^2) + \log\sigma$.

In canonical form,

$$\begin{aligned} p_\eta(x) &= \exp\left\{\eta^\top \begin{pmatrix} x \\ x^2 \end{pmatrix} - A(\eta)\right\} \\ A(\eta) &= -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right) \end{aligned}$$

Example 3.1.4. Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then their joint density is

$$\begin{aligned} p_\theta(x_1, \ldots, x_n) &= \prod_{i=1}^n p_\theta^{(1)}(x_i) \\ &= \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i-\mu)^2}{(2\sigma^2)}\right\}\right] \\ &= \exp\left\{\sum_{i=1}^n \left[\frac{\mu}{\sigma^2}x_i - \frac{1}{2\sigma^2}x_i - \left(\frac{\mu}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)\right]\right\} \\ &= \exp\left\{\frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 - n\left(\frac{\mu}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)\right\}. \end{aligned}$$

These densities also form a two-parameter exponential family with $\eta(\theta) = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}$, $T(x) = \begin{pmatrix} \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2 \end{pmatrix}$, $B(\theta) = nB^{(1)}(\theta)$.

Generally, suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\eta^{(1)}(x) = \exp\{\eta^\top T(x) - A(\eta)\}h(x)$. Then

$$X \sim \prod_{i=1}^n p_\eta^{(1)}(x_i) = \prod_{i=1}^n \exp\left\{\eta^\top T(x_i) - A(\eta)\right\} h(x_i)$$
$$= \exp\left\{\eta^\top \underbrace{\sum_{i=1}^n T(x_i)}_{\text{sufficient statistic}} - \underbrace{nA(\eta)}_{\text{cgf}}\right\} \underbrace{\prod_{i=1}^n h(x_i)}_{\text{carrier density}}.$$

Suppose $X \in \mathcal{X}$ follows an exponential family. Then $T(X)$ also follows a closely related exponential family. $T(X) \in \mathcal{T} \subseteq \mathbb{R}^s$. If $X \sim p_\eta(x) = \exp\left\{\eta^\top T(x) - A(\eta)\right\}$ (WLOG assume $h(x) = 1$) with respect to μ.

For a set $B \subseteq \mathcal{T}$, define $\nu(B) = \mu(T^{-1}(B))$. Then $T(X) \sim q_\eta(t) = \exp\left\{\eta^\top t - A(\eta)\right\}$ with respect to ν.

Discrete case:

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

Thus, $T \sim \exp\{\eta^\top t - A(\eta)\}$ with respect to ν.

Example 3.1.5 (Binomial)**.** Let $X \sim \text{Binom}(n, \theta)$. n is fixed and so the parameter is $\theta \in [0, 1]$. Then

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

with natural parameter $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$ called the *log odds ratio.*

Example 3.1.6 (Poisson)**.** Let $X \sim \text{Poisson}(\lambda)$. Then

$$\begin{aligned}
p_\lambda(x) &= \frac{\lambda^x e^{-\lambda}}{x!} \qquad i = 0, 1, 2, \dots \\
&= \exp\{x \log(\lambda) - \lambda\} \frac{1}{x!}
\end{aligned}$$

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

Example 3.1.7 (Non-exponential families)**.** Uniform distribution and Cauchy distribution are non-exponential families.

3.2 Differential Identities

Write

$$e^{A(\eta)} = \int \exp\left\{\eta^\top T(x)\right\} h(x) d\mu(x).$$

Theorem 3.2.1.
For $f : \mathcal{X} \to \mathbb{R}$, let

$$\Xi_f = \left\{\eta \in \mathbb{R}^s : \int |f| \exp\{\eta^\top T\} h d\mu < \infty\right\}.$$

Then the function

$$g(\eta) = \int f \exp\{\eta^\top T\} h d\mu$$

has continuous partial derivatives of all orders for $\mu \in \Xi_f^\circ$ (interior of Ξ_f), which can be computed by differentiating under the integral.

Differentiating $e^{A(\eta)}$ once:

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

Thus, we have

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

Differentiating it again:

$$\begin{aligned}
\frac{\partial^2}{\partial \eta_j \partial \eta_k} e^{A(\eta)} &= \int \frac{\partial^2}{\partial \eta_j \partial \eta_k} e^{\eta^\top T(x)} h(x) d\mu(x) \\
\cancel{e^{A(\eta)}} \left(\frac{\partial^2 A}{\partial_i \partial \eta_k} + \underbrace{\frac{\partial A}{\partial \eta_j}}_{\mathbb{E}_\eta[T_j]} \cdot \underbrace{\frac{\partial A}{\partial \eta_k}}_{\mathbb{E}_\eta[T_k]} \right) &= \underbrace{\int T_j T_k e^{\eta^\top T - A(\eta)} h d\mu}_{\mathbb{E}_\eta[T_j T_k]} \\
\frac{\partial^2 A}{\partial \eta_j \partial \eta_k}(\eta) &= \mathbb{E}_\eta[T_j T_k] - \mathbb{E}_\eta[T_j]\mathbb{E}_\eta[T_k] \\
&= \mathrm{Cov}_\eta(T_j, T_k).
\end{aligned}$$

Thus, we have

$$\boxed{\nabla^2 A(\eta) = \mathrm{Var}_\eta(T(X))}.$$

Remark.

- Log-likelihood function of observing $\{x_i\}_{i\in[n]} \sim P_\eta$ gives

$$\begin{aligned}\log \prod_{i=1}^{n} P_\eta(x_i) &= \sum_{i=1}^{n} \log P_\eta(x_i) \\ &= \sum_{i=1}^{n} \left(\eta' T(x_i) - A(\eta) + \log h(x_i)\right).\end{aligned}$$

  MLE satisfies gradient of log-likelihood being equal to zero:

$$\frac{1}{n}\nabla_\eta \left(\log \prod_{i=1}^{n} P_\eta(x_i)\right) = \frac{1}{n}\sum_{i=1}^{n} T(x_i) - \mathbb{E}_\eta[T(X)] = 0.$$

  MLE $\hat{\eta}_{\mathrm{ML}}$ satisfies **expectation matching equation**:

$$\mathbb{E}_{\hat{\eta}_{\mathrm{ML}}}[T(X)],$$

  i.e. the expectation of T over $P_{\hat{\eta}_{\mathrm{ML}}}$ is equal to the expectation of T over samples.

- The Hessian of log-likelihood gives

$$\begin{aligned}\nabla_\eta^2 \log \prod_{i=1}^{n} P_\eta(x_i) &= -n\nabla_\eta^2 A(\eta) \\ &= -n\mathrm{Var}_\eta(T(X)) \\ &\preccurlyeq 0,\end{aligned}$$

  which implies that the log-likelihood function is concave. Hence, gradient ascent converges to MLE $\hat{\eta}_{\mathrm{ML}}$.

3.2.1 Moment Generating Functions

Differentiating $e^{A(\eta)}$ repeatedly, we get

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

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

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

4 Sufficiency

4.1 Sufficient Statistics

Motivation: Not all data is useful for a decision problem. Sometimes there is redundancy in data and we want to remove the redundancy.

Example 4.1.1. Coin flipping. Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$. Then the vector

$$X \sim \prod_i \theta^{x_i}(1-\theta)^{1-x_i} \text{ on } \{0,1\}^n,$$

and

$$T(X) = \sum_i X_i \sim \text{ Binom}(n, \theta)$$

with density

$$\binom{n}{t}\theta^t(1-\theta)^{n-t} \text{ on } \{0, 1, \dots, n\}.$$

The map $(X_1, \dots, X_n) \mapsto T(X)$ is throwing away data because we do not know if heads come first or tails come first. How do we justify this? Why does it not matter?

Definition 4.1.2 (Sufficient Statistics). Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a statistical model for data X. Then $T(X)$ is **sufficient** for $\mathcal{P}$ if $P_\theta(X \mid T)$ does not depend on θ, i.e. $\forall \theta, \theta' \in \Theta$,

$$P_\theta(X \mid T) = P_{\theta'}(X \mid T).$$

Example 4.1.3 (Cont'd).

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

So given $T(x) = t$, X is uniform on all sequences with $\sum_i x_i = t$.

4.1.1 Interpretations of Sufficiency

Recall we only care about X in the first place because it is (indirectly) informative about θ. Sufficiency means only $T(X)$ is informative We can think of the data as being generated in two stages:

1. Generate $T(X) \sim P_\theta(T(X))$ (Pick a slice of X, depends on θ)

2. Generate $X \sim P(X \mid T)$ (Generate within the slice, does not depend on θ)

So we only care about the first step if $T(X)$ is sufficient.

Theorem 4.1.4.
Suppose $X \sim P_\theta$ and T is sufficient. Let $L(\theta, \delta)$ be an arbitrary loss function. Then for any estimator δ, there is a (possibly randomized) estimator $\tilde{\delta}$ that only depends on $T(X)$ such that

$$R(\theta; \delta) = R(\theta; \tilde{\delta}), \quad \forall \theta \in \Theta.$$

Remark. $\tilde{\delta}(T(x))$ is defined as follows:

- Sample $\tilde{x} \mid T(x) \sim P_(x \mid T(x))$

- Let $\tilde{\delta}(T(x)) = \delta(\tilde{x})$.

Claim. Under $x \sim P_\theta(x)$ and $\tilde{x} \mid T(x) \sim P(x \mid T(x))$, we have

$$\tilde{\delta}(T(x)) \overset{D}{=} \delta(x).$$

Proof. $x \mid T(x) \overset{D}{=} \tilde{x} \mid T(x) \implies x \overset{D}{=} \tilde{x} \implies \delta(x) \overset{D}{=} \delta(\tilde{x}) = \tilde{\delta}(T(x)).$ □

4.2 Sufficiency Principle

Theorem 4.2.1 (Sufficiency Principle)**.**
If $T(X)$ is sufficient for $\mathcal{P}$, then any statistical procedure should depend on X only through $T(X)$.

In fact, we could throw away X and generate a new $\widetilde{X} \sim P(X \mid T)$ and it would be just as good as X, i.e. $\widetilde{X} \overset{D}{=} X$ implies $\delta(\widetilde{X}) \overset{D}{=} \delta(X)$.

4.3 Factorization Theorem

There is a very convenient way to verify sufficiency of a statistic based only on the density:

Theorem 4.3.1 (Factorization Theorem).
Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of distributions dominated by μ. T is sufficient for $\mathcal{P}$ iff there exist functions $g_\theta, h \geq 0$ such that the densities p_θ for the family satisfy

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

for almost every x under μ.

"Proof". $(\Longleftarrow)$:

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

which does not depend on θ and so T is sufficient.

$(\Longrightarrow)$: Take

$$\begin{aligned} g_\theta(t) &= \mathbb{P}_\theta(T(x) = t) = \int_{T(x)=t} p_\theta(x) d\mu(x) \\ h(x) &= \mathbb{P}_{\theta_0}(X = x \mid T(X) = t) = \frac{p_{\theta_0}(x)}{\int_{T(z)=t} p_{\theta_0}(z) d\mu(z)}. \end{aligned}$$

Then

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

□

Example 4.3.2 (Exponential Families).

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

Example 4.3.3. $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta^{(1)}$ for any model $\mathcal{P}^{(1)} = \{P_\theta^{(1)} : \theta \in \Theta\}$ on $\mathcal{X} \subseteq \mathbb{R}$. P_θ is invariant to permuting $X = (X_1, \ldots, X_n)$. Thus, the **order statistics** $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ (where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$) with $X_{(k)}$ being the kth smallest value (counting repeats) are sufficient.

Remark. $(X_1, \ldots, X_n) \mapsto (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ loses information about the original ordering.

For more general $\mathcal{X}$, we say the **empirical distribution** $\hat{P}_n(\cdot) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(\cdot)$ is sufficient where $\delta_{x_i}(A) = \mathbf{1}\{x_i \in A\}$.

4.4 Minimal Sufficiency

Consider $\overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$. Then

$$\begin{aligned} T(X) &= \sum_{i=1}^{n} X_i \\ \overline{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ S(X) &= (X_{(1)}, \dots, X_{(n)}) \\ X &= (X_1, \dots, X_n) \end{aligned}$$

are all sufficient statistics.

Proposition 4.4.1. If $T(X)$ is sufficient and $T(X) = f(S(X))$, then $S(X)$ is sufficient.

Proof. By Factorization Theorem,

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

Our new $g'_\theta = (g_\theta \circ f)$ and $T'(x) = S(x)$. Thus, $S(x)$ is sufficient. □

Definition 4.4.2 (Minimal Sufficient). $T(X)$ is **minimal sufficient** if $T(X)$ is sufficient and for any other sufficient statistic $S(X)$, $T(X) = f(S(X))$ for some f (a.s. in $\mathcal{P}$).

If $\mathcal{P}$ has densities $p_\theta(x)$ with respect to μ, then the log-likelihood function (denoted by $\ell(\theta; X)$) is the log-density function reframed as a *random* function of θ.

If $T(X)$ is sufficient, then

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

Theorem 4.4.3.
Assume $\mathcal{P}$ has densities p_θ and $T(X)$ is sufficient for $\mathcal{P}$. If

$$\frac{P_\theta(x)}{P_{\theta'}(x)} = \frac{P_\theta(y)}{P_{\theta'}(y)}, \quad \forall \theta, \theta' \in \Theta \implies T(x) = T(y),$$

i.e.

$$\left\{(x, y) : \frac{P_\theta(x)}{P_{\theta'}(x)} = \frac{P_\theta(y)}{P_{\theta'}(y)}, \quad \forall \theta, \theta' \in \Theta\right\} \subseteq \{(x, y) : T(x) = T(y)\},$$

then $T(X)$ is minimal sufficient.

Proof. For the sake of contradiction, suppose S is sufficient and there is no f such that $f(S(X)) = T(X)$. Then there exist x, y with $S(x) = S(y)$, $T(x) \neq T(y)$. $L(\theta; x) = g_\theta(S(x))h(x) \propto_\theta g_\theta(S(y))h(y) = L(\theta; y)$, a contradiction since we must have $T(x) = T(y)$ but we don't. □

Remark. The key takeaway is that if a sufficient statistic determine the likelihood shape in a one-to-one way, then we can recover it from the likelihood shape and so it's minimal sufficient.

Example 4.4.4. $p_\theta(x) = e^{\eta(\theta)^\top T(x) - B(\theta)} h(x)$. Is $T(x)$ minimal?

Answer. Assume $L(\theta; x) \propto_\theta L(\theta; y)$. We want to show that $T(x) = T(y)$. For any θ, we have

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

To get rid of $a(x, y)$, we can use arbitrary θ_1, θ_2 to get

$$(\eta(\theta_1) - \eta(\theta_2))^\top (T(x) - T(y)) = 0.$$

This implies that $\eta(\theta_1) - \eta(\theta_2)$ and $T(x) - T(y)$ are orthogonal to each other, which is equivalent to saying that

$$T(x) - T(y) \perp \text{span} \{\eta(\theta_1) - \eta(\theta_2) : \theta_1, \theta_2 \in \Theta\}.$$

Unfortunately, we are not able to conclude that $T(x)$ is minimal. However, if

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

then $T(x) - T(y) = 0$ as desired.

4.5 Completeness

Definition 4.5.1 (Complete). $T(X)$ is **complete** for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if $\mathbb{E}_\theta[f(T(X))] = 0 \; \forall \theta$ implies

$$f(T(X)) \stackrel{a.s.}{=} 0.$$

Remark. There is no non-trivial unbiased estimator of 0 based on $T(X)$.

Suppose there are two unbiased estimator of θ, $\delta(T(X)), \tilde{\delta}(T(X))$ where $T(X)$ is complete. Since

$$\mathbb{E}_\theta[\delta(T(X))] = \theta = \mathbb{E}_\theta[\tilde{\delta}(T(X))] \quad \forall \theta \in \Theta,$$

we have

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

and so we must have

$$\delta(T(X)) = \tilde{\delta}(T(X)) \quad \forall T.$$

Example 4.5.2. Consider $\{x_i\}_{i \in [n]} \stackrel{i.i.d.}{\sim} p_\theta(x) = \frac{1}{2} \exp\{-|x - \theta|\}$ and $S(x) = (x_{(1)}, x_{(2)}, \ldots, x_{(3)})$ is minimal sufficient. Let's check if S is complete. Let

$$f(S(x)) = \overline{x} - \text{median}(x).$$

Then

$$\mathbb{E}_\theta[f(S(X))] = \theta - \theta = 0 \quad \forall \theta \in \Theta,$$

which implies S is not complete.

Remark. Thus, a minimal sufficient statistic may not necessarily be complete.

Example 4.5.3. Consider $\{x_i\}_{i\in[n]} \overset{i.i.d.}{\sim} U[0,\theta]$ where $\theta \in (0,\infty)$ and $T(x) = \max_i x_i$ is minimal sufficient. Then for $t \in (0,\theta)$

$$P_\theta(T(x) \le t) = \left(\frac{t}{\theta}\right)^n.$$

Then for $t \in (0,\theta)$,

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

Suppose $\mathbb{E}_\theta[f(T)] = 0$ for any $\theta > 0$. Then we have

$$\begin{aligned} \frac{n}{\theta^n}\int_0^\theta f(t)t^{n-1}dt &= 0 \\ \int_0^\theta f(t)t^{n-1}dt &= 0 \\ f(t)t^{n-1} &= 0 \\ f(t) &= 0. \end{aligned}$$

Thus, $T(X)$ is complete.

Definition 4.5.4 (Full-rank)**.** Let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ be an exponential family of densities (with respect to μ),

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

Assume WLOG that there does not exist $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^s$ with $\beta^\top T(X) \overset{a.s.}{=} \alpha$. If

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

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

Theorem 4.5.5.
If $\mathcal{P}$ is full rank, then $T(X)$ is complete sufficient.

Proof. Proof in Lehmann & Romano, Theorem 4.3.1. □

Theorem 4.5.6.
If $T(X)$ is complete and sufficient for $\mathcal{P}$, then $T(X)$ is minimal sufficient.

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

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

does not depend on θ by sufficiency of S. Define $g(t) = t - \mu(f(t))$. Then

$$\begin{aligned} \mathbb{E}_\theta[g(T(X))] &= \mathbb{E}_\theta[T(X)] - \mathbb{E}_\theta[\mu(S(X))] \\ &= \mathbb{E}_\theta[T(X)] - \mathbb{E}_\theta[\mathbb{E}[T \mid S]] \\ &= 0. \end{aligned}$$

Thus, $g(T(X)) \overset{a.s.}{=} 0$ by completeness of T. Hence,

$$T(X) \overset{a.s.}{=} \mu(S(X)).$$

Therefore, T is minimal sufficient. □

4.6 Ancillarity

Definition 4.6.1 (Ancillary)**.** A statistic $V(X)$ is **ancillary** for $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ if its distribution does not depend on θ (in other words, $V(X)$ carries no information about θ):

$$P_\theta(V(x) = v) = P_{\theta'}(V(x) = v) \quad \forall \theta, \theta' \in \Theta.$$

Remark. For sufficient statistic, $X \mid T(X)$ carries no information about θ.

Theorem 4.6.2 (Basu's Theorem)**.**
If $T(x)$ is complete sufficient and $V(x)$ is ancillary for $\mathcal{P}$, then

$$V(X) \perp\!\!\!\perp T(X)$$

under P_θ for all $\theta \in \Theta$.

Proof. We want to show that for any $\theta \in \Theta$, and any set A and B,

$$P_\theta(V(x) \in A \mid T(x) = t) = P_\theta(V(x) \in A).$$

□

4.7 Convex Loss and the Rao-Blackwell Theorem

Recall the following property of sufficient statistics:

Suppose T is sufficient for $\mathcal{P}$. For any estimator $\delta(x)$ and any loss function $L(\theta, \delta)$, there exists a possibly random estimator $\tilde{\delta}(T(x))$ such that $\forall \theta \in \Theta$, then

$$\begin{aligned} R(\theta, \tilde{\delta}) &= \mathbb{E}_{\theta,\tilde{\delta}}[L(\theta, \tilde{\delta}(T(x))] \\ &\leq \mathbb{E}_\theta[L(\theta, \delta(x))] \\ &= R(\theta, \delta). \end{aligned}$$

4.7.1 Two steps construction of a random $\tilde{\delta}(t)$

1. Sample $\tilde{x} s \mathbb{P}(x \mid T(x) = t)$ (This step is random).
2. Define $\tilde{(t)} = \delta(x)$.

Remark. SO we only need to consider estimators that are based on sufficient statistics $T(x)$. But we may possibly need to consider *random* estimators based on $T(x)$.

Definition 4.7.1 (Convex). A function $f : X \to \mathbb{R}$ is **convex** if $\forall x, y \in X$, $c \in (0, 1)$

$$f(cx + (1-c)y) \leq cf(x) + (1-c)f(y).$$

Theorem 4.7.2 (Jensen's Inequality).
If f is convex, then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$. If f is strictly convex, then the inequality is strict unless X is almost surely constant.

If $L(\theta, \delta)$ is convex in δ (e.g. $L(\theta, \delta) = (\theta - \delta)^2$), then we would want to add extra noise to δ. Let $\tilde{\delta}(x) = \delta(x) + \epsilon$ where $\mathbb{E}[\epsilon \mid X] = 0$. Then

$$\begin{aligned} \mathbb{E}_\theta[L(\theta, \delta(X)] &= \mathbb{E}_\theta[L(\theta, \mathbb{E}[\tilde{\delta}(X)])] \\ &\leq \mathbb{E}_{\theta,\tilde{\delta}}[L(\theta, \tilde{\delta}(X)]. \end{aligned}$$

Theorem 4.7.3 (Rao-Blackwell Theorem).
Assume $T(X)$ is sufficient for $\mathcal{P}$ and $\delta(X)$ is an estimator. Let

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

If $L(\theta, \delta)$ is convex in δ, then

$$R(\theta, \bar{\delta}(T(\cdot)) \leq R(\theta, \delta).$$

If L is strictly convex, then the inequality is strict unless $\bar{\delta}(T(X)) = \delta(X)$ almost surely.

Proof. Recall $\tilde{\delta}(t) = \delta(x')$ with $x' \sim P(x \mid T(x) = t)$ then $\overline{\delta}(t) = \mathbb{E}[\tilde{\delta}(t)]$.
By Jensen's inequality, we have

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

□

Remark. When the loss function is convex, it is sufficient to consider deterministic estimators of sufficient statistics.

5 Unbiased Estimation

5.1 Uniform minimum risk unbiased estimator

Recall that the problem of using $R(\theta, \delta)$ for comparing estimators is that they are not always comparable.

Definition 5.1.1 (U-estimable). $g(\theta)$ is **U-estimable** if there exists an unbiased estimator.

Example 5.1.2. Consider $X \sim \text{Ber}(\theta)$. Then $\mathbb{E}_\theta[\delta(X)] = \theta\delta(1) + (1-\theta)\delta(0)$. So θ^2 is not U-estimable. Any function of which is U-estimable must be of form $a\theta + b$.

Definition 5.1.3 (UMRUE). We say that $\delta(X)$ is **uniformly minimum risk unbiased estimator** if $\delta(X)$ is unbiased, and for any other unbiased $\tilde{\delta}(X)$, we have

$$R(\theta, \delta) \leq R(\theta, \tilde{\delta}, \quad \forall \theta \in \Theta.$$

Definition 5.1.4 (UMVUE). Take $R(\theta, \delta) = \mathbb{E}_\theta[(\theta - \delta(X))^2] = \text{Var}_\theta[\delta(X)]$.

Theorem 5.1.5 (Lehmann-Scheffe Theorem).
Suppose $T(X)$ is complete sufficient and $g(\theta)$ is U-estimable with an unbiased estimator $\delta_0(X)$. Define

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

Then

1. δ is the only function of T that is unbiased for $g(\theta)$.
2. $\delta(T(X))$ is an UMRUE under any convex L.
3. $\delta(T(X))$ is the unique UMRUE under any strict convex L.
4. $\delta(T(X))$ is the unique UMVUE.

Proof.

1.

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

By completeness, δ is the only function of T that is unbiased of $g(\theta)$.

2. Consider $\tilde{\delta}(X)$ unbaised for $g(\theta)$. Define

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

Then $\mathbb{E}[\overline{\delta}(T(X))] = g(\theta) = \mathbb{E}[\delta(T(X))]$. By completeness of T, $\overline{\delta}(T) = \delta(t)$ almost everywhere. Thus, by Rao-Blackwell/Jensen's

$$R(\theta, \delta) = R(\theta, \overline{\delta}) \leq R(\theta, \tilde{\delta}),$$

so δ is UMRUE.

3. If $L(\theta, \delta)$ is strictly convex in δ, then $R(\theta, \overline{\delta}) < R(\theta, \tilde{\delta})$ unless $\tilde{\delta}(X) = \overline{\delta}(T(X)) = \delta(T(X))$. Thus, $\delta(T(X))$ is the unique UMRUE.

□

Remark. Lehmann-Scheffe gives us a way to find UMRUE:

1. Find an unbiased estimator δ_0 for $g(\theta)$.

2. Find the complete sufficient statistics $T(X)$.

3. Rao-Blackwellize it,

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

Remark. UMRUE could be inadmissible.

6 Equivariant Estimation

6.1 Location Invariance

Definition 6.1.1 (Location invariant model)**.** A family of densities $\mathcal{P} = \{f_\theta^{\bigotimes n} : \theta \in \Theta\}$ is **location invariant** if

$$f_\theta(x) = f_{\theta+c}(x+c) \quad \forall \theta, x, c.$$

Example 6.1.2. Gaussian, exponential, Laplace, Cauchy are some common location invariant models.

Definition 6.1.3 (G-invariant model)**.** G group *acts* on a set Θ if $g \in G$, $\theta \in \Theta$, $g \star \theta \in \Theta$. $\mathcal{P}$ is **G-invariant** if

$$f_\theta(x) = f_{g\star\theta}(g \star x).$$

Definition 6.1.4 (Invariant loss function)**.** A loss function L for location θ is called **invariant** if

$$L(\theta, d) = L(\theta - c, d - c), \quad \forall \theta, c, d.$$

Defining $\rho(x) = L(0, x)$, L is **invariant** if

$$L(\theta, d) = \rho(d - \theta), \quad \forall \theta, d.$$

6.2 Location Equivariance

Definition 6.2.1 (Location equivariant estimator)**.** An estimator δ is **location equivariant** if

$$\delta(x_1, \ldots, x_n) + c = \delta(x_1 + c, \ldots, x_n + c).$$

Example 6.2.2. $\delta(x) = \overline{x}$ is location equivariant.

The following theorem simplifies the search of optimal location equivariant estimator.

Theorem 6.2.3.
If δ is location equivariant, and $\mathcal{P}$ and L are location invariant, then the bias, risk, variance of δ has no θ dependence.

Proof. Proof for risk:

$$\begin{aligned} R(\theta,\delta) &= \mathbb{E}_\theta[L(\theta,\delta(X))] \\ &= \mathbb{E}_\theta[L(\theta,\delta(X-\theta)+\theta)] && \text{(equivariance of } \delta) \\ &= \mathbb{E}_\theta[L(0,\delta(X-\theta))] && \text{(invariance of } L) \\ &= \mathbb{E}_0[L(0,\delta(x))] && \text{(invariance of } \mathcal{P}\text{: } X \sim P_\theta \implies X-\theta \sim P_0). \end{aligned}$$

We have thus eliminated the θ dependence. □

Remark. With location equivariance structure, we are able to compare any two equivariant risk functions $R(\theta,\delta)$ since $R(\theta,\delta)$ doesn't depend on θ anymore.

6.3 Minimum Risk Equivariant Estimator

Definition 6.3.1 (MRE). δ is a **minimum risk equivariant estimator (MREE)** if it is location equivariant and $\forall \tilde{\delta}$ location equivariant,

$$R(\theta,\delta) \le R(\theta,\tilde{\delta}), \quad \forall \theta \in \Theta.$$

Lemma 6.3.2. Let δ_0 be location equivariant. Then δ is location equivariant if and only if there exists v such that

$$\delta(x_1,\ldots,x_n) = \delta_0(x_1,\ldots,x_n) + v(y_1,\ldots,y_{n-1}),$$

where $y_i = x_i - x_n$.

Proof. We want to show that

$$\delta(x_1+c,\ldots,x_n+c) = \delta(x_1,\ldots,x_n) + c.$$

$$\begin{aligned} \delta(x_1+c,\ldots,x_n+c) &= \delta_0(x_1+c,\ldots,x_n+c) + v(x_1+c-(x_n+c),\ldots,x_{n-1}+c-(x_n+c)) \\ &= \delta_0(x_1+c,\ldots,x_n+c) + v(y_1,\ldots,y_{n-1}) \\ &= \delta_0(x_1,\ldots,x_n) + c + v(y_1,\ldots,y_{n-1}) \\ &= \delta(x_1,\ldots,x_n) + c. \end{aligned}$$

For the other direction, define

$$v(y_1,\ldots,y_{n-1}) = \delta(y_1,\ldots,y_{n-1},0) - \delta_0(y_1,\ldots,y_{n-1},0).$$

□

Theorem 6.3.3.
If $(\mathcal{P},L)$ is location invariant and δ_0 is location equivariant with finite risk. If

$$v^*(y) = \arg\min_v \mathbb{E}_0[\rho(\delta(X)-v(y)) \mid Y=y], \quad \forall y.$$

Then the MREE is $\delta(X) = \delta_0(X) - v^*(Y)$, where $Y = (X_1-X_n, X_2-X_n, \cdots, X_{n-1}-X_n)$.

Corollay 6.3.4. If ρ is convex and non-monotone. Then an MREE exists. If ρ is strictly convex, then MREE is unique.

Example 6.3.5. Consider $(X_i)_{i\in[n]} \overset{i.i.d.}{\sim} \text{Exp}(\theta, b)$ with θ unknown, b known. Then

$$p(x_i;\theta) = \frac{1}{b}\exp\left\{-\frac{(x_i-\theta}{b}\right\}\mathbf{1}\{x_i > \theta\}.$$

Let $\delta_0(X_1,\dots,X_n) = \min_{i\in[n]}\{X_i\}$ (complete sufficient) . Then δ_0 is location equivariant. Note that $Y = (X_1 - X_n, \dots, X_{n-1} - X_n$ is ancillary. Then by Basu's theorem δ_0 and Y are independent and so

$$\min_v \mathbb{E}_0[\rho(\delta(X) - v) \mid Y = y] = \min_v \mathbb{E}_0[\rho(\delta(X) - v)].$$

So

$$v^*(y) = \arg\min_v \mathbb{E}_0[\rho(\delta(X) - v) \mid Y = y] = \arg\min_v \mathbb{E}_0[\rho(\delta(X) - v)] = v_*.$$

Case 1: $\rho(t) = t^2$. Then

$$v_* = \arg\min_v \mathbb{E}_0\left[(\min_{i\in[n]}(X_i) - v)^2\right] = \mathbb{E}_0\left[\min_{i\in[n]}(X_i)\right] = \frac{b}{n}.$$

Then the MREE is

$$\delta(X) = \delta_0(X) - v_* = \min_{i\in[n]} X_i - \frac{b}{n}.$$

Case 2: $\rho(t) = |t|$. MREE is

$$\delta(X) = \min_{i\in[n]} X_i - \frac{b}{n}\log 2.$$

6.4 The Pitman estimator of location

Take $\rho(t) = t^2$, there is an explicit form of MREE. Recall that if δ_0 is location equivariant, then

$$v^*(y) = \arg\min_v \mathbb{E}_0[(\delta_0(X) - v)^2 \mid Y = y] = \mathbb{E}_0[\delta_0(X) \mid Y = y].$$

Then the MREE is

$$\delta^*(X) = \delta_0(X) - \mathbb{E}_0[\delta_0(X) \mid Y = y].$$

Theorem 6.4.1.
Consider $\mathcal{P}$ invariant, $\rho(t) = t^2$, then MREE δ^* gives

$$\delta^*(X_1,\dots,X_n) = \frac{\int_{-\infty}^{\infty} u\prod_{i=1}^n f(X_i - u)du}{\int_{-\infty}^{\infty}\prod_{i=1}^n f(X_i - u)du}$$

where $f(X) = P_0(X)$ is the density of X when $\theta = 0$.

Theorem 6.4.2.
Under square loss,

(i) If $\delta(X)$ is location equivariant with bias b, then $\delta(X) - b$ is unbiased and location equivariant with smaller risk than $\delta(X)$.

(ii) The unique MREE is unbiased.

(iii) If UMVUE exists and is location equivariant, then it is also MREE.

Remark. MREE can be inadmissible.