STAT 5380
Advanced Mathematical Statistics I
(Lecture Notes – Spring 2024)[1]

Alex Trindade
Department of Mathematics & Statistics
Texas Tech University

[1]Based primarily on TPE (Lehmann & Casella, 1998) and TSH (Lehmann & Romano, 2005).

ABSTRACT. According to Efron & Hastie (2016):

> "Optimality theories – statements of best possible results – are marks of maturity in applied mathematics. Classical statistics achieved two such theories: for unbiased or asymptotically unbiased estimation, and for hypothesis testing."

This course covers the details of this **optimality** story. You already know the basics from your introductory mathematical statistics course: it is possible to find *optimal* (uniformly smallest variance) estimators if one restricts attention to the class of unbiased estimators. For finite samples these are the UMVUEs, and for infinte samples the UMVUEs are MLEs (maximum likelihood estimators). But there are other forms of optimality (if we do not restrict ourselves to the unbiased class), leading to:

- **MREs:** minimum risk equivariant estimators, minimize risk under the principle of *equivariance* (invariance under location-scale transforms);
- **Bayes estimators:** minimize the *Bayes risk*, an integrated risk weighted by the prior; and
- **minimax estimators** minimize the *maximum* risk.

The equivalent *optimal* hypothesis tests are uniformly most powerful (UMP) and UMP unbiased (UMPU).

An important sobering message is that this *optimal* inference is infeasible in most practical applications, and so one usually settles for the sub-optimal and "automatic" MLE, and accompanying Likelihood Ratio, Wald, or Score test. All the details left out in earlier courses (probability measure-theoretic and otherwise) are covered here.

The term *classical statistics* refers to the 20th century dominant theme whereby the number of parameters to estimate is smaller than the available sample size ($p \ll n$). The 21st century *bigdata* era has reversed this situation, but at the moment there is no comparable optimality theory when $p \gg n$.... The course will bring you to this frontier and provide you with the essential tools and knowledge to go beyond it.

Contents

CHAPTER 1

Preliminaries

1.1. Conditional Expectation

Definition. Let $(\mathcal{X}, \mathcal{A}, P)$ be a probability space. If $X \in L^1(\mathcal{A}, P)$ and $\mathcal{G}$ is a sub-σ-field of $\mathcal{A}$, then $E(X|\mathcal{G})$ is a random variable such that

(i) $E(X|\mathcal{G}) \in \mathcal{G}$ (i.e. is $\mathcal{G}$ measurable)
(ii) $E(I_G X) = E(I_G E(X|\mathcal{G})),\ \forall G \in \mathcal{G}$

Construction. For $X \geq 0$, $\mu(G) = E(I_G X)$ is a measure on $\mathcal{G}$ and $P(G) = 0 \Rightarrow \mu(G) = 0$, so by the Radon-Nikodym theorem there exists a $\mathcal{G}$-measurable function $E(X|\mathcal{G})$ such that $\mu(G) = \int_G E(X|\mathcal{G})dP$, i.e.(ii) is satisfied. This shows the existence of $E(X^+|\mathcal{G})$ and $E(X^-|\mathcal{G})$. Then we define $E(X|\mathcal{G}) = E(X^+|\mathcal{G}) - E(X^-|\mathcal{G})$.

Remark 1.1.1. (ii) generalizes to $E(YX) = E(YE(X|\mathcal{G}))\ \ \forall Y \in \mathcal{G}$ such that $E|YX| < \infty$.

The *conditional probability of A given $\mathcal{G}$* is defined for all $A \in \mathcal{A}$ as $P(A|\mathcal{G}) = E(I_A|\mathcal{G})$.

Remark 1.1.2. If $X \in L^2(\mathcal{A}, P)$, then $E(X|\mathcal{G})$ is the orthogonal projection in $L^2(\mathcal{A}, P)$ of X onto the closed linear subspace $L^2(\mathcal{G}, P)$ of $L^2(\mathcal{A}, P)$ since

(i) $E(X|\mathcal{G}) \in L^2(\mathcal{G}, P)$ and
(ii) $E(Y(X - E(X|\mathcal{G}))) = 0,\ \forall Y \in L^2(\mathcal{G}, P)$.

Conditioning on a Statistic

Let X be a r.v. defined on $(\mathcal{X}, \mathcal{A}, P)$ with $E|X| < \infty$ and let T be a measurable function (not necessarily real-valued) from $(\mathcal{X}, \mathcal{A})$ into $(\mathcal{T}, \mathcal{F})$.

$$(\mathcal{X}, \mathcal{A}, P) \xrightarrow{T} (\mathcal{T}, \mathcal{F}, P^T)$$

Such a T is called a *statistic* (and is not necessarily real-valued). The *σ-field of subsets of $\mathcal{X}$ induced by T* is

$$\sigma(T) = \{T^{-1}S,\ S \in \mathcal{F}\} = T^{-1}\mathcal{F}$$

Definition 1.1.3. $E(X|T) \equiv E(X|\sigma(T))$

Recall that a real-valued function f on $\mathcal{X}$ is $\sigma(T)$ measurable $\Leftrightarrow f = g \circ T$ for some $\mathcal{F}$-measurable g on $\mathcal{T}$, i.e. $f(x) = g(T(x))$ as shown below.

$$\mathcal{X} \xrightarrow{T} \mathcal{T} \xrightarrow{g} \mathbb{R}$$

.

This implies that $E(X|T)$ is expressible as $E(X|T) = h(T)$ for some function $h \in \mathcal{F}$ which is unique a.e. P^T.

$$\mathcal{X} \xrightarrow{T} \mathcal{T} \xrightarrow{h} \mathbb{R}$$

DEFINITION 1.1.4. $E(X|t) \equiv h(t)$

EXAMPLE 1.1.5. Suppose (X, T) has probability density $p(x, t)$ w.r.t. Lebesgue measure on $\mathbb{R}^2$ and $E|X| < \infty$. Then $E(X|\sigma(T)) = h(T)$ where $h(t) = E(X|T = t) = \frac{\int xp(x,t)\ dx}{\int p(x,t)\ dx} I_{p^T(t)>0}(t)$, $a.s.\ P^T$.

PROOF

(i) R.S. is Borel measurable in t (by Fubini)
(ii) $G \in \sigma(T) \Rightarrow G = T^{-1}F$ for some $F \in \mathcal{F} \Rightarrow I_G = I_F(T)$

$$\begin{aligned} \therefore E(I_G E(X|\sigma(T))) &= E(I_G X) = \int I_G X\ dP \\ &= \int\int xI_F(t)p(x,t)\ dxdt = \int I_F(t)h(t)p^T(t)\ dt \\ &= E[I_F(T)h(T)] = E[I_G h(T)] \end{aligned}$$

□

Properties of Conditional Expectation

If T is a statistic, X is the identity function on $\mathcal{X}$ and f_n, f, g are integrable, then

(i) $E[af(X) + bg(X)|T] = aE[f(X)|T] + bE[g(X)|T] \quad a.s.$
(ii) $a \le f(X) \le b \quad a.s. \Rightarrow a \le E[f(X)|T] \le b \quad a.s.$
(iii) $|f_n| \le g, f_n(x) \to f(x) \quad a.s. \Rightarrow E[f_n(X)|T] \to E[f(X)|T] \quad a.s.$
(iv) $E[E(f(X)|T)] = Ef(X)$.
(v) If $E|h(T)f(X)| < \infty$, then $E[h(T)f(X)|T] = h(T)E[f(X)|T] \quad a.s.$
(vi) If $\mathcal{G}_1$ and $\mathcal{G}_2$ are sub-σ-fields of $\mathcal{G}$ with $\mathcal{G}_1 \subset \mathcal{G}_2$, then $E[E(X|\mathcal{G}_1)|\mathcal{G}_2] = E(X|\mathcal{G}_2)$.

1.2. Sufficiency

Set up

X: random observable quantity (the identity function on $(\mathcal{X}, \mathcal{A}, \mathcal{P})$)
$\mathcal{X}$: sample space, the set of possible values of X
$\mathcal{A}$: σ-algebra of subsets of $\mathcal{X}$

$\mathcal{P}$: $\{P_\theta,\, \theta \in \Omega\}$ is a family of probability measures on $\mathcal{A}$ (distributions of X)
T: $\mathcal{X} \to \mathcal{T}$ is an $\mathcal{A}/\mathcal{F}$ measurable function and $T(X)$ is called a statistic.

$$\text{probability space } (\mathcal{X}, \mathcal{A}, \mathcal{P}) \xrightarrow{X} \text{ sample space } (\mathcal{X}, \mathcal{A}, \mathcal{P}) \xrightarrow{T} (\mathcal{T}, \mathcal{F}, \mathcal{P}^{\mathcal{T}})$$

We adopt this notation because sometimes we wish to talk about $T(X(\cdot))$ the *random variable* and sometimes about $T(X(x)) = T(x)$, a *particular element* of $\mathcal{T}$. We shall also use the notation $P(A|T(x))$ for $P(A|T = T(x))$ and $P(A|T)$ for the random variable $P(A|T(\cdot))$ on $\mathcal{X}$.

DEFINITION 1.2.1. The statistic T is *sufficient* for θ(or $\mathcal{P}$) iff the conditional distribution of X given $T = t$ is independent of θ for all t, i.e. there exists an $\mathcal{F}$ measurable $P(A|T = \cdot)$ such that $P(A|T = t) = P_\theta(A|T = t)$ a.s. P_θ^T for all $A \in \mathcal{A}$ and all $\theta \in \Omega$.

EXAMPLE 1.2.2.

$$X = (X_1, \ldots, X_n) \text{ iid with pdf } f_\theta(x) \; w.r.t. \, dx$$
$$\mathcal{P} = P_\theta(dx_1, \ldots, dx_n) = f_\theta(x_1) \cdots f_\theta(x_n)\, dx_1 \cdots dx_n$$
$$T(X) = (X_{(1)}, \ldots, X_{(n)}) \quad \text{where } X_{(i)} \text{ is the } i^{th} \text{ order statistic.}$$

The probability mass function of X given $T = t$ is

$$p_\theta^{X|T=t}(x|t) = \frac{\delta_{t_1}(x_{(1)}) \cdots \delta_{t_n}(x_{(n)})}{n!}$$

i.e. it assigns point mass $\frac{1}{n!}$ to each x such that $x_{(1)} = t_1, \cdots, x_{(n)} = t_n$. This is independent of θ, indicating that T contains all the information about θ contained in the sample.

The Factorization Criterion

DEFINITION 1.2.3. A family of probability measure's $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ is *equivalent* to a p.m. λ if

$$\lambda(A) = 0 \iff P_\theta(A) = 0 \quad \forall \theta \in \Omega.$$

We also say that $\mathcal{P}$ is *dominated* by a σ-finite measure μ on $(\mathcal{X}, \mathcal{A})$ if

$$P_\theta \ll \mu \text{ for all } \theta \in \Omega.$$

It is clear that equivalence to λ implies domination by λ.

THEOREM 1.2.4. *Let $\mathcal{P}$ be dominated by a p.m. λ where*

$$\lambda = \sum_{i=0}^{\infty} c_i P_{\theta_i} \qquad (c_i \geq 0,\ \sum c_i = 1).$$

Then the statistic T (with range $(\mathcal{T}, \mathcal{F})$) is sufficient for $\mathcal{P} \iff$ there exists an $\mathcal{F}$-measurable function $g_\theta(\cdot)$ such that

$$dP_\theta(x) = g_\theta(T(x))\; d\lambda(x) \quad \forall \theta \in \Omega.$$

PROOF. ($\Rightarrow$) Suppose T is sufficient for $\mathcal{P}$. Then

$$P_\theta(A|T(x)) = P(A|T(x)) \ \forall\theta.$$

Throughout this part of the proof X will denote the indicator function of a subset of $\mathcal{X}$. The preceding equality then implies that

$$E_\theta(X|T) = E(X|T) \quad \forall X \in \mathcal{A}, \ \ \forall\theta.$$

Hence for all $\theta \in \Omega,\ X \in \mathcal{A},\ G \in \sigma(T)$, we have

$$E_\theta(I_G E(X|T)) = E_\theta(E_\theta(I_G X|T)) = E_\theta(I_G X).$$

Set $\theta = \theta_i$, multiply by c_i and sum over $i = 0, 1, 2, \ldots,$ to get

$$E_\lambda(I_G E(X|T)) = E_\lambda(I_G X) \ \ \forall X \in \mathcal{A}, \ \ \forall G \in \sigma(T).$$

This implies that $E(X|T) = E_\lambda(X|T) \ \ \forall X \in \mathcal{A}$, and hence

$$E_\theta(X|T) = E(X|T) = E_\lambda(X|T) \quad \forall X \in \mathcal{A}, \ \ \forall\theta.$$

Now define $g_\theta(T(\cdot))$ to be the Radon-Nikodym derivative of P_θ with respect to λ, with both regarded as measures on $\sigma(T)$. We know this exists since λ dominates every P_θ. We also know it is $\sigma(T)$ measurable, so it can be written in the form $g_\theta(T(\cdot))$, and we know that $E_\theta(X) = E_\lambda(g_\theta(T)X)$ for all $X \in \sigma(T)$. We need to establish however that this last relation holds for all $X \in \mathcal{A}$. We do this as follows.

$$\begin{aligned} X \in \mathcal{A} \Rightarrow E_\theta(X) &= E_\theta[E(X|T)] \\ &= E_\lambda[g_\theta(T)E(X|T)] \\ &= E_\lambda[E(g_\theta(T)X|T)] \\ &= E_\lambda[E_\lambda(g_\theta(T)X|T)] \\ &= E_\lambda[g_\theta(T)X]. \end{aligned}$$

This shows that $g_\theta(T(x)) = \frac{dP_\theta}{d\lambda}(x)$ when P_θ and λ are regarded as measures on $\mathcal{A}$.

($\Leftarrow$) Suppose that for each θ, $\frac{dP_\theta}{d\lambda}(x) = g_\theta(T(x))$ for some g_θ. We shall then show that the conditional probability $P_\lambda(A|t)$ is a version of $P_\theta(A|t)\ \forall\theta$.

$$\begin{aligned} A \in \mathcal{A},\ G \in \sigma(T) \Rightarrow \int_G I_A \, dP_\theta &= \int_G P_\theta(A|T) \, dP_\theta \\ &= \int_G P_\theta(A|T) g_\theta(T) \, d\lambda \end{aligned}$$

and

$$\begin{aligned}\int_G I_A dP_\theta &= \int_G I_A g_\theta(T)\ d\lambda \\ &= \int_G E_\lambda[I_A g_\theta(T)|T]\ d\lambda \\ &= \int_G E_\lambda[I_A|T] g_\theta(T)\ d\lambda \\ \Rightarrow \quad P_\theta(A|T) g_\theta(T) &= E_\lambda(I_A|T) g_\theta(T) \quad a.s.\ \lambda\end{aligned}$$

and hence a.s. $P_\theta\ \forall\theta$. Also $g_\theta(T) \neq 0,\ a.s.\ P_\theta$, since $dP_\theta = g_\theta(T)\ d\lambda$. Hence $P_\theta(A|T) = E_\lambda(I_A|T) = P_\lambda(A|T)\ a.s.\ P_\theta$ and the R.S. is independent of θ. □

THEOREM 1.2.5. *(Theorem A.4.2 in appendix of TSH) If $\mathcal{P} = \{P_\theta,\ \theta \in \Omega\}$ is dominated by a σ-finite measure μ, then it is equivalent to $\lambda = \sum_{i=0}^{\infty} c_i P_{\theta_i}$ for some countable subcollection $P_{\theta_i} \in \mathcal{P},\ i = 0, 1, 2, \ldots,$ with $c_i \geq 0$ and $\sum c_i = 1$.*

PROOF. μ is $\sigma-$finite, $\Rightarrow \exists A_n \in \mathcal{A}$ with $A_1, A_2, \ldots$ disjoint, and $\cup A_i = \mathcal{X}$ such that $0 < \mu(A_i) < \infty,\ i = 1, 2, \ldots$. Set

$$\mu^*(A) = \sum_{i=1}^{\infty} \frac{\mu(A \cap A_i)}{2^i \mu(A_i)}$$

Then, μ^* is a probability measure equivalent to μ. Hence we can assume without loss of generality that the dominating measure μ is a probability measure Let

$$f_\theta = \frac{dP_\theta}{d\mu}$$

and set

$$S_\theta = \{x\colon f_\theta(x) > 0\}$$

Then

$$P_\theta(A) = P_\theta(A \cap S_\theta) = 0 \ \text{ iff } \mu(A \cap S_\theta) = 0. \tag{1.2.1}$$

(Since $P_\theta \ll \mu$ and since $\mu(A \cap S_\theta) > 0,\ f_\theta > 0$ on $A \cap S_\theta \Rightarrow P_\theta(A \cap S_\theta) > 0$.) A set $A \in \mathcal{A}$ is a *kernel* if $A \subseteq S_\theta$ for some θ; a finite or countable union of kernels is called a *chain*. Set

$$\alpha = \sup_{\text{chains } C} \mu(C)$$

Then $\alpha = \mu(C)$ for some chain $C = \cup_{n=1}^{\infty} A_n,\ A_n \subseteq S_{\theta_n}$. (since $\exists\, \{C_n\}$ such that $\mu(C_n) \uparrow \alpha$ and for this sequence $\mu(\cup C_n) = \alpha$.)
It follows from the following Lemma that $\mathcal{P}$ is dominated by $\lambda(\cdot) = \sum_{n=1}^{\infty} \frac{1}{2^n} P_{\theta_n}(\cdot)$. Since

$$\begin{aligned}\lambda(A) = 0 \ &\Rightarrow \ P_{\theta_n}(A) = 0\ \forall n \\ &\Rightarrow \ P_\theta(A) = 0\ \forall\theta \text{ (by the Lemma)},\end{aligned}$$

it is obvious that

$$P_\theta(A) = 0\ \forall\,\theta \Rightarrow \lambda(A) = 0$$

Hence $\mathcal{P}$ is equivalent to $\lambda(\cdot) = \sum_{n=1}^{\infty} \frac{1}{2^n} P_{\theta_n}(\cdot)$. □

LEMMA 1.2.6. *If $\{\theta_n\}$ is the sequence used in the construction of C, then $\{P_\theta,\ \theta \in \Omega\}$ is dominated by $\{P_{\theta_n},\ n = 1, 2, \ldots\}$, i.e.*

$$P_{\theta_n}(A) = 0\ \forall n \Rightarrow P_\theta(A) = 0\ \forall \theta$$

PROOF.

$$\begin{aligned} P_{\theta_n}(A) = 0\ \forall n &\Rightarrow \mu(A \cap S_{\theta_n}) = 0\ \forall n\ (by\ 1.2.1) \\ &\Rightarrow^{(C \subseteq \cup S_{\theta_n})} \mu(A \cap C) = 0 \\ &\Rightarrow^{(P_\theta \ll \mu)} P_\theta(A \cap C) = 0\ \forall \theta \end{aligned}$$

If $P_\theta(A) > 0$ for some θ then, since $P_\theta(A) = P_\theta(A \cap C) + P_\theta(A\ \cap C^c)$,

$P_\theta(A \cap C^c) = P_\theta(A \cap C^c \cap S_\theta) > 0$

$\Rightarrow A \cap C^c \cap S_\theta$ is a kernel disjoint from C

$\Rightarrow C \cup (A \cap C^c \cap S_\theta)$ is a chain with $\mu > \alpha$, $(P_\theta(A) > 0 \Rightarrow \mu(A) > 0)$

contradicting the definition of α.

Hence, $P_\theta(A) = 0\ \forall\, \theta$. □

THEOREM 1.2.7. ***The Factorization Theorem***
Let μ be a σ-finite measure which dominates $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ and let

$$p_\theta = \frac{dP_\theta}{d\mu}.$$

Then the statistic T is sufficient for $\mathcal{P}$ if and only if there exists a non negative $\mathcal{F}$-measurable function $g_\theta : \mathcal{T} \to \mathbb{R}$ and an $\mathcal{A}$-measurable function $h : \mathcal{X} \to \mathbb{R}$ such that

$$p_\theta(x) = g_\theta(T(x))\, h(x) \quad a.e.\ \mu. \tag{1.2.2}$$

PROOF. By theorem 1.2.5, $\mathcal{P}$ is equivalent to

$$\lambda = \sum_i c_i P_{\theta_i},\ \text{where } c_i \geq 0,\ \sum_i c_i = 1.$$

If T is sufficient for $\mathcal{P}$,

$$\begin{aligned} p_\theta(x) &= \frac{dP_\theta(x)}{d\mu(x)} = \frac{dP_\theta(x)}{d\lambda(x)} \cdot \frac{d\lambda(x)}{d\mu(x)} \\ &= g_\theta(T(x))\, h(x) \qquad \text{by theorem 1.2.4.} \end{aligned}$$

On the other hand, if equation (1.2.2) holds,

$$\begin{aligned} d\lambda(x) &= \sum c_i\ dP_{\theta_i}(x) = \sum c_i p_{\theta_i}(x)\ d\mu(x) \\ &= \sum_{i=1}^{\infty} c_i g_{\theta_i}(T(x))\, h(x)\, d\mu(x) \\ &= K(T(x))\, h(x)\, d\mu(x). \end{aligned} \tag{1.2.3}$$

Thus,

$$\begin{aligned} dP_\theta(x) &= p_\theta(x)\ d\mu(x) \qquad \text{by the definition of } p_\theta(x) \\ &= \frac{g_\theta(T(x))\,h(x)}{K(T(x))\,h(x)}\,d\lambda(x) \qquad \text{by equations (1.2.2) and (1.2.3)} \\ &= \tilde{g}_\theta(T(x))\ d\lambda(x) \qquad \text{where } \tilde{g}_\theta(T(x)) := 0 \text{ if } K(T(x)) = 0. \end{aligned}$$

Hence T is sufficient for $\mathcal{P}$ by theorem 1.2.4. □

REMARK 1.2.8. If $f_\theta(x)$ is the density of X with respect to Lebesgue measure then T is sufficient for $\mathcal{P}$ iff

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

where h is independent of θ.

EXAMPLE 1.2.9. Let $X_1, X_2, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$, and write $X = (X_1, X_2, \cdots, X_n)$. A σ-finite dominating measure on $\mathcal{B}^n$ is Lebesgue measure with

$$\begin{aligned} p_{\mu,\sigma^2}(x) &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}\exp\left(\frac{-1}{2\sigma^2}\sum_1^n x_i^2 + \frac{\mu}{\sigma^2}\sum x_i - \frac{n\mu^2}{2\sigma^2}\right) \\ &= g_{\mu,\sigma^2}\left(\sum x_i, \sum x_i^2\right). \end{aligned}$$

Therefore $T(X) = (\sum X_i, \sum X_i^2)$ is sufficient for $\mathcal{P} = \{P_{\mu,\sigma^2}\}$.

REMARK 1.2.10. $T^*(X) = \left(\bar{X}, S^2\right)$ is also sufficient for $\mathcal{P} = \{P_{\mu,\sigma^2}\}$, since

$$g_{\mu,\sigma^2}\left(\sum x_i, \sum x_i^2\right) = g^*_{\mu,\sigma^2}\left(\bar{x}, S^2\right)$$

T and T^* are equivalent in the following sense.

DEFINITION 1.2.11. Two statistics T and S are *equivalent* if they induce the same σ-algebra up to $\mathcal{P}$-null sets. i.e. if there exists a $\mathcal{P}$-null set $\mathcal{N}$ and functions f and g such that

$$T(x) = f(S(x)) \qquad \text{and} \qquad S(x) = g(T(x)) \qquad \text{for all } x \in \mathcal{N}^c.$$

EXAMPLE 1.2.12. Let $X_1, \ldots, X_n$ be iid $U(0, \theta)$, $\theta > 0$ and $X = (X_1, \ldots, X_n)$.

$$\begin{aligned} p_\theta(x) &= \frac{1}{\theta^n}\prod_1^n I_{[0,\infty)}(x_i) I_{(-\infty,\theta]}(x_i) \\ &= \frac{1}{\theta^n} I_{[0,\infty)}(x_{(1)}) I_{(-\infty,\theta]}(x_{(n)}) \\ &= g_\theta(x_{(n)})h(x) \\ \Rightarrow T(X) &= X_{(n)} \text{ is sufficient for } \theta. \end{aligned}$$

EXAMPLE 1.2.13. $X_1, \ldots, X_n$ iid $N(0, \sigma^2)$, $\Omega = \{\sigma^2 \colon \sigma^2 > 0\}$. Define

$$\begin{aligned} T_1(X) &= (X_1, \ldots, X_n) \\ T_2(X) &= (X_1^2, \ldots, X_n^2) \\ T_3(X) &= (X_1^2 + \cdots + X_m^2,\ X_{m+1}^2 + \cdots + X_n^2) \\ T_4(X) &= X_1^2 + \cdots + X_n^2 \\ p_\theta(x) &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_1^n X_i^2\right) \end{aligned}$$

Each $T_i(X)$ is sufficient. However $\sigma(T_4) \subseteq \sigma(T_3) \subseteq \sigma(T_2) \subseteq \sigma(T_1)$.
(since functions of T_4 are functions of T_3, functions of T_3 are functions of T_2 and functions of T_2 are functions of T_1.)

REMARK 1.2.14. If T is sufficient for θ and $T = H(S)$ where S is some statistic, then S is also sufficient since

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

Since $\sigma(T) = S^{-1}H^{-1}\mathcal{B}_\mathcal{T} \subset S^{-1}\mathcal{B}_\mathcal{S}$ $((\mathcal{X}, \mathcal{A}) \xrightarrow{S} (\mathcal{S}, \mathcal{B}_\mathcal{S}) \xrightarrow{H} (\mathcal{T}, \mathcal{B}_\mathcal{T}))$, T provides a greater reduction of the data than S, strictly greater unless H is one to one, in which case S and T are equivalent.

DEFINITION 1.2.15. T is a *minimal sufficient* statistic, if for any sufficient statistic S, there exists a measurable function H such that

$$T = H(S) \qquad \text{a.s. } \mathcal{P}.$$

THEOREM 1.2.16. *If $\mathcal{P}$ is dominated by a σ-finite measure μ, then the statistic U is sufficient iff for every fixed θ and θ_0, the ratio of the densities p_θ and p_{θ_0} with respect to μ, defined to be 1 when both densities are zero, satisfies*

$$\frac{p_\theta(x)}{p_{\theta_0}(x)} = f_{\theta,\theta_0}(U(x)) \ \text{ a.s. } \mathcal{P} \text{ for some measurable } f_{\theta,\theta_0}.$$

PROOF. HW problem (TPE Ch 1 Problem 6.6). □

THEOREM 1.2.17. *Let $\mathcal{P}$ be a finite family with densities $\{p_0, p_1, \ldots, p_k\}$, all having the same support (i.e. $S = \{x \colon p_i(x) > 0\}$ is independent of i). Then*

$$T(x) = \left(\frac{p_1(x)}{p_0(x)}, \frac{p_2(x)}{p_0(x)}, \cdots, \frac{p_k(x)}{p_0(x)}\right)$$

is minimal sufficient. (Also true for a countable collection of densities with no change in the proof.)

PROOF. First T is sufficient by theorem (1.2.16) since $\frac{p_i(x)}{p_j(x)}$ is a function of $T(x)$ for all i and j (need common support here.) If U is a sufficient statistic then by theorem

(1.2.16),

$$\begin{aligned} & \tfrac{p_i(x)}{p_0(x)} \text{ is a function of } U \text{ for each } i \\ \Rightarrow \quad & T \text{ is a function of } U \\ \Rightarrow \quad & T \text{ is minimal sufficient.} \end{aligned}$$

□

REMARK 1.2.18. The theorem 1.2.17 extends to uncountable collections under further conditions. It also extends to countable collections without common support (Prob. 1.6.11).

THEOREM 1.2.19. *Let $\mathcal{P}$ be a family with common support and suppose $\mathcal{P}_0 \subseteq \mathcal{P}$. If T is minimal sufficient for $\mathcal{P}_0$ and sufficient for $\mathcal{P}$, then T is minimal sufficient for $\mathcal{P}$.*

PROOF.

U is sufficient for $\mathcal{P} \Rightarrow U$ is sufficient for $\mathcal{P}_0$ by Definition 1.2.1.
T is minimal sufficient for $\mathcal{P}_0 \Rightarrow T(x) = H(U(x))$ a.s. $\mathcal{P}_0$.
But since $\mathcal{P}$ has common support, $T(x) = H(U(x))$ a.s. $\mathcal{P}$.

□

Note the following points.

(1) Minimal sufficient statistics for uncountable families $\mathcal{P}$ can often be obtained by combining the above theorems.
(2) Minimal sufficient statistics exist under weak assumptions (but not always). In particular they exist if $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^n, \mathcal{B}^n)$ and $\mathcal{P}$ is dominated by a σ-finite measure.
(3) A generalization of the above results for establishing minimality, which accommodates uncountable families without common support, is Theorem 1.2.20.

THEOREM 1.2.20. *Let $\mathcal{P} = \{p_\theta(x) : \theta \in \Theta\}$ be a family of densities dominated by a σ-finite measure. If there exists a measurable function $T : \mathcal{X} \to \mathcal{T}$ such that $T(x) = T(y)$ if and only if $y \in \mathcal{D}(x)$, where*

$$\mathcal{D}(x) = \{y \in \mathcal{X} : p_\theta(y) = p_\theta(x)h(x,y),\ \forall\theta \text{ and some } h(x,y) > 0\},$$

then $T(X)$ is a minimal sufficient statistic.

PROOF. Schervish (1995), Theorem 2.29. □

EXAMPLE 1.2.21. $\mathcal{P}_0 : (X_1, \ldots, X_n)$ *iid* $N(\theta, 1)$, $\theta \in \{\theta_0, \theta_1\}$.

$\mathcal{P}$: $(X_1, \ldots, X_n)$ *iid* $N(\theta, 1)$, $\theta \in \mathbb{R}$.

$$\begin{aligned}\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} &= \exp\left\{-\frac{1}{2}\left[\sum(x_i-\theta_1)^2-\sum(x_i-\theta_0)^2\right]\right\}\\ &= \exp\left\{-\frac{1}{2}\left[\sum 2x_i(\theta_0-\theta_1)+n\theta_1^2-n\theta_0^2\right]\right\}\end{aligned}$$

This is a function of $\bar{x}$, hence $\bar{X}$ is minimal sufficient for $\mathcal{P}_0$ by Theorem 1.2.17. Since $\bar{X}$ is sufficient for $\mathcal{P}$ (by the factorization theorem), $\bar{X}$ is minimal sufficient for $\mathcal{P}$.

EXAMPLE 1.2.22. $\mathcal{P}$: $(X_1,\ldots,X_n)$ *iid* $U(0,\theta)$, $\theta>0$.

To show that $X_{(n)}$ is minimal sufficient via Theorem 1.2.20, suppose that

$$\frac{p_\theta(x)}{p_\theta(y)}=\frac{\theta^{-n}I_{(0,\theta)}(x_{(n)})}{\theta^{-n}I_{(0,\theta)}(y_{(n)})}=\frac{I_{(0,\theta)}(x_{(n)})}{I_{(0,\theta)}(y_{(n)})}=h(x,y),\quad\forall\theta.$$

This is true if and only if $x_{(n)}=y_{(n)}$, in which case $h(x,y)=1$ and $\mathcal{D}(x)=\{y: y_{(n)}=x_{(n)}\}$, whence $T=X_{(n)}$ is minimal sufficient.

EXAMPLE 1.2.23. Logistic

We'll show the order statistics are minimal sufficient via Theorems 1.2.17 and 1.2.19 (but this could also be be accomplished via Theorem 1.2.20).

$\mathcal{P}$: $(X_1,\ldots,X_n)$ *iid* $L(\theta,1)$, $\theta\in\mathbb{R}$.

$\mathcal{P}_0$: $(X_1,\ldots,X_n)$ *iid* $L(\theta,1)$, $\theta\in\{0,\theta_1,\ldots,\theta_n\}$.

$$\begin{aligned}p_\theta(x) &= \frac{\exp\left[-\sum(x_i-\theta)\right]}{\prod_{i=1}^n\left\{1+\exp\left[-(x_i-\theta)\right]\right\}^2},\\ \text{so } T &= (T_1(X),\ldots,T_n(X)) \quad \text{is minimal sufficient,}\end{aligned}$$

where

$$T_i(x)=\frac{p_{\theta_i}(x)}{p_0(x)}=e^{n\theta_i}\prod_{j=1}^n\frac{(1+e^{-x_j})^2}{(1+e^{-(x_j-\theta_i)})^2}.$$

We will show that $T(X)$ is equivalent to $(X_{(1)},\ldots,X_{(n)})$, by showing that

$$T(x)=T(y)\Leftrightarrow x_{(1)}=y_{(1)},\cdots,x_{(n)}=y_{(n)}.$$

PROOF. ($\Leftarrow$) Obvious from the expression for $T_i(x)$.
($\Rightarrow$) Suppose that $T_i(x)=T_i(y)$ for $i=1,2,\ldots,n$,

$$\text{i.e. } \prod_{j=1}^n\frac{(1+e^{-x_j})^2}{(1+e^{-(x_j-\theta_i)})^2}=\prod_{j=1}^n\frac{(1+e^{-y_j})^2}{(1+e^{-(y_j-\theta_i)})^2},\qquad i=1,\ldots,n,$$

$$\text{i.e. } \prod_{j=1}^n\frac{1+u_j\omega}{1+u_j}=\prod_{j=1}^n\frac{1+v_j\omega}{1+v_j},\qquad \omega=\omega_1,\ldots,\omega_n,$$

where $u_j = e^{-x_j},\ v_j = e^{-y_j}$ and $\omega_i = e^{\theta_i}$. Here we have two polynomials in ω of degree n which are equal for $n+1$ distinct values, $1, \omega_1, \ldots, \omega_n$, of ω and hence for all ω.

$$\omega = 0 \Rightarrow \prod_{j=1}^{n}(1+u_j) = \prod_{j=1}^{n}(1+v_j)$$

$$\therefore \prod_{j=1}^{n}(1+u_j\omega) = \prod_{j=1}^{n}(1+v_j\omega) \qquad \forall\omega$$

$\therefore$ the zero sets of both these polynomials are the same

$\therefore$ x and y have the same order statistics.

By theorem 1.2.17, the order statistics are therefore minimal sufficient for $\mathcal{P}_0$. They are also sufficient for $\mathcal{P}$, so by theorem 1.2.19, the order statistics are minimal sufficient for $\mathcal{P}$. There is not much reduction possible here! This is fairly typical of location families, the normal, uniform and exponential distributions providing happy exceptions. □

Ancillarity

DEFINITION 1.2.24. A statistic V is said to be *ancillary* for $\mathcal{P}$ if the distribution, P_θ^V, of V does not depend on θ. It is called *first order ancillary* if $E_\theta V$ is independent of θ.

EXAMPLE 1.2.25. In Example 1.2.23, $X_{(2)} - X_{(1)}$ is ancillary since $Y_1 = X_1 - \theta, \ldots, Y_n = X_n - \theta$ are iid P_0 (the standard member of the family with $\theta = 0$) and $X_{(2)} - X_{(1)} = Y_{(2)} - Y_{(1)}$.

EXAMPLE 1.2.26.

$$\mathcal{P} : (X_1, \ldots, X_n)\ iid\ N(\theta, 1),\ \theta \in \mathbb{R}.$$

$$S^2 = \sum (X_i - \bar{X})^2 \text{ is ancillary}$$

since

$$S^2 = \sum (Y_i - \bar{Y})^2 \qquad \text{where } Y_i = X_i - \theta,\ i = 1, 2, \ldots, \text{are } iid\ N(0,1).$$

REMARK 1.2.27. Ancillary statistics by themselves contain no information about θ, however minimal sufficient statistics may contain ancillary components. For example, in 1.2.23, $T = (X_{(1)}, \cdots, X_{(n)})$ is equivalent to $T^* = (X_{(1)}, X_{(2)} - X_{(1)}, \cdots, X_{(n)} - X_{(1)})$, whose last $(n-1)$ components are ancillary. You can't drop them as $X_{(1)}$ is not even sufficient.

Complete Statistic

A sufficient statistic should bring about the best reduction of the data if it contains as little ancillary material as possible. This suggests requiring that no non-constant function of T be ancillary, or not even first order ancillary, i.e. that

$$E_\theta f(T) = c \text{ for all } \theta \in \Omega \qquad \Rightarrow \qquad f(T) = c \text{ a.s. } \mathcal{P}$$

or equivalently that

$$E_\theta f(T) = 0 \text{ for all } \theta \in \Omega \qquad \Rightarrow \qquad f(T) = 0 \text{ a.s. } \mathcal{P}.$$

DEFINITION 1.2.28. A statistic T is *complete* if

$$E_\theta f(T) = 0 \text{ for all } \theta \in \Omega \qquad \Rightarrow \qquad f(T) = 0 \text{ a.s. } \mathcal{P} \tag{1.2.4}$$

T is said to be *boundedly complete* if equation (1.2.4) holds for all bounded measurable functions f.

Since complete sufficient statistics are intended to give a good reduction of the data, it is not unreasonable to expect them to be minimal. We shall prove a slightly weaker result.

THEOREM 1.2.29. *Let U be a complete sufficient statistic. If there exists a minimal sufficient statistic, then U is minimal sufficient.*

PROOF. Let T be a minimal sufficient statistic and let ψ be a bounded measurable function. We will show that

$$\psi(U) \in \sigma(T) \quad \text{by showing that} \quad E(\psi(U)|T) = \psi(U) \ \ a.s.$$

Now

$E(\psi(U)|T) = g(U)$ for some measurable g since T is minimal and U is sufficient.

Let $h(U) = E(\psi(U)|T) - \psi(U)$, then $E_\theta h(U) = 0 \quad \forall\, \theta$ so $h(U) = 0$ *a.s.* $\mathcal{P}$ since U is complete. Hence $\psi(U) = E(\psi(U)|T) \in \sigma(T)$. Hence U-measurable bounded functions are T-measurable, i.e. $\sigma(U) \subset \sigma(T)$, i.e. U is minimal sufficient. □

REMARK 1.2.30.

(1) If $\mathcal{P}$ is dominated by a σ-finite measure and $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^n, \mathcal{B}^n)$, the existence of a minimal sufficient statistic does not need to be assumed.
(2) A minimal sufficient statistic is not necessarily complete. See the next example.

EXAMPLE 1.2.31.

$$\begin{aligned} \mathcal{P} &= \{N(\theta, \theta^2),\ \theta > 0\} \\ p_\theta(x) &= \frac{1}{\theta\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\theta)^2}{\theta^2}} = \frac{1}{\theta\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x}{\theta}-1)^2} \end{aligned}$$

The single observation X is minimal sufficient but not complete since

$$E_\theta[I_{(0,\infty)}(X) - \Phi(1)] = P_\theta(X > 0) - \Phi(1) = 0 \quad \forall\, \theta$$

however $P_\theta(I_{(0,\infty)}(X) - \Phi(1) = 0) = 0\ \forall \theta$.

THEOREM 1.2.32. (Basu's theorem) *If T is complete and sufficient for $\mathcal{P}$, then any ancillary statistic is independent of T.*

PROOF. If S is ancillary, then $\mathcal{P}_\theta(S \in B) = p_B$, independent of θ.
Sufficiency of $T \Rightarrow P_\theta(S \in B|T) = h(T)$, independent of θ.

$$\begin{aligned}
&\therefore E_\theta(h(T) - p_B) = 0 \\
&\Rightarrow h(T) = p_B \ a.s.\, \mathcal{P} \qquad \text{by completeness} \\
&\Rightarrow S \text{ is independent of } T
\end{aligned}$$

□

1.3. Exponential Families.

DEFINITION 1.3.1. A family of probability measure's $\{P_\theta : \theta \in \Omega\}$ is said to be an *s-parameter exponential family* if there exists a σ-finite measure μ such that

$$p_\theta(x) = \frac{dP_\theta(x)}{d\mu(x)} = \exp\left(\sum_1^s \eta_i(\theta) T_i(x) - B(\theta)\right) h(x),$$

where η_i, T_i and B are real-valued.

REMARK 1.3.2.

(1) P_θ, $\theta \in \Omega$ are equivalent (since $\{x\colon p_\theta(x) > 0\}$ is independent of θ).
(2) The factorization theorem implies that $T = (T_1, \cdots, T_s)$ is sufficient.
(3) If we observe $X_1, \ldots, X_n$, iid with marginal distributions P_θ then $\sum_{j=1}^n T(X_j)$ is sufficient for θ .

THEOREM 1.3.3. *If $\{1, \eta_1, \ldots, \eta_s\}$ is LI, then $T = (T_1, \ldots, T_s)$ is minimal sufficient. (Linear independence of $\{1, \eta_1, \ldots, \eta_s\}$ means $c_1\eta_1(\theta) + \cdots + c_s\eta_s(\theta) + d = 0\ \forall\theta \Rightarrow c_1 = \cdots = c_s = d = 0$. Equivalently we can say that $\{\eta_i\}$ is* **affinely independent** *or AI since the set of points $\{(\eta_1(\theta), \ldots, \eta_s(\theta)), \theta \in \Omega\}$ then lie in a proper affine subspace of $\mathbb{R}^s$.)*

PROOF. Fix $\theta_0 \in \Omega$ and consider

$$\frac{dP_\theta}{dP_{\theta_0}}(x) = \frac{p_\theta(x)}{p_{\theta_0}(x)} = \exp\{B(\theta_0) - B(\theta)\} \exp\left\{\sum_1^s (\eta_i(\theta) - \eta_i(\theta_0)) T_i(x)\right\}. \tag{1.3.1}$$

If $\{1, \eta_1, \ldots, \eta_s\}$ is LI then so is $\{1, \eta_1 - \eta_1(\theta_0), \ldots, \eta_s - \eta_s(\theta_0)\}$.

Set $S = \{(\eta_1(\theta) - \eta_1(\theta_0), \ldots, \eta_s(\theta) - \eta_s(\theta_0)),\ \theta \in \Omega\} \subseteq \mathbb{R}^s$. Then $span(S)$ is a *linear subspace* of $\mathbb{R}^s$.

If $\dim(span(S)) < s$, then there exists a non-zero vector $v = (v_1, \ldots, v_s)$ s.t.

$$v_1(\eta_1(\theta) - \eta_1(\theta_0)) + \cdots + v_s(\eta_s(\theta) - \eta_s(\theta_0)) = 0 \quad \forall\theta$$

contradicting the linear independence of $\{1, \eta_i - \eta_i(\theta_0)\}$. Hence

$$\dim(span(S)) = s \quad i.e.\ \exists\ \theta_1, \ldots, \theta_s \in \Omega\ s.t.$$

$$\{(\eta_1(\theta_i) - \eta_1(\theta_0), \cdots, \eta_s(\theta_i) - \eta_s(\theta_0)),\ i = 1, \cdots, s\} \text{ is LI.} \tag{1.3.2}$$

From 1.3.1,

$$\sum_{j=1}^{s} (\eta_j(\theta_i) - \eta_j(\theta_0)) T_j(x) = \ln \frac{p_{\theta_i}(x)}{p_{\theta_0}(x)} + (B(\theta_i) - B(\theta_0)), \quad i = 1, \ldots, s.$$

Since the matrix $[\eta_j(\theta_i) - \eta_j(\theta_0)]_{i,j=1}^{s}$ is non-singular, $T_j(x)$ can be expressed uniquely in terms of $\ln \frac{p_{\theta_i}(x)}{p_{\theta_0}(x)}$, $i = 1, \ldots, s$.

But $\frac{p_{\theta_i}(x)}{p_{\theta_0}(x)},\ i = 1, \ldots, s$ is minimal sufficient for $\mathcal{P}_0 = \{P_{\theta_j}, j = 0, 1, \cdots, s\}$ by theorem 1.2.17. Hence T is minimal sufficient by theorem 1.2.19. □

EXAMPLE 1.3.4.

$$p_\theta(x) = \sqrt{\frac{\theta}{2\pi}} \exp\{-\frac{1}{2}\theta x^2 + \theta x - \frac{\theta}{2}\}.$$

$$\eta_1(\theta) = -\frac{1}{2}\theta,\ \eta_2(\theta) = \theta,\ T(x) = (x^2, x) \text{ is sufficient but not minimal}$$

since rewriting the model as $p_\theta(x) = \sqrt{\frac{\theta}{2\pi}} \exp\{-\frac{1}{2}\theta(x-1)^2\},$ we see that

$$T^*(x) = (x-1)^2 \text{ is minimal sufficient.}$$

REMARK 1.3.5. The exponential family can always be rewritten in such a way that the functions $\{T_i\}$ and $\{\eta_i\}$ are AI. If there exist constants $c_1, \ldots, c_s, d$, not all zero, such that

$$c_1 T_1(x) + \cdots + c_s T_s(x) = d \quad a.s.\ \mathcal{P}$$

then one of the T_i's can be expressed in terms of the others (or is constant). After reducing the number of functions T_i as far as possible, the same can be done with their coefficients until the new functions $\{T_i\}$ and $\{\eta_i\}$ are AI.

DEFINITION 1.3.6. (**Order** of the exponential family.) If the functions $\{T_i, i = 1, \ldots, s\}$ on $\mathcal{X}$ and $\{\eta_i, i = 1, \ldots, s\}$ on Ω are both AI, then s is the *order* of the exponential family

$$p_\theta(x) = \frac{dP_\theta}{d\mu}(x) = \exp\left(\sum_1^s \eta_i(\theta) T_i(x) - B(\theta)\right) h(x).$$

PROPOSITION 1.3.7. *The order is well-defined.*

PROOF. We shall show that

$$s + 1 = dim(V)$$

where V is the set of functions on $\mathcal{X}$ defined by $V = span\{1, \ln \frac{dP_\theta}{dP_{\theta_0}}(\cdot),\ \theta \in \Omega\}$ (independent of the dominating measure μ and the choice of $\{\eta_i\}$, $\{T_i\}$).

$$\ln \frac{dP_\theta}{dP_{\theta_0}}(x) = \sum_{i=1}^{s} (\eta_i(\theta) - \eta_i(\theta_0))T_i(x) + B(\theta_0) - B(\theta)$$

so that

$$V \subseteq span\{1, T_i(\cdot),\ i = 1, \ldots, s\} \quad \therefore dim(V) \leq s + 1$$

On the other hand, since $\{1, \eta_i,\ i = 1, \ldots, s\}$ is LI, each $T_j(x)$ can be expressed as a linear combination of $1, \ln \frac{dP_{\theta_i}}{dP_{\theta_0}}(x),\ i = 1, \ldots, s$, as in the proof of the previous theorem,

$$\therefore span\{1, T_i(\cdot),\ i = 1, \ldots, s\} \subseteq V$$
$$\therefore s + 1 \leq dim(V)$$

□

Definition 1.3.8. (**Canonical Form**) For any s-parameter exponential family (not necessarily of *order* s) we can view the vector $\eta(\theta) = (\eta_1(\theta), \ldots, \eta_s(\theta))'$ as the parameter rather than θ. Then the density with respect to μ can be rewritten as

$$p(x, \eta) = \exp[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)]h(x), \quad \eta \in \eta(\Omega).$$

Since $p(\cdot, \eta)$ is a probability density with respect to μ,

$$e^{A(\eta)} = \int e^{\sum_1^s \eta_i T_i(x)} h(x) d\mu(x). \tag{1.3.3}$$

Definition 1.3.9. (**The Natural Parameter Set**) This is a possibly larger set than $\{\eta(\theta),\ \theta \in \Omega\}$. It is the set of all s-vectors for which, by suitable choice of $A(\eta)$, $p(\cdot, \eta)$ *can* be a probability density, i.e.

$$\mathcal{N} = \{\eta = (\eta_1, \cdots, \eta_s) \subseteq \mathbb{R}^s \colon \int e^{\sum_1^s \eta_i T_i(x)} h(x) d\mu(x) < \infty\}$$

Theorem 1.3.10. *$\mathcal{N}$ is a convex set, and $A(\eta)$ is a convex function.*

Proof. Suppose $\alpha = (\alpha_1, \ldots, \alpha_s)$ and $\beta = (\beta_1, \ldots, \beta_s) \in \mathcal{N}$. Then,

$$\begin{aligned}
&\int e^{p\sum_1^s \alpha_i T_i(x) + (1-p)\sum_1^s \beta_i T_i(x)} h(x)\ d\mu(x) \\
&\leq \left[\int e^{\sum \alpha_i T_i(x)} h(x)\ d\mu(x)\right]^p \cdot \left[\int e^{\sum \beta_i T_i(x)} h(x)\ d\mu(x)\right]^{1-p} \qquad \text{(Holder's Inequality)} \\
&< \infty
\end{aligned}$$

The convexity of $A(\eta)$ follows similarly (Bickel & Doksum, 2015, Theom 1.6.3). □

THEOREM 1.3.11. *$T = (T_1, \cdots, T_s)$ has density*

$$p_\eta(t) = \exp(\eta \cdot t - A(\eta))$$

relative to $\nu = \tilde{\mu} \circ T^{-1}$ where $d\tilde{\mu}(x) = h(x)\, d\mu(x)$. (Note: this introduces the Jacobian term $k(t)d\nu$ in the density over Euclidean s-space, where $k(t) = h(T^{-1}(t))$.)

PROOF. If f:$\mathcal{T} \to \mathbb{R}$ is a bounded measurable function,

$$\begin{aligned} Ef(T) &= \int f(T(x)) e^{\eta \cdot T(x)} e^{-A(\eta)} d\tilde{\mu}(x) \\ &= \int f(t) e^{\eta \cdot t} e^{-A(\eta)} d\tilde{\mu} \circ T^{-1}(t) \end{aligned}$$

□

DEFINITION 1.3.12. The family of densities

$$p_\eta(t) = \exp(\eta \cdot t - A(\eta)), \quad \eta \in \eta(\Omega),$$

n is called an *s-dimensional* or *s-parameter* **standard exponential family**. (Defined on $\mathbb{R}^s$, not $\mathcal{X}$.)

THEOREM 1.3.13. *Let $\{p_\eta(x)\}$ be the s-parameter exponential family,*

$$p_\eta(x) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\right) h(x)), \quad \eta \in \eta(\Omega),$$

and suppose

$$\int \phi(x)\, e^{\sum_1^s \eta_j T_j(x)} d\mu(x) \tag{1.3.4}$$

exists and is finite for some ϕ and all $\eta_j = a_j + ib_j$ such that $a \in \mathcal{N}$ (=natural parameter space). Then

(i) *$\int \phi(x)\, e^{\sum_1^s \eta_j T_j(x)} d\mu(x)$ is an analytic function of each η_i on $\{\eta : \Re(\eta) \in int(\mathcal{N})\}$ and*
(ii) *the derivative of all orders with respect to the η_i's of $\int \phi(x)\, e^{\sum_1^s \eta_j T_j(x)} d\mu(x)$ can be computed by differentiating under the integral sign.*

PROOF. Let $a^0 = (a_1^0, \ldots, a_s^0)$ be in $int(\mathcal{N})$ and let $\eta_1^0 = a_1^0 + ib_1^0$. Then

$$\phi(x) e^{\sum_2^s \eta_j T_j(x)} = h_1(x) - h_2(x) + i(h_3(x) - h_4(x))$$

where h_1 and h_2 are the positive and negative parts of the real part and h_3 and h_4 are the positive and negative parts of the imaginary part.
Then $\int \phi(x)\, e^{\sum_1^s \eta_j T_j(x)} d\mu(x)$ can be expressed as

$$\int e^{\eta_1 T_1(x)}\, d\mu_1(x) - \int e^{\eta_1 T_1(x)}\, d\mu_2(x) + i \int e^{\eta_1 T_1(x)}\, d\mu_3(x) - i \int e^{\eta_1 T_1(x)}\, d\mu_4(x),$$

where $d\mu_i(x) = h_i(x)\, d\mu(x),\ i = 1, \ldots, 4$. Hence it suffices to prove (i) and (ii) for

$$\psi(\eta_1) = \int e^{\eta_1 T_1(x)} d\mu(x).$$

Since $a^0 \in int(\mathcal{N})$, there exists $\delta > 0$ s.t. $\psi(\eta_1)$ exists and is finite for all η_1 with $|a_1 - a_1^0| < \delta$. Now consider the difference quotient

$$(*) \qquad \frac{\psi(\eta_1) - \psi(\eta_1^0)}{\eta_1 - \eta_1^0} = \int e^{\eta_1^0 T_1(x)} \frac{e^{(\eta_1 - \eta_1^0)T_1(x)} - 1}{\eta_1 - \eta_1^0} \mu(dx) \ \bigcap \text{with } |\eta_1 - \eta_1^0| < \delta/2.$$

Observe that

$$\begin{aligned} |e^{zt} - 1| = |\sum_1^\infty \frac{(zt)^j}{j!}| &\leq \sum_1^\infty \frac{|zt|^j}{j!} = e^{|zt|} - 1 \\ &\leq |zt| e^{|zt|} \\ \Rightarrow |\frac{e^{zt} - 1}{z}| &\leq |t| e^{|zt|} \end{aligned}$$

The integrand in (*) is therefore bounded in absolute value by $|T_1(x)| e^{(a_1^0 + \frac{\delta}{2})|T_1(x)|}$, where $a_1^0 = Re(\eta_1^0)$ and $\int |T_1(x)| e^{(a_1^0 + \frac{\delta}{2})|T_1(x)|} \mu(dx) < \infty$ since

$$|T_1| e^{(a_1^0 + \frac{\delta}{2})|T_1|} = \begin{cases} \underbrace{|T_1| e^{-\frac{\delta}{4} T_1}}_{\text{bounded}} \underbrace{e^{(a_1^0 + \frac{3\delta}{4})T_1}}_{\text{integrable}} & \text{if } T_1 > 0 \\ \underbrace{|T_1| e^{\frac{\delta}{4} T_1}}_{} \underbrace{e^{(a_1^0 + \frac{\delta}{4})T_1}}_{} & \text{if } T_1 < 0 \end{cases}$$

(independent of η_1).

Letting $\eta_1 \to \eta_1^0$ in (*) and using the dominated convergence theorem therefore gives

$$\phi'(\eta_1^0) = \int T_1(x) e^{\eta_1^0 T_1(x)} \mu(dx), \tag{1.3.5}$$

where the integral exists and is finite $\forall \eta_1^0$ which is the first component of some η^0 for which $Re(\eta^0) \in \mathcal{N}$.

Applying the same argument to (1.3.5) which we applied to (1.3.4) $\Rightarrow$ existence of all derivatives $\Rightarrow$ (i) and (ii). $\square$

THEOREM 1.3.14. *For an exponential family of order s in canonical form and $\eta \in int(\mathcal{N})$, where $\mathcal{N}$ is the natural parameter space,*

(i) $E_\eta(T) = \frac{\partial A}{\partial \eta} = \left(\frac{\partial A}{\partial \eta_1}, \cdots, \frac{\partial A}{\partial \eta_s} \right)^T$, and

(ii) $\text{Cov}_\eta(T) = \frac{\partial^2 A}{\partial \eta \partial \eta^T} = \left[\frac{\partial^2 A}{\partial \eta_i \partial \eta_j} \right]_{i,j=1}^s$.

PROOF. From theorem 1.3.11

$$e^{A(\eta)} = \int e^{\eta \cdot t} \nu(dt) = \int e^{\eta \cdot T(x)} h(x) \mu(dx)$$

so

(i) $\frac{\partial A}{\partial \eta_i} e^{A(\eta)} = \int T_i(x) e^{\eta \cdot T(x)} h(x) \mu(dx)$
whence $E_\eta T_i = \frac{\partial A}{\partial \eta_i}$.

(ii) $\frac{\partial^2 A}{\partial \eta_i \partial \eta_j} e^{A(\eta)} + \frac{\partial A}{\partial \eta_i} \frac{\partial A}{\partial \eta_j} e^{A(\eta)} = \int T_i(x) T_j(x) e^{\eta \cdot T(x)} h(x) \mu(dx)$
i.e. $\frac{\partial^2 A}{\partial \eta_i \partial \eta_j} = E_\eta(T_i T_j) - E_\eta(T_i) E_\eta(T_j) = Cov_\eta(T_i, T_j)$

□

Higher order moments of $T_1, \cdots, T_s$ are frequently required, e.g.

$$\begin{aligned} \alpha_{r_1,\ldots,r_s} &= E(T_1^{r_1} \cdots T_s^{r_s}) \\ \mu_{r_1,\ldots,r_s} &= E[(T_1 - E(T_1))^{r_1} \cdots (T_s - E(T_s))^{r_s}] \end{aligned}$$

etc. These can often be obtained readily from the MGF:

$$M_T(u_1, \ldots, u_s) := E(e^{u_1 T_1 + \cdots + u_s T_s})$$

If M_T exists in some neighborhood of $\underline{0}$ ($\sum u_i^2 < \delta$), then all the moments $\alpha_{r_1,\cdots,r_s}$ exist and are the coefficients in the power series expansion

$$M_T(u_1, \ldots, u_s) = \sum_{r_1,\ldots,r_s}^{\infty} \alpha_{r_1,\cdots,r_s} \frac{u_1^{r_1} \cdots u_s^{r_s}}{r_1! \cdots r_s!}$$

The cumulant generating function, CGF, sometimes more convenient for calculations (especially for sums of independent random vectors), is defined as

$$K_T(u_1, \ldots, u_s) := \log M_T(u_1, \ldots, u_s).$$

If M_T exists in a neighborhood of $\underline{0}$, then so does K_T and

$$K_T(u_1, \ldots, u_s) = \sum_{r_1,\ldots,r_s=0}^{\infty} \kappa_{r_1,\ldots,r_s} \frac{u_1^{r_1} \cdots u_s^{r_s}}{r_1! \cdots r_s!},$$

where the coefficients $\kappa_{r_1,\ldots,r_s}$ are called the *cumulants* of T.
The moments and cumulants can be found from each other by formal comparison of the two series, and can be retrived from their respective generating functions as follows:

$$\alpha_{r_1 \cdots r_s} = \left. \frac{\partial^{r_1+\cdots+r_s} M_T(u_1, \ldots, u_s)}{\partial u_1^{r_1} \cdots \partial u_s^{r_s}} \right|_{\underline{0}}, \qquad \kappa_{r_1 \cdots r_s} = \left. \frac{\partial^{r_1+\cdots+r_s} K_T(u_1, \ldots, u_s)}{\partial u_1^{r_1} \cdots \partial u_s^{r_s}} \right|_{\underline{0}}.$$

For an exponential family, computation of these generating functions is particularly easy.

THEOREM 1.3.15. *If X has the density*

$$p_\eta(x) = \exp[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)]h(x)$$

w.r.t. some σ-finite measure μ, then for any $\eta \in int(\mathcal{N})$ the MGF and CGF of T exist in a neighborhood of $\underline{0}$ and

$$\begin{aligned} K_T(u) &= A(\eta+u) - A(\eta) \\ M_T(u) &= e^{A(\eta+u)-A(\eta)} \end{aligned}$$

PROOF. HW problem. □

Summary on Exponential Families. The family of probability measures $\{P_\theta\}$ with densities relative to some σ-finite measure μ,

$$p_\theta(x) = \frac{dP_\theta}{d\mu}(x) = \exp\{\sum_{1}^{s} \eta_i(\theta)T_i(x) - B(\theta)\}h(x),\ \theta \in \Omega, \tag{1.3.6}$$

is an ***s*-parameter exponential family**

By redefining the functions $T_i(\cdot)$ and $\eta_i(\cdot)$ if necessary, we can always arrange for both sets of functions to be affinely independent. The number of summands in the exponent is then the **order** of the exponential family.

If $\{1, \eta_1, \ldots, \eta_s\}$ and $\{1, T_1, \ldots, T_s\}$ are both L.I., then the family is said to be **minimal** (but this does not imply minimal sufficiency), and

$$\begin{aligned} s &= dim(span\{1, \log \frac{dp_\theta}{dp_{\theta_0}}(\cdot), \theta \in \Omega\}) - 1 \\ &= \text{order of the exponential family} \end{aligned}$$

REMARK 1.3.16. Since (1.3.6) is by definition a probability density w.r.t. μ for each $\theta \in \Omega$, we have

$$\int \exp\left\{\sum \eta_i(\theta)T_i(x) - B(\theta)\right\} h(x)\mu(dx) = 1$$

$$\therefore \exp B(\theta) = \int \exp\left\{\sum \eta_i(\theta)T_i(x)\right\} h(x)\mu(dx)$$

which shows that the dependence of B on θ is through $\eta(\theta) = (\eta_1(\theta), \ldots, \eta_s(\theta))$ only, i.e. $B(\theta) = A(\eta(\theta))$.

REMARK 1.3.17. The previous note implies that each member of the family (1.3.6) is a member of the family.

$$\pi_\xi(x) = \exp\{\sum_{1}^{s} \xi_i T_i(x) - A(\xi)\}h(x),\ \xi = (\xi_1, \ldots, \xi_s) \in \eta(\Omega) \tag{1.3.7}$$

(in fact $p_\theta(x) = \pi_{\eta(\theta)}(x)$).

The family of densities $\{\pi_\xi, \xi \in \eta(\Omega)\}$ defined by (1.3.7) is the **canonical family associated with (1.3.6)**. It is the same family parameterized by the natural parameter, ξ =vector of coefficients of $T_i(x),\ i = 1, \ldots, s$.

REMARK 1.3.18. Instead of restricting ξ to the set $\eta(\Omega)$, it is natural to extend the family (1.3.7) to allow *all* $\xi \in \mathbb{R}^s$ for which we can choose a value of $A(\xi)$ to make (1.3.7) a probability density, i.e. for which

$$\int \exp\{\sum \xi_i T_i(x)\} h(x)\mu(dx) < \infty \tag{1.3.8}$$

$\mathcal{N} = \{\xi \in \mathbb{R}^s \colon (1.3.8)\ \textit{holds}\}$ is the **natural parameter space** of the family (1.3.7).

REMARK 1.3.19. $\mathcal{N} \supseteq \eta(\Omega)$ since (1.3.7) is by definition a family of probability densities.

DEFINITION 1.3.20. (**Full rank family**) As with the original parameterization, we can always redefine ξ to ensure that $\{T_1, \ldots, T_s\}$ is A.I. If $\eta(\Omega)$ **contains an s-dimensional rectangle and $\{T_1(\cdot), \ldots, T_s(\cdot)\}$ is A.I.**, then T is minimal sufficient and we say the family (1.3.7) is of full rank. (A full rank family is clearly minimal.)

REMARK 1.3.21. Since $\mathcal{N} \supseteq \eta(\Omega)$, full rank $\Rightarrow int(\mathcal{N}) \neq \phi$ and this is important in view of the consequence of theorem 1.3.13 that

$$e^{A(\xi)} = \int \exp(\sum_{i=1}^{s} \xi_i T_i(x)) h(x)\mu(dx)$$

is analytic in each ξ_i on the set of s-dimensional complex vectors, $\xi \colon Re(\xi) \in int(\mathcal{N})$. (So derivatives of $e^{A(\xi)}$ w.r.t. $\xi_i,\ i = 1, \ldots, s$ of all orders can be obtained by differentiation under the integral, yielding explicit expressions for the moments of T for all values of the canonical parameter vector $\xi \in int(\mathcal{N})$.)

EXAMPLE 1.3.22. **Multinomial** $X \sim M(\theta_0, \ldots, \theta_s;\ n) = (X_0, \ldots, X_s)$, where $X_i =$ number of outcomes of type i in n independent trials where $\theta_i,\ i = 0, \ldots, s$, is the probability of an outcome of type i on any one trial.

$$\Omega = \{\theta \colon \theta_0 \geq 0, \cdots, \theta_s \geq 0, \quad \theta_0 + \cdots + \theta_s = 1\}$$

(1) Probability density with respect to counting measure on $\mathbb{Z}_+^{s+1}$

$$\begin{aligned} p_\theta(x) &= \frac{n!}{x_0! \cdots x_s!} \theta_0^{x_0} \cdots \theta_s^{x_s} \prod_{i=0}^{s} I_{[0,\,n]}(x_i) I_{\{n\}}(\sum x_i) \\ &= \exp\{\sum_{i=0}^{s} x_i \log \theta_i\} h(x), \quad \theta \in \Omega. \end{aligned}$$

This is an $(s+1)$-parameter exponential family with $T_i(x) = x_i, \eta_i(\theta) = \log\theta_i$. The vectors $\eta(\theta),\ \theta \in \Omega$, are not confined to a proper affine subspace of $\mathbb{R}^s$, so T is minimal sufficient.

(2) $\{T_0, \ldots, T_s\}$ is not A.I. since $T_0 + \cdots + T_s = n$. Setting $T_0(x) = x_0 = n - x_1 - \cdots - x_n$ gives

$$p_\theta(x) = h(x)\exp\{n\log\theta_0 + \sum_{i=1}^{s} x_i \log\frac{\theta_i}{\theta_0}\}$$

Redefining $\eta(\theta) = (\log\frac{\theta_1}{\theta_0}, \cdots, \log\frac{\theta_s}{\theta_0})$, we now have an s-parameter representation in which $\{T_1, \ldots, T_s\}$ is A.I., since the vectors $(x_1, \cdots, x_s), x \in \mathcal{X}$, are subject only to the constraints $x_i \geq 0$ and $\sum_{i=1}^{s} x_i \leq n$.

(3) Furthermore the new parameter vectors, $\eta(\theta) = (\log\frac{\theta_1}{\theta_0}, \cdots, \log\frac{\theta_s}{\theta_0})$, $\theta \in \Omega$, are not confined to any proper affine subspace of $\mathbb{R}^s$, since for any $x \in \mathbb{R}^s\ \exists\, \theta_0, \ldots, \theta_s$ such that $\eta(\theta) = x$ and so $\eta(\Omega) = \mathbb{R}^s$. Hence $T(x) = (x_1, \ldots, x_s)$ is minimal sufficient for $\mathcal{P}$ and the order of the family is s.

(4) The canonical representation of the family (2) is

$$\pi_\xi(x) = \exp\{\sum_{1}^{s} \xi_i x_i - A(\xi)\}h(x), \quad \xi \in \eta(\Omega) = \{(\log\frac{\theta_1}{\theta_0}, \cdots, \log\frac{\theta_s}{\theta_0}) \colon \theta \in \Omega\}$$

We know from remark 1.3.16 before that $B(\theta) = A(\eta(\theta))$ for some function $A(\cdot)$. Although it is not necessary, we can verify this directly in this example, since from the representation (2) we have

$$B(\theta) = -n\log\theta_0$$

and

$$\begin{aligned} \theta_0 = 1 - \theta_1 - \cdots - \theta_s \Rightarrow \frac{1}{\theta_0} &= 1 + \frac{\theta_1}{\theta_0} + \cdots + \frac{\theta_s}{\theta_0} \\ &= 1 + e^{\eta_1(\theta)} + \cdots + e^{\eta_s(\theta)} \\ \Rightarrow B(\theta) &= n\log(1 + e^{\eta_1(\theta)} + \cdots + e^{\eta_s(\theta)}) \\ \Rightarrow A(\xi) = n\log(1 + e^{\xi_1} + \cdots + e^{\xi_s}) & \end{aligned}$$

$A(\xi)$ is of course also determined by

$$e^{A(\xi)} = \int \exp\{\sum_{1}^{s} \xi_i x_i\}h(x)d\mu(x)$$

.

(5) The natural parameter space in this case is $\mathcal{N} = \mathbb{R}^s$, since we know that $\mathcal{N} \supseteq \eta(\Omega)$ and $\eta(\Omega) = \mathbb{R}^s$ by (3) above. Clearly $\mathcal{N}$ contains an s-dimensional rectangle and $\{T_1, \ldots, T_s\}$ is A.I., hence $\{\pi_\xi(x), \xi \in \mathcal{N}\}$ is of full rank.

(6) Moments of $T(X) = (X_1, \ldots, X_s)$

$$\text{Theorem 1.3.14} \Rightarrow E_\xi T_i = \frac{\partial A}{\partial \xi_i} \quad \forall \xi \in \mathbb{R}^s$$

$$= \frac{ne^{\xi_i}}{1 + e^{\xi_1} + \cdots + e^{\xi_s}}$$

$$= \frac{n\theta_i/\theta_0}{1 + \frac{\theta_1}{\theta_0} + \cdots + \frac{\theta_s}{\theta_0}}$$

$$= n\theta_i$$

$$\text{and} \quad Cov(T_i, T_j) = \frac{\partial^2 A}{\partial \xi_i \partial \xi_j}$$

$$= \begin{cases} \frac{-ne^{\xi_i} e^{\xi_j}}{(1+e^{\xi_1}+\cdots+e^{\xi_s})^2} = -n\theta_i\theta_j & i \neq j \\ \frac{ne^{\xi_i}}{(1+\cdots+e^{\xi_s})} - \frac{ne^{2\xi_i}}{(1+\cdots+e^{\xi_s})^2} = n\theta_i(1-\theta_i) & i = j \end{cases}$$

(Moments exist $\forall \xi \in int(\mathcal{N}) = \mathbb{R}^s$)

THEOREM 1.3.23. **(Sufficient condition for completeness of T)** *If*

$$\pi_\xi(x) = \exp\left(\sum_{i=1}^{s} \xi_i T_i(x) - A(\xi)\right) h(x), \quad \xi \in \eta(\Omega)$$

is a minimal canonical representation of the exponential family $\mathcal{P} = \{p_\theta : \theta \in \Omega\}$ and $\eta(\Omega)$ contains an open subset of $\mathbb{R}^s$, then $T = (T_1, \ldots, T_s)$is complete for $\mathcal{P}$.

PROOF. Suppose $E_\xi(f(T)) = 0 \; \forall \xi \in \eta(\Omega)$. Then,

$$E_\xi f^+(T) = E_\xi f^-(T) \; \forall \xi \in \eta(\Omega). \tag{1.3.9}$$

Choose $\xi_0 \in int(\eta(\Omega))$ and $r > 0$ such that

$$N(\xi_0, r) := \{\xi : ||\xi - \xi_0|| < r\} \subseteq \eta(\Omega).$$

Now define the probability measures,

$$\lambda^+(A) = \frac{\int_A f^+ e^{\xi_0 \cdot t} \nu(dt)}{\int_{\mathcal{J}} f^+ e^{\xi_0 \cdot t} \nu(dt)}, \quad \nu = \tilde{\mu} \circ T^{-1}, \; d\tilde{\mu}(x) = h(x)\mu(dx),$$

$$\lambda^-(A) = \frac{\int_A f^- e^{\xi_0 \cdot t} \nu(dt)}{\int_{\mathcal{J}} f^- e^{\xi_0 \cdot t} \nu(dt)},$$

where we have assumed that $\nu(\{t\colon f(t) \neq 0\}) > 0$, since otherwise $f = 0$ *a.s.* $\mathcal{P}_T$ and we are done.

Observe now that

$$\int e^{\delta \cdot t} \lambda^+(dt) = \int e^{\delta \cdot t} \lambda^-(dt) \quad \forall \delta \in \mathbb{R}^s \; with \; ||\delta|| < r \tag{1.3.10}$$

since

$$\begin{aligned} L.S. &= \int_{\mathcal{J}} f^+(t)e^{(\xi_0+\delta)\cdot t}\nu(dt) / \int_{\mathcal{J}} f^+(t)e^{\xi_0\cdot t}\nu(dt) \\ &= \int_{\mathcal{J}} f^-(t)e^{(\xi_0+\delta)\cdot t}\nu(dt) / \int_{\mathcal{J}} f^-(t)e^{\xi_0\cdot t}\nu(dt) \end{aligned}$$

by (1.3.9)

Now consider each side of (1.3.10) as a function of the complex argument $\delta = \delta_0 + i\theta$, $\theta \in \mathbb{R}^s$. Then

$$L(\delta) = R(\delta) \quad \forall \delta = \delta_0 + i \cdot \theta$$

with $||\delta_0|| < r$, since (by Theorem 1.3.13 (i)) both sides are analytic in each component of δ on the set where $\mathcal{R}e(\xi_0 + \delta) \in \mathcal{N}$ and they are equal when δ is real. In particular,

$$L(i\theta) = \int e^{i\theta\cdot t}\lambda^+(dt) = R(i\theta) = \int e^{i\theta\cdot t}\lambda^-(dt)$$

for all $\theta \in \mathbb{R}^s$. Hence λ^+ and λ^- have the same characteristic function $\Rightarrow \lambda^+ = \lambda^- \Rightarrow f^+ = f^-$ *a.s.*, contradicting $\nu(f \neq 0) > 0$. So $f = 0$ *a.s.* ν. □

EXAMPLE 1.3.24. $X_1, \ldots, X_n$ *iid* $N(\mu, \sigma^2)$, with σ^2 known.

$$p_\mu(x) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\{\frac{\mu}{\sigma^2}\sum x_i - \frac{1}{2\sigma^2}\sum x_i^2 - \frac{n}{2}\},$$

$$\eta(\mu) = \frac{\mu}{\sigma^2}, \quad T(x) = \sum x_i$$

Since $\eta(\Omega) = \mathbb{R}$ contains a 1-dim rectangle in $\mathbb{R}$, $T(x) = \sum x_i$ is complete (and sufficient, or CSS).

EXAMPLE 1.3.25. $X_1, \ldots, X_n$ *iid* $N(\sigma, \sigma^2)$

$$p_\sigma(x) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\{-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{1}{\sigma}\sum x_i - \frac{n}{2}\},$$

$$\eta_1(\sigma) = \frac{1}{2\sigma^2}, \quad T_1(x) = -\sum x_i^2$$

$$\eta_2(\sigma) = \frac{1}{\sigma}, \quad T_2(x) = \sum x_i$$

Since $\eta(\Omega)$ does not contain a 2-dim rectangle in $\mathbb{R}^2$, the theorem is silent about completeness. In fact, we can show that $T(x) = (\sum x_i^2, \sum x_i)$ is not complete since

$$E_\theta\left[\sum x_i^2 - \frac{2}{n+1}(\sum x_i)^2\right] = n(2\sigma^2) - \frac{2}{n+1}(n\sigma^2 + n^2\sigma^2) = 0, \quad \forall\sigma,$$

but there exists no $\mathcal{P}$-null set N such that $\sum x_i^2 - \frac{2}{n+1}(\sum x_i)^2 = 0$ on N^c.

1.4. Convex Loss Function

LEMMA 1.4.1. *Let ϕ be a convex function on $(-\infty,\infty)$ which is bounded below and suppose that ϕ is not monotone. Then, ϕ takes on its minimum value c and $\phi^{-1}(c)$ is a closed interval and is a singleton when ϕ is strictly convex.*

PROOF. Since ϕ is convex and not monotone,

$$\lim_{x\to\pm\infty}\phi(x)=\infty.$$

Since ϕ is continuous, ϕ attains its minimum value c. $\phi^{-1}(\{c\})$ is closed by continuity, and is an interval by convexity. The interval must have zero length if ϕ is strictly convex. □

THEOREM 1.4.2. *Let ρ be a convex function defined on $(-\infty,\infty)$ and X a random variable such that $\phi(a)=E(\rho(X-a))$ is finite for some a. If ρ is not monotone, $\phi(a)$takes on its minimum value and $\phi^{-1}(a)$ is a closed set and is a singleton when ρ is strictly convex.*

PROOF. By the lemma, we only need to show that ϕ is convex and not monotone. Because $\lim_{t\to\pm\infty}\rho(t)=\infty$ and $\lim_{a\to\pm\infty}x-a=\pm\infty$,

$$\lim_{a\to\pm\infty}\phi(a)=\infty$$

so that ϕ is not monotone.

The convexity comes from

$$\begin{aligned}\phi(pa+(1-p)b) &= E\rho(p(X-a)+(1-p)(X-b))\\ &\leq E(p\rho(X-a)+(1-p)\rho(X-b))\\ &= p\phi(a)+(1-p)\phi(b).\end{aligned}$$

□

1.5. Model Selection

Throughout the course we assume the family $\mathcal{P}$ is known *a priori*, so that the model to be fitted to the data $\{X_1,\ldots,X_n\}$ is correct.

CHAPTER 2

Unbiasedness

2.1. UMVU estimators.

Notation. $\mathcal{P}$=$\{P_\theta,\, \theta \in \Omega\}$ is a family of probability measures on $\mathcal{A}$ (distributions of X).

T:$\mathcal{X} \to \mathbb{R}$ is an $\mathcal{A}/\mathcal{B}$ measurable function and T (or $T(X)$) is called a statistic.

$g : \Omega \to \mathbb{R}$ is a function on Ω whose value at θ is to be estimated.

$$(\mathcal{X}, \mathcal{A}, P_\theta) \xrightarrow{X} (\mathcal{X}, \mathcal{A}, P_\theta) \xrightarrow{T} \left(\mathbb{R}, \mathcal{B}, P_\theta^T\right)$$

DEFINITION 2.1.1. A statistic T (or $T(X)$) is called an unbiased estimator of $g(\theta)$ if

$$E_\theta(T(X)) = g(\theta) \text{ for all } \theta \in \Omega.$$

Objectives of point estimation. In order to specify what we mean by a good estimator of $g(\theta)$, we need to specify what we mean when we say that $T(X)$ is close to $g(\theta)$. A fairly general way of defining this is to specify a *loss function:*

$L(\theta, d)$ = cost of concluding that $g(\theta) = d$, when the parameter value is θ.
$L(\theta, d) \geqslant 0$ and $L(\theta, g(\theta)) = 0$.

Since $T(X)$ is a random variable, we measure the performance of $T(X)$ for estimating $g(\theta)$ in terms of its expected (or long-term average) loss

$$R(\theta, T) = E_\theta L(\theta, T(X)),$$

known as the *risk function.*

Choice of a loss function will depend on the problem and the purpose of the estimation. For many estimation problem, the conclusion is not particularly sensitive to the choice of loss function within a reasonable range of alternatives. Because of this and especially because of its mathematical convenience, we often choose (and will do so in this chapter) the *squared-error loss function*

$$L(\theta, d) = (g(\theta) - d)^2$$

with corresponding risk function

$$R(\theta, T) = E_\theta(T(X) - g(\theta))^2 \tag{2.1.1}$$

Ideally we would like to choose T to minimize (2.1.1) uniformly in θ. Unfortunately this is impossible since the estimator T defined by

$$T(x) = g(\theta_0) \quad \forall x \in \mathcal{X} \tag{2.1.2}$$

(where θ_0 is some fixed parameter value in Ω) has the risk function,

$$R(\theta, T) = \begin{cases} 0 & if \quad \theta = \theta_0 \\ (g(\theta) - g(\theta_0))^2 & if \quad \theta \neq \theta_0 \end{cases}$$

An estimator which simultaneously minimized $R(\theta, T)$ for all $\theta \in \Omega$ would necessarily have $R(\theta, T) = 0 \; \forall \theta \in \Omega$ and this is impossible except in trivial cases.

Why consider the class of unbiased estimators? There is nothing intrinsically good about unbiased estimators. The only criterion for goodness is that $R(\theta, T)$ should be small. The hope is that by restricting attention to a class of estimators which excludes (2.1.2), we may be able to minimize $R(\theta, T)$ uniformly in θ and that the resulting estimator will give small values of $R(\theta, T)$. This programme is frequently successful if we attempt to minimize $R(\theta, T)$ with T restricted to the class of unbiased estimators of $g(\theta)$.

DEFINITION 2.1.2. $g(\theta)$ is *U-estimable*, if there exists an unbiased estimator of $g(\theta)$.

EXAMPLE 2.1.3. $X_1, \ldots, X_n$ iid Bernoulli(p), $p \in (0, 1)$. $g(p) = p$ is U-estimable, since $E\bar{X}_n = p \;\; \forall p \in (0, 1)$, while $h(p) = \frac{1}{p}$ is not U-estimable, since if

$$\sum T(x) p^{\sum x_i} (1-p)^{n - \sum x_i} = \frac{1}{p} \;\; \forall p \in (0, 1),$$

$\lim_{p \to 0} RS = \infty$ and $\lim_{p \to 0} LS = T(0)$. So $T(0) = \infty$, but this is not possible since then $E_p T(X) = \infty \neq \frac{1}{p} \;\; \forall p \in (0, 1)$.

REMARK 2.1.4. $\frac{\sum_{i=1}^n X_i}{n} \xrightarrow{a.s.} p$ and $\frac{n}{\sum_{i=1}^n X_i} \xrightarrow{a.s.} p^{-1} \quad \forall p \in (0, 1)$. Hence $\frac{n}{\sum X_i}$ is a reasonable estimate of p^{-1} even though it is not unbiased.

THEOREM 2.1.5. *If T_0 is an unbiased estimator of $g(\theta)$ then the totality of unbiased estimators of $g(\theta)$ is given by*

$$\{T_0 - U : E_\theta U = 0 \text{ for all } \theta \in \Omega\}.$$

PROOF. If T is unbiased for $g(\theta)$, then $T = T_0 - (T_0 - T)$ where $E_\theta(T_0 - T) = 0 \; \forall \theta \in \Omega$. Conversely if $T = T_0 - U$ where $E_\theta U = 0 \; \forall \theta \in \Omega$, then $E_\theta T = E_\theta T_0 = g(\theta) \; \forall \theta \in \Omega$. $\square$

REMARK 2.1.6. For *squared error loss*, $L(\theta, d) = (d - g(\theta))^2$, the risk $R(\theta, T)$ is

$$\begin{aligned} R(\theta, T) &= E_\theta((T(X) - g(\theta))^2) \\ &= Var_\theta(T(X)) \text{ if } T \text{ is unbiased} \\ &= Var_\theta(T_0(X) - U) \\ &= E_\theta[(T_0(X) - U)^2] - g(\theta)^2 \end{aligned}$$

and hence the risk is minimized by minimizing $E_\theta[(T_0(X) - U)^2]$ with respect to U, i.e. by taking any *fixed* unbiased estimator of $g(\theta)$ and finding the unbiased estimator of zero which minimizes $E_\theta[(T_0(X) - U)^2]$. Then if U does not depend on θ we shall have found a uniformly minimum risk estimator of $g(\theta)$, while if U depends on θ, there is no uniformly minimum risk estimator. Note that for unbiased estimators and squared error loss, the risk is the same as the variance of the estimator, so uniformly minimum risk unbiased is the same as uniformly minimum variance unbiased in this case.

EXAMPLE 2.1.7. $P(X = -1) = p,\ P(X = k) = q^2 p^k,\ k = 0, 1, \ldots$, where $q = 1 - p$.

$$\begin{aligned} T_0(X) &= I_{\{-1\}}(X) \text{ is unbiased for } p,\ 0 < p < 1 \\ T_1(X) &= I_{\{0\}}(X) \text{ is unbiased for } q^2, \end{aligned}$$

U is unbiased for 0

$$\begin{aligned} \Leftrightarrow 0 &= \sum_{k=-1}^{\infty} U(k) P(X = k) = pU(-1) + \sum_{k=0}^{\infty} U(k) q^2 p^k \quad \forall p \\ &= U(0) + \sum_{k=1}^{\infty} (U(k) - 2U(k-1) + U(k-2)) p^k \\ \Leftrightarrow U(k) &= -kU(-1) = ka \text{ for some } a \\ &\quad \text{(comparing coefficients of } p^k,\ k = 0, 1, 2, \ldots) \end{aligned}$$

So an unbiased estimator of p with minimum risk (i.e. variance) is $T_0(X) - a_0^* X$ where a_0^* is the value of a which minimizes

$$E_p(T_0(X) - aX)^2 = \sum P_p(X = k)[T_0(k) - ak]^2$$

Similarly an unbiased estimator of q^2 with minimum risk (i.e. variance) is $T_1(X) - a_1^* X$ where a_1^* is the value of a which minimizes

$$E_p(T_1(X) - aX)^2 = \sum P_p(X = k)[T_1(k) - ak]^2$$

Some straightforward calculations give

$$a_0^* = \frac{-p}{p + q^2 \sum_1^\infty k^2 p^k} \quad \text{and} \quad a_1^* = 0$$

Since a_1^* is independent of p, the estimator $T_1(X)$ of q^2 is minimum variance unbiased for all p, i.e. UMVU. However a_0^* does depend on p and so the estimator $T_0^*(X) = T_0(X) - a_0^* X$ is only **locally minimum variance unbiased** at p. (We are using estimator

in a generalized sense here since $T_0^*(X)$ depends on p. We shall continue to use this terminology.) An UMVU estimator of p does not exist in this case.

DEFINITION 2.1.8. Let $V(\theta) = \inf_T Var_\theta(T)$ where the inf is over all unbiased estimators of $g(\theta)$. If an unbiased estimator T of $g(\theta)$ satisfies

$$Var_\theta(T) = V(\theta) \quad \forall \theta \in \Omega \qquad \text{it is called } UMVU$$

If

$$Var_{\theta_0} T = V(\theta_0) \quad \text{for some } \theta_0 \in \Omega \qquad T \text{ is called } LMVU \text{ at } \theta_0$$

REMARK 2.1.9. Let $\mathcal{H}$ be the Hilbert space of functions on $\mathcal{X}$ which are square integrable with respect to $\mathcal{P}$ (i.e. with respect to every $P_\theta \in \mathcal{P}$), and let $\mathcal{U}$ be the set of all unbiased estimators of 0. If T_0 is an unbiased estimator of $g(\theta)$ in $\mathcal{H}$, then a LMVU estimator in $\mathcal{H}$ at θ_0 is $T_0 - P_{\mathcal{U}}(T_0)$, where $P_{\mathcal{U}}$ denotes orthogonal projection on $\mathcal{U}$ in the inner product space $L^2(P_{\theta_0})$, i.e. $P_{\mathcal{U}}(T_0)$ is the unique element of $\mathcal{U}$ such that

$$T_0 - P_{\mathcal{U}}(T_0) \perp \mathcal{U} \quad (in\ L^2(P_{\theta_0})).$$

$T_0 - P_{\mathcal{U}}(T_0)$ is LMVU since $P_{\mathcal{U}}(T_0) = \arg\min_{U \in \mathcal{U}} E_{\theta_0}(T_0 - U)^2$.

NOTATION 2.1.10. We denote the set of all estimators T with $E_\theta T^2 < \infty$ for all $\theta \in \Omega$ by Δ and the set of all unbiased estimators of 0 in Δ by $\mathcal{U}$.

THEOREM 2.1.11. *An unbiased estimator $T \in \Delta$ of $g(\theta)$ is UMVU iff*

$$E_\theta(TU) = 0 \text{ for all } U \in \mathcal{U} \text{ and for all } \theta \in \Omega.$$

(i.e. $Cov_\theta(T, U) = 0$ since $E_\theta U = 0$ for all θ and $E_\theta T = g(\theta)$ for all $\theta \in \Omega$.)

PROOF. ($\Rightarrow$) Suppose T is UMVU. For $U \in \mathcal{U}$, let $T' = T + \lambda U$ with λ real. Then T' is unbiased and, by definition of T,

$$Var_\theta(T') = Var_\theta(T) + \lambda^2 Var_\theta(U) + 2\lambda Cov_\theta(T, U) \geqslant Var_\theta(T)$$

therefore, $\lambda^2 Var_\theta(U) + 2\lambda Cov_\theta(T, U) \geqslant 0$. Setting $\lambda = -\frac{Cov_\theta(T,U)}{Var_\theta(U)}$ gives a contradiction to this inequality unless $Cov_\theta(T, U) = 0$. Hence $Cov_\theta(T, U) = 0$.

($\Leftarrow$) If $E_\theta(TU) = 0\ \forall U \in \mathcal{U}$ and $\forall \theta \in \Omega$, let T' be any other unbiased estimator. If $Var_\theta(T') = \infty$, then $Var_\theta(T) < Var_\theta(T')$, so suppose $Var_\theta(T') < \infty$.
Then $T' = T - U$, for some U which is unbiased for 0 (by Theorem 2.1.5).

$$\begin{aligned} U = T - T' \Rightarrow E_\theta U^2 &= E_\theta(T' - T)^2 \\ &\leqslant 2E_\theta T'^2 + 2E_\theta T^2 < \infty \\ &\Rightarrow U \in \mathcal{U} \end{aligned}$$

Hence

$$\begin{aligned} Var_\theta(T') &= Var_\theta(T-U) \\ &= Var_\theta(T) + Var_\theta(U) - 2Cov_\theta(T,U) \\ &\geqslant Var_\theta(T) \text{ since } Cov_\theta(T,U) = 0, \\ &\Rightarrow T \text{ is UMVU.} \end{aligned}$$

□

Unbiasedness and sufficiency. Suppose now that $T \in \Delta$ is unbiased for $g(\theta)$ and S is sufficient for $\mathcal{P} = \{P_\theta,\ \theta \in \Omega\}$. Consider

$$T' = E_\theta(T|S) = E(T|S) \quad \text{independent of } \theta$$

Then

(a)

$$E_\theta T' = E_\theta E(T|S) = E_\theta(T) = g(\theta)\ \forall\theta.$$

(b)

$$\begin{aligned} Var_\theta(T) &= E_\theta(T - E(T|S) + E(T|S) - g(\theta))^2 \\ &= E_\theta((T-E(T|S))^2) + Var_\theta(T') + 2E_\theta[(T-E(T|S))(E(T|S)-g(\theta))] \\ &\geqslant Var_\theta(T'). \end{aligned}$$

On the second line we used the fact that $T - E(T|S)$ is orthogonal to $\sigma(S)$. The inequality on the third line is strict for all $\theta \Leftrightarrow T = E(T|S)$ *a.s.* $\mathcal{P}$.

THEOREM 2.1.12. *If S is a complete sufficient statistic for $\mathcal{P}$, then every U-estimable function $g(\theta)$ has one and only one unbiased estimator which is a function of S.*

PROOF.

$$\begin{aligned} T \text{ unbiased} &\Rightarrow E(T|S) \text{ is unbiased and a function of } S \\ T_1(S), T_2(S) \text{ unbiased} &\Rightarrow E_\theta(T_1(S) - T_2(S)) = 0 \quad \forall\theta \\ &\Rightarrow T_1(S) = T_2(S) \quad a.s.\ \mathcal{P} \text{ (completeness)} \end{aligned}$$

□

THEOREM 2.1.13. (**Rao-Blackwell**) *Suppose S is a complete sufficient statistic for $\mathcal{P}$. Then*

(i) If $g(\theta)$ is U-estimable, there exists an unbiased estimator which uniformly minimizes the risk for any loss function $L(\theta, d)$ which is convex in d.
(ii) The $UMVU$ in (i) is the unique unbiased estimator which is a function of S; it is the **unique** unbiased estimator with minimum risk provided the risk is finite and L is strictly convex in d.

PROOF. (i) $L(\theta, d)$ convex in d means

$$L(\theta, pd_1 + (1-p)d_2) \leqslant pL(\theta, d_1) + (1-p)L(\theta, d_2),\ 0 < p < 1.$$

Let T be any unbiased estimator of $g(\theta)$ and let $T' = E(T \mid S)$, another unbiased estimator of $g(\theta)$. Then

$$\begin{aligned} R(\theta, T') &= E_\theta[L(\theta, E(T \mid S))] \\ &\leqslant E_\theta[E_\theta(L(\theta, T) \mid S)], \text{ by Jensen's inequality for conditional expectation,} \\ &= E_\theta L(\theta, T) = R(\theta, T)\ \forall\theta. \end{aligned}$$

If T_2 is any other unbiased estimator then

$$T_2' = E(T_2 \mid S) = T' \quad \text{a.s. } \mathcal{P} \text{ by Theorem 2.1.12.}$$

Hence starting from any unbiased estimator and conditioning on the CSS S gives a uniquely defined unbiased estimator which is UMVU and is the unique function of S which is unbiased for $g(\theta)$.

(ii) The first statement was established at the end of the proof of (i).

If T is UMVU then so is $T' = E(T \mid S)$ as shown in (i); We will show that T is necessarily the uniquely determined unbiased function of S, by showing that T is a function of S a.s. $\mathcal{P}$.

The proof is by contradiction. Suppose that "T is a function of S a.s. $\mathcal{P}$" is false. Then there exists θ and a set of positive P_θ measure where

$$T' := E(T \mid S) \neq T$$

But this implies that

$$\begin{aligned} R(\theta, T') &= E_\theta(L(\theta, E(T \mid S))) \\ &< E_\theta(E_\theta(L(\theta, T) \mid S)) \\ &\quad \text{(Jensen's inequality is strict unless } E(T \mid S) = T \text{ a.s. } \mathcal{P}_\theta) \\ &= R(\theta, T) \end{aligned}$$

contradicting the UMVU property of T.

□

THEOREM 2.1.14. *If $\mathcal{P}$ is an exponential family of full rank (i.e. $\{\eta_1, \ldots, \eta_s\}$ and $\{T_1, \ldots, T_s\}$ are A.I. and $\eta(\Omega)$ contains an open subset of $\mathbb{R}^s$) then the Rao-Blackwell theorem applies to any U-estimable $g(\theta)$ with $S = T$.*

PROOF. T is complete sufficient for $\mathcal{P}$.
[Some obvious U-estimable $g(\theta)$'s are

$$E_\theta T_i(X) = \frac{\partial A}{\partial \xi_i}\Big|_{\xi=\eta(\theta)}, \quad \{\theta : \eta(\theta) \in int(\mathcal{N})\},$$

where $\pi_\xi(x) = e^{\sum \xi_i T_i(x) - A(\xi)} h(x)$ is the canonical representation of $p_\theta(x)$.] □

Two methods for finding UMVU's

Method 1. Search for a function $\delta(T)$, where T is a CSS, such that

$$E_\theta \delta(T) = g(\theta),\ \forall \theta \in \Omega.$$

EXAMPLE 2.1.15. $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2),\ \mu \in \mathbb{R},\ \sigma^2 > 0.$

$$T = (\bar{X}, S^2) \text{ is CSS.}$$
$$E_{\mu,\sigma^2} \bar{X} = \mu$$
$$\bar{X} \text{ is UMVU for } \mu.$$

Method 2. Search for an unbiased $\delta(X)$ and a CSS T. Then

$$S = E(\delta(X) \mid T) \ is\ UMVU$$

EXAMPLE 2.1.16. $X_1, \ldots, X_n$ *iid* $U(0, \theta),\ \theta > 0$

$$g(\theta) = \frac{\theta}{2}$$
$$\delta_1(X) = X_1 \text{ is unbiased}$$
$$X_{(n)} \text{ is CSS}$$
$$\therefore\ S = E(X_1 \mid X_{(n)}) \text{ is UMVU}$$

To compute S we note that given $X_{(n)} = x$,

$$X_1 = x\ w.p.\ \frac{1}{n}$$
$$X_1 \backsim U(0, x)\ w.p.\ 1 - \frac{1}{n}$$
$$\therefore\ S(x) = \frac{x}{n} + (1 - \frac{1}{n})\frac{x}{2} = \frac{n+1}{n}\frac{x}{2}$$
$$\therefore\ S(X_{(n)}) = \frac{1}{2}\frac{n+1}{n} X_{(n)} \text{ is UMVU for } \frac{\theta}{2}$$
$$\Rightarrow \frac{n+1}{n} X_{(n)} \text{ is UMVU for } \theta$$

REMARK 2.1.17.

(a) Convexity of $L(\theta, \cdot)$ is crucial to the Rao-Blackwell theorem.
(b) Large-sample theory tends to support the use of convex $L(\theta, \cdot)$.

Heuristically if $X_1, \ldots, X_n$ are iid, then as $n \to \infty$ the error in estimating $g(\theta) \to 0$ for any reasonable estimates (in some probabilistic sense). Thus only the behavior of $L(\theta, d)$ for d close to $g(\theta)$ is relevant for large samples.

A Taylor expansion around $d = g(\theta)$ gives

$$L(\theta, d) = a(\theta) + b(\theta)(d - g(\theta)) + c(\theta)(d - g(\theta))^2 + Remainder$$

But

$$L(\theta, g(\theta)) = 0 \Rightarrow a(\theta) = 0$$
$$L(\theta, d) \geq 0 \Rightarrow b(\theta) = 0$$

Hence locally, $L(\theta, d) \backsim c(\theta)(d - g(\theta))^2$, a convex weighted squared error loss function.

EXAMPLE 2.1.18. Observe $X_1, \ldots, X_m$, iid $N(\xi, \sigma^2)$, and $Y_1, \ldots, Y_n$, iid $N(\eta, \tau^2)$, independent of $X_1, \ldots, X_m$.

(i) For the 4-parameter family $\mathcal{P} = \{P_{\xi,\eta,\sigma^2,\tau^2}\}$, $(\bar{X}, \bar{Y}, S_X^2, S_Y^2)$ is a CSS since the exponential family is of full rank. Hence $\bar{X}$ and S_X^2 are UMVU for ξ and σ^2 respectively and $\bar{Y}$ and S_Y^2 are UMVU for η and τ^2.

(ii) For the 3-parameter family $\mathcal{P} = \{P_{\xi,\eta,\sigma^2,\sigma^2}\}$, $(\bar{X}, \bar{Y}, SS)$ is a CSS, where $SS := (m-1)S_X^2 + (n-1)S_Y^2$. Hence $\bar{X}, \bar{Y}$ and $\frac{SS}{m+n-2}$ are UMVU for ξ, η and σ^2 respectively.

(iii) For the 3-parameter family with $\xi = \eta$, $\sigma^2 \neq \tau^2$ (which arises when estimating a mean from 2 sets of readings with different accuracies), $(\bar{X}, \bar{Y}, S_X^2, S_Y^2)$ is minimal sufficient but not complete, since $\bar{X} - \bar{Y} \neq 0$ *a.s.* $\mathcal{P}$, but $E_\theta(\bar{X} - \bar{Y}) = 0 \; \forall \, \theta$.

To deal with Case (iii) we shall first show the following: If $\frac{\sigma^2}{\tau^2} = r$ for some fixed r, i.e.

$$\mathcal{P}^* = \{P_{\xi,\xi,r\tau^2,\tau^2}\}$$

then $T^* = (\sum X_i + r\sum Y_j, \sum X_i^2 + r\sum Y_j^2)$ is CSS

PROOF.

$$p_{\xi,\tau^2}(x,y) = \frac{1}{(2\pi)^{\frac{m+n}{2}}} \frac{1}{(r\tau^2)^{\frac{m}{2}}} \frac{1}{(\tau^2)^{\frac{n}{2}}}$$
$$\times \exp\left\{-\frac{1}{2r\tau^2}\sum x_i^2 + \frac{1}{r\tau^2} m\xi\bar{x} - \frac{m\xi^2}{2r\tau^2} - \frac{1}{2\tau^2}\sum y_i^2 + \frac{1}{\tau^2} n\xi\bar{y} - \frac{n\xi^2}{2\tau^2}\right\}$$
$$= \exp\left\{-A(\xi, \tau^2)\right\} * \exp\left\{-\frac{1}{2r\tau^2}(\sum x_i^2 + r\sum y_i^2) + \frac{\xi}{r\tau^2}(\sum x_i + r\sum y_i)\right\}$$

□

Since T^* is a CSS for $\mathcal{P}^*$ and since $T_1 = \frac{\sum X_i + r\sum Y_i}{m+rn}$ is unbiased for ξ, it is UMVU for ξ in $\mathcal{P}^*$.

T_1 is also unbiased for ξ in $\mathcal{P} = \{P_{\xi,\xi,\sigma^2,\tau^2}\}$

$$\therefore V(\xi_0, \sigma_0^2, \tau_0^2) \leqslant Var_{\xi_0,\sigma_0^2,\tau_0^2}(T_1) = \frac{\sigma_0^2\tau_0^2}{m\tau_0^2 + n\sigma_0^2}, \text{ where } \quad \frac{\sigma_0^2}{\tau_0^2} = r.$$

(V is the smallest variance of all unbiased estimators of ξ for $\mathcal{P}$ evaluated at $\xi_0, \sigma_0^2, \tau_0^2$.)

On the other hand, every T which is unbiased for ξ in $\mathcal{P}$ is also unbiased in $\mathcal{P}^*$. Hence if T is unbiased for ξ in $\mathcal{P}$, then

$$Var_{\xi_0,\sigma_0^2,\tau_0^2}(T) \geqslant Var_{\xi_0,\sigma_0^2,\tau_0^2}(\frac{\sum X_i + r\sum Y_i}{m+rn}), \quad where\ r = \frac{\sigma_0^2}{\tau_0^2},$$

and the inequality continues to hold with the left-hand side replaced by $V(\xi_0, \sigma_0^2, \tau_0^2)$. So $V(\xi_0, \sigma_0^2, \tau_0^2) = \frac{\sigma_0^2\tau_0^2}{m\tau_0^2+n\sigma_0^2}$ and the LMVU estimator at $(\xi_0, \sigma_0^2, \tau_0^2)$ is

$$\frac{\sum X_i + \frac{\sigma_0^2}{\tau_0^2}\sum Y_i}{m + \frac{\sigma_0^2}{\tau_0^2}n}.$$

Since this estimate depends on the ratio $r = \frac{\sigma_0^2}{\tau_0^2}$, an UMVU for ξ does not exist in $\mathcal{P}$.
A natural estimate for ξ is

$$\hat{\xi} = \frac{\sum X_i + \frac{S_X^2}{S_Y^2}\sum Y_i}{m + \frac{S_X^2}{S_Y^2}n}.$$

(See Graybill & Deal, 1959, for its properties.)

2.2. Non-parametric families

Consider $X = (X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are *iid* F, where $F \in \mathcal{F}$, a family of distribution functions, and $\mathcal{P}$ is the corresponding product measure on $(\mathbb{R}^n, \mathcal{B}^n)$. For example,

$$\mathcal{F}_0 = \text{df's with density relative to Lebesgue measure,}$$
$$\mathcal{F}_1 = \text{df's with } \int |x|F(dx) < \infty,$$
$$\mathcal{F}_2 = \text{df's with } \int x^2F(dx) < \infty, \quad etc.$$

The estimand is $g\colon \mathcal{F} \to \mathbb{R}$. For example,

$$g(F) = \int xF(dx) = \mu_F$$
$$g(F) = \int x^2F(dx)$$
$$g(F) = F(a)$$
$$g(F) = F^{-1}(p)$$

PROPOSITION 2.2.1. *If $\mathcal{F}_0$ is defined as above, then $(X_{(1)}, \ldots, X_{(n)})$ is complete sufficient for $\mathcal{F}_0$ (i.e. for the corresponding family of probability measures $\mathcal{P}$).*

PROOF. We know that $T(X) = (X_{(1)}, \ldots, X_{(n)})$ is sufficient for $\mathcal{P}$. It remains to show (by problem 1.6.32, p.72) that T is complete and sufficient for a family $\mathcal{P}_0 \subseteq \mathcal{P}$ such that each member of $\mathcal{P}_0$ has positive density on $\mathbb{R}^n$. Choose $\mathcal{P}_0$ to be the set of probability measures on $\mathcal{B}^n$ with densities relative to Lebesgue measure,

$$C(\theta_1, \cdots, \theta_n) \exp\{\theta_1 \sum x_i + \theta_2 \sum_{i<j} x_i x_j + \cdots + \theta_n x_1 \cdots x_n - \sum x_i^{2n}\}\}$$

This is an exponential family whose natural parameter set $\mathcal{N}$ contains an open set ($\mathcal{N} = \mathbb{R}^n$). So $S(x) = (\sum x_i, \sum_{i<j} x_i x_j, \cdots, x_1 \cdots x_n)$ is complete. But S is equivalent to T (consider the n^{th} degree polynomial whose zeroes are $x_{(1)}, \cdots, x_{(n)}$), so T is complete for $\mathcal{F}_0$. □

Measurable functions of the order statistics. If $T(x) := (x_{(1)}, \ldots, x_{(n)})$ then

$$\delta(X_1, \ldots, X_n) \in \sigma(T) \Leftrightarrow \delta(X_1, \ldots, X_n) = \delta(X_{\pi_1}, \ldots, X_{\pi_n})$$

for every permutation $(\pi_1, \ldots, \pi_n)$ of $(1, \ldots, n)$. Since T is a CSS for $\mathcal{F}_0$, this enables us to identify UMVU estimators of estimands g for which they exist.

EXAMPLE 2.2.2. $g(F) = F(a)$. An obvious unbiased estimator of $F(a)$ is

$$T_1(X) := \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, a]}(X_i)$$

and $T_1 \in \sigma(T)$ so T_1 is UMVU for $F(a)$, $F \in \mathcal{F}_0$.

EXAMPLE 2.2.3. $g(F) = \int x dF,\ F \in \mathcal{F}_0 \cap \mathcal{F}_2$. Let

$$T_2(x) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then $T_2 \in \sigma(T)$ and, since T is also complete for $\mathcal{F}_0 \cap \mathcal{F}_2$, it is therefore UMVU for μ_F.

EXAMPLE 2.2.4. $g(F) = \sigma_F^2,\ F \in \mathcal{F}_0 \cap \mathcal{F}_4$. Let

$$T_3(x) = S(x)^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum (x_{(i)} - \frac{1}{n} \sum x_{(i)})^2}{n-1}$$

$T_3 \in \sigma(T)$ and is unbiased for σ_F^2. Since T is complete for $\mathcal{F}_0 \cap \mathcal{F}_4$, T_3 is UMVU for σ_F^2.

REMARK 2.2.5. T complete for $\mathcal{F}$ does not imply generally that T is complete for $\mathcal{F}^* \subseteq \mathcal{F}$. In fact the reverse is true. Completeness for $\mathcal{F}^*$ implies completeness for $\mathcal{F}$. However the same argument used in the proof of Proposition 2.2.1 shows that

T is complete for $\mathcal{F}_0 \cap \mathcal{F}_2$ (used in example 2.2.3) and

T is complete for $\mathcal{F}_0 \cap \mathcal{F}_4$ (used in example 2.2.4).

EXAMPLE 2.2.6. $g(F) = \mu_F^2,\ F \in \mathcal{F}_0 \cap \mathcal{F}_4$

$$T_4(X) = \frac{1}{n}\sum X_{(i)}^2 - S^2(X) \text{ is UMVU for } g(F).$$

This result could also be obtained by observing that X_1X_2 is unbiased for $\mu_F^2,\ F \in \mathcal{F}_0 \cap \mathcal{F}_4$, therefore $E(X_1X_2 \mid X_{(1)}, \ldots, X_{(n)})$ is UMVU. But conditioned on $X_{(1)}, \cdots, X_{(n)}$,

$$X_1X_2 = X_{(i)}X_{(j)} \quad w.p.\ \frac{2}{n(n-1)} \text{ for each subset } \{i,j\} \text{ of } \{1,\ldots,n\} \text{ with } i<j$$

$$\begin{aligned}
\therefore E(X_1X_2 \mid X_{(1)}, \cdots, X_{(n)}) &= \frac{1}{n(n-1)}\sum_{i\neq j} X_{(i)}X_{(j)} \\
&= \frac{1}{n(n-1)}((\sum X_i)^2 - \sum X_i^2) \\
&= \frac{1}{n}\sum X_i^2 - \frac{1}{n-1}(\sum X_i^2 - \frac{1}{n}(\sum X_i)^2) \\
&= T_4(X)
\end{aligned}$$

More generally suppose $g(F)$ is U-estimable in $\mathcal{F}_0$. Then

$$\exists\ \delta(X_1,\ldots,X_m) \text{ such that } E_F\delta(X_1,\ldots,X_m) = g(F) \quad \forall F \in \mathcal{F}_0.$$

Suppose also that $\delta(X_1,\ldots,X_m)$ has finite second moment for $F \in \mathcal{F}_0 \cap \mathcal{F}_k$ for some positive integer k. We can assume δ is symmetric in $X_1,\ldots,X_m$, since if not we can redefine δ as

$$\delta^*(X_1,\ldots,X_m) = \frac{1}{m!}\sum_{\text{permutations } \pi \text{ of } (1,\ldots,m)} \delta(X_{\pi_1},\ldots,X_{\pi_m})$$

which is also unbiased and symmetric.
Now we define the **U-statistic** (Serfling, 1980),

$$T = \frac{1}{\binom{n}{m}}\sum_{1\leqslant i_1<i_2<\cdots<i_m\leqslant n} \delta^*(X_{i_1},\ldots,X_{i_m})$$

This is symmetric in $X_1,\ldots,X_n$ and unbiased, and therefore UMVU for $g(F),\ F \in \mathcal{F}_0 \cap \mathcal{F}_k$.

Questions

(1) Which $g(F)$ are U-estimable?
(2) If g is U-estimable, what is the smallest value of m for which there exists a U-statistic for g of the form T? This number is called the **degree** of g.

PROPOSITION 2.2.7. *If g is of degree 1, then for any $F_1, F_2 \in \mathcal{F}_0$, $g(\alpha F_1 + (1-\alpha)F_2)$ is linear in α.*

PROOF. If g is of degree 1, there exists $\delta(X_1)$ such that

$$\int \delta(x)F(dx) = g(F)$$

$$\therefore\ g(\alpha F_1 + (1-\alpha)F_2) = \alpha \int \delta(x)F_1(dx) + (1-\alpha)\int \delta(x)F_2(dx)$$
$$= \alpha g(F_1) + (1-\alpha)g(F_2).$$

□

Generalization. If g is of degree s, then $g(\alpha F_1 + (1-\alpha)F_2)$ is a polynomial in α of degree $\leqslant s$ (since $dF(x_1,\ldots,x_s) = \alpha^s dF_1(x_1)\cdots dF_1(x_s)$ if F_1 is replaced by αF_1.)

EXAMPLE 2.2.8. $g(F) = \sigma_F^2$ is of degree 2 in $\mathcal{F}_0 \cap \mathcal{F}_2$.

PROOF. Let $\delta(X_1, X_2) = \frac{1}{2}(X_1 - X_2)^2$, then

$$E_F\delta = E_F X_1^2 - E_F(X_1X_2) = \sigma_F^2$$

so $\deg(g) \leqslant 2$. To show $\deg(g) \neq 1$, consider

$$g(\alpha F_1 + (1-\alpha)F_2) = \sigma^2_{\alpha F_1 + (1-\alpha)F_2}$$
$$= \alpha \int x^2 dF_1(x) + (1-\alpha)\int x^2 dF_2(x) - [\alpha\mu_{F_1} + (1-\alpha)\mu_{F_2}]^2$$

and this is linear in α. $\Leftrightarrow \mu_{F_1} = \mu_{F_2}$. But this is not the case for every $F_1, F_2 \in \mathcal{F}_0 \cap \mathcal{F}_2$. Hence $\deg(g) = 2$. □

EXAMPLE 2.2.9. $g(F) = \sigma_F$ is not U-estimable in $\mathcal{F}_0$, since $g(\alpha F_1 + (1-\alpha)F_2)$ is not a polynomial.

2.3. The Information Inequality

For any estimator $T \in \Delta$ of $g(\theta)$ and any function $\psi(X,\theta)$ such that $E_\theta|\psi(X,\theta)|^2 < \infty$, we have the inequality

$$\mathrm{Var}_\theta(T) \geq \frac{|\mathrm{Cov}_\theta(T,\psi)|^2}{\mathrm{Var}_\theta(\psi(X,\theta))}. \tag{2.3.1}$$

However, this will not in general provide a useful lower bound for $\mathrm{Var}_\theta T$ since the RHS depends on T. It can be useful however when the RHS depends on T in a simple way, in particular when it depends on T only through $E_\theta T$.

THEOREM 2.3.1. *$Cov_\theta(T,\psi)$ depends on T only through $E_\theta T$ iff*

$$Cov_\theta(U,\psi) = 0 \text{ for all } U \in \mathcal{U} \cap \Delta \text{ (unbiased square-integrable estimators of } 0).$$

PROOF. ($\Leftarrow$) Suppose $Cov_\theta(U,\psi) = 0$ for all $U \in \mathcal{U} \cap \Delta$ and that T_1, T_2 are two estimators with finite variance and

$$E_\theta T_1 = E_\theta T_2 \; \forall \theta \in \Omega.$$

Then $T_1 - T_2 \in \mathcal{U}$, so $Cov_\theta(T_1,\psi) = Cov_\theta(T_2,\psi)$.

($\Rightarrow$) If $Cov_\theta(T,\psi)$ depends on T only through $E_\theta T$ and if $U \in \mathcal{U}$, then

$$Cov_\theta(T+U,\psi) = Cov_\theta(T,\psi)$$
$$\therefore Cov_\theta(U,\psi) = 0$$

□

Hammersley-Chapman-Robbins Inequality

Suppose X has probability density $p(x,\theta)$ $\theta \in \Omega$, where $p(x,\theta) > 0$ $\forall x$ and θ. Suppose $\exists\, \theta, \theta+\delta$ s.t. $g(\theta) \neq g(\theta+\delta)$. Then,

$$\psi(x,\theta) = \frac{p(x,\theta+\delta)}{p(x,\theta)} - 1$$

satisfies the conditions of the previous theorem, i.e. $Cov_\theta(U,\psi) = 0 \; \forall U \in \mathcal{U} \cap \Delta$, since $E_\theta\psi(X,\theta) = 0$ and $E_\theta(U\psi) = \int U(x)(p(x,\theta+\delta) - p(x,\theta))d\mu(x)$.

For *any* statistic $S \in \Delta$,

$$\begin{aligned} Cov_\theta(S,\psi) = E_\theta(S\psi) &= \int S[p(x,\theta+\delta) - p(x,\theta)]\mu(dx) \\ &= E_{\theta+\delta}S - E_\theta S \\ &= \begin{cases} 0 & if \quad S \in \mathcal{U}, \\ g(\theta+\delta) - g(\theta) & if \quad S \text{ is unbiased for } g(\theta.) \end{cases} \end{aligned}$$

Hence from (2.3.1), if $T \in \Delta$ is unbiased for $g(\theta)$,

$$Var_\theta(T) \geqslant \frac{(g(\theta+\delta) - g(\theta))^2}{E_\theta[(\frac{p(X,\theta+\delta)}{p(X,\theta)} - 1)^2]} \quad \forall \delta.$$

Hence we obtain the

HCR bound

$$Var_\theta(T) \geqslant \sup_\delta \frac{(g(\theta+\delta) - g(\theta))^2}{Var_\theta(\frac{p(X,\theta+\delta)}{p(X,\theta)})},$$

if T is unbiased for $g(\theta)$.

Letting $\delta \to 0$ in the HCR bound gives

$$\begin{aligned} Var_\theta T &\geq \lim_{\delta \to 0} \frac{(\frac{g(\theta+\delta)-g(\theta)}{\delta})^2}{E_\theta(\frac{1}{\delta}\frac{p(X,\theta+\delta)-p(X,\theta)}{p(X,\theta)})^2} \\ &= \frac{g'(\theta)^2}{E_\theta(\frac{\partial p}{\partial \theta}/p)^2} \\ &= \frac{g'(\theta)^2}{E_\theta(\frac{\partial \log p(X,\theta)}{\partial \theta})^2} \end{aligned}$$

provided g is differentiable and we can differentiate under the expectation. These steps are legitimized under the conditions of the following theorem.

THEOREM 2.3.2. **(Cramer-Rao Lower Bound CRLB)** *Suppose that the density of the sample $p(x,\theta) > 0$ and $\frac{\partial}{\partial \theta} \log p(x,\theta)$ exists for all x and θ, and that for each θ there exists δ such that*

$$|\phi - \theta| < \delta \Rightarrow \begin{cases} P_\phi \ll P_\theta \text{ and} \\ \frac{1}{|\phi-\theta|}\left|\frac{p(x,\phi)}{p(x,\theta)} - 1\right| \leq G(x,\theta) \end{cases}$$

(where G is independent of ϕ and $E_\theta G(X,\theta)^2 < \infty$ for all θ). Then for any unbiased estimator T of $g(\theta)$,

$$Var_\theta T \geq \frac{g'(\theta)^2}{I(\theta)},$$

where

$$\begin{cases} I(\theta) = E_\theta\left(\frac{\partial \log p(x,\theta)}{\partial \theta}\right)^2, & = \textbf{Fisher Information} \\ (g'(\theta))^2 = \limsup_{\phi \to \theta}\left(\frac{g(\phi)-g(\theta)}{\phi-\theta}\right)^2. \end{cases}$$

PROOF. By the HCR bound

$$Var_\theta(T) \int \frac{1}{|\phi-\theta|^2}(\frac{p(x,\phi)}{p(x,\theta)} - 1)^2 \, p(x,\theta) \, \mu(dx) \geq (\frac{g(\phi)-g(\theta)}{\phi-\theta})^2$$

Let $\{\phi_n\}$ be a sequence such that $\phi_n \to \theta$ and

$$(\frac{g(\phi_n)-g(\theta)}{\phi_n-\theta})^2 \to (g'(\theta))^2.$$

Then setting $\phi = \phi_n$ in the above inequality and letting $n \to \infty$, gives (by DC)

$$Var_\theta(T) \; E_\theta(\frac{\partial}{\partial \theta} \log p(X,\theta))^2 \geq g'(\theta)^2$$

□

COROLLARY 2.3.3. *If $X_1, \ldots, X_n$ are iid $P_{1,\theta}$ (the marginal distribution of X_i) and the corresponding marginal density $p_1(x, \theta)$ satisfies the conditions of the Cramer-Rao Lower Bound theorem, then for any unbiased estimator $T(X_1, \cdots, X_n)$ of $g(\theta)$,*

$$Var_\theta(T) \geq \frac{g'(\theta)^2}{nI_1(\theta)} \qquad \text{where } I_1(\theta) = E_\theta\left(\frac{\partial \log p_1(X_1, \theta)}{\partial \theta}\right)^2.$$

PROOF. The sample space is $\mathcal{X}^n$ and

$$\frac{dP_\theta}{d\mu^n}(\underline{x}) = \prod_{i=1}^n p_1(x_i, \theta) \quad \text{where } \mu^n = \mu \otimes \mu \otimes \cdots \otimes \mu.$$

The Fisher information for P_θ is

$$\begin{aligned}
&\int \left(\frac{\partial}{\partial \theta} \log \prod_{i=1}^n p_1(x_i, \theta)\right)^2 p_1(x_1, \theta) \cdots p_1(x_n, \theta) \mu^n(dx) \\
&= \int \left(\sum \frac{\partial \log p_1(x_i, \theta)}{\partial \theta}\right)^2 p_1(x_1, \theta) \cdots p_1(x_n, \theta) \mu^n(dx) \\
&= \int \sum \left(\frac{\partial \log p_1(x_i, \theta)}{\partial \theta}\right)^2 p_1(x_1, \theta) \cdots p_1(x_n, \theta) \mu^n(dx) = nI(\theta)
\end{aligned}$$

since

$$E_\theta\left[\frac{\partial \log}{\partial \theta} p_1(X_i, \theta) \frac{\partial \log}{\partial \theta} p_1(X_j, \theta)\right] = \left[E_\theta \frac{\partial \log}{\partial \theta} p_1(X_i, \theta)\right]^2 \text{ (by independence)}$$

and

$$E(\frac{\partial \log}{\partial \theta} p_1(X_i, \theta)) = 0.$$

(We can differentiate under the integral sign by DC and the assumptions on $p_1(x_i, \theta)$).

The statement of the Corollary now follows if we can show that the assumptions on $p_1(x_i, \theta)$ carry over to $p(\underline{x}, \theta)$, i.e. that $|\phi - \theta| < \delta \Rightarrow$

$$\left| \frac{p_1(x_1, \phi) \cdots p_1(x_n, \phi)}{p_1(x_1, \theta) \cdots p_1(x_n, \theta)} - 1 \right| / |\phi - \theta| \leq \tilde{G}(\underline{x}, \theta)$$

where $E_\theta \tilde{G}^2(\underline{X}, \theta) < \infty$.

Now

$$\begin{aligned}
\frac{p_1(x_i, \phi)}{p_1(x_i, \theta)} &\leq 1 + |\phi - \theta| G(x_i, \theta) \\
&\leq 1 + \delta G(x_i, \theta)
\end{aligned}$$

and

$$|a_1 \cdots a_n - 1| \leq |a_1 \cdots a_n - a_2 \cdots a_n| + |a_2 \cdots a_n - a_3 \cdots a_n| + \cdots \text{ by the triangle inequality,}$$
$$\leq |a_1 - 1||a_2 \cdots a_n| + |a_2 - 1||a_3 \cdots a_n| + \cdots + |a_n - 1|.$$

Setting $a_i = \frac{p_1(x_i,\phi)}{p_1(x_i,\theta)}$ gives

$$\begin{aligned} \left| \frac{p_1(x_1,\phi)\cdots p_1(x_n,\phi)}{p_1(x,\theta)\cdots p_1(x_n,\theta)} - 1 \right| &\leq |\phi-\theta| G(x_1,\theta) \prod_{i>1}(1+\delta G(x_i,\theta)) \\ &\quad + |\phi-\theta| G(x_2,\theta) \prod_{i>2}(1+\delta G(x_i,\theta)) \\ &\quad + \cdots \\ &\quad + |\phi-\theta| G(x_n,\theta) \\ &:= |\phi-\theta| \tilde{G}(\underline{x},\theta) \end{aligned}$$

and $E_\theta \tilde{G}(\underline{X},\theta)^2 < \infty$ since $X_1, \ldots, X_n$ are independent and $E_\theta G(X_i,\theta)^2 < \infty$. (No $G(X_i,\theta)$ is raised to a power greater than 2.) □

COROLLARY 2.3.4. *Suppose $p(\underline{x},\theta)$ satisfies the conditions of theorem 2.3.2 and*

$$E_\theta T(X) = g(\theta) + b(\theta)$$

i.e., $T(X)$ has bias $b(\theta)$ for estimating $g(\theta)$. Then

$$\begin{aligned} MSE_\theta(T) &= E_\theta \left(T(X) - g(\theta)\right)^2 \\ &= b^2(\theta) + V_\theta(T) \\ &\geq b^2(\theta) + \frac{c(\theta)}{I(\theta)} \end{aligned}$$

where

$$c(\theta) = \limsup_{\phi \to \theta} \left(\frac{g(\phi) + b(\phi) - g(\theta) + b(\theta)}{\phi - \theta} \right)^2$$

PROOF. T is unbiased for $g(\theta) + b(\theta)$, so

$$E_\theta \left(T(X) - g(\theta) - b(\theta)\right)^2 \geq \frac{c(\theta)}{I(\theta)}$$

Hence $MSE = E_\theta(T(X) - g(\theta))^2 = \mathrm{Var}_\theta T(X) + [E_\theta(T(X) - g(\theta))]^2 \geq b^2(\theta) + \frac{c(\theta)}{I(\theta)}$.

□

Behavior of $I(\cdot)$ under reparameterization. Suppose $\alpha = h(\theta)$ reparametrizes $\{P_\theta : \theta \in \Omega\}$ to $\{P^*_\alpha : \alpha \in h(\Omega)\}$. Then

$$p^*(x,\alpha) = p\left(x, h^{\leftarrow}(\alpha)\right),$$

where $h^{\leftarrow}$ denotes the inverse mapping and, by definition,

$$\begin{aligned} I^{*}(\alpha) &= E_{\alpha}\left(\frac{\partial \log}{\partial \alpha} p^{*}(X,\alpha)\right)^{2} \\ &= E_{\alpha}\left(\frac{\partial \log p(X,\theta)}{\partial \theta}\Big|_{\theta=h^{\leftarrow}(\alpha)} \frac{d}{d\alpha} h^{\leftarrow}(\alpha)\right)^{2} \\ &= I\left(h^{\leftarrow}(\alpha)\right)\left(\frac{dh^{\leftarrow}(\alpha)}{d\alpha}\right)^{2} \\ &= \frac{I(\theta)}{\left(h'(\theta)\right)^{2}}\Big|_{\theta=h^{\leftarrow}(\alpha)}. \end{aligned}$$

An alternative expression for $I(\theta)$. Provided $\frac{\partial^2}{\partial\theta^2}\log p(x,\theta)$ exists for all x, θ and if

$$\int \frac{\partial^2}{\partial\theta^2} p(x,\theta)\, d\mu(x) = \frac{\partial^2}{\partial\theta^2}\int p(x,\theta)\, d\mu(x) = 0, \tag{2.3.2}$$

then

$$I(\theta) = -E_{\theta}\left[\frac{\partial^2}{\partial\theta^2}\log p(x,\theta)\right].$$

PROOF.

$$\frac{\partial^2}{\partial\theta^2}\log p(x,\theta) = \frac{\frac{\partial^2}{\partial\theta^2}p(x,\theta)}{p(x,\theta)} - \frac{(\frac{\partial p(x,\theta)}{\partial\theta})^2}{p(x,\theta)^2}$$

and $E_{\theta}\left[p(X,\theta)^{-1}\frac{\partial^2}{\partial\theta^2}p(X,\theta)\right] = 0$ by (2.3.2). □

THEOREM 2.3.5. **One-parameter exponential family.** *Suppose*

$$p(x,\theta) = \exp\left(T(x)\,\eta(\theta) - B(\theta)\right)\cdot h(x),$$

where $\theta = E_{\theta}T$ *and* $\eta(\theta) \in Int(\mathcal{N})$. *Then*

$$I(\theta) = \frac{1}{Var_{\theta}T}, \qquad \textit{and} \qquad I(\eta) = Var_{\eta}T.$$

PROOF.

$$\int e^{\eta T(x) - A(\eta)} h(x)\mu(dx) = 1$$

and $A'(\eta) = E_{\eta}T$, $A''(\eta) = Var_{\eta}T$. Hence $\theta = A'(\eta)$.
Now $\pi(x,\eta) = e^{T(x)\eta - A(\eta)}h(x)$ (This is the canonical representation of the density of the exponential family. See (1.3.7)). Hence $\frac{\partial \log \pi(x,\eta)}{\partial\eta} = T(x) - A'(\eta)$, so

$$I^{*}(\eta) = E_{\eta}(T(X) - A'(\eta))^2 = Var_{\eta}T$$

Since $\theta = A'(\eta) = h^{-1}(\eta)$, and we have $I^*(\eta) = I(\theta(\eta))(A''(\eta))^2$. So

$$\begin{aligned} I(\theta) &= I^*(\eta(\theta))/A''(\eta(\theta))^2 \\ &= Var_\theta T/(Var_\theta T)^2 \\ &= \frac{1}{Var_\theta T} \end{aligned}$$

□

REMARK 2.3.6. T attains the CR lower bound in this case. A converse result also holds: under some regularity conditions, attainment of the CR lower bound implies that T is the natural sufficient statistic of some exponential family $\{P_\theta\}$.

EXAMPLE 2.3.7. **Poisson family.** Suppose that $X_1, \ldots, X_n$ are iid Poisson. Then

$$\begin{aligned} p(x,\theta) &= e^{-n\theta}\ \theta^{\sum x_i}\ \frac{1}{\prod_1^n x_i!} \\ &= e^{-n\theta + n\log\theta \frac{\sum x_i}{n}} \frac{1}{\prod_1^n x_i!} \qquad \eta = n\log\theta,\ \theta = e^{\eta/n} \\ \pi(x,\eta) &= e^{-ne^{\eta/n} + \eta\frac{\sum x_i}{n}} \frac{1}{\prod_1^n x_i!} \\ T(x) &= \frac{\sum x_i}{n} \text{ is UMVU for } \theta \\ A'(\eta) &= e^{\eta/n} = \theta \\ A''(\eta) &= \frac{1}{n}e^{\eta/n} = \frac{\theta}{n} \\ I^*(\eta) &= Var_\eta T = \frac{\theta}{n} \\ I(\theta) &= \frac{1}{Var_\theta T} = \frac{n}{\theta} \end{aligned}$$

The information on $n\log\theta$ increases with θ. The information on θ decreases with θ.

THEOREM 2.3.8. **Alternative version of CRLB theorem.**
Suppose Ω is an open interval, $A = \{x\colon p(x,\theta) > 0\}$ is independent of θ, $\frac{\partial p}{\partial\theta}(x,\theta)$ is finite for all $x \in A$ and for all $\theta \in \Omega$, $E_\theta \frac{\partial}{\partial\theta}\log p(x,\theta) = 0$, $\frac{\partial}{\partial\theta}(E_\theta T) = \int T(x)\frac{\partial p}{\partial\theta}(x,\theta)\mu(dx)$. Then

$$Var_\theta(T(X)) \geq \frac{(\frac{\partial}{\partial\theta}E_\theta T)^2}{I(\theta)}$$

PROOF. Choose $\psi(x,\theta) = \frac{\partial}{\partial\theta}\log p(x,\theta)$ in (2.3.1).

$$\begin{aligned} Cov_\theta\left(T, \frac{\partial}{\partial\theta}\log p(X,\theta)\right) &= E_\theta\left(T\frac{\partial}{\partial\theta}\log p(X,\theta)\right) \\ &= \int T(x)\frac{\partial p(X,\theta)}{\partial\theta}\mu(dx) \\ &= \frac{\partial}{\partial\theta}(E_\theta T) \\ Var_\theta\left(\frac{\partial}{\partial\theta}\log p(X,\theta)\right) &= I(\theta). \end{aligned}$$

□

2.4. Multiparameter Case

Here we consider the generalization to vector $\theta = (\theta_1, \ldots, \theta_s) \in \Omega$. The estimator T may be either scalar (when estimating a scalar function of θ) or vector. (In Ch. 4 we will consider generalizations to estimation of vector $g(\theta)$ based on vector T.)

THEOREM 2.4.1. *For $T(X), \psi_1(X,\theta), \cdots, \psi_s(X,\theta)$ functions with finite 2nd moments under P_θ, we have the multiparameter analogue of (2.3.1),*

$$Var_\theta(T) \geq \gamma^T C^{-1}\gamma,$$

where $\gamma^T = (\gamma_1, \cdots, \gamma_s)$, $\gamma_i = Cov_\theta(T, \psi_i)$ and $C = [Cov_\theta(\psi_i, \psi_j)]_{i,j=1}^s$

PROOF. Let $\hat{Y}$ denote the minimum mean squared error linear predictor, of $Y = T - E_\theta T$ in terms of $\Psi = \begin{bmatrix} \psi_1 - E_\theta\psi_1 \\ \cdots \\ \psi_s - E_\theta\psi_s \end{bmatrix}$. Then

$$\hat{Y} = \mathbf{a}^T\Psi$$

where $E_\theta\left((Y - \hat{Y})(\psi_j - E_\theta\psi_j)\right) = 0,\ j = 1, \ldots, s$, i.e.

$$C\mathbf{a} = \gamma = Cov_\theta(T, \Psi)$$

These equations have a solution and hence

$$\mathbf{a} = C^{-1}\gamma$$

where C^{-1} is any generalized inverse of $C = E_\theta(\Psi\Psi')$. So $\hat{Y} = \gamma^T C^{-1}\Psi$.
Also

$$\begin{aligned} E(Y-\hat{Y})^2 &= EY^2 - E\hat{Y}^2 \qquad \text{since } \hat{Y} \perp Y-\hat{Y} \text{ in } L^2(P_\theta) \\ &= Var_\theta T - E(\gamma^T C^{-1}\Psi\Psi^T C^{-1}\gamma) \\ &= Var_\theta T - \gamma^T C^{-1}\gamma \\ \therefore E(Y-\hat{Y})^2 &= Var_\theta T - \gamma^T C^{-1}\gamma \geq 0 \\ \therefore Var_\theta T &\geq \gamma^T C^{-1}\gamma \end{aligned}$$

Notice that the right hand side is the same for any generalized inverse C^{-1} of C. □

Generalization of the Information Inequality

Assume that

$$\begin{cases} \Omega \text{ is an open interval in } \mathbb{R}^s \\ A = \{x\colon p(x,\theta) > 0\} \text{ is independent of } \theta \\ \frac{\partial p}{\partial \theta_i}(x,\theta) \text{ is finite } \forall x \in A,\ \forall \theta \in \Omega,\ i = 1,\ldots,s \\ E_\theta \frac{\partial}{\partial \theta_i} \log p(X,\theta) = 0,\ i = 1,\ldots,s \end{cases}$$

DEFINITION 2.4.2. **The information matrix.**

$$\begin{aligned} I(\theta) &:= \left[E_\theta \left(\frac{\partial}{\partial \theta_i} \log p(X,\theta) \frac{\partial}{\partial \theta_j} \log p(X,\theta) \right) \right]_{i,j=1}^{s} \\ &= \left[Cov_\theta \left(\frac{\partial}{\partial \theta_i} \log p(X,\theta) \frac{\partial}{\partial \theta_j} \log p(X,\theta) \right) \right]_{i,j=1}^{s} \end{aligned}$$

$I(\theta)$ is strictly positive definite if $\{\frac{\partial}{\partial \theta_i} \log p(X,\theta), i = 1,\ldots,s\}$ is linearly independent *a.s.* P_θ.

THEOREM 2.4.3. *Under the previous assumptions, if $I(\theta)$ is strictly positive definite and if $T(X)$ satisfies*

$$E_\theta T(X)^2 < \infty \quad \forall \theta$$

and

$$\frac{\partial}{\partial \theta_j} E_\theta T(X) = \int T(x) \frac{\partial}{\partial \theta_j} p(x,\theta) \mu(dx)$$

then

$$Var_\theta T(X) \geq \gamma^T I(\theta)^{-1} \gamma$$

where $\gamma = \begin{bmatrix} \frac{\partial}{\partial \theta_1} E_\theta T(X) \\ \cdots \\ \frac{\partial}{\partial \theta_s} E_\theta T(X) \end{bmatrix}$

PROOF. This is a direct application of Theorem 2.4.1 with the functions ψ_i defined by

$$\psi_i(x,\theta) = \frac{\partial}{\partial\theta_i}\log p(x,\theta)$$

□

Reparameterization. If $\theta_i = f_i(\alpha_1,\ldots,\alpha_s),\ i=1,\ldots,s$, then we have:

$$\begin{aligned} I^*(\alpha) &= \left[E\left[\frac{\partial \log p^*}{\partial\alpha_i}(X,\alpha)\frac{\partial \log p^*}{\partial\alpha_j}(X,\alpha)\right]\right]_{i,j=1}^{s} \\ &= \left[\sum_n\sum_m \frac{\partial\theta_m}{\partial\alpha_i}E\left(\frac{\partial\log p}{\partial\theta_m}(X,\theta)\frac{\partial\log p}{\partial\theta_n}(X,\theta)\right)\frac{\partial\theta_n}{\partial\alpha_j}\right]_{i,j=1}^{s} \\ &= JI(\theta)J^T \end{aligned}$$

where $J = \left[\frac{\partial\theta_j}{\partial\alpha_i}\right]_{i,j=1}^{s}$.

COROLLARY 2.4.4. *If $I(\theta)$ is strictly positive definite, the elements of $T=(T_1,\ldots,T_n)$ are finite variance unbiased for the respective elements of $g(\theta)=(g_1(\theta),\ldots,g_n(\theta))$ and each T_i satisfies the conditions of theorem 2.4.3, then, $E_\theta T = g(\theta)$ and,*

$$Cov_\theta T \geq (\frac{\partial g}{\partial\theta})I^{-1}(\theta)(\frac{\partial g}{\partial\theta})^T, \tag{2.4.1}$$

where $A \geq B$ means $a^T(A-B)a \geq 0,\ \forall a\in\mathbb{R}^n$ and $\frac{\partial g}{\partial\theta} := \begin{bmatrix} \frac{\partial g_1}{\partial\theta_1}\cdots\frac{\partial g_1}{\partial\theta_s} \\ \cdots \\ \frac{\partial g_n}{\partial\theta_1}\cdots\frac{\partial g_n}{\partial\theta_s}\end{bmatrix}$.

PROOF. Since $a^T(Cov_\theta T)a = Var(a^TT)$, (2,4,1) is equivalent to

$$Var_\theta(a^TT) \geq a^T(\frac{\partial g}{\partial\theta})I^{-1}(\theta)(\frac{\partial g}{\partial\theta})^Ta \quad \forall a\in\mathbb{R}^n.$$

But the above inequality follows at once by applying theorem 2.4.3 to the *real-valued* statistic a^TT for which

$$\gamma = \begin{bmatrix} \frac{\partial}{\partial\theta_1}E_\theta(a^TT) \\ \cdots \\ \frac{\partial}{\partial\theta_s}E_\theta(a^TT)\end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial\theta_1}E_\theta(T^T) \\ \cdots \\ \frac{\partial}{\partial\theta_s}E_\theta(T^T)\end{bmatrix} a = (\frac{\partial g}{\partial\theta})^T a.$$

□

COROLLARY 2.4.5. *If $T_1,\cdots,T_s$ are unbiased for*

$\theta_1,\cdots,\theta_s$, then

$$\mathrm{Cov}_\theta(T) \geq I(\theta)^{-1}.$$

PROOF. Apply corollary 2.4.4 with $g_i(\theta)=\theta_i,\ i=1,\ldots,s$ (where $\frac{\partial g}{\partial\theta} = I_{s\times s}$). □

REMARK 2.4.6. Suppose we wish to estimate θ_1. If $\theta_2, \ldots, \theta_s$ are known then the CR lower bound is

$$\left[E_\theta\left(\frac{\partial \log p}{\partial \theta_1}(X,\theta)\right)^2\right]^{-1}.$$

If $\theta_2, \ldots, \theta_s$ are not known, then the CR bound for estimating θ_1 is the (1,1) component of $I^{-1}(\theta)$, denoted by $I^{-1}(\theta)_{1,1}$ (by Corollary 2.4.4 with $T = \theta_1$ and $g(\theta) = \theta_1$). Naturally we expect

$$\left[E_\theta\left(\frac{\partial log}{\partial \theta_1}p(X,\theta)\right)^2\right]^{-1} \leq I^{-1}(\theta)_{1,1}$$

By the general formula for the inverse of a partitioned matrix,

$$A^{-1} = \begin{bmatrix} D & -DA_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}D & A_{22}^{-1}A_{21}DA_{12}A_{22}^{-1} \end{bmatrix},$$

where $D = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$, we find that

$$I^{-1}(\theta)_{1,1} = \frac{1}{a - b^T A^{-1} b} \geq \frac{1}{a}.$$

THEOREM 2.4.7. **(Order-s exponential family)**. *Suppose that*

$$p(x,\theta) = \exp\left(\sum_1^s \eta_i(\theta)\, T_i(x) - B(\theta)\right) h(x), \qquad \theta \in \Omega$$

is an order-s exponential family parameterized by

$$\theta = E_\theta T$$

and that $\eta(\Omega)$ contains an open subset of $\mathbb{R}^s$. Then, if $C = Cov(T)$, we have

$$I(\theta) = C^{-1} \qquad \text{and} \qquad I^*(\eta) = C.$$

PROOF. By theorem 1.3.14 of chapter 1, we know that for $\eta \in int(\mathcal{N})$

$$\frac{\partial^2 A}{\partial \eta_i \partial \eta_j} = cov(T_i, T_j), \quad cov(T) = \frac{\partial A^2}{\partial \eta \partial \eta^T} := \left[\frac{\partial^2 A}{\partial \eta_i \partial \eta_j}\right]_{i,j=1}^s$$

We also know that if $\pi(x,\eta)$ is the canonical form of the density

$$\frac{\partial \log \pi(x,\eta)}{\partial \eta_i}\frac{\partial \log \pi(x,\eta)}{\partial \eta_j} = (T_i - \frac{\partial A}{\partial \eta_i})(T_j - \frac{\partial A}{\partial \eta_j})$$

$$\therefore\ I^*(\eta) = cov(T) = \frac{\partial^2 A}{\partial \eta \partial \eta^T}$$

Moreover, $\theta = E_\theta T = \frac{\partial A}{\partial \eta} = \begin{bmatrix} \partial A/\partial \eta_1 \\ \cdots \\ \partial A/\partial \eta_s \end{bmatrix}$, and hence by the reparameterization formula

$$I^*(\eta) = \frac{\partial^2 A}{\partial \eta \partial \eta^T} I(\theta) \frac{\partial^2 A}{\partial \eta \partial \eta^T}$$

But the left hand side is $\frac{\partial^2 A}{\partial\eta\partial\eta^T}$ (a symmetric matrix) and hence

$$I(\theta) = \left[\frac{\partial^2 A}{\partial\eta\partial\eta^T}\right]^{-1} = C^{-1}.$$

□

REMARK 2.4.8. Note that for a random sample of size n from $p(x,\theta)$, the CSS is $(\sum T_1(X_i),\ldots,\sum T_s(X_i))$, and the new "A" function is $nA(\eta)$.

Examples of Information Matrices.

EXAMPLE 2.4.9. $X \sim N(\xi,\Sigma)$, $\xi \in \mathbb{R}^s$, Σ fixed and non-singular. Then

$$p(x,\xi) = \frac{1}{(\sqrt{2\pi})^s |det\Sigma|^{1/2}} exp\left\{-\frac{1}{2}(x-\xi)^T\Sigma^{-1}(x-\xi)\right\}$$

Writing $\Sigma = [\sigma_{ij}]_{i,j}^s$, $\Sigma^{-1} = [\gamma_{ij}]_{i,j}^s$, we have

$$\begin{aligned}
\frac{\partial \log p}{\partial \xi_i}(X,\xi) &= \sum_{k=1}^{s}\gamma_{ik}(x_k-\xi_k)\\
\frac{\partial \log p}{\partial \xi_j}(X,\xi) &= \sum_{m=1}^{s}\gamma_{jm}(x_m-\xi_m)\\
\therefore\ E\left[\frac{\partial \log p}{\partial \xi_i}(X,\xi)\frac{\partial \log p}{\partial \xi_j}(X,\xi)\right] &= \sum_{k=1}^{s}\sum_{m=1}^{s}\gamma_{ik}E(X_k-\xi_k)(X_m-\xi_m)\gamma_{jm}\\
&= \sum_{k=1}^{s}\sum_{m=1}^{s}\gamma_{ik}\sigma_{km}\gamma_{mj}\\
&= \Sigma^{-1}\Sigma\Sigma^{-1}\\
&= \Sigma^{-1}.
\end{aligned}$$

EXAMPLE 2.4.10. **The order-two exponential family** $\{N(\xi,\sigma)\}$, $\xi \in \mathbb{R}; \sigma > 0\}$.

$$\begin{aligned}
p^*(x,\psi) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\xi)^2}\\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}x^2+\frac{\xi}{\sigma^2}x-\frac{\xi^2}{2\sigma^2}-\log\sigma}
\end{aligned}$$

where $\psi = \begin{bmatrix}\xi\\ \sigma\end{bmatrix}$. By Theorem 2.4.7, if we let

$$\theta = \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix} = \begin{bmatrix}\xi\\ \xi^2+\sigma^2\end{bmatrix} = E_\theta\begin{bmatrix}X\\ X^2\end{bmatrix}$$

then we can rewrite $p^*(x,\psi)$ as

$$p(x,\theta) = \frac{1}{\sqrt{2\pi}} exp\left\{\eta_1(\theta)x + \eta_2(\theta)x^2 - B(\theta)\right\}$$

and since $E_\theta \begin{bmatrix} X \\ X^2 \end{bmatrix} = \theta = E_\theta T$, $T = \begin{bmatrix} X \\ X^2 \end{bmatrix}$,

$$I(\theta) = [Cov(T)]^{-1}$$

where $Cov(T) = \begin{bmatrix} \sigma^2 & 2\sigma^2\xi \\ 2\sigma^2\xi & 2\sigma^4 + 4\sigma^2\xi^2 \end{bmatrix}$ (check from MGF).

We can now find $I^*(\psi)$ from the reparameterization formula,

$$I^*(\psi) = JI(\theta)J^T$$

where $J = \begin{bmatrix} \frac{\partial\theta_1}{\partial\psi_1} & \cdots & \frac{\partial\theta_s}{\partial\psi_1} \\ \vdots & & \vdots \\ \frac{\partial\theta_1}{\partial\psi_s} & \cdots & \frac{\partial\theta_s}{\partial\psi_s} \end{bmatrix} = \begin{bmatrix} 1 & 2\xi \\ 0 & 2\sigma \end{bmatrix}$, $\theta = \begin{bmatrix} \xi \\ \xi^2 + \sigma^2 \end{bmatrix}$.

$$\begin{aligned}
\therefore I^*(\psi)^{-1} &= (J^{-1})^T I^{-1}(\theta) J^{-1} \quad \text{where } J^{-1} = \begin{bmatrix} 1 & -\xi/\sigma \\ 0 & 1/2\sigma \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 \\ -\xi/\sigma & 1/2\sigma \end{bmatrix} \begin{bmatrix} \sigma^2 & 2\sigma^2\xi \\ 2\sigma^2\xi & 2\sigma^4 + 4\sigma^2\xi^2 \end{bmatrix} \begin{bmatrix} 1 & -\xi/\sigma \\ 0 & 1/2\sigma \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2/2 \end{bmatrix} \\
\therefore I^*(\xi,\sigma) &= \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{bmatrix} \quad \text{(could also be obtained directly from } p^*\text{)}
\end{aligned}$$

EXAMPLE 2.4.11. **Location-scale families.**
Suppose that f is a probability density with respect to Lebesgue measure, that $f(x)$ is strictly positive for all x and that

$$p(x,\mu,\sigma) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma}),\ \mu \in \mathbb{R},\ \sigma > 0.$$

Then

$$\begin{aligned}
\frac{\partial \log p}{\partial \mu}(x,\mu,\sigma) &= -\frac{1}{\sigma}\frac{f'(\frac{x-\mu}{\sigma})}{f(\frac{x-\mu}{\sigma})} \\
\frac{\partial \log p}{\partial \sigma} &= -\frac{1}{\sigma} - \frac{x-\mu}{\sigma^2}\frac{f'(\frac{x-\mu}{\sigma})}{f(\frac{x-\mu}{\sigma})} \\
I_{11}(\mu,\sigma) &= \frac{1}{\sigma^2}\int \left(\frac{f'(\frac{x-\mu}{\sigma})}{f(\frac{x-\mu}{\sigma})}\right)^2 \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})dx \\
&= \frac{1}{\sigma^2}\int \left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx \\
I_{22}(\mu,\sigma) &= \frac{1}{\sigma^2}\int \left(1+\frac{x-\mu}{\sigma}\frac{f'(\frac{x-\mu}{\sigma})}{f(\frac{x-\mu}{\sigma})}\right)^2 \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})dx \\
&= \frac{1}{\sigma^2}\int \left(1+x\frac{f'(x)}{f(x)}\right)^2 f(x)dx \\
I_{12}(\mu,\sigma) &= \frac{1}{\sigma^2}\int \frac{f'(\frac{x-\mu}{\sigma})}{f(\frac{x-\mu}{\sigma})}\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})dx + \frac{1}{\sigma^2}\int \frac{x-\mu}{\sigma}\left(\frac{f'(\frac{x-\mu}{\sigma})}{f(\frac{x-\mu}{\sigma})}\right)^2 \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})dx \\
&= 0 + \frac{1}{\sigma^2}\int x\left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx
\end{aligned}$$

Thus I is a diagonal matrix if f is symmetric about 0. Example 2.4.9 is a special case of this.

When the CR bound in theorem 2.4.3 is not sharp, it can be improved by using higher order derivatives of ψ. This is the content of the following theorem.

THEOREM 2.4.12. *(**Bhattacharya bounds**) Suppose $p(x,\theta)$ have common support for any θ. Let T be unbiased for $g(\theta)$. Further assume*

$$\int T(x)\frac{\partial^i}{\partial\theta^i}p(x,\theta)\mu(dx) = g^{(i)}(\theta),\ i=1,\ldots,k$$

and

$$\int \frac{\partial^i}{\partial\theta^i}p(x,\theta)\mu(dx) = 0,\ i=1,\ldots,k$$

Then

$$Var_\theta(T) \geq \left[g^{(1)}(\theta),\ldots,g^{(k)}(\theta)\right] V^{-1} \left[g^{(1)}(\theta),\ldots,g^{(k)}(\theta)\right]^T$$

where $V = Cov\left[\left(\frac{1}{p(x,\theta)}\frac{\partial}{\partial\theta}p(x,\theta),\ldots,\frac{1}{p(x,\theta)}\frac{\partial^k}{\partial\theta^k}p(x,\theta)\right)^T\right]$, *and V is assumed to be non-singular.*

PROOF. This follows immediately from theorem 2.4.1 by taking

$$\psi_i(x,\theta) = \frac{1}{p(x,\theta)}\frac{\partial^i}{\partial\theta^i}p(x,\theta),\ i = 1,\dots,k$$

giving

$$\begin{aligned}\gamma_i &= Cov(T,\psi_i) = E(T\psi_i)\\ &= \int T(x)\frac{\partial^i}{\partial\theta^i}p(x,\theta)\mu(dx) = g^{(i)}(\theta).\end{aligned}$$

□

EXAMPLE 2.4.13. $X_1,\dots,X_n$ iid Poisson(θ), $\theta > 0$, and $g(\theta) = e^{-\theta} = P(X_i = 0)$. Then $I_0(X)$ is unbiased for $g(\theta)$. Here

$$p(x,\theta) = e^{-n\theta+(\sum x_i)\log\theta}\prod\frac{1}{x_i!},\ \theta > 0,$$

is an exponential family of full rank so $T(X) = \sum X_i$ is complete and sufficient. So the UMVU estimator of $g(\theta$ is $E\left[I_0(X_1)|T(X)\right]$ and

$$\begin{aligned}E\left[I_0(X_1)|T(X) = t\right] &= P\left(X_1 = 0|T(X) = t\right)\\ &= \frac{P\left(X_1 = 0 \cap T(X) = t\right)}{P\left(T(X) = t\right)}\\ &= \frac{e^{-\theta}e^{-(n-1)\theta}\left((n-1)\theta\right)^t/t!}{e^{-n\theta}(n\theta)^t/t!}\\ &= (1-\frac{1}{n})^t,\end{aligned}$$

so $(1-\frac{1}{n})^{T(X)}$ is UMVU.

Noting that $T(X) \sim P(n\theta)$, we can write its probability generating function as $Ez^{T(X)} = e^{-n\theta(1-z)}$, and hence

$$\begin{aligned}E_\theta(1-\frac{1}{n})^{T(X)} &= e^{-\theta}\ \text{ and}\\ E_\theta(1-\frac{1}{n})^{2T(X)} &= \exp(-n\theta(1-(1-1/n)^2))\\ &= e^{-2\theta+\frac{\theta}{n}}.\\ \therefore Var_\theta(1-\frac{1}{n})^{T(X)} &= e^{-2\theta}(e^{\frac{\theta}{n}}-1).\end{aligned}$$

For this problem $CRLB = \frac{g'(\theta)^2}{nI_1(\theta)}$, where $g(\theta) = e^{-\theta}$ and

$$I_1(\theta) = E\left(-1+\frac{X_1}{\theta}\right)^2 = \frac{1}{\theta^2}Var_\theta X = \frac{1}{\theta} \tag{2.4.2}$$

$$\therefore CRLB = \frac{\theta e^{-2\theta}}{n} < e^{-2\theta}(e^{\frac{\theta}{n}}-1) = e^{-2\theta}(\frac{\theta}{n}+\frac{1}{2}\frac{\theta^2}{n^2}+\cdots)$$

(Alternatively T is the CSS for a full-rank exponential family with $E_\theta(T) = \theta$, so $I(\theta) = \frac{1}{Var_\theta T} = \frac{n}{\theta} = nI_1(\theta)$.)

The Bhattacharya bound with $k = 2$,

$$\begin{aligned}
\begin{bmatrix} g^{(1)}(\theta) \\ g^{(2)}(\theta) \end{bmatrix} &= \begin{bmatrix} -e^{-\theta} \\ e^{-\theta} \end{bmatrix} \\
\psi_1(x,\theta) &= \frac{1}{p}\frac{\partial p}{\partial \theta} = \frac{\partial}{\partial \theta}\log p = -n + \frac{\sum x_i}{\theta} \\
\psi_2(x,\theta) &= \frac{1}{p}\frac{\partial^2 p}{\partial \theta^2} = \frac{\partial^2}{\partial \theta^2}\log p + (\frac{1}{p}\frac{\partial p}{\partial \theta})^2 = -\frac{\sum x_i}{\theta^2} + (\frac{\sum x_i}{\theta} - n)^2 \\
Cov_\theta \psi(X,\theta) &= \begin{bmatrix} n/\theta & 0 \\ 0 & 2n^2/\theta^2 \end{bmatrix}
\end{aligned}$$

Hence the Bhattacharya bound is

$$\begin{aligned}
Var_\theta T &\geq \begin{bmatrix} -e^{-\theta} & e^{-\theta} \end{bmatrix} \begin{bmatrix} n/\theta & 0 \\ 0 & 2n^2/\theta^2 \end{bmatrix} \begin{bmatrix} -e^{-\theta} \\ e^{-\theta} \end{bmatrix} \\
&= e^{-2\theta}(\frac{\theta}{n} + \frac{\theta^2}{2n^2}) > CRLB
\end{aligned}$$

but less than $Var_\theta T$. By taking more derivatives the bound can be make arbitrarily close to $Var_\theta T$. Extends to the multiparameter case also.

CHAPTER 3

Equivariance

3.1. Equivariance for Location family

In chapter 2 we introduced unbiasedness as a constraint to eliminate estimators which may do well at a particular parameter value at the cost of poor performance elsewhere. Within this limited class we could then sometimes determine uniformly minimum risk estimators for any convex loss function. Equivariance is a more physically motivated restriction. We start by considering how equivariance enters as a natural constraint on statistics used to estimate *location parameters*.

Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)$ has density

$$f(\boldsymbol{x} - \boldsymbol{\xi}) = f(x_1 - \xi, \cdots, x_n - \xi)$$

where f is known, ξ is a location parameter to be estimated and $L(\xi, d)$ is the loss when ξ is estimated as d. Suppose we have settled on $T(\boldsymbol{x})$ as a reasonable estimator of ξ as measured by

$$R(\xi, T) = E_\xi(L(\xi, T(\boldsymbol{x})))$$

Suppose another statistician B wants to measure the data using a different origin. So instead of recording $X_1, X, \cdots, X_n$, B records $X_1' = X_1 + 273, \cdots, X_n' = X_n + 273$ say (This would be the case if we measured temperatures in °C and B measured them in °K.) Then

$$X' = X + a. \tag{3.1.1}$$

On the new scale the location parameter becomes

$$\xi' = \xi + a \tag{3.1.2}$$

and the joint density of $\boldsymbol{X}' = (X_1', \cdots, X_n')$ is

$$f(\boldsymbol{x}' - \xi') = f(x_1' - \xi', \cdots, x_n' - \xi').$$

The estimated value d on the original scale becomes

$$d' = d + a \tag{3.1.3}$$

on the new one and the loss resulting from its use is $L(\xi', d')$. **The problem of estimating ξ is said to be *invariant* under the transformations (3.1.1),(3.1.2), and (3.1.3) if**

$$L(\xi + a, d + a) = L(\xi, d). \tag{3.1.4}$$

This condition on L is equivalent to the assumption that L has the functional form,

$$L(\xi, d) = \rho(d - \xi)$$

for some function ρ. This is called an **invariant loss function**.

Suppose we chose $T(\boldsymbol{X})$ as a good estimator of ξ in the original scale. Then since estimation of ξ' in terms of $X_1', \cdots, X_n'$ is exactly the same problem, we should use

$$T'(X_1', \cdots, X_n') = T(X_1, \cdots, X_n) + a$$

as our estimator of $\xi' = \xi + a$. If

$$T(X_1 + a, \cdots, X_n + a) = T(X_1, \cdots, X_n) + a$$

then we say that the estimator T is *equivariant* under the transformations (3.1.1),(3.1.2), and (3.1.3) or ***location equivariant***. Let $\mathcal{E}$ denote the class of all equivariant estimators.

REMARK 3.1.1. The mean, median, weighted average of order statistics (with $\sum w_i = 1$) and the MLE of ξ for the family $f(x - \xi)$ are all location equivariant.

THEOREM 3.1.2. *If $\boldsymbol{X}$ has density $f(\boldsymbol{x} - \xi\mathbf{1})$ with respect to Lebesgue measure μ and T is equivariant for ξ with loss*

$$L(\xi, d) = \rho(\xi - d).$$

Then the bias, risk and variance of T are independent of ξ.

PROOF. We give the proof for the bias. The other proofs are similar.

$$\begin{aligned}
\mathrm{b}(\xi) &= E_\xi T(\boldsymbol{x}) - \xi \\
&= \int (T(\boldsymbol{x}) - \xi) f(\boldsymbol{x} - \xi\mathbf{1}) \mu(d\boldsymbol{x}) \\
&= \int (T(\boldsymbol{x} + \xi\mathbf{1}) - \xi) f(\boldsymbol{x}) \mu(d\boldsymbol{x}) \text{ by shift-invariance of } \mu \\
&= E_0 T(\boldsymbol{X}') - \xi \\
&= E_0 (T(\boldsymbol{X}) + \xi) - \xi \\
&= E_0 T(\boldsymbol{X}).
\end{aligned}$$

□

REMARK 3.1.3. Since the risk of an equivariant estimator is independent of θ, the determination of a uniformly minimum risk equivariant estimator reduces to finding the equivariant estimator with minimum (for every θ) risk - such an estimator typically exists - and is called a *minimum risk equivariant* (MRE) estimator. Our first step is to find a representation of all location equivariant estimators (Just as we found a representation of all unbiased estimators in Chap 2)

LEMMA 3.1.4. *If T_0 is any location-equivariant estimator then*

$$T \text{ is equivariant} \iff T(\boldsymbol{x}) = T_0(\boldsymbol{x}) - U(\boldsymbol{x}) \tag{3.1.5}$$

where $U(\boldsymbol{x})$ is any function such that

$$U(\boldsymbol{x}+a\mathbf{1}) = U(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \text{ and } a. \tag{3.1.6}$$

PROOF. If T is location-equivariant, set

$$U(\boldsymbol{x}) = T_0(\boldsymbol{x}) - T(\boldsymbol{x}).$$

Then

$$\begin{aligned} U(\boldsymbol{x}+a\mathbf{1}) &= T_0(\boldsymbol{x}+a\mathbf{1}) - T(\boldsymbol{x}+a\mathbf{1}) \\ &= T_0(\boldsymbol{x}) + a - T(\boldsymbol{x}) - a \\ &= U(\boldsymbol{x}). \end{aligned}$$

Conversely if equation (3.1.5) and (3.1.6) hold, then

$$\begin{aligned} T(\boldsymbol{x}+a\mathbf{1}) &= T_0(\boldsymbol{x}+a\mathbf{1}) - U(\boldsymbol{x}+a\mathbf{1}) \\ &= T_0(\boldsymbol{x}) + a - U(\boldsymbol{x}) \\ &= T(\boldsymbol{x}) + a. \end{aligned}$$

□

LEMMA 3.1.5. *U satisfies*

$$U(\boldsymbol{x}+a\mathbf{1}) = U(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \text{ and } a$$

if and only if

$$U(\boldsymbol{x}) = v(x_1 - x_n, \cdots, x_{n-1} - x_n) \text{ for some function } v.$$

PROOF. $\Leftarrow$)

$$U(\boldsymbol{x}+a\mathbf{1}) = v((x_1+a)-(x_n+a), \cdots, (x_{n-1}+a)-(x_n+a)) = U(\boldsymbol{x}).$$

$\Rightarrow$) Choosing $a = -x_n$, we have

$$\begin{aligned} U(\boldsymbol{x}) = U(\boldsymbol{x}+a\mathbf{1}) &= U(x_1 - x_n, \cdots, x_{n-1} - x_n, 0) \\ &= v(x_1 - x_n, \cdots, x_{n-1} - x_n). \end{aligned}$$

Combining these two lemmas gives the following theorem. □

THEOREM 3.1.6. *If T_0 is any location-equivariant estimator, then a necessary and sufficient condition for T to be equivariant is that there is a function v of $n-1$ variables such that*

$$T(\boldsymbol{x}) = T_0(\boldsymbol{x}) - v(\boldsymbol{y}) \text{ for all } \boldsymbol{x},$$

where $\boldsymbol{y} = (x_1 - x_n, \cdots, x_{n-1} - x_n)$.

EXAMPLE 3.1.7. If $n = 1$, then only equivariant estimators are $X + c$ for $c \in \mathbb{R}$.

Now we can determine the location-equivariant estimator with minimum risk.

THEOREM 3.1.8. *Let $\boldsymbol{x}$ have a density function $f(\boldsymbol{x}-\theta)$ with respect to Lebesgue measure and let $\boldsymbol{y}=(y_1,\cdots,y_{n-1})^T$ where $y_i=x_i-x_n$. Suppose that the loss function is given by $L(\theta,d)=\rho(d-\theta)$ and that there exists an equivariant estimator T_0 with finite risk. Assume that for each $\boldsymbol{y}$ there exists a number $v(\boldsymbol{y})=v^*(\boldsymbol{y})$ which minimizes*

$$E_0(\rho(T_0(\boldsymbol{X})-v(\boldsymbol{Y}))|\boldsymbol{Y}=\boldsymbol{y}) \tag{3.1.7}$$

then there exists an MRE estimator T^* of θ given by

$$T^*(\boldsymbol{x})=T_0(\boldsymbol{x})-v^*(\boldsymbol{y}).$$

PROOF. If T is equivariant then

$$T(\boldsymbol{X})=T_0(\boldsymbol{X})-v(\boldsymbol{Y})$$

for some v. So to find the MRE, we need to find v to minimize

$$R(\theta,T)=E_\theta(\rho(T-\theta))$$

and we calculate:

$$\begin{aligned}
R(\theta,T) &= E_\theta(\rho(T_0(\boldsymbol{X})-v(\boldsymbol{Y})-\theta)) \\
&= E_0(\rho(T_0(\boldsymbol{X})-v(\boldsymbol{Y}))) \text{ by theorem (3.1.2)} \\
&= \int E_0(\rho(T_0(\boldsymbol{X})-v(\boldsymbol{Y}))|\boldsymbol{Y}=\boldsymbol{y})\,dP_0(\boldsymbol{y}) \\
&\geq \int E_0(\rho(T_0(\boldsymbol{X})-v^*(\boldsymbol{Y}))|\boldsymbol{Y}=\boldsymbol{y})\,dP_0(\boldsymbol{y}) \\
&= R(0,T^*).
\end{aligned}$$

The risk is finite since $R(0,T^*)\leq R(0,T_0)<\infty$ by assumption. □

COROLLARY 3.1.9. *If ρ is convex and not monotone, then an MRE exists and is unique if ρ is strictly convex (under the conditions of theorem (3.1.8)).*

PROOF. Let

$$\phi(c)=E(\rho(T_0(\boldsymbol{X})-c)|\boldsymbol{Y}=\boldsymbol{y})$$

and apply theorem (1.4.2). □

COROLLARY 3.1.10. *The following results hold:*

(1) $\rho(d-\theta)=(d-\theta)^2 \quad\Rightarrow\quad v^*(\boldsymbol{y})=E_0(T_0(\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y})$
(2) $\rho(d-\theta)=|d-\theta| \quad\Rightarrow\quad v^*(\boldsymbol{y})=\text{med}_0(T_0(\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y})$

PROOF. (1) $E_0(\rho(T_0(\boldsymbol{X})-c)|\boldsymbol{Y}=\boldsymbol{y}) = E_0\left((T_0(\boldsymbol{X})-c)^2|\boldsymbol{Y}=\boldsymbol{y}\right)$ is minimized at $c=E_0(T_0(\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y})$.
(2) $E_0(|T_0(\boldsymbol{X})-c||\boldsymbol{Y}=\boldsymbol{y})$ is minimized at $c=\text{med}_0(T_0(\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y})$.

□

EXAMPLE 3.1.11. MRE's can exist also for non-convex ρ. Suppose ρ is given by

$$\rho(d-\theta) = \begin{cases} 1 & \text{if } |d-\theta| > c \\ 0 & \text{otherwise,} \end{cases}$$

where c is fixed. Then for $n=1$, using $T_0(X) = X$, v will minimize

$$E_0\rho(X-v) = P_0(|X-v| > c)$$

if and only if it maximizes

$$P_0(|X-v| \le c).$$

If f is symmetric and unimodal then $v^* = 0$ and hence

$$T_0(X) - 0 = X \text{ is MRE.}$$

On the other hand if f is U-shaped, say $f(x) = (x^2+1) I_{[-L,L]}$ where $c < L$, then

$$P_0(|X-v| \le c) = P_0(v-c \le X \le v+c)$$

is maximized at $v+c = L$ and $v-c = -L$. Thus there are two MREE's, $X-L+c$ and $X+L-c$.

EXAMPLE 3.1.12. Let $X_1, \cdots, X_n$ be iid $N(\mu, \sigma^2)$ where σ^2 is known. Because $\overline{X}$ is complete sufficient and $\boldsymbol{Y} = (X_1 - X_n, \cdots, X_{n-1} - X_n)$ is ancillary, $T_0(\boldsymbol{X}) = \overline{X}$ is independent of $\boldsymbol{Y}$. Therefore, by minimizing the expression (3.1.7), we find that

$$v^*(\boldsymbol{y}) = \text{argmin } E_0\left(\rho\left(\overline{X} - v\right)\right).$$

If ρ is convex and even, then $\phi(v) := E_0\rho\left(\overline{X} - v\right)$ is convex and even. Therefore, $v^*(\boldsymbol{y}) = 0$ and $\overline{X}$ is MRE. ($\overline{X}$ is also MRE when ρ is the non-convex function of Example 3.1.11.)

THEOREM 3.1.13. (Least favorable property of the normal distribution) *Let $\mathcal{F}$ be the set of all distributions with pdf relative to Lebesgue measure and with variance 1. Let $X_1, \cdots, X_n$ be iid with pdf $f(x-\xi)$, where $\xi = EX_i$. If $r_n(F)$ is the risk of the MRE estimator of ξ with squared error loss, then $r_n(F)$ has its maximum value over $\mathcal{F}$ when F is normal.*

PROOF. By the previous example, the MREE in the normal case is $\overline{X}$ with corresponding risk,

$$r_n = E_0\overline{X}^2 = \frac{1}{n}.$$

However, $\overline{X}$ is an equivariant estimator of ξ for all $F \in \mathcal{F}$, and the risk of $\overline{X}$ is

$$E_\xi\left(\overline{X} - \xi\right)^2 = \frac{1}{n}, \quad \text{for all } F \in \mathcal{F}.$$

Therefore, we must have

$$r_n(F) \le \frac{1}{n}, \quad \text{for all } F \in \mathcal{F}.$$

□

REMARK 3.1.14. From corollary (3.1.10), the MRE in general in the previous theorem is

$$\overline{X} - E_0\left(\overline{X}|\boldsymbol{Y} = \boldsymbol{y}\right).$$

But for $n \geq 3$, $E_0\left(\overline{X}|\boldsymbol{Y} = \boldsymbol{y}\right) = 0$ if and only if F is normal (Kagan, Linnik, Rao (1965, 1973)). So the MRE for $n \geq 3$ is $\overline{X}$ if and only if F is normal.

EXAMPLE 3.1.15. Let $X_1, \cdots, X_n$ be iid with

$$F(x) = \begin{cases} 1 - e^{-(x-\theta)/b} & x \geq \theta \\ 0 & x < \theta, \end{cases}$$

where b is known and $\theta \in \mathbb{R}$. Then $T_0 = X_{(1)}$ is equivariant, complete and sufficient for θ (CHECK) and T_0 is independent of the ancillary statistic

$$\boldsymbol{Y} = (X_1 - X_n, \cdots, X_{n-1} - X_n) = (X_1 - \theta - (X_n - \theta), \cdots, X_{n-1} - \theta - (X_n - \theta)).$$

Therefore $X_{(1)} - v^*$ is MRE if v^* minimizes

$$E_0\rho\left(X_{(1)} - v\right).$$

(In general we have to minimize $E_0\left(\rho\left(T_0 - v\left(y\right)\right)|\boldsymbol{Y} = \boldsymbol{y}\right)$ for each $\boldsymbol{y}$ by Theorem 3.1.8 but here the complete sufficiency of T_0 and the ancillarity of $\boldsymbol{Y}$ implies that $v^*\left(\boldsymbol{y}\right)$ is independent of $\boldsymbol{y}$ since the distribution of T_0 is the same for all $\boldsymbol{y}$.)

We now consider some special cases:

(1) If $\rho\left(d - \theta\right) = \left(d - \theta\right)^2$, then $v = E_0\left(X_{(1)}\right) = b/n$ so that MRE is $X_{(1)} - \frac{b}{n}$.
(2) If $\rho\left(d - \theta\right) = |d - \theta|$, then $v = \text{med}_0\left(X_{(1)}\right) = b\log 2/n$ so that MRE is $X_{(1)} - \frac{b\log 2}{n}$. (Because $F_{X_{(1)}}\left(x\right) = 1 - e^{-(x-\theta)n/b} = \frac{1}{2}$ implies $\left(x - \theta\right)n/b = \log 2$.)
(3) If

$$\rho\left(d - \theta\right) = \begin{cases} 1 & \text{if } |d - \theta| > c, \\ 0 & \text{if } |d - \theta| \leq c, \end{cases}$$

then v is the center of the interval I of length $2c$ which maximizes $P_0\left(X_{(1)} \in I\right)$ so that $v = c$ and the MREE is $X_{(1)} - c$.

THEOREM 3.1.16. (**Pitman Estimator**) *Under the assumptions of Theorem 3.1.8, if* $L\left(\theta, d\right) = \left(d - \theta\right)^2$,

$$T^* = \frac{\int uf\left(x_1 - u, \cdots, x_n - u\right)du}{\int f\left(x_1 - u, \cdots, x_n - u\right)du}$$

is the MREE of θ.

REMARK 3.1.17. This coincides with the Bayes estimator corresponding to an improper flat prior for the location parameter. (i.e. the conditional expectation of Θ given $\mathbf{X} = \mathbf{x}$ under the assumed joint "density" $f\left(\boldsymbol{x} - \theta\mathbf{1}\right)$ of $\mathbf{X}$ and Θ.)

PROOF. Corollary (3.1.10) implies

$$T^*\left(\boldsymbol{X}\right) = T_0\left(\boldsymbol{X}\right) - E_0\left(T_0|\boldsymbol{Y}\right),$$

where T_0 is any equivariant estimator. Let $T_0(\boldsymbol{X}) = X_n$. Because

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{n-1} \\ X_n \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 & -1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & -1 \\ 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_{n-1} \\ X_n \end{bmatrix},$$

we have

$$f_{Y_1,\cdots,Y_{n-1},X_n}(y_1,\cdots,y_{n-1},x_n) = f_{X_1,\cdots,X_n}(y_1+x_n,\cdots,y_{n-1}+x_n,x_n)$$

and

$$\begin{aligned} E_0(X_n|\boldsymbol{Y}=\boldsymbol{y}) &= \frac{\int x f(y_1+x,\cdots,y_{n-1}+x,x)\,dx}{\int f(y_1+x,\cdots,y_{n-1}+x,x)\,dx} \\ &= \frac{\int x f(x_1-x_n+x,\cdots,x_{n-1}-x_n+x,x)\,dx}{\int f(y_1-x_n+x,\cdots,x_{n-1}-x_n+x,x)\,dx} \\ &= x_n - \frac{\int u f(x_1-u,\cdots,x_{n-1}-u,x_n-u)\,du}{\int f(x_1-u,\cdots,x_{n-1}-u,x_n-u)\,du}. \end{aligned}$$

Substituting in the expression for T^* completes the proof. □

EXAMPLE 3.1.18. Suppose $f(x) = I_{(-1/2,1/2)}(x)$ and $X_1,\cdots,X_n$ are iid with density $f(x-\theta) = I_{(\theta-1/2,\theta+1/2)}(x)$. Then

$$f(x_1,\cdots,x_n) = \begin{cases} 1 & \text{if } -\frac{1}{2} \le x_{(1)} \text{ and } x_{(n)} \le \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f(\boldsymbol{x}-u\mathbf{1}) = \begin{cases} 1 & \text{if } u-\frac{1}{2} \le x_{(1)} \text{ and } x_{(n)} \le u+\frac{1}{2} \\ 0 & \text{otherwise,} \end{cases}.$$

Therefore, since $x_{(n)} - x_{(1)} \le 1$,

$$\begin{aligned} T^*(\boldsymbol{x}) &= \frac{\int_{x_{(n)}-\frac{1}{2}}^{x_{(1)}+\frac{1}{2}} u\,du}{\int_{x_{(n)}-\frac{1}{2}}^{x_{(1)}+\frac{1}{2}} du} \\ &= \frac{\frac{1}{2}\left(\left(x_{(1)}+\frac{1}{2}\right)^2 - \left(x_{(n)}-\frac{1}{2}\right)^2\right)}{x_{(1)}-x_{(n)}} \\ &= \frac{1}{2}\left(x_{(1)}+x_{(n)}\right). \end{aligned}$$

REMARK 3.1.19. (**UMVU vs. MRE**)

- MRE estimators often exist for more than just convex loss functions.
- For convex loss functions MRE estimators generally vary with the loss function (unlike UMVUE's).

- UMVUE's are frequently inadmissible (i.e., there exists an estimator with uniformly smaller risk). The Pitman estimator is admissible under mild assumptions.
- The principal application of UMVUE's is to exponential families.
- For location problems, UMVUE's typically do not exist.
- An MRE estimator is not necessarily unbiased. The following lemma examines this connection.

LEMMA 3.1.20. *Suppose* $L(d,\theta) = (d-\theta)^2$, *and* $f(\mathbf{x}-\theta\mathbf{1})$, $\theta \in \mathbb{R}$, *are densities with respect to Lebesgue measure.*

(1) *If* T *is equivariant with bias* b, *then* $T-b$ *is equivariant, unbiased, with smaller risk than* T.
(2) *The unique MRE estimator is unbiased.*
(3) *If there exists an UMVU estimator and it is equivariant, then it is MRE.*

PROOF. (1) It is clear that $T-b$ is equivariant and unbiased. For the smaller risk part,

$$R(\theta, T-b) = R(0, T-b) = E_0(T-b)^2 = \mathrm{Var}T \leq E_0\left(T^2\right) = R(0,T) = R(\theta,T).$$

(2) The MRE estimator is unique by corollary (3.1.9). It is unbiased by (1), otherwise its risk could be improved by using the equivariant estimator $T-b$.
(3) The UMVUE is the unique MR estimator in $\mathcal{U}$. If it falls in $\mathcal{E}$ it is the MR estimator in $\mathcal{U}\cap\mathcal{E}$. But the MREE is the MR estimator in $\mathcal{U}\cap\mathcal{E}$ since it is necessarily unbiased. Hence they are the same.

$$\mathcal{U} = \{\text{unbiased estimators of zero}\}, \quad \mathcal{E} = \{\text{equivariant estimators}\}.$$

□

DEFINITION 3.1.21. An estimator T of $g(\theta)$ is *risk-unbiased* if

$$E_\theta L(\theta, T) \leq E_\theta L(\theta', T) \text{ for all } \theta \neq \theta'.$$

i.e. T is "closer" to $g(\theta)$ on average than to any false value $g(\theta')$.

EXAMPLE 3.1.22. (mean unbiasedness) If $L(\theta,d) = (d-g(\theta))^2$ and T is risk-unbiased, then

$$E_\theta(T-g(\theta))^2 \leq E_\theta(T-g(\theta'))^2 \text{ for all } \theta \neq \theta'.$$

This means (assuming that $E_\theta T^2 < \infty$ and $E_\theta T \in g(\Omega)$) that $E_\theta(T-g(\theta'))^2$ is minimized by $g(\theta') = E_\theta T$.

Hence $g(\theta) = E_\theta T$ for all θ i.e. T is unbiased for g in the sense defined in chapter 2.

EXAMPLE 3.1.23. (Median unbiasedness) If $L(\theta,d) = |d-g(\theta)|$ and T is risk-unbiased, then

$$E_\theta|T-g(\theta)| \leq E_\theta|T-g(\theta')| \text{ for all } \theta \neq \theta'.$$

But the right hand side is minimized when $g(\theta') = \mathrm{med}_\theta T$.

Hence

$$g(\theta) = \text{med}_\theta T \text{ for all } \theta. \tag{3.1.8}$$

(assuming that $E_\theta |T| < \infty$ and there exists a $\text{med}_\theta T$ in $g(\Omega)$ for all θ.) An estimator satisfying equation (3.1.8) is said to be *median-unbiased for* $g(\theta)$.

THEOREM 3.1.24. *Suppose* $\boldsymbol{X}$ *has density* $f(\mathbf{x} - \theta\mathbf{1})$ *with respect to Lebesgue measure. If* T *is an MRE estimator with* $L(\theta, d) = \rho(d - \theta)$ *then* T *is risk-unbiased (for* θ*).*

PROOF. By Theorem 3.1.2,

$$E_\theta \rho(T - \theta) = E_0 \rho(T).$$

If $\theta \neq \theta'$, then $T - \theta'$ is equivariant and by definition of MRE, we have $E_0\rho(T) \leq E_0\rho(T - \theta')$ for all θ'

$$\begin{aligned} &\Rightarrow \quad E_0\rho(T) \leq E_\theta\rho(T - \theta' - \theta) \quad \text{for all } \theta' \\ &\Rightarrow \quad E_0\rho(T) \leq E_\theta\rho(T - \theta') \quad \text{for all } \theta' \\ &\Rightarrow \quad E_\theta\rho(T - \theta) \leq E_\theta\rho(T - \theta') \quad \text{for all } \theta'. \end{aligned}$$

□

3.2. The General Equivariant Framework

Notation

X: data.
Ω: parameter space.
$\mathcal{P}$: $= \{P_\theta : \theta \in \Omega\}$.
$\mathcal{G}$: a group of measurable bijective transformations of $\mathcal{X} \to \mathcal{X}$.

REMARK 3.2.1. The operation associated with the group $\mathcal{G}$ is function composition. i.e.

$$fg(x) = f \circ g(x) = f(g(x)).$$

DEFINITION 3.2.2. We say that g *leaves* $\mathcal{P}$ *invariant* if for all $\theta \in \Omega$, there exists $\theta' \in \Omega$ such that

$$X \sim P_\theta \Rightarrow gX \sim P_{\theta'}$$

and there exists $\theta^* \in \Omega$ such that

$$X \sim P_{\theta^*} \Rightarrow gX \sim P_\theta.$$

If $\mathcal{C}$ is a class of transformations which leave $\mathcal{P}$ invariant then

$$\mathcal{G}(\mathcal{C}) = \left\{ g_1^{\pm 1} g_2^{\pm 1} \cdots g_m^{\pm 1} : g_i \in \mathcal{C},\, m \in \mathbb{N} \right\}$$

is a group (the group generated by $\mathcal{C}$), each of whose member leaves $\mathcal{P}$ invariant.

If each member of a group $\mathcal{G}$ leaves $\mathcal{P}$ invariant we say that $\mathcal{G}$ leaves $\mathcal{P}$ invariant. If $\mathcal{G}$ leaves $\mathcal{P}$ invariant and $P_\theta \neq P_{\theta'}$ for $\theta \neq \theta'$ then there exists a unique $\theta' \in \Omega$ such that

$$X \sim P_\theta \Rightarrow gX \sim P_{\theta'}.$$

This defines a bijection $\overline{g} : \Omega \to \Omega$, via the relation

$$\overline{g}(\theta) = \theta'.$$

where $P_{\overline{g}(\theta)}$ is the distribution of $g(\boldsymbol{X})$ under θ.

DEFINITION 3.2.3. Under the preceding conditions we define

$$\overline{\mathcal{G}} := \{\overline{g} : g \in \mathcal{G}\}.$$

It is clear that $\overline{\mathcal{G}}$ is also a group.

REMARK 3.2.4. For $g \in \mathcal{G}$,

$$E_\theta f(gX) = E_{\overline{g}(\theta)} f(X)$$

since

$$\begin{aligned} E_\theta f(gX) &= \int f(g(x)) P_\theta(dx) \\ &= \int f(x) P_\theta \circ g^{-1}(dx) \\ &= \int f(x) P_{\overline{g}(\theta)}(dx) \\ &= E_{\overline{g}(\theta)} f(X). \end{aligned}$$

Equivariant Estimation

Let $T : \mathcal{X} \to \mathcal{D}$ be an estimator of $h(\theta)$. Instead of observing X, suppose we observe

$$X' = g(X),$$

where X' is a sample from $P_{\overline{g}(\theta)}$. Suppose that for any $\overline{g} \in \overline{\mathcal{G}}$, $h(\overline{g}\theta)$ depends on θ only through $h(\theta)$, i.e.

$$h(\theta_1) = h(\theta_2) \Rightarrow h(\overline{g}\theta_1) = h(\overline{g}\theta_2). \tag{3.2.1}$$

Then we denote

$$g^* h(\theta) = h(\overline{g}\theta).$$

It is clear that

$$\mathcal{G}^* := \left\{g^* : \overline{g} \in \overline{\mathcal{G}}\right\}$$

is a group and each g^* is a 1 to 1 mapping from $\mathcal{H} = h(\Omega)$ to $\mathcal{H}$.

DEFINITION 3.2.5. Under the conditions prescribed for existence of the groups of mappings $\overline{\mathcal{G}}$ and $\mathcal{G}^*$, if

$$L\left(\overline{g}\theta, g^*d\right) = L\left(\theta, d\right) \text{ for all } g \in \mathcal{G}, \tag{3.2.2}$$

then we say that L **is invariant under** $\mathcal{G}$. (g if we remove 'for all $g \in \mathcal{G}$'). If conditions (3.2.1) and (3.2.2) hold, we say that **the problem of estimating** $h\left(\theta\right)$ is **invariant under the group of transformations** $\mathcal{G}$ (under g if we remove "for all $g \in \mathcal{G}$"). An estimator $T\left(\boldsymbol{X}\right)$ of $h\left(\theta\right)$ is said to be **equivariant under** $\mathcal{G}$ if

$$g^*T\left(\boldsymbol{X}\right) = T\left(g\boldsymbol{X}\right) \text{ for all } g \in \mathcal{G}. \tag{3.2.3}$$

If (3.2.2) and (3.2.3) hold and $T\left(\boldsymbol{X}\right)$ is a good estimator of $h\left(\theta\right)$ based on $\boldsymbol{X}$ then $T\left(g\boldsymbol{X}\right)$ will be a good estimator of $g^*(h\left(\theta\right))$ based on $g(\mathbf{X})$.

EXAMPLE 3.2.6. (Location parameter) Let $h\left(\theta\right) = \theta$ and $g\left(\boldsymbol{X}\right) = \boldsymbol{X} + a$.

$$\begin{array}{ll} \boldsymbol{X} \to \boldsymbol{X} + a\mathbf{1} & \\ \boldsymbol{X} + a \sim P_{\theta+a} & \theta \to \overline{g}\theta = \theta + a \\ h\left(\overline{g}\theta\right) = \theta + a & g^*(h(\theta)) = h(\theta) + a. \end{array}$$

Then, the problem of estimating θ is location-invariant if

$$L(\overline{g}\theta, g^*d) = L\left(\theta + a, d + a\right) = L\left(\theta, d\right).$$

An estimator of $h(\theta)$ is equivariant if

$$T(\mathbf{X} + a\mathbf{1}) = T(\mathbf{X}) + a,$$

e.g. X_1, median$(\mathbf{X})$, $n^{-1}\sum_{i=1}^{n} X_i$, etc.

THEOREM 3.2.7. *If T is equivariant and g leaves $\mathcal{P}$ invariant and*

$$L\left(\theta, d\right) = L\left(\overline{g}\theta, g^*d\right),$$

then

$$R\left(\theta, T\right) = R\left(\overline{g}\left(\theta\right), T\right) \text{ for all } \theta,$$

where

$$R\left(\theta, T\right) := E_\theta L\left(\theta, T\left(\boldsymbol{X}\right)\right).$$

PROOF.

$$\begin{aligned} E_\theta L\left(\theta, T\left(\boldsymbol{X}\right)\right) &= E_\theta L\left(\overline{g}\left(\theta\right), g^*T\left(\boldsymbol{X}\right)\right) \\ &= E_\theta L\left(\overline{g}\left(\theta\right), T\left(g\left(\boldsymbol{X}\right)\right)\right) \\ &= E_{\overline{g}(\theta)} L\left(\overline{g}\left(\theta\right), T\left(\boldsymbol{X}\right)\right) \\ &= R\left(\overline{g}\left(\theta\right), T\right). \end{aligned}$$

□

COROLLARY 3.2.8. *If $\overline{\mathcal{G}}$ is transitive over Ω (i.e. if $\theta_1, \theta_2 \in \Omega$ and $\theta_1 \neq \theta_2$ then there exists $\overline{g} \in \overline{\mathcal{G}}$ such that $\overline{g}\left(\theta_1\right) = \theta_2$) then $R\left(\theta, T\right)$ is constant for every equivariant T.*

PROOF. Fix $\theta_0 \in \Omega$. If $\theta \neq \theta_0$, there exists $\overline{g}$ such that $\overline{g}(\theta_0) = \theta$. Hence

$$R(\theta_0, T) = R(\overline{g}(\theta_0), T) = R(\theta, T).$$

□

REMARK 3.2.9. $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ is invariant relative to a transitive group of transformations if and only if $\mathcal{P}$ is a *group family* (see TPE section 1.4.1). These are families generated by subjecting a r.v. with fixed distribution to a group of transformations. We can then index $\mathcal{P}$ using $\overline{\mathcal{G}}$. Thus

$$\mathcal{P} = \left\{P_{\overline{g}} : \overline{g} \in \overline{\mathcal{G}}\right\},$$

where $\theta = \overline{g}(\theta_0)$.

THEOREM 3.2.10. *Suppose $\overline{\mathcal{G}}$ is transitive and $\mathcal{G}^*$ is commutative. If T is MRE, then T is risk unbiased.*

PROOF. Let T be MRE. For $\theta \neq \theta'$, there exists $\overline{g}$ such that $\overline{g}(\theta') = \theta$. Consequently,

$$\begin{aligned} E_\theta\left(L\left(\theta', T(\boldsymbol{X})\right)\right) &= E_\theta L\left(\overline{g}^{-1}(\theta), T(\boldsymbol{X})\right) \\ &= E_\theta L\left(\theta, g^* T(\boldsymbol{X})\right) \\ &\geq E_\theta L(\theta, T). \end{aligned}$$

(In the 2nd equality, note that if T is equivariant then so is g^*T, since letting $h^* \in \mathcal{G}^*$, we have

$$h^* g^* T = g^* h^* T = g^* T(h),$$

by commutativity.) The inequality then follows because T is MRE. □

EXAMPLE 3.2.11. Suppose $X \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$, and we want to estimate $h(\theta) = \mu$.

(1) Let $\mathcal{G}_1 = \{g : gx = x + c,\ c \in \mathbb{R}\}$, then $\mathcal{P}$ is invariant under $\mathcal{G}_1$. $Tx = x + c$, $c \in \mathbb{R}$, are the only equivariant functions since

$$\begin{aligned} & T(x + a) = T(x) + a \quad \text{for all } a \\ \Rightarrow \quad & \tfrac{T(x+a) - T(x)}{a} = 1 \quad \text{for all } a \\ \Rightarrow \quad & T'(x) = 1 \\ \Rightarrow \quad & T(x) = x + c. \end{aligned}$$

(a) Suppose $L(d, \theta) = (d - \mu)^2 / \sigma^2$ (squared loss function measured in units of σ). Then X is MRE under $\mathcal{G}_1$, i.e.

$$\begin{aligned} E_\theta L(\theta, X) &= E_\theta \frac{(X - \mu)^2}{\sigma^2} \\ &= 1 \\ &= \min\left\{E_0 \frac{(T - \mu)^2}{\sigma^2} : T \text{ equivariant with respect to } \mathcal{G}_1\right\}. \end{aligned}$$

Because

$$\overline{g}\left(\left(\mu,\sigma^2\right)\right)=\left(\mu+c,\sigma^2\right),$$

$\overline{\mathcal{G}}_1$ is not transitive and X is not risk unbiased. To see this, for θ fixed choose $\theta'=(\mu,10\sigma^2)$, whence

$$E_\theta L\left(\theta',X\right)=E_\theta\frac{\left(X-\mu\right)^2}{10\sigma^2}=\frac{1}{10}<E_\theta L\left(\theta,X\right)=1.$$

Here, $\overline{\mathcal{G}}_1$ is not transitive since there exists no $\overline{g}$ such that $\overline{g}\left(\mu,10\sigma^2\right)=\overline{g}\left(\mu,\sigma^2\right)$.

(b) Suppose $L'\left(d,\theta\right)=\left(d-\mu\right)^2$. X is MRE and risk-unbiased since

$$E_\theta L'\left(\theta',X\right)=E_\theta\left(X-\mu'\right)^2\geq E_\theta\left(X-\mu\right)^2\text{ for all }\theta\neq\theta'.$$

$\overline{\mathcal{G}}_1$ transitive is therefore not *necessary* in theorem (3.2.10)

(2) Let $\mathcal{G}_2=\{g:\,gx=ax+c,\,a>0,\,c\in\mathbb{R}\}$. Then, $\overline{\mathcal{G}}_2=\{\overline{g}:\,\overline{g}\left(\mu,\sigma^2\right)=\left(a\mu+c,a^2\sigma^2\right)\}$ since $gX\sim N\left(a\mu+c,a^2\sigma^2\right)$. Also $h\left(\theta\right)=\mu$ so that $h\left(\overline{g}\left(\theta\right)\right)=h\left(a\mu+c,a^2\sigma^2\right)=ah\left(\theta\right)+c$. Thus,

$$g^*h\left(\theta\right)=a\mu+c\text{ and }\mathcal{G}_2^*=\left\{g^*:\,g^*\left(d\right)=ad+c\right\}.$$

$\mathcal{P}$ is invariant under $\mathcal{G}_2$ and $T\left(x\right)=x$ is equivariant under $\mathcal{G}_2$ since

$$Tgx=T\left(ax+c\right)=ax+c=g^*\left(T\left(x\right)\right)$$

and

$$g^*(Tx)=g^*(x)=ax+c.$$

X is MRE under $\mathcal{G}_1\subset\mathcal{G}_2$ and is therefore MRE under $\mathcal{G}_2$. (There exist no $\mathcal{G}_1$-equivariant estimators with smaller risk so there exist no $\mathcal{G}_2$-estimators with smaller risk since $\mathcal{G}_2$-equivariance is a more severe restriction than $\mathcal{G}_1$-equivariance.) But X is not risk unbiased with respect to $L\left(d,\theta\right)=\left(d-\mu\right)^2/\sigma^2$. $\overline{\mathcal{G}}_2$ is transitive in this case but $\mathcal{G}_2^*$ is not commutative. So, theorem (3.2.10) cannot be applied here.

3.3. Location-Scale Families

Suppose

$$\boldsymbol{X}=\left(X_1,\cdots,X_n\right)^T\sim\sigma^{-n}f\left(\frac{x_1-\mu}{\sigma},\cdots,\frac{x_n-\mu}{\sigma}\right),$$

where f is known. The density is with respect to Lebesgue measure and

$$\theta=(\mu,\sigma)\in\Omega=\mathbb{R}\times\mathbb{R}^+,$$

where μ is the **location parameter** and σ is the **scale parameter**.

Estimation of μ. Our first objective is to estimate the location parameter. Suppose

$$\mathcal{X} = \mathbb{R}^n,\ \Omega = \mathbb{R} \times \mathbb{R}^+,\ \mathcal{D} = \mathbb{R}.$$

Define $g_{a,b} : \mathbb{R}^n \to \mathbb{R}^n$ by

$$g_{a,b}(\boldsymbol{x}) = (ax_1 + b, \cdots, ax_n + b).$$

Then $\overline{g}_{a,b} : \Omega \to \Omega$ is defined by

$$\overline{g}_{a,b}(\theta) = (a\mu + b, a\sigma)$$

and $g^*_{a,b} : \mathbb{R} \to \mathbb{R}$ is defined by

$$g^*_{a,b}(d) = ad + b.$$

since $g^* h(\theta) = g^* \mu = h(\overline{g}\theta) = a\mu + b$.

LEMMA 3.3.1. *$L(\theta, d)$ is invariant under $\mathcal{G}$ if and only if*

$$L((\mu, \sigma), d) = \rho\left(\frac{d - \mu}{\sigma}\right)$$

(i.e. if L is a function of error measured in terms of σ).

PROOF. $\Leftarrow$) If $\rho\left(\frac{d-\mu}{\sigma}\right) = L((\mu, \sigma), d)$, then

$$\begin{aligned} L(\overline{g}\theta, g^* d) &= L((a\mu + b, a\sigma), ad + b) \\ &= \rho\left(\frac{ad + b - a\mu - b}{a\sigma}\right) \\ &= L(\theta, d). \end{aligned}$$

$\Rightarrow$) We need to show if μ, σ, d and μ', σ', d' satisfy $\frac{d-\mu}{\sigma} = \frac{d'-\mu'}{\sigma'}$ and if L is invariant then

$$L((\mu, \sigma), d) = L((\mu', \sigma'), d').$$

But this holds since

$$\begin{cases} d = \frac{\sigma}{\sigma'} d' + \mu - \frac{\sigma}{\sigma'} \mu' \\ \sigma = \frac{\sigma}{\sigma'} \sigma' \\ \mu = \frac{\sigma}{\sigma'} \mu' + \mu - \frac{\sigma}{\sigma'} \mu' \end{cases}$$

and L is invariant under this transformation by assumption. □

Now observe that $\mathcal{P}$ is invariant under each $g \in \mathcal{G}$ (check!)

$$\begin{aligned} T \text{ is equivariant} &\iff g^* T(\boldsymbol{x}) = Tg(\boldsymbol{x}) \\ &\iff aT(\boldsymbol{x}) + b = T(a\boldsymbol{x} + b). \end{aligned}$$

Since $\overline{\mathcal{G}}$ is transitive (given any (μ, σ) we can find a, b such that $(a\mu + b, a\sigma) = (\mu', \sigma')$), every equivariant estimator of μ has constant risk by corollary (3.2.8).

PROPOSITION 3.3.2. *If T_0 is an equivariant estimator of μ and $\delta_1 \neq 0$, and if*

$$\delta_1(a\boldsymbol{X}+b) = a\delta_1(\boldsymbol{X}) \text{ for all } a>0,\, b\in\mathbb{R}$$

then T is equivariant if and only if

$$T(\boldsymbol{X}) = T_0(\boldsymbol{X}) - W(\boldsymbol{Z})\,\delta_1(\boldsymbol{X})$$

for some W where

$$\boldsymbol{Z} = (Z_1,\cdots,Z_{n-2},Z_{n-1}) = \left(\frac{X_1-X_n}{X_{n-1}-X_n},\cdots,\frac{X_{n-2}-X_n}{X_{n-1}-X_n},\frac{X_{n-1}-X_n}{|X_{n-1}-X_n|}\right).$$

NOTE 3.3.3.
(1) We have assumed that $X_i \neq X_j$ for all $i \neq j$. This is ok since $\{\boldsymbol{X} : X_i = X_j \text{ for some } i \neq j\}$ has measure 0.
(2) $\boldsymbol{Z}$ is ancillary for μ so if T is a function of a CSS then T is independent of $\boldsymbol{Z}$.

PROOF.
(1) T is equivariant (i.e. $T(a\boldsymbol{x}+b) = aT(\boldsymbol{x})+b) \Longleftrightarrow U := \frac{T-T_0}{\delta_1}$ satisfies $U(a\boldsymbol{x}+b) = U(\boldsymbol{x})$ for all $a>0$,$b\in\mathbb{R}$. (check the details - simple algebra)
(2) If $U = \frac{T-T_0}{\delta_1} = W(\boldsymbol{Z})$ for some W then it is easy to see that $U(a\boldsymbol{x}+b) = U(\boldsymbol{x})$ and hence by (1), T is equivariant.
(3) If T is equivariant then $U(a\boldsymbol{x}+b) = U(\boldsymbol{x})$ for all $a>0$, $b\in\mathbb{R}$. Setting

$$a = \frac{1}{|X_{n-1}-X_n|} \quad b = -\frac{X_n}{|X_{n-1}-X_n|},$$

$$\begin{aligned}
\Rightarrow U(\boldsymbol{X}) &= U\left(\frac{X_1-X_n}{|X_{n-1}-X_n|},\cdots,\frac{X_{n-1}-X_n}{|X_{n-1}-X_n|},0\right)\\
&= U(Z_1Z_{n-1},\ldots,Z_{n-2}Z_{n-1},Z_{n-1},0)\\
&=: V(Z_1,\ldots,Z_{n-1}).
\end{aligned}$$

□

THEOREM 3.3.4. *Suppose T_0 is equivariant with finite risk and δ_1 and $\boldsymbol{Z}$ are as defined in Proposition 3.3.2. If*

$$E_{(0,1)}\left(\rho\left(T_0(\boldsymbol{X}) - W(\boldsymbol{Z})\,\delta_1(\boldsymbol{X})\right)|\boldsymbol{Z}\right)$$

is minimized when $W(\boldsymbol{Z}) = W^*(\boldsymbol{Z})$ then

$$T(\boldsymbol{X}) = T_0(\boldsymbol{X}) - W^*(\boldsymbol{Z})\,\delta_1(\boldsymbol{X})$$

is MRE.

PROOF. For any equivariant $T = T_0 - W(\boldsymbol{Z})\delta_1$,

$$\begin{aligned} R((0,1),T) &= E_{(0,1)}\rho\left(\frac{T(\boldsymbol{X})-0}{1}\right) \\ &= E_{(0,1)}\left(E_{(0,1)}\left(\rho\left(T_0(\boldsymbol{X}) - W(\boldsymbol{Z})\delta_1(\boldsymbol{X})\right)|\boldsymbol{Z}\right)\right) \\ &\geq E_{(0,1)}\left(E_{(0,1)}\left(\rho\left(T_0(\boldsymbol{X}) - W^*(\boldsymbol{Z})\delta_1(\boldsymbol{X})\right)|\boldsymbol{Z}\right)\right) \\ &= E_{(0,1)}\rho(T^*) \\ &= R((0,1),T^*). \end{aligned}$$

Since every equivariant estimator has constant risk, T^* is MRE (for all θ). □

EXAMPLE 3.3.5. Suppose $X_1, \cdots, X_n$ are iid with common pdf with respect to Lebesgue measure,

$$\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}e^{-\frac{x-\mu}{\sigma}}I_{(\mu,\infty)}(x)$$

and

$$L((\mu,\sigma),d) = \rho\left(\frac{d-\mu}{\sigma}\right) = \left(\frac{d-\mu}{\sigma}\right)^2.$$

Let $T_0 = X_{(1)}$, $\delta_1(\boldsymbol{X}) = \sum_{i=2}^n \left(X_{(i)} - X_{(1)}\right)$, $Z_i = \frac{X_i - X_n}{X_{n-1} - X_n}$ $(i = 1, \cdots, n-2)$, $Z_{n-1} = \frac{X_{n-1} - X_n}{|X_{n-1} - X_n|}$.

We first show that T_0, δ_1, and $\boldsymbol{Z}$ are independent (see TPE example (2.2.5)). $\left(X_{(1)}, \delta_1\right)$ is complete sufficient for (μ, σ) and $\boldsymbol{Z}$ is already ancillary so that $\left(X_{(1)}, \delta_1\right)$ is independent of $\boldsymbol{Z}$ by Basu's theorem. Also $nX_{(1)}$, $(n-1)\left(X_{(2)} - X_{(1)}\right), \cdots, X_{(n)} - X_{(n-1)}$ are iid $E(1)$ so that $X_{(1)}$ is independent of
$\delta_1 = (n-1)\left(X_{(2)} - X_{(1)}\right) + \cdots + \left(X_{(n)} - X_{(n-1)}\right)$. Thus, $X_{(1)}$, δ_1, and $\boldsymbol{Z}$ are independent.

Consequently,

$$\begin{aligned} &E_{(0,1)}\left(\left(T_0(\boldsymbol{X}) - W(\boldsymbol{Z})\delta_1(\boldsymbol{X})\right)^2|\boldsymbol{Z}\right) \\ &\qquad = E_{(0,1)}T_0^2 - 2W(\boldsymbol{Z})E_{(0,1)}T_0E_{(0,1)}\delta_1 + W(\boldsymbol{Z})^2E_{(0,1)}\delta_1^2 \end{aligned}$$

is minimized with respect to $W(\boldsymbol{Z})$ if $W(\boldsymbol{Z}) = W^*(\boldsymbol{Z})$ where

$$\begin{aligned} W^*(\boldsymbol{Z}) &= \frac{E_{(0,1)}T_0E_{(0,1)}\delta_1}{E_{(0,1)}\delta_1^2} \\ &= \frac{\frac{1}{n}(n-1)}{n-1+(n-1)^2} \\ &= \frac{1}{n^2} \end{aligned}$$

since $\delta_1 \sim \Gamma(n-1, 1)$.

Hence the MRE estimator of μ is

$$T^* = T_0 - W^*\delta_1 = X_{(1)} - \frac{1}{n^2}\sum_{i=2}^{n}\left(X_{(i)} - X_{(1)}\right).$$

NOTE 3.3.6. (1) T^* is not unbiased and the bias depends on σ since

$$\begin{aligned} E_{(\mu,\sigma)}T^*(\boldsymbol{X}) &= E_{(0,1)}T^*(\sigma\boldsymbol{X}+\mu) \\ &= \mu + \sigma E_{(0,1)}T^*(\boldsymbol{X}) \\ &= \mu + \sigma\left(\frac{1}{n} - \frac{n-1}{n^2}\right) \\ &= \mu + \frac{\sigma}{n^2} \end{aligned}$$

(2) Because

$$\begin{aligned} R(\theta, T^*) &= E_{(0,1)}\left(\frac{T^*-0}{1}\right)^2 \\ &= \left(\frac{1}{n^2}\right)^2 + \mathrm{Var}_{(0,1)}T^*, \end{aligned}$$

where

$$\begin{aligned} \mathrm{Var}_{(0,1)}T^* &= \mathrm{Var}_{(0,1)}X_{(1)} + \frac{1}{n^4}\mathrm{Var}_{(0,1)}\delta_1 \\ &= \frac{1}{n^2} + \frac{n-1}{n^4}, \end{aligned}$$

we have

$$R(\theta, T^*) = \frac{n+1}{n^3}.$$

(3) The UMVUE of μ is

$$T(\boldsymbol{X}) = X_{(1)} - \frac{1}{n(n-1)}\sum_{2}^{n}\left(X_{(i)} - X_{(1)}\right)$$

with corresponding

$$\begin{aligned} R(\theta, T) &= E_{(0,1)}T^2 \\ &= \frac{1}{n^2} + \frac{n-1}{n^2(n-1)^2} + 0^2 = \frac{1}{n(n-1)} \\ &> \frac{n+1}{n^3} = R(\theta, T^*). \end{aligned}$$

Estimation of σ^r for some constant r. Assume $L(\theta, d) = \rho\left(\frac{d}{\sigma^r}\right)$ and $\mathcal{P}$ is invariant under $\mathcal{G} = \{g : g\boldsymbol{x} = a\boldsymbol{x} + b\}$. Define $g_{a,b} : \mathbb{R}^n \to \mathbb{R}^n$ by

$$g_{a,b}(\boldsymbol{x}) = a\boldsymbol{x} + b = (ax_1 + b, \cdots, ax_n + b).$$

Then $\overline{g}_{a,b} : \Omega \to \Omega$ is defined by

$$\overline{g}_{a,b}(\theta) = (a\mu + b, a\sigma)$$

and

$$g^*_{a,b} h(\theta) = h\left(\overline{g}_{a,b}(\theta)\right) = a^r \sigma^r = a^r h(\theta)$$

,where

$$h(\theta) = h((\mu, \sigma)) = \sigma^r.$$

Thus,

$$g^*_{a,b}(d) = a^r d.$$

We can check invariance of loss function (i.e. $L(\overline{g}\theta, g^* d) \stackrel{?}{=} L(\theta, d)$) since

$$L(\overline{g}\theta, g^* d) = \rho\left(\frac{a^r d}{a^r \sigma^r}\right) = \rho\left(\frac{d}{\sigma^r}\right) = L(\theta, d)$$

as required. Thus, the condition for equivariant of T is

$$T(g(\boldsymbol{x})) = g^* T(\boldsymbol{x})$$

or

$$T(a\boldsymbol{x} + b) = a^r T(\boldsymbol{x}).$$

PROPOSITION 3.3.7. *Let T_0 be any positive equivariant estimator of σ^r. Then T is equivariant if and only if*

$$T(\boldsymbol{X}) = W(\boldsymbol{Z}) T_0(\boldsymbol{X})$$

for some W, where $\boldsymbol{Z}$ is defined in proposition (3.3.2) .

PROOF. If $T = W(\boldsymbol{Z}) T_0(\boldsymbol{X})$, then

$$\begin{aligned} T(a\boldsymbol{X} + b) &= W(\boldsymbol{Z}) a^r T_0(\boldsymbol{X}) \\ &= a^r T(\boldsymbol{X}). \end{aligned}$$

Conversely, if T is equivariant, then

$$U(\boldsymbol{X}) := \frac{T(\boldsymbol{X})}{T_0(\boldsymbol{X})}$$

satisfies $U(a\boldsymbol{X} + b) = U(\boldsymbol{X})$ for all $a > 0$, $b \in \mathbb{R}$ and so by part (3) of proof of proposition (3.3.2),

$$U(\boldsymbol{X}) = W(\boldsymbol{Z})$$

for some W. □

THEOREM 3.3.8. *Let T_0 be a particular equivariant estimator of σ^r with finite risk. Suppose*

$$E_{(0,1)}\left(\rho\left(W(\boldsymbol{Z}) T_0(\boldsymbol{X})\right) | \boldsymbol{Z}\right)$$

is minimized when $W(\boldsymbol{Z}) = W^*(\boldsymbol{Z})$. Then

$$T^*(\boldsymbol{X}) = W^*(\boldsymbol{Z}) T_0(\boldsymbol{X})$$

is MRE.

PROOF. If T is any equivariant estimator then

$$\begin{aligned} R(\theta, T) &= R((0,1), T) \\ &= E_{(0,1)}\rho(W(\boldsymbol{Z})\, T_0(\boldsymbol{X})) \\ &= E_{(0,1)}\left(E_{(0,1)}\left(\rho(W(\boldsymbol{Z})\, T_0(\boldsymbol{X}))\,|\boldsymbol{Z}\right)\right) \\ &\geq E_{(0,1)}\left(E_{(0,1)}\left(\rho(W^*(\boldsymbol{Z})\, T_0(\boldsymbol{X}))\,|\boldsymbol{Z}\right)\right) \\ &= E_{(0,1)}\rho(T^*) \\ &= R(\theta, T^*). \end{aligned}$$

□

EXAMPLE 3.3.9. Let $X_1, \cdots, X_n$be iid with density $\frac{1}{\sigma}e^{-(x-\mu)/\sigma}I_{(\mu,\infty)}(x)$ and suppose we wish to estimate σ by minimizing the risk under the (invariant) squared fractional error loss function,

$$L((\mu,\sigma), d) = \rho\left(\frac{d}{\sigma}\right) = \left(\frac{d}{\sigma} - 1\right)^2.$$

Let

$$T_0(\boldsymbol{X}) = \sum_{i=1}^{n} \left|X_i - X_{(1)}\right| = \sum_{i=2}^{n} \left(X_{(i)} - X_{(1)}\right)$$

which is independent of $\boldsymbol{Z}$ (by the argument in example (3.3.5).)

Now, noting that T_0 is equivariant ($T(a\boldsymbol{X} + b) = aT(\boldsymbol{X})$), by Theorem 3.3.8 we need to minimize

$$\begin{aligned} E_{(0,1)}\left(\rho(W(\boldsymbol{Z})\, T_0(\boldsymbol{X}))\,|\boldsymbol{Z}\right) &= E_{(0,1)}\left((W(\boldsymbol{Z})\, T_0(\boldsymbol{X}) - 1)^2\,|\boldsymbol{Z}\right) \\ &= W(\boldsymbol{Z})^2 E_{(0,1)}\left(T_0(\boldsymbol{X})^2\,|\boldsymbol{Z}\right) - 2W(\boldsymbol{Z})\, E_{(0,1)}\left(T_0(\boldsymbol{X})\,|\boldsymbol{Z}\right) + 1 \end{aligned}$$

which is minimized for

$$W(\boldsymbol{Z}) = W^*(\boldsymbol{Z}) = \frac{E_{(0,1)}\left(T_0(\boldsymbol{X})\,|\boldsymbol{Z}\right)}{E_{(0,1)}\left(T_0(\boldsymbol{X})^2\,|\boldsymbol{Z}\right)} = \frac{E_{(0,1)}T_0(\boldsymbol{X})}{E_{(0,1)}T_0(\boldsymbol{X})^2} = \frac{n-1}{(n-1)\, n} = \frac{1}{n}.$$

Hence,

$$T^*(\boldsymbol{X}) = W^*(\boldsymbol{Z})\, T_0(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}\left(X_{(i)} - X_{(1)}\right)$$

is MRE.

NOTE 3.3.10. (1) Once again we notice that the MRE is biased:

$$E_{(\mu,\sigma)}T^*(\boldsymbol{X}) = E_{(0,1)}T^*(\sigma\boldsymbol{X} + \mu) = \sigma E_{(0,1)}T^*(\boldsymbol{X}) = \sigma\frac{n-1}{n} \neq \sigma.$$

(2) The UMVUE is (TPE example (2.2.5)):

$$T(\boldsymbol{X}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{(i)} - X_{(1)}\right).$$

Scale Equivariance

For the special case when only the **scale** paramter σ is unknown, i.e., the density of the sample consitutes a scale family:

$$\boldsymbol{X} = (X_1, \cdots, X_n)^T \sim \sigma^{-n} f\left(\frac{x_1}{\sigma}, \cdots, \frac{x_n}{\sigma}\right),$$

see the discussion culminating in Theorem 3.3 in TPE. The summary points are as follows.

- Estimand: $h(\theta) = \sigma^r$.
- Invariant loss function: $L(\theta, d) = \rho(d/\sigma^r)$.
- Equivariant T satisfies: $T(aX) = a^r T(X)$.
- MRE $T^*(\boldsymbol{X})$ of $h(\theta)$ is then found as follows:
  - let T_0 be equivariant for σ^r with finite risk,
  - define
  $$\boldsymbol{Z} = (Z_1, \cdots, Z_{n-1}, Z_n) = \left(\frac{X_1}{X_n}, \cdots, \frac{X_{n-1}}{X_n}, \frac{X_n}{|X_n|}\right),$$
  - find
  $$W^*(\boldsymbol{Z}) = \arg\min_W E_1\left[\rho\left(\frac{T_0(\boldsymbol{X})}{W(\boldsymbol{Z})}\right) \mid \boldsymbol{Z}\right],$$
  - and finally
  $$T^*(\boldsymbol{X}) = \frac{T_0(\boldsymbol{X})}{W^*(\boldsymbol{Z})}.$$

CHAPTER 4

Average-Risk Optimality

Thus far we have focused on finding estimators T which minimize the risk $R(\theta, T)$ at every value of θ. This was possible by restricting the class of estimators to be either unbiased (Ch. 2) or equivariant (Ch. 3). We now drop these restrictions, thus bringing all estimators into play, but will therefore have to sacrifice *uniform minimum risk* for other optimality criteria which make $R(\theta, T)$ small in some overall sense. Two specific versions of this type of alternative optimality are:

- minimize *(weighted) average risk*, which leads to **Bayes estimates**, e.g., hierarchical Bayes and Empirical Bayes, and is discussed in detail in TPE Ch. 4.
- minimize *maximum risk*, which leads to **minimax estimates**, is discussed in detail in TPE Ch. 5.

4.1. Bayes Estimation

The main factor contributing to the recent explosion of interest in Bayes estimation is its ability to handle extremely complicated practical problems. Some other factors which make Bayes estimation attractive are as follows.

(1) The mathematical structure is very nice.
(2) It permits the incorporation of prior information (although there is a lot of debate about how this should be done).
(3) It provides a systematic approach to the determination of minimax estimators.

In the Bayesian framework, we consider the parameter and observation vectors to be jointly distributed on $\Omega \times \mathcal{X}$. We shall suppose that the parameter vector Θ has the marginal distribution Λ and that the conditional distribution of the observation vector $\mathbf{X}$ given $\Theta = \theta$ is P_θ. For any **particular** value θ of Θ, we define the risk of the estimator T' at θ as $R(\theta, T') := E[L(\Theta, T')|\Theta = \theta] = \int_{\mathcal{X}} L(\theta, T'(\mathbf{x}))dP_\theta(\mathbf{x})$ as before.

DEFINITION 4.1.1. For *any* estimator T', the integral

$$\int_\Omega R(\theta, T')\, d\Lambda(\theta) = \int_\Omega \int_{\mathcal{X}} L(\theta, T'(\mathbf{x}))dP_\theta(\mathbf{x})d\Lambda(\theta).$$

is called the **Bayes risk of T' with respect to the prior distribution** Λ. The Bayes risk of T' is thus

$$r_\Lambda := E(R(\Theta, T')) = E(L(\Theta, T')).$$

An estimator T is a **Bayes estimator with respect to the distribution** Λ if

$$\int_{\Omega} R(\theta, T)\, d\Lambda(\theta) = \inf_{T'} \int_{\Omega} R(\theta, T')\, d\Lambda(\theta).$$

THEOREM 4.1.2. *Suppose* $\Theta \sim \Lambda$ *and* $\mathbf{X}$ *given* $\Theta = \theta$ *has distribution* P_θ. *Then if there exists* $T(\cdot)$ *which minimizes*

$$E(L((\Theta, T(\mathbf{X})))|\mathbf{X} = \mathbf{x}) = \int_{\Omega} L(\theta, T(\mathbf{x}))\, dP(\theta|\mathbf{X} = \mathbf{x})$$

for each $\mathbf{x}$, *where* $P(\cdot|\mathbf{X} = \mathbf{x})$ *is the conditional (or posterior) distribution of* Θ *given* $\mathbf{X} = \mathbf{x}$, *then* $T(\mathbf{X})$ *is Bayes with respect to* Λ.

PROOF. For any estimator T',

$$E(L(\Theta, T'(\mathbf{X}))|\mathbf{X}) \geq E(L(\Theta, T(\mathbf{X}))|\mathbf{X})$$

and so, taking expectations of each side,

$$EL(\Theta, T'(\mathbf{X})) \geq EL(\Theta, T(\mathbf{X})).$$

□

EXAMPLE 4.1.3.

(1) Let $L(\theta, d) = (d - g(\theta))^2$. The Bayes estimator T of $g(\Theta)$ minimizes
$$E\left((T(\mathbf{X}) - g(\Theta))^2 |\mathbf{X} = \mathbf{x}\right) \ \forall\, \mathbf{x}.$$
Hence
$$T(\mathbf{x}) = E(g(\Theta)|\mathbf{X} = \mathbf{x}),$$
which is the posterior mean.

(2) Let $L(\theta, d) = |d - g(\theta)|$. The Bayes estimator T of $g(\Theta)$ minimizes
$$E(|T(\mathbf{X} - g(\Theta))||\mathbf{X} = \mathbf{x}) \ \forall\, \mathbf{x}.$$
Hence
$$T(\mathbf{x}) = \text{med}(g(\Theta)|\mathbf{X} = \mathbf{x}),$$
which is the posterior median.

(3) Let $L(\theta, d) = w(\theta)(d - g(\theta))^2$. The Bayes estimator T of $g(\Theta)$ minimizes
$$E\left(w(\Theta)(T(\mathbf{X}) - g(\Theta))^2 |\mathbf{X} = \mathbf{x}\right) \ \forall\, \mathbf{x}.$$
It can therefore be obtained by solving
$$\begin{array}{rl} \frac{d}{dT(\mathbf{x})} \int (T(\mathbf{x}) - g(\theta))^2 w(\theta)\, dP(\theta|\mathbf{X} = \mathbf{x}) & \\ = \int 2(T(\mathbf{x}) - g(\theta)) w(\theta)\, dP(\theta|\mathbf{X} = \mathbf{x}) & = 0. \end{array}$$
Hence
$$T(\mathbf{x}) = \frac{E(w(\Theta) g(\Theta)|\mathbf{X} = \mathbf{x})}{E(w(\Theta)|\mathbf{X} = \mathbf{x})}.$$

EXAMPLE 4.1.4. Suppose $X \sim \text{bin}(n,p)$ and $p \sim B(a,b)$. Thus,

$$d\Lambda(p) = \frac{\Gamma(a+b)}{(\Gamma(a)\Gamma(b))} p^{a-1}(1-p)^{b-1}\,dp, \quad 0<p<1,\ a,b>0.$$

The posterior distribution of p given $X=x$ is $B(a+x, b+n-x)$ since

$$\begin{aligned} f_{P|X}(p|x) &= \frac{f_{X|P}(x|p) f_P(p)}{f_X(x)} \\ &= \frac{\binom{n}{x} p^x (1-p)^{n-x} K p^{a-1}(1-p)^{b-1}}{f_X(x)} \\ &= c(x)\, p^{a+x-1}(1-p)^{b+n-x-1} \end{aligned}$$

which is a beta density. (Without doing any calculations it is clear that $c(x)$ must be $\Gamma(a+b+n)/(\Gamma(a+x)\Gamma(b+n-x))$.)

Let $L(p,d) = (d-p)^2$. Then the Bayes estimator of p is

$$\begin{aligned} T(x) = \int_0^1 p f_{P|X}(p|x)\,dp &= \frac{a+x}{a+b+n} \\ &= \frac{a+b}{a+b+n} \cdot \underbrace{\frac{a}{a+b}}_{\text{prior mean}} + \frac{n}{a+b+n} \cdot \underbrace{\frac{x}{n}}_{\text{the usual etimator}}. \end{aligned}$$

Thus $T(X) - X/n \to 0$ a.s. as $n \to \infty$ with a and b fixed and as $a+b \to 0$ with n fixed.

Clearly X/n is not a Bayes estimator for any beta prior (i.e. for any $a>0$ and $b>0$). However if Λ is concentrated on the two-point set $\{0,1\}$ then X/n is Bayes as the following argument shows. If $P(P=1) = 1-\pi$ and $P(P=0)=\pi$, then X is either 0 or n with probability 1 and

$$\begin{cases} P(P=1|X=n) = \frac{P(X=n,P=1)}{P(X=n)} = \frac{1-\pi}{1-\pi} & = 1 \\ P(P=0|X=n) & = 0 \end{cases}.$$

Hence the Bayes estimator satisfies $T(n)=1$ and a similar argument shows that $T(0)=0$. Hence $T(x) = x/n,\ x=0,n$. Notice that this two-point distribution is the limit in distribution of Beta(a,b) as $a+b \to 0$ with $a/(a+b) = \pi$.

THEOREM 4.1.5. *Let $L(\theta,d) = (d-g(\theta))^2$. Then no unbiased estimator T of $g(\theta)$ can be Bayes unless*

$$E(T(\mathbf{X}) - g(\Theta))^2 = 0$$

i.e. unless $T(X) = g(\Theta)$ with probability 1 and the Bayes risk of T is zero.

PROOF. If T is Bayes with respect to some Λ and unbiased for $g(\theta)$ then

$$\begin{aligned} E(T(\mathbf{X})|\Theta) &= g(\Theta), \text{ and} \\ E(g(\Theta)|\mathbf{X}=\mathbf{x}) &= T(\mathbf{x}) \end{aligned}$$

Hence

$$\begin{aligned} E\left(T\left(\mathbf{X}\right)g\left(\Theta\right)\right) &= E\left(E\left(T\left(\mathbf{X}\right)g\left(\Theta\right)\right)|\mathbf{X}\right) \\ &= ET\left(\mathbf{X}\right)^2 \\ &= E\left(E\left(T\left(\mathbf{X}\right)g\left(\Theta\right)\right)|\Theta\right) \\ &= Eg\left(\Theta\right)^2 \end{aligned}$$

and so

$$\begin{aligned} E\left(T\left(\mathbf{X}\right)-g\left(\Theta\right)\right)^2 &= ET\left(\mathbf{X}\right)^2-2E\left(T\left(\mathbf{X}\right)g\left(\Theta\right)\right)+E\left(g\left(\Theta\right)\right)^2 \\ &= 0. \end{aligned}$$

□

EXAMPLE 4.1.6. Given $\Theta=\theta$, let $X_1,\cdots,X_n$ be iid $N\left(\theta,\sigma^2\right)$ with σ^2 known and suppose that $\Theta\sim N\left(\mu,\tau^2\right)$ with μ,τ^2 known. Then the joint density of Θ and $\mathbf{X}$ is

$$f_{\Theta,\mathbf{X}}\left(\theta,\mathbf{x}\right)=\frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}e^{-\frac{1}{2\sigma^2}\sum_1^n\left(x_i-\theta\right)^2}\frac{1}{\tau\sqrt{2\pi}}e^{-\frac{1}{2\tau^2}\left(\theta-\mu\right)^2}$$

and the posterior density of Θ is

$$\begin{aligned} f_{\Theta|\mathbf{X}}\left(\theta|\mathbf{x}\right) &= \frac{f_{\Theta,\mathbf{X}}\left(\theta,\mathbf{x}\right)}{f_{\mathbf{X}}\left(\mathbf{x}\right)} \\ &= c\left(\mathbf{x}\right)\exp\left(-\frac{n}{2\sigma^2}\theta^2+\frac{\sum x_i}{\sigma^2}\theta-\frac{\theta^2}{2\tau^2}+\frac{\mu\theta}{\tau^2}\right) \\ &= N\left(\frac{\frac{n\overline{x}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}},\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\right). \end{aligned}$$

For squared error loss, the Bayes estimator of Θ is

$$E\left(\Theta|\mathbf{X}\right)=\frac{n\sigma^{-2}}{n\sigma^{-2}+\tau^{-2}}\overline{X}+\frac{\tau^{-2}}{n\sigma^{-2}+\tau^{-2}}\mu=T_\Lambda\left(\mathbf{X}\right)$$

and

$$\begin{aligned} \mathrm{Var}\left(\Theta|\mathbf{X}\right) &= \frac{1}{n\sigma^{-2}+\tau^{-2}}=E\left(\left(\Theta-T_\Lambda\left(\mathbf{X}\right)\right)^2|\mathbf{X}\right) \\ \mathrm{r}_\Lambda &= EL\left(\Theta,T_\Lambda\right)=E\left(\Theta-T_\Lambda\left(\mathbf{X}\right)\right)^2=\frac{1}{n\sigma^{-2}+\tau^{-2}}. \end{aligned}$$

For large n, the Bayes estimator is close to $\overline{X}$ in the sense that $T_\Lambda(\mathbf{X})-\overline{X}\to 0$ a.s. as $n\to\infty$ with τ and σ fixed. Also $T(\mathbf{X})-\overline{X}\to 0$ as $\tau\to\infty$ with n and σ fixed. However, $\overline{X}$ is not Bayes since the prior probability distribution $N\left(\mu,\tau^2\right)$ does not converge to a probability measure as $\tau\to\infty$.

However one can formally obtain $\overline{X}$ as a Bayesian estimator with respect to the **improper prior distribution**, $\Lambda(d\theta) = d\theta$. Suppose

$$p(\mathbf{x}|\theta) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{-\frac{1}{2\sigma^2}\sum_1^n (x_i-\theta)^2}$$

with σ^2 known. Setting $\Lambda(d\theta) = d\theta$ we find that the joint density of $(\mathbf{X}, \Theta)$ with respect to Lebesgue measure is

$$p(\mathbf{x}, \theta) = p(\mathbf{x}|\theta).$$

The posterior distribution of θ is therefore

$$\begin{aligned} p(\theta|\mathbf{x}) &= k(\mathbf{x}) \exp\left(-\frac{1}{2\sigma^2}\left(n\theta^2 - 2\theta\sum_{i=1}^n x_i\right)\right) \\ &= k^*(\mathbf{x}) \exp\left(-\frac{n}{2\sigma^2}(\theta - \overline{x})^2\right). \end{aligned}$$

Hence,

$$p(\theta|\overline{x}) = N\left(\overline{x}, \frac{\sigma^2}{n}\right)$$

and so $\overline{X}$is the **generalized Bayes estimator** of Θ with respect to $L(\theta, d) = (d-\theta)^2$ and the **improper Lebesgue prior** $\Lambda(d\theta) = d\theta$ for Θ. The improper prior densities $I_{(-\infty<\infty)}$ and $I_{[0,\infty)}$ are frequently used to account for total ignorance of parameters with values in $\mathbb{R}$ and $\mathbb{R}^+$ respectively.

(**Note:** unless otherwise stated, all results from now on apply to **Bayes estimators**, not *generalized Bayes.*)

Conjugate Priors

If there exists a parametric family of prior distributions such that the posterior distribution also belongs to the same parametric family, then the family is called **conjugate**.

EXAMPLE 4.1.7. Conditional on $\Sigma^2 = \sigma^2$, let $X_1, \cdots, X_n$ be iid $N(0, \sigma^2)$ and define $\Xi = (2\Sigma^2)^{-1}$.

$$f_{\mathbf{X}|\Xi}(\mathbf{x}|\xi) = c\xi^r e^{-\xi\sum_1^n x_i^2} \text{ where } r = \frac{n}{2}.$$

and let the prior distribution for Ξ be $\Gamma\left(g, \frac{1}{\alpha}\right)$ with density

$$\lambda(\xi) = \frac{\alpha^g}{\Gamma(g)}\xi^{g-1}e^{-\alpha\xi}, \quad \xi \geq 0.$$

We note that

$$\begin{aligned} E(\Xi) &= \frac{g}{\alpha}, & E\left(\Xi^2\right) &= \frac{g(g+1)}{\alpha^2} \\ E\left(\Xi^{-1}\right) &= \frac{\alpha}{g-1} & E\left(\Xi^{-2}\right) &= \frac{\alpha^2}{(g-1)(g-2)}. \end{aligned}$$

Then, the posterior becomes

$$\begin{aligned} f_{\Xi|\mathbf{X}}(\xi|\mathbf{x}) &= c(\mathbf{x})\,\xi^{r+g-1}e^{-\xi\left(\sum x_i^2+\alpha\right)} \\ &= \text{density at } \xi \text{ of } \Gamma\left(r+g, \frac{1}{\sum x_i^2+\alpha}\right), \end{aligned}$$

so that the gamma distribution family is a conjugate family for the normal distribution.

If the loss is squared error then the Bayes estimator of $\sigma^2 = (2\xi)^{-1}$ is

$$\begin{aligned} T(\mathbf{X}) = \int \frac{1}{2\xi} f_{\Xi|\mathbf{X}}(\xi|\mathbf{x})\,d\xi &= \frac{\alpha+\sum x_i^2}{2(r+g-1)} \\ &= \frac{\alpha+\sum x_i^2}{n+2g-2}. \end{aligned}$$

As $\alpha \to 0$ with $g = 1$, the prior density/α converges pointwise to the improper prior density $I_{[0,\infty)}$ and the Bayes estimator $T(\mathbf{X})$ satisfies

$$T(\mathbf{X}) - \Sigma_{i=1}^n X_i^2/n \to 0 \text{ a.s.}.$$

s-Parameter Exponential Families

Calculation of Bayes estimators simplifies under an s-parameter exponential family, where for a random sample $\boldsymbol{x} = (x_1, \ldots, x_n)$, the density is given by (the canonical form):

$$p_{\boldsymbol{\eta}}(\boldsymbol{x}) = \exp\{\sum_{i=1}^s \eta_i T_i(\boldsymbol{x}) - A(\boldsymbol{\eta})\}h(\boldsymbol{x}). \tag{4.1.1}$$

THEOREM 4.1.8. *If $\boldsymbol{X}$ has density (4.1.1) and $\boldsymbol{\eta}$ has prior density $\pi(\boldsymbol{\eta})$, then for $j = 1, \ldots, n$:*

$$\mathbb{E}\left[\sum_{i=1}^s \eta_i \frac{\partial T_i(\boldsymbol{x})}{\partial x_j}\Big|\boldsymbol{x}\right] = \frac{\partial}{\partial x_j}\log m(\boldsymbol{x}) - \frac{\partial}{\partial x_j}\log h(\boldsymbol{x}), \tag{4.1.2}$$

where

$$m(\boldsymbol{x}) = \int p_\eta(\boldsymbol{x})\pi(\boldsymbol{\eta})d\boldsymbol{\eta} = \text{ marginal of } \boldsymbol{X}.$$

PROOF. Letting $f(\boldsymbol{\eta}|\boldsymbol{x}) = p_{\boldsymbol{\eta}}(\boldsymbol{x})\pi(\boldsymbol{\eta})/m(\boldsymbol{x})$, note that for any integrable function $g(\boldsymbol{\eta}, \boldsymbol{X})$,

$$\mathbb{E}[g(\boldsymbol{\eta}, \boldsymbol{X})|\boldsymbol{x}] = \int g(\boldsymbol{\eta}, \boldsymbol{x})f(\boldsymbol{\eta}|\boldsymbol{x})d\boldsymbol{\eta},$$

whence upon substituting $g(\boldsymbol{\eta}, \boldsymbol{X}) = \sum_{i=1}^{s} \eta_i \partial T_i(\boldsymbol{x})/\partial x_j$, the LHS of (4.1.2) becomes:

$$\begin{aligned}
\mathbb{E}[g(\boldsymbol{\eta}, \boldsymbol{X})|\boldsymbol{x}] &= \frac{1}{m(\boldsymbol{x})} \int \sum_i \left(\eta_i \frac{\partial T_i}{\partial x_j} \right) e^{\sum \eta_i T_i - A(\boldsymbol{\eta})} h(\boldsymbol{x}) \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} \\
&= \frac{1}{m(\boldsymbol{x})} \int \left[\frac{\partial}{\partial x_j} e^{\sum \eta_i T_i} \right] e^{-A(\boldsymbol{\eta})} h(\boldsymbol{x}) \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} \\
&= \frac{1}{m(\boldsymbol{x})} \int \left[\frac{\partial}{\partial x_j} \left(e^{\sum \eta_i T_i} h(\boldsymbol{x}) \right) - e^{\sum \eta_i T_i} \frac{\partial h(\boldsymbol{x})}{\partial x_j} \right] e^{-A(\boldsymbol{\eta})} \pi(\boldsymbol{\eta}) d\boldsymbol{\eta},
\end{aligned}$$

where the last equality follows by bringing $h(\boldsymbol{x})$ inside the round brackets and using the chain rule. Finally, switching integration and differentiation gives,

$$\begin{aligned}
\mathbb{E}[g(\boldsymbol{\eta}, \boldsymbol{X})|\boldsymbol{x}] &= \frac{1}{m(\boldsymbol{x})} \frac{\partial}{\partial x_j} \int e^{\sum \eta_i T_i - A(\eta)} h(\boldsymbol{x}) \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} \\
&\qquad - \frac{\partial h(\boldsymbol{x})/\partial x_j}{h(\boldsymbol{x})} \frac{1}{m(\boldsymbol{x})} \int e^{\sum \eta_i T_i - A(\boldsymbol{\eta})} h(\boldsymbol{x}) \pi(\boldsymbol{\eta}) d\boldsymbol{\eta} \\
&= \frac{\partial m(\boldsymbol{x})/\partial x_j}{m(\boldsymbol{x})} - \frac{\partial h(\boldsymbol{x})/\partial x_j}{m(\boldsymbol{x})} \\
&= \frac{\partial}{\partial x_j} \log m(\boldsymbol{x}) - \frac{\partial}{\partial x_j} \log h(\boldsymbol{x}).
\end{aligned}$$

□

COROLLARY 4.1.9. *If in Theorem 4.1.8* $\boldsymbol{X} = (X_1, \ldots, X_s)$ *has the density* $p_{\boldsymbol{\eta}}(\boldsymbol{x})$ *with* $T_i(\boldsymbol{x}) = x_i$, *then the Bayes estimator of* $\boldsymbol{\eta}$ *under the loss* $L(\boldsymbol{\eta}, \boldsymbol{\delta}) = \sum(\eta_i - \delta_i)^2$ *is given by:*

$$\mathbb{E}(\eta_j|\boldsymbol{x}) = \mathbb{E}\left[\sum_{i=1}^{s} \eta_i \frac{\partial T_i(\boldsymbol{x})}{\partial x_j} \Big| \boldsymbol{x} \right] = \frac{\partial}{\partial x_j} \log m(\boldsymbol{x}) - \frac{\partial}{\partial x_j} \log h(\boldsymbol{x}),$$

for $j = 1, \ldots, s$.

PROOF. Problem 4.3.3 using Theorem 4.1.8 with $T_i = X_i$. □

EXAMPLE 4.1.10 (Multiple Normal Model). $X_i|\theta_i \sim$ indep. $N(\theta_i, \sigma^2)$ and $\Theta_i \sim N(\mu, \tau^2)$, where σ^2, μ, τ^2, are all known, and $i = 1, \ldots, s$.

$$\begin{aligned}
p_{\boldsymbol{\eta}}(\boldsymbol{x}) &= (2\pi\sigma^2)^{-s/2} \exp\left\{ -\frac{\sum(x_i - \theta_i)^2}{2\sigma^2} \right\} \\
&= \exp\left\{ -\sum \underbrace{\frac{\theta_i}{\sigma^2}}_{\eta_i} x_i - \underbrace{\sum \frac{\theta_i^2}{2\sigma^2}}_{A(\boldsymbol{\eta})} \right\} \underbrace{(2\pi\sigma^2)^{-s/2} e^{-\frac{1}{2\sigma^2}\sum x_i^2}}_{h(\boldsymbol{x})}.
\end{aligned}$$

By Corollary 4.1.9 the Bayes estimator of θ_i is

$$\begin{aligned}\mathbb{E}(\Theta_i|\boldsymbol{x}) &= \sigma^2\mathbb{E}(\eta_i|\boldsymbol{x}) = \sigma^2\left[\frac{\partial \log m(\boldsymbol{x})}{\partial x_i} - \frac{\partial \log h(\boldsymbol{x})}{\partial x_i}\right] \\ &= \frac{\tau^2}{\tau^2+\sigma^2}x_i + \frac{\sigma^2}{\tau^2+\sigma^2}\mu, \qquad (\underline{\text{Claim}}).\end{aligned}$$

To prove the Claim, note that one term is easily handled:

$$\frac{\partial}{\partial x_i}\log h(\boldsymbol{x}) = \frac{\partial}{\partial x_i}\left(-\sum \frac{x_j^2}{2\sigma^2}\right) = -\frac{x_i}{\sigma^2}.$$

For the other term, we compute $m(\boldsymbol{x})$ to within multiplicative constants (free of $\boldsymbol{x}$):

$$\begin{aligned}m(\boldsymbol{x}) &= \int_{\mathbb{R}^s} p_{\boldsymbol{\eta}}(\boldsymbol{x})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \int (2\pi\sigma^2)^{-s/2}\exp\left\{-\frac{\sum(x_i-\theta_i)^2}{2\sigma^2} - \frac{\sum(\theta_i-\mu)^2}{2\tau^2}\right\}(2\pi\tau^2)^{-s/2}d\boldsymbol{\theta} \\ &= (4\pi^2\sigma^2\tau^2)^{-s/2}e^{-\frac{1}{2\sigma^2}\sum x_i^2}\int \exp\left\{-\frac{1}{2(\sigma^2+\tau^2)}\sum\theta_i^2 + \sum\left(\frac{x_i}{\sigma^2}+\frac{\mu}{\tau^2}\right)\theta_i\right\}d\boldsymbol{\theta}.\end{aligned}$$

Completing the square on the exp term in the integrand:

$$\exp = -\frac{\sum(\theta_i-c_i)^2}{2(\sigma^2\tau^2/\lambda^2)} - \frac{\sum(x_i-\mu)^2}{2\lambda^2}, \qquad c_i = \frac{\tau^2x_i+\mu\sigma^2}{\lambda^2},\ \lambda^2 = \sigma^2+\tau^2.$$

Thus,

$$m(\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2\lambda^2}\sum(x_i-\mu)^2\right\}\underbrace{\int \exp\left\{-\frac{1}{2(\sigma^2+\tau^2)}\sum(\theta_i-c_i)^2\right\}d\boldsymbol{\theta}}_{\text{constant (free of } \boldsymbol{x})},$$

from which we conclude that, marginally, $X_i \sim$ iid $\mathrm{N}(\mu,\lambda^2)$, whence

$$\frac{\partial \log m(\boldsymbol{x})}{\partial x_i} = -\frac{\partial}{\partial x_i}\sum\frac{(x_j-\mu)^2}{2\lambda^2} = -(x_i-\mu)/\lambda^2,$$

and therefore

$$\sigma^2\left[\frac{\partial \log m(\boldsymbol{x})}{\partial x_i} - \frac{\partial \log h(\boldsymbol{x})}{\partial x_i}\right] = \sigma^2\left[-\frac{x_i-\mu}{\lambda^2}+\frac{x_i}{\sigma^2}\right] = \frac{\tau^2}{\tau^2+\sigma^2}x_i + \frac{\sigma^2}{\tau^2+\sigma^2}\mu.$$

Having obtained the Bayes estimator by simply differentiating apropriate functions, the next obvious question concerns the computation of its risk.

THEOREM 4.1.11. *The risk of the Bayes estimator in Corollory 4.1.9 is:*

$$R(\boldsymbol{\eta}, \mathbb{E}(\boldsymbol{\eta}|\boldsymbol{X})) = R\left(\boldsymbol{\eta}, -\frac{\partial \log h(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) + \sum_{i=1}^{s}\mathbb{E}\left[2\frac{\partial^2 \log m(\boldsymbol{X})}{\partial X_i^2} + \left(\frac{\partial \log m(\boldsymbol{X})}{\partial X_i}\right)^2\right].$$

PROOF. Stein's Identity is applicable to an exponential family in canonical form (as in Theorem 4.1.8), and states that for any differentiable real-valued function g with $\mathbb{E}|g'(\boldsymbol{X})| < \infty$, we have

$$\mathbb{E}\left\{\left[\frac{\partial \log h(\boldsymbol{X})}{\partial X_j} + \sum_{i=1}^{s} \eta_i \frac{\partial T_i(\boldsymbol{X})}{\partial X_j}\right] g(\boldsymbol{X})\right\} = -\mathbb{E}\frac{\partial g(\boldsymbol{X})}{\partial X_j}, \qquad \text{for } j = 1, \ldots, n.$$

provided the support of each X_j is all of $\mathbb{R}$. (If the support is a bounded interval, then the above holds if $\exp\{\sum \eta_i T_i(x)\}h(x) \to 0$, as x approaches the boundaries.) Applying the Identity with $g(\boldsymbol{x}) = 1$, and noting that in our case $T_i = X_i$, we get:

$$\mathbb{E}_{\boldsymbol{\eta}}\left[\frac{\partial h(\boldsymbol{X})/\partial X_j}{h(\boldsymbol{X})} + \sum_{i=1}^{s} \eta_i \underbrace{\frac{\partial X_i}{\partial X_j}}_{\delta_{ij}}\right] = -\mathbb{E}(0) = 0,$$

which leads to

$$-\mathbb{E}_{\boldsymbol{\eta}}\left[\frac{\partial \log h(\boldsymbol{X})}{\partial X_j}\right] = \eta_j, \ \forall j \qquad \Longleftrightarrow \qquad -\mathbb{E}_{\boldsymbol{\eta}}\left[\frac{\partial \log h(\boldsymbol{X})}{\partial \boldsymbol{X}}\right] = \boldsymbol{\eta}.$$

Thus $-\partial \log h(\boldsymbol{X})/\partial \boldsymbol{X}$ is an unbiased estimate of $\boldsymbol{\eta}$ with risk:

$$R\left(\boldsymbol{\eta}, -\frac{\partial \log h(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) = \mathbb{E}_{\boldsymbol{\eta}} \sum_{i=1}^{s}\left[\eta_i + \frac{\partial \log h(\boldsymbol{X})}{\partial X_i}\right]^2.$$

Now, the risk of the Bayes estimator is given by:

$$\begin{aligned}
& R(\boldsymbol{\eta}, \mathbb{E}(\boldsymbol{\eta}|\boldsymbol{X})) = \mathbb{E}\sum_{i=1}^{s}\left[\eta_i - \mathbb{E}(\eta_i|\boldsymbol{X})\right]^2 \\
= \ & \mathbb{E}\sum_{i=1}^{s}\left[\eta_i - \left(\frac{\partial \log m(\boldsymbol{X})}{\partial X_i} - \frac{\partial \log h(\boldsymbol{X})}{\partial X_i}\right)\right]^2 \\
= \ & \underbrace{\mathbb{E}\sum_{i=1}^{s}\left[\eta_i + \frac{\partial \log h(\boldsymbol{X})}{\partial X_i}\right]^2}_{R\left(\boldsymbol{\eta}, -\frac{\partial \log h(\boldsymbol{X})}{\partial \boldsymbol{X}}\right)} - 2\sum_{i=1}^{s} \underbrace{\mathbb{E}\left[\left(\eta_i + \frac{\partial \log h(\boldsymbol{X})}{\partial X_i}\frac{\partial \log m(\boldsymbol{X})}{\partial X_i}\right)\right]}_{-\mathbb{E}\left(\frac{\partial^2 \log m(\boldsymbol{X})}{\partial X_i^2}\right),\ \text{Stein's Id. with } g(x)=\partial \log m/\partial x_i} + \sum_{i=1}^{s}\mathbb{E}\left(\frac{\partial \log m(\boldsymbol{X})}{\partial X_i}\right)^2 \\
= \ & R\left(\boldsymbol{\eta}, -\frac{\partial \log h(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) + \sum_{i=1}^{s}\mathbb{E}\left[2\frac{\partial^2 \log m(\boldsymbol{X})}{\partial X_i^2} + \left(\frac{\partial \log m(\boldsymbol{X})}{\partial X_i}\right)^2\right].
\end{aligned}$$

□

EXAMPLE 4.1.12 (Multiple Normal Model (continued)). From before, $-\partial \log h(\boldsymbol{x})/\partial x_i = x_i/\sigma^2$, which as we saw in the above proof, is unbiased for $\eta_i = \theta_i/\sigma^2$. Thus, and since $\boldsymbol{X} = (X_1, \ldots, X_s)$ is CSS, we have that $-\partial \log h(\boldsymbol{x})/\partial \boldsymbol{X} = (X_1/\sigma^2, \ldots, X_s/\sigma^2)$ is UMVU

for $\boldsymbol{\eta} = (\theta_1/\sigma^2, \ldots, \theta_s/\sigma^2)$. Now, and from the above proof, the risk of the UMVUE is:

$$R\left(\boldsymbol{\eta}, -\frac{\partial \log h(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) = \mathbb{E}_{\boldsymbol{\eta}} \sum_{i=1}^{s} \left[\eta_i + \frac{\partial \log h(\boldsymbol{X})}{\partial X_i}\right]^2 = \sum_{i=1}^{s} \mathbb{E}_{\theta_i} \left(\frac{X_i - \theta_i}{\sigma^2}\right)^2 = \frac{s}{\sigma^2},$$

since $X_i|\theta_i \sim N(\theta_i, \sigma^2)$ implies $\mathbb{E}_{\theta_i}(X_i - \theta_i)^2 = \sigma^2$. Thus, the risk of the Bayes estimator is given by:

$$\begin{aligned}
R(\boldsymbol{\eta}, \mathbb{E}(\boldsymbol{\eta}|\boldsymbol{X})) &= R\left(\boldsymbol{\eta}, -\frac{\partial \log h(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) + \sum_{i=1}^{s} \mathbb{E}_{\boldsymbol{\eta}} \left[2\frac{\partial^2 \log m(\boldsymbol{X})}{\partial X_i^2} + \left(\frac{\partial \log m(\boldsymbol{X})}{\partial X_i}\right)^2\right] \\
&= \frac{s}{\sigma^2} + \sum_{i=1}^{s} \mathbb{E}_{\boldsymbol{\eta}} \left[\frac{(x_i - \mu)^2}{(\sigma^2 + \tau^2)^2} - \frac{2}{\sigma^2 + \tau^2}\right] \\
&= \frac{s\tau^4}{\sigma^2(\sigma^2 + \tau^2)^2} + \left(\frac{\sigma^2}{\sigma^2 + \tau^2}\right)^2 \sum_{i=1}^{s} \underbrace{\left(\eta_i - \frac{\mu}{\sigma^2}\right)^2}_{a_i^2}, \qquad \text{(by Problem 4.3.6)} \\
&= \frac{s}{\sigma^2}\frac{\tau^4}{(\sigma^2 + \tau^2)^2} < \frac{s}{\sigma^2}, \qquad \text{if } \sum a_i^2 = 0.
\end{aligned}$$

Summary:

- UMVUE of $\boldsymbol{\eta}$ is $(X_1/\sigma^2, \ldots, X_s/\sigma^2)$, with risk $R\left(\boldsymbol{\eta}, -\partial \log h(\boldsymbol{X})/\partial \boldsymbol{X}\right) = s/\sigma^2$.
- Bayes estimator of $\boldsymbol{\eta}$ is

  $$\left(\frac{\tau^2/\sigma^2}{\sigma^2 + \tau^2} X_1 + \frac{\mu}{\sigma^2 + \tau^2}, \ldots, \frac{\tau^2/\sigma^2}{\sigma^2 + \tau^2} X_s + \frac{\mu}{\sigma^2 + \tau^2}\right),$$

  with risk

  $$R\left(\boldsymbol{\eta}, \mathbb{E}(\boldsymbol{\eta}|\boldsymbol{X})\right) = \frac{s\tau^4}{\sigma^2(\sigma^2 + \tau^2)^2} + \frac{\sigma^4 \sum a_i^2}{(\sigma^2 + \tau^2)^2},$$

  which is smaller than the risk of the UMVUE if, e.g., $\eta_i = \mu/\sigma^2, \ \forall i$.

Empirical Bayes

General Bayes setup thus far has been:

$$\begin{cases} X_i|\theta \sim f(x|\theta), & i = 1, \ldots, n \\ \Theta|\gamma \sim \pi(\theta|\gamma), & \gamma \text{ is known.} \end{cases}$$

The Empirical Bayes idea is to treat γ as unknown, and use frequentist methods to estimate it based on the marginal distribution of $\boldsymbol{X}$:

$$m(\boldsymbol{x}|\gamma) = \int \prod_{i=1}^{n} f(x_i|\theta)\pi(\theta|\gamma)d\theta.$$

For example, one can use $m(\boldsymbol{x}|\gamma)$ to produce the MLE $\hat{\gamma}(\boldsymbol{x})$. Then, the empirical Bayes estimator, $\hat{T}(\boldsymbol{x})$, minimizes the (empirical) posterior risk:

$$\int L(\theta, \hat{T}(\boldsymbol{x}))\pi(\theta|\boldsymbol{x}, \hat{\gamma}(\boldsymbol{x}))d\theta.$$

The empirical Bayes estimator is best suited to situations in which there are many problems that can be modeled simultaneously in a common way.

EXAMPLE 4.1.13 (Multiple Binomial Model). For the k-th treatment group, $k = 1, \ldots, K$, we measure X_k, the number of successes out of n trials, and model it as

$$X_k \sim \text{Bin}(n, p_k).$$

We attach the same prior to each p_k (this is appropriate since the treatments all correspond to the same disease):

$$p_k \sim \text{Beta}(a, b).$$

Now, from Example 4.1.4, the Bayes estimator of p_k is:

$$T(\boldsymbol{x}) = \frac{a + x_k}{a + b + n}.$$

These were easy calculations for fixed (a, b) and conjugacy. To compute the empirical Bayes estimator, we first need the marginal of $\boldsymbol{X} = (X_1, \ldots, X_K)$:

$$\begin{aligned} m(\boldsymbol{x}|a, b) &= \int_0^1 \cdots \int_0^1 \prod_{k=1}^{K} \binom{n}{x_k} p_k^{x_k}(1 - p_k)^{n-x_k} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} p_k^{a-1}(1 - p_k)^{b-1} dp_k \\ &= \prod_{k=1}^{K} \binom{n}{x_k} \frac{\Gamma(a + b)\Gamma(a + x_k)\Gamma(n - x_k + b)}{\Gamma(a)\Gamma(b)\Gamma(a + b + n)}. \end{aligned}$$

There is no closed-form solution to this, but one can compute the MLEs $(\hat{a}, \hat{b})$ numerically, leading to the empirical Bayes estimator:

$$\hat{T}(\boldsymbol{x}) = \frac{\hat{a} + x_k}{\hat{a} + \hat{b} + n}.$$

It turns out that the Bayes risk of the empirical Bayes estimator is often just slightly higher than that of the Bayes estimator (which therefore enjoys a certain degree of *robustness*).

For estimation of the canonical parameter in exponential families, the empirical Bayes estimator can be expressed in the same form as the Bayes estimator.

THEOREM 4.1.14. , *For the situation of Corollary 4.1.9 with prior $\pi(\boldsymbol{\eta}|\lambda)$ where λ is a hyperparameter, the empirical Bayes estimator of η_i is:*

$$\mathbb{E}(\eta_i|\boldsymbol{x}, \hat{\lambda}) = \begin{cases} \frac{\partial \log m(\boldsymbol{x}|\lambda)}{\partial x_i}\Big|_{\lambda=\hat{\lambda}(\boldsymbol{x})} - \frac{\partial \log h(\boldsymbol{x})}{\partial x_i}, & \textit{using the MLE } \hat{\lambda}(\boldsymbol{x}), \\ \frac{\partial \log m(\boldsymbol{x}|\hat{\lambda}(\boldsymbol{x}))}{\partial x_i} - \frac{\partial \log h(\boldsymbol{x})}{\partial x_i}, & \textit{using general estimate } \hat{\lambda}(\boldsymbol{x}). \end{cases}$$

PROOF. Straightforward from Corollary 4.1.9 with with $m(\boldsymbol{x}) \mapsto m(\boldsymbol{x}|\hat{\lambda})$. For the MLE-based simplification, apply the chain-rule to the general estimate:

$$\frac{\partial \log m(\boldsymbol{x}|\hat{\lambda}(\boldsymbol{x}))}{\partial x_i} = \underbrace{\left.\frac{\partial \log m(\boldsymbol{x}|\lambda)}{\partial \lambda}\right|_{\lambda=\hat{\lambda}(\boldsymbol{x})}}_{=0 \text{ since } \hat{\lambda}(\boldsymbol{x}) \text{ is the MLE}} \cdot \frac{\partial \hat{\lambda}(\boldsymbol{x})}{\partial x_i} + \left.\frac{\partial \log m(\boldsymbol{x}|\lambda)}{\partial x_i}\right|_{\lambda=\hat{\lambda}(\boldsymbol{x})}.$$

□

EXAMPLE 4.1.15 (Multiple Normal Model (continued)).

$$X_i|\theta_i \sim \text{ indep. } N(\theta_i, \sigma^2), \quad \Theta_i \sim \text{ indep. } N(\mu, \tau^2), \quad i = 1, \ldots, s,$$

with μ unknown, $\{\sigma^2, \tau^2\}$ known. Note the marginal of $\boldsymbol{X}$ from before:

$$m(\boldsymbol{x}|\mu) = [2\pi(\sigma^2+\tau^2)]^{-s/2} \exp\left\{-\frac{\sum(x_i-\mu)^2}{2(\sigma^2+\tau^2)}\right\},$$

which implies $\hat{\mu} = \bar{X}$ is the MLE of μ. By Problem 4.6.10(a) and Example 4.1.10, we thus obtain:

$$\begin{aligned} \text{empirical Bayes estimator of } \theta_i &= \frac{\tau^2}{\sigma^2+\tau^2} x_i + \frac{\sigma^2}{\sigma^2+\tau^2}\hat{\mu} \\ &= \text{Bayes estimator of } \theta_i \text{ under a } N(\hat{\mu}, \sigma^2) \text{ prior.} \end{aligned}$$

NOTE 4.1.16. Using Theorem 4.1.11, we can express risk of Bayes estimator in Theorem 4.1.14 as:

$$R\left(\boldsymbol{\eta}, \mathbb{E}(\boldsymbol{\eta}|\boldsymbol{X}, \hat{\lambda})\right) = R\left(\boldsymbol{\eta}, -\frac{\partial \log h(\boldsymbol{X})}{\partial \boldsymbol{X}}\right) + \sum_{i=1}^{s} \mathbb{E}_{\boldsymbol{\eta}}\left[2\frac{\partial^2 \log m(\boldsymbol{X}|\hat{\lambda})}{\partial X_i^2} + \left(\frac{\partial \log m(\boldsymbol{X}|\hat{\lambda})}{\partial X_i}\right)^2\right].$$

For the Multiple Normal Model we get the result in Problem 4.6.10(b).

4.2. Minimax Estimation

DEFINITION 4.2.1. A statistic T is said to be **minimax** if T satisfies

$$\min_{T'} \sup_{\theta \in \Omega} R(\theta, T') = \sup_{\theta} R(\theta, T).$$

We have seen that many estimation problems allow the determination of UMVU, MRE or Bayes estimators. Minimax estimators however are usually much harder to find.

A minimax estimator minimizes the maximum risk. i.e., a minimax estimator minimizes the risk in the worst case. This suggests a possible connection with Bayes estimation under the worst possible prior.

Given Λ on Ω, **let r_Λ denote the Bayes risk of the Bayes estimator** T_Λ, i.e.

$$r_\Lambda := \int R(\theta, T_\Lambda)\, d\Lambda(\theta).$$

DEFINITION 4.2.2. A **prior distribution Λ is least favorable** if

$$r_\Lambda \geq r_{\Lambda'} \text{ for all other priors } \Lambda'.$$

THEOREM 4.2.3. *Suppose a prior Λ satisfies*

$$r_\Lambda = \sup_\theta R(\theta, T_\Lambda).$$

Then

(1) *T_Λ is minimax.*
(2) *If T_Λ is the unique Bayes estimator under Λ, then it is the unique minimax estimator.*
(3) *Λ is least favorable.*

PROOF. (1) For any estimator T,

$$\begin{aligned}
\sup_{\theta\in\Omega} R(\theta, T) &\geq \int R(\theta, T)\, d\Lambda(\theta) \\
&\geq \int R(\theta, T_\Lambda)\, d\Lambda(\theta) \text{ (since } T_\Lambda \text{ is Bayes for } \Lambda.) \\
&= \sup_{\theta\in\Omega} R(\theta, T_\Lambda) \text{ by hypothesis.}
\end{aligned}$$

(2) If T_Λ is the unique Bayes solution then the second inequality in (1) becomes strict. (i.e., $\geq \to >$.)
(3) If Λ' is another prior distribution then

$$\begin{aligned}
r_{\Lambda'} &= \int R(\theta, T_{\Lambda'})\, \Lambda'(d\theta) \\
&\leq \int R(\theta, T_\Lambda)\, d\Lambda'(\theta) \text{ since } T_{\Lambda'} \text{ is Bayes for } \Lambda' \\
&\leq \sup_\theta R(\theta, T_\Lambda) \\
&= r_\Lambda \text{ by hypothesis.}
\end{aligned}$$

□

COROLLARY 4.2.4. *If T_Λ has constant risk then it is minimax.*

PROOF. If $R(\theta, T_\Lambda)$ is independent of θ then

$$\int R(\theta, T_\Lambda)\, d\Lambda(\theta) = \sup_\theta R(\theta, T_\Lambda).$$

□

EXAMPLE 4.2.5. Let $X \sim \text{bin}(n,p)$, $L(\theta,d) = (d-\theta)^2$, and $\theta = p$. We shall show below that X/n is not minimax. Let us use Corollary 4.2.4 to derive a minimax estimator. Suppose $P \sim B(a,b)$. Then we have from example (4.1.4) that

$$\begin{aligned} f_{P|X}(p|x) &= c(x)\, p^{a+x-1}(1-p)^{b+n-x-1} \text{ and} \\ T_\Lambda &= \frac{a+x}{a+b+n}. \end{aligned}$$

Thus we obtain

$$\begin{aligned} R(p,T_\Lambda) &= \frac{1}{(a+b+n)^2} E_p\left(a+X-p(a+b+n)\right)^2 \\ &= \frac{npq+(qa-pb)^2}{(a+b+n)^2}. \end{aligned}$$

Now, choose a and b to make the risk independent of p. Since the coefficient of p^2 is $-n+(a+b)^2$ and the coefficient of p is $n-2a(a+b)$, setting these equal to zero gives

$$a = b = \frac{\sqrt{n}}{2}.$$

Hence

$$T_{\Lambda_0} = \frac{x+\frac{\sqrt{n}}{2}}{n+\sqrt{n}}$$

is Bayes with constant risk and is therefore minimax.

Since T_{Λ_0} is the unique Bayes solution with respect to $\Lambda_0 \sim B\left(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}\right)$, it follows from theorem 4.2.3 part (2) that T_{Λ_0} is the unique minimax estimator of p with respect to squared error loss and the risk is

$$\begin{aligned} r_{\Lambda_0} &= E(T_{\Lambda_0}-p)^2 = \int R(p,T_{\Lambda_0})\,\Lambda_0(dp) \\ &= \frac{1}{(n+\sqrt{n})^2}\int \frac{n}{4}\Lambda_0(dp) \\ &= \frac{1}{\beta_n} = R(p,T_{\Lambda_0}), \qquad \text{where } \beta_n = 4\left(1+\sqrt{n}\right)^2. \end{aligned}$$

The risk of the usual estimator with squared error loss, $T(X) = X/n$, is

$$R(p,T) = E_p\left(\frac{X}{n}-p\right)^2 = \frac{p(1-p)}{n}.$$

Thus, since the risks of T and T_{Λ_0} coincide at $\frac{1}{2} \pm c_n$, where $c_n = \sqrt{\beta_n^2 - 4n\beta_n}/(2\beta_n)$, we have:

$$\begin{aligned} R(p,T_{\Lambda_0}) &< R(p,T), \qquad \text{for } \left|p-\frac{1}{2}\right| < c_n, \\ R(p,T_{\Lambda_0}) &> R(p,T), \qquad \text{for } \left|p-\frac{1}{2}\right| > c_n. \end{aligned}$$

Now, $c_n \to 0$ as $n \to \infty$, but c_n is larger for small n. For $n = 1$,

$$T_{\Lambda_0}(x) = \begin{cases} \frac{1}{4} & \text{if } x = 0 \\ \frac{3}{4} & \text{if } x = 1. \end{cases}$$

REMARK 4.2.6. Squared error loss might not be best for estimating p. The penalty should perhaps be larger at the endpoints $p = 0$ and $p = 1$. If we use

$$L(p, d) = \frac{(d-p)^2}{pq} = \frac{(d-p)^2}{p(1-p)},$$

then X/n has constant risk and is Bayes with respect to $U(0, 1)$ prior (check!).

DEFINITION 4.2.7. Let Λ_k be a sequence of priors, and suppose

$$r_{\Lambda_k} = \int R(\theta, T_k)\, d\Lambda_k(\theta) \to r \text{ as } k \to \infty,$$

where T_k is Bayes with respect to Λ_k for each k. We say that the **sequence $\{\Lambda_k\}$ is least favorable** if

$$r_\Lambda \leq r \text{ for all } \Lambda,$$

i.e. if the **limit** of the minimum Bayes risk is at least as bad as the minimum Bayes risk for any prior. (Compare Definition 4.2.2.)

THEOREM 4.2.8. *If there exists an estimator T and a sequence of prior distributions $\{\Lambda_k\}$ such that*

$$\sup_\theta R(\theta, T) = \lim_{k\to\infty} r_{\Lambda_k}$$

then

(1) *T is minimax (but not necessarily unique.)*
(2) *$\{\Lambda_k\}$ is least favorable.*

PROOF. (1) If T' is any estimator, then

$$\begin{aligned} \sup_\theta R(\theta, T') &\geq \int R(\theta, T')\, d\Lambda_k \\ &\geq r_{\Lambda_k}, \quad \forall k \end{aligned}$$

and hence

$$\sup_\theta R(\theta, T') \geq \lim r_{\Lambda_k} = \sup_\theta R(\theta, T).$$

(2) If Λ is any prior, then

$$\begin{aligned} r_\Lambda &= \int R(\theta, T_\Lambda)\, d\Lambda(\theta) \\ &\leq \int R(\theta, T)\, d\Lambda(\theta) \\ &\leq \sup_\theta R(\theta, T) \\ &= \lim_{k\to\infty} r_{\Lambda_k}. \end{aligned}$$

□

EXAMPLE 4.2.9. Suppose that $X_1, \cdots, X_n \sim$ iid $N(\theta, \sigma^2)$.

(1) σ^2 **known:**
$\Theta \sim N(\mu, k^2) = \Lambda_k$. In example (4.1.6), we saw that the Bayes estimator for squared error loss is

$$E(\Theta|\mathbf{X}) = \frac{n\sigma^{-2}}{n\sigma^{-2} + k^{-2}}\overline{X} + \frac{k^{-2}}{n\sigma^{-2} + k^{-2}}\mu = T_{\Lambda_k}(\mathbf{X})$$

with

$$\begin{aligned} r_{\Lambda_k} &= E(T_{\Lambda_k} - \Theta)^2 \\ &= E\left(E(\Theta - E(\Theta|\mathbf{X}))^2 |\mathbf{X}\right) \\ &= E\mathrm{Var}(\Theta|\mathbf{X}) \\ &= \frac{1}{n/\sigma^2 + 1/k^2}. \end{aligned}$$

As $k \to \infty$,

$$r_\Lambda \to \frac{\sigma^2}{n} = R\left(\theta, \overline{X}\right).$$

Hence, $\overline{X}$ is minimax (for squared error loss) by theorem (4.2.8).

(2) σ^2 **unknown:**
Since $\sup_{\theta,\sigma^2} R((\theta, \sigma^2), T) = \infty$ for $T(\mathbf{X}) = \overline{X}$ and $\overline{X}$ is minimax for each fixed σ, we restrict σ^2 to satisfy $\sigma^2 \leq m < \infty$. If T is minimax on $\theta \in \mathbb{R}$, $\sigma^2 \leq m$ then

$$\sup_{\theta,\sigma^2=m} R\left(\left(\theta, \sigma^2\right), T\right) \leq \sup_{\theta,\sigma^2\leq m} R\left(\left(\theta, \sigma^2\right), \overline{X}\right) = \sup_{\theta,\sigma^2=m} R\left(\left(\theta, \sigma^2\right), \overline{X}\right).$$

But $\overline{X}$ is minimax on $\sigma^2 = m$, hence this is an equality. Hence $\overline{X}$ is minimax on $\theta \in \mathbb{R}$, $\sigma^2 \leq m$. Although the restriction $\sigma^2 \leq m$ was necessary to make the minimax problem meaningful, the minimax estimator $\overline{X}$ does not depend on m.

4.3. Minimaxity and Admissibility in Exponential families

Given two estimators T, T', such that

$$R(\theta, T) \leq R(\theta, T') \text{ for all } \theta$$

then T **is preferable to** T' **on the basis of risk**.

DEFINITION 4.3.1. T' is **inadmissible** (with respect to the loss function L) if there exists T such that

$$R(\theta, T) \leq R(\theta, T') \text{ for all } \theta$$

with strict inequality for some θ. (In this case, we also say that T dominates T'.)

An estimator is **admissible** if it is not inadmissible.

In general, it is difficult to determine whether or not an estimator is admissible...

(**Note:** We will naturally restrict our attention to *non-constant* estimators, since any $T = \theta_0$, where $\theta_0 \in \Omega$, is trivially admissible.)

THEOREM 4.3.2. *If T is a unique Bayes estimator with respect to some Λ, then T is admissible. (Uniqueness means that any two Bayes estimators T and T' differ only a set $\mathcal{D}$ where $P_\theta(\mathcal{D}) = 0$, $\forall\theta$.)*

PROOF. If T is not admissible then there exists T' such that

$$R(\theta, T) \geq R(\theta, T') \text{ for all } \theta$$

and

$$R(\theta, T) > R(\theta, T') \text{ for some } \theta.$$

Now $\int R(\theta, T)\, d\Lambda \geq \int R(\theta, T')\, d\Lambda$ implies that T' is Bayes with respect to Λ. Hence by uniqueness $R(\theta, T') = R(\theta, T)$ for all θ and we obtain the contradiction. □

EXAMPLE 4.3.3. Suppose that $X_1, \cdots, X_n \sim$ iid $N(\theta, \sigma^2)$ with σ^2 known. Let $\Theta \sim N(\mu, \tau^2)$ and $L(\theta, d) = (d - \theta)^2$. Then

$$T = \frac{n\tau^2}{\sigma^2 + n\tau^2}\overline{X} + \frac{\sigma^2}{\sigma^2 + n\tau^2}\mu \qquad (*)$$

is the unique Bayes estimator of Θ and is therefore admissible.

This example shows that $a\overline{X} + b$ is admissible for all $a \in (0, 1)$ and $b \in \mathbb{R}$ since any $a \in (0, 1)$ and $b \in \mathbb{R}$ can be obtained in (*) by suitable choice of μ and τ^2 and hence $a\overline{X} + b$ is unique Bayes for some Λ.

THEOREM 4.3.4. *If $X \sim N(\theta, \sigma^2)$, σ^2 is known and $L(\theta, d) = (d - \theta)^2$ then $aX + b$ is inadmissible for θ if*

(1) $a > 1$,
(2) $a < 0$, *or*
(3) $a = 1$ *and* $b \neq 0$.

PROOF. We first calculate

$$\begin{aligned} R(\theta, aX + b) &= E_\theta (aX + b - \theta)^2 \\ &= E_\theta (a(X - \theta) + \theta(a - 1) + b)^2 \\ &= a^2\sigma^2 + ((a - 1)\theta + b)^2. \end{aligned}$$

(1) If $a > 1$,

$$R(\theta, aX + b) > R(\theta, X) = \sigma^2 \text{ for all } \theta.$$

(2) If $a < 0$, then $(a-1)^2 > 1$ and

$$\begin{aligned} R(\theta, aX+b) &\geq ((a-1)\theta + b)^2 \\ &= (a-1)^2 \left(\theta + \frac{b}{a-1}\right)^2 \\ &> R\left(\theta, 0 \cdot X - \frac{b}{a-1}\right). \end{aligned}$$

(3) If $a = 1$ and $b \neq 0$, then

$$R(\theta, X+b) = \sigma^2 + b^2 > \sigma^2 = R(\theta, X).$$

□

COROLLARY 4.3.5. *Suppose that $X_1, \cdots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. Then $a\overline{X} + b$ is inadmissible if*

(1) $a > 1$,
(2) $a < 0$, *or*
(3) $a = 1$ *and* $b \neq 0$,

and admissible if $0 \leq a < 1$.

PROOF. It remains only to establish admissibility when $a = 0$. This is trivial since for the estimator $T(X) = b$ corresponding to $a = 0$, the risk at $\theta = b$ is $R(b, b) = 0$ and every other estimator has positive risk at b since if $P_\theta(T'(X) \neq b) > 0$ for some θ then $T'^{-1}(\{b\}^c)$ has positive Lebesque measure and this implies that $P_b(T'(X) \neq b) > 0$ and hence that $E_b(T' - b)^2 > 0$. □

PROPOSITION 4.3.6. *Suppose that $X_1, \cdots, X_n \sim$ iid $N(\theta, \sigma^2)$ with σ^2 known. Then, $\overline{X}$ is admissible.*

PROOF. We give two proofs.

(1) (Limiting Bayes method) Assume without loss of generality that $\sigma^2 = 1$. If $\overline{X}$ is inadmissible then there exists T^* such that $R(\theta, T^*) \leq \frac{1}{n}$ for all θ and $R(\theta_0, T^*) < \frac{1}{n}$ for some θ_0.

$$R(\theta, T^*) = E_\theta (T^* - \theta)^2 = \int (T^*(\mathbf{x}) - \theta)^2 \prod_{i=1}^n \frac{e^{-\frac{1}{2}(x_i - \theta)^2}}{\sqrt{2\pi}} d\mathbf{x}$$

is continuous in a neighborhood of θ_0. Hence there exist a and b such that $a < \theta_0 < b$ and $c > 0$ such that

$$R(\theta, T^*) < \frac{1}{n} - c \text{ for all } \theta \in (a, b).$$

Suppose that $\Theta \sim N(0,\tau^2) =: \Lambda$. Then we shall obtain a contradiction by showing that

$$r_\Lambda^* := ER(\Theta, T^*) = \frac{1}{\tau\sqrt{2\pi}} \int R(\theta, T^*)\, e^{-\frac{\theta^2}{2\tau^2}} d\theta$$

is smaller than the minimum Bayes risk for Λ, i.e.

$$r_\Lambda = ER(\Theta, T_\Lambda) = \frac{1}{n\sigma^{-2}+\tau^{-2}} = \frac{\tau^2}{n\tau^2+1}.$$

(As in example (4.1.6) ,

$$E\left((\Theta - T_\Lambda(\mathbf{X}))^2 |\mathbf{X}\right) = \frac{\tau^2}{n\tau^2+1} = \mathrm{Var}(\Theta|\mathbf{X})$$

so that

$$r_\Lambda = E(\Theta - T_\Lambda(\mathbf{X}))^2 = E\left(E\left((\Theta - T_\Lambda(\mathbf{X}))^2 |\mathbf{X}\right)\right) = \frac{\tau^2}{n\tau^2+1}.)$$

Now

$$\begin{aligned}
\frac{\frac{1}{n} - r_\Lambda^*}{\frac{1}{n} - r_\Lambda} &= \frac{\frac{1}{\tau\sqrt{2\pi}} \int \left(\frac{1}{n} - R(\theta, T^*)\right) e^{\frac{-\theta^2}{2\tau^2}}}{\frac{1}{n} - \frac{\tau^2}{n\tau^2+1}} d\theta \\
&\geq \frac{n(n\tau^2+1)}{\tau\sqrt{2\pi}} \int_a^b \left(\frac{1}{n} - R(\theta, T^*)\right) e^{-\frac{\theta^2}{2\tau^2}} d\theta \\
&> c\frac{n(n\tau^2+1)}{\tau\sqrt{2\pi}} \int_a^b e^{-\frac{\theta^2}{2\tau^2}} d\theta
\end{aligned}$$

where $c > 0$ is independent of τ. Since the integral in the last line converges to $b-a$ as $\tau \to \infty$ by DCT, the ratio goes to infinity as $\tau \to \infty$. Thus, for all sufficiently large τ_0,

$$r^*_{\Lambda(\mu,\tau_0)} < r_{\Lambda(\mu,\tau_0)} \ ,$$

which contradicts the fact that $r_{\Lambda(\mu,\tau_0)}$ is minimum Bayes risk.

(2) (Via the information inequality) If T is any estimator of θ with finite second moment under each P_θ, then $E_\theta T = b(\theta) + \theta$ and

$$\begin{aligned}
R(\theta, T) &= \mathrm{Var}_\theta T + b^2(\theta) \\
&\geq \frac{(1+b'(\theta))^2}{I(\theta)} + b^2(\theta) \\
&= \frac{(1+b'(\theta))^2}{n} + b^2(\theta)
\end{aligned}$$

($b'(\theta)$ exists by theorem (1.3.13) and we assume without loss of generality $\sigma = 1$.) Hence if T is risk-preferable to $\overline{X}$,

$$R(\theta, T) \leq \frac{1}{n} \text{ for all } \theta, \qquad (*)$$

i.e.

$$b^2(\theta) + \frac{(1+b'(\theta))^2}{n} \leq \frac{1}{n} \text{ for all } \theta \qquad (**)$$

and so

$$|b(\theta)| \leq \frac{1}{\sqrt{n}} \text{ for all } \theta \qquad (***)$$

and

$$\begin{aligned} & (1+b'(\theta))^2 \leq 1 \\ \Rightarrow \quad & -2 \leq b'(\theta) \leq 0 \\ \Rightarrow \quad & b \text{ is non-increasing.} \end{aligned}$$

Now, we claim $b'(\theta_k) \to 0$ for some sequence $\theta_k \to \infty$. If this is not the case, then $\overline{\lim}_{\theta\to\infty} b'(\theta) < 0$ and there exists θ_0 and $\epsilon > 0$ such that $b'(\theta) < -\varepsilon$ for all $\theta > \theta_0$. Then,

$$\begin{aligned} b(\theta) &= \int_{\theta_0}^{\theta} b'(y)\,dy + b(\theta_0) \\ &\leq (\theta - \theta_0)(-\varepsilon) + b(\theta_0) \to -\infty \text{ as } \theta \to \infty, \end{aligned}$$

contradicting $(***)$ and thus proving the claim. Similarly, there exists $\theta_j^* \to -\infty$ such that $b'\left(\theta_j^*\right) \to 0$.
Now, $(**)$ implies that $b\left(\theta_j^*\right) \to 0$ and $b(\theta_k) \to 0$. But since b is non-increasing, this implies that $b(\theta) = 0$ for all θ and hence that $b'(\theta) = 0$ for all θ. Hence, by the information inequality, $R(\theta, T) \geq \frac{1}{n}$, and so by $(*)$,

$$R(\theta, T) = \frac{1}{n} = R\left(\theta, \overline{X}\right)$$

so that $\overline{X}$ is admissible.

□

The above argument also shows that $\overline{X}$ is minimax since there is no estimator whose maximum risk is less than $1/n$. In fact, $\overline{X}$ is the unique minimax estimator by the following theorem.

PROPOSITION 4.3.7. *Suppose that T has constant risk and is admissible. Then T is minimax. If in addition $L(\theta, \cdot)$ is strictly convex, then T is the unique minimax estimator.*

PROOF. By the admissibility of T, if there is another estimator T' with $\sup_\theta R(\theta, T') \leq R(\theta, T)$ then $R(\theta, T') = R(\theta, T)$ for all θ, since the risk of T' can't go strictly below that of T for any θ. This proves that T is minimax. If the loss function is strictly convex and T' is a minimax estimator such that $P_\theta(T' \neq T) > 0$, then if $T^* = \frac{1}{2}(T + T')$,

$$R(\theta, T^*) < \frac{1}{2}\left(R(\theta, T) + R(\theta, T')\right) = R(\theta, T),$$

which contradicts the admissibility of T. □

Exponential Families (with $s = 1$). Suppose that the probability density of X with respect to the σ-finite measure μ is

$$e^{\theta T(x)-\varphi(\theta)}h(x)$$

where

$$\int e^{\theta T(x)}h(x)\,d\mu(x) = e^{\varphi(\theta)}$$

Then

$$E_\theta T(\mathbf{X}) = \varphi'(\theta) = g(\theta).$$

Suppose the natural parameter space Ω is an interval with end-points θ_L, θ_U, $-\infty \leq \theta_L \leq \theta_U \leq \infty$ and

$$L(\theta, d) = (d - g(\theta))^2.$$

The same argument used in the proof of theorem (4.3.4) shows that $aT+b$ is inadmissible for (1) $a < 0$, (2) $a > 1$, or (3) $a = 1$ and $b \neq 0$.

If $a = 0$, then $aT + b$ is admissible since $g\left(\hat{\theta}\right) = b$ is the only estimator with zero risk at b. To deal with the remaining cases consider

$$\frac{1}{1+\lambda}T + \frac{r\lambda}{1+\lambda}\ , \quad 0 \leq \lambda < \infty,\ r \in \mathbb{R} \quad \text{(i.e. } 0 < a \leq 1\text{)}.$$

THEOREM 4.3.8 (Karlin's Theorem). *The estimator*

$$\frac{1}{1+\lambda}T + \frac{r\lambda}{1+\lambda}, \quad 0 \leq \lambda < \infty,\ r \in \mathbb{R},$$

is admissible for $g(\theta) = \varphi'(\theta) = E_\theta T$ if for some (and hence for all) $\theta_0 \in (\theta_L, \theta_U)$

$$\int_{\theta_L}^{\theta_0} e^{-r\lambda\theta+\lambda\varphi(\theta)}d\theta = \infty$$

and

$$\int_{\theta_0}^{\theta_U} e^{-r\lambda\theta+\lambda\varphi(\theta)}d\theta = \infty.$$

PROOF. Recall that

$$\begin{aligned}
\varphi'(\theta) &= E_\theta T \\
\varphi''(\theta) &= \mathrm{Var}_\theta(T) \\
I(\theta) &= E_\theta\left(\frac{\partial \log p(x,\theta)}{\partial\theta}\right)^2 \\
&= E_\theta(T - \varphi'(\theta))^2 = \varphi''(\theta).
\end{aligned}$$

Suppose there exists $\delta(X)$ such that

$$E_\theta(\delta(X) - \varphi'(\theta))^2 \leq E_\theta\left(\frac{T + r\lambda}{1+\lambda} - \varphi'(\theta)\right)^2 \text{ for all } \theta. \qquad (*)$$

We have that

$$\begin{aligned} E_\theta\left(\delta\left(X\right)-\varphi'\left(\theta\right)\right)^2 &= \mathrm{Var}_\theta\delta\left(X\right)+b^2\left(\theta\right) \\ &\geq \frac{\left(\frac{d}{d\theta}\left(b\left(\theta\right)+\varphi'\left(\theta\right)\right)\right)^2}{I\left(\theta\right)}+b^2\left(\theta\right) \\ &= b^2\left(\theta\right)+\frac{\left(b'\left(\theta\right)+I\left(\theta\right)\right)^2}{I\left(\theta\right)}. \end{aligned}$$

($b'\left(\theta\right)$ exists since $E_\theta\left|\delta\left(X\right)\right|<\infty$.) So by (*),

$$\frac{I\left(\theta\right)}{\left(1+\lambda\right)^2}+\frac{\lambda^2\left(r-\varphi'\left(\theta\right)\right)^2}{\left(1+\lambda\right)^2}\geq b^2\left(\theta\right)+\frac{\left(b'\left(\theta\right)+I\left(\theta\right)\right)^2}{I\left(\theta\right)} \qquad (**).$$

Letting

$$h\left(\theta\right)=b\left(\theta\right)-\frac{\lambda}{1+\lambda}\left(r-\varphi'\left(\theta\right)\right)=b\left(\theta\right)-\text{bias}\left(\frac{T}{1+\lambda}+\frac{r\lambda}{1+\lambda}\right),$$

$$h'\left(\theta\right)=b'\left(\theta\right)+\frac{\lambda}{1+\lambda}\varphi''\left(\theta\right).$$

(**) is exactly equivalent to

$$\begin{aligned} 0 &\geq h^2\left(\theta\right)+2h\left(\theta\right)\frac{\lambda}{1+\lambda}\left(r-\varphi'\left(\theta\right)\right)+\frac{\left(h'\left(\theta\right)+\frac{1}{1+\lambda}\varphi''\left(\theta\right)\right)^2}{\varphi''\left(\theta\right)}-\frac{\phi''\left(\theta\right)}{\left(1+\lambda\right)^2} \qquad (***) \\ &= h^2\left(\theta\right)-2h\left(\theta\right)\frac{\lambda}{1+\lambda}\left(\varphi'\left(\theta\right)-r\right)+\frac{2h'\left(\theta\right)}{1+\lambda}\left(+\frac{h'\left(\theta\right)^2}{\varphi''\left(\theta\right)}\right). \end{aligned}$$

Letting $k\left(\theta\right)=h\left(\theta\right)e^{r\lambda\theta-\lambda\varphi(\theta)}$, (***) becomes

$$\begin{aligned} k^2\left(\theta\right)e^{r\lambda\theta-\lambda\varphi(\theta)}+\frac{2}{1+\lambda}k'\left(\theta\right) &\leq 0 \qquad (****) \\ \Rightarrow k'\left(\theta\right)\leq 0 &\text{ for all } \theta. \end{aligned}$$

Hence, $\underline{k\left(\theta\right)\text{ is decreasing.}}$ To prove

$$k\left(\theta\right)\geq 0 \text{ for all } \theta,$$

suppose $k\left(\theta_0\right)<0$.Then $k\left(\theta\right)<0$ for all $\theta>\theta_0$. From (****), we also have

$$\frac{d}{d\theta}\left(\frac{1}{k\left(\theta\right)}\right)=-\frac{k'\left(\theta\right)}{k\left(\theta\right)^2}\geq\frac{1+\lambda}{2}e^{r\lambda\theta-\lambda\varphi(\theta)} \text{ for all } \theta>\theta_0 \qquad (*****).$$

Integrating both sides of (*****) from θ_0 to $\theta_1>\theta_0$,

$$\frac{1}{k\left(\theta_1\right)}-\frac{1}{k\left(\theta_0\right)}\geq\frac{1+\lambda}{2}\int_{\theta_0}^{\theta_1}e^{r\lambda\theta-\lambda\varphi(\theta)}d\theta.$$

As $\theta_1\to\theta_U$, this integral converges to ∞ by assumption. But the left hand side is less than $-\frac{1}{k(\theta_0)}$ so that we obtain the contradiction. Hence

$$k\left(\theta\right)\geq 0 \text{ for all } \theta.$$

Similarly,

$$\int_{\theta_L}^{\theta_0} e^{r\lambda\theta-\lambda\varphi(\theta)}d\theta=\infty\Rightarrow k(\theta)\leq 0 \text{ for all } \theta.$$

Hence,

$$k(\theta)=0\Rightarrow h(\theta)=0 \text{ for all } \theta\Rightarrow h'(\theta)=0 \text{ for all } \theta.$$

Thus, $h'(\theta)=0$ and $h(\theta)=0$ implies equality in (***), equality in (**), and finally equality in (*). (RS of (*)=LS of (**) $\geq$ LS of (*) $\geq$RS of (**)). Since $\frac{T+r\lambda}{1+\lambda}$ has the same risk as δ, $\frac{T+r\lambda}{1+\lambda}$ is admissible. □

NOTE 4.3.9. **The case $\lambda=0$ is of particular interest, i.e. T is admissible for $E_\theta T$ provided $\theta_L=-\infty$ and $\theta_U=\infty$.**

EXAMPLE 4.3.10. Suppose $X\sim b(n,p)$ and $\theta:=\log\frac{p}{1-p}$, $-\infty<\theta<\infty$. Then

$$p(x,\theta)=\binom{n}{x}p^x(1-p)^{n-x}=\binom{n}{x}e^{\theta x-n\log\left(1+e^\theta\right)}$$

and $T(X)=X$ is admissible for

$$\phi'(\theta)=\frac{ne^\theta}{1+e^\theta}=np$$

since $\theta_L=-\infty$ and $\theta_U=\infty$.

EXAMPLE 4.3.11. Suppose $X_1,\cdots,X_n\sim$iid $N(\theta,\sigma^2)$ with σ^2 known.

$$\begin{aligned}p(\mathbf{x},\theta) &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}\exp\left(-\frac{\sum x_i^2}{2\sigma^2}\right)\exp\left(\theta\frac{\sum x_i}{\sigma^2}-\frac{n\theta^2}{2\sigma^2}\right)\\ T(\mathbf{X}) &= \frac{\sum X_i}{\sigma^2}\\ \phi'(\theta) &= \frac{n\theta}{\sigma^2}.\end{aligned}$$

$T(\mathbf{X})$ is admissible for $n\theta/\sigma^2$ since $\theta_L=-\infty$ and $\theta_U=\infty$.

4.4. Shrinkage Estimators and Bigdata

The idea behind "shrinkage" is to deliberately introduce bias in unbiased (or nearly unbiased) estimators (UMVUE, MLE) in order to reduce their risk. James & Stein (1961) were the first to do this. While this was seen as unusual and irrelevant to the applications of the time, it has become crucially important in the bigdata era. Before telling this story we discuss extensions to our earlier esults.

Simultaneous estimation and extensions of earlier results. So far our estimand, $g(\boldsymbol{\theta})$, has been scalar. In general we want to estimate an r-dimensional function $\boldsymbol{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_r(\boldsymbol{\theta}))^T$, based on an r-dimensional estimator $\boldsymbol{T} = (T_1, \ldots, T_r)^T$. The first issue is how to define risk with multivariate arguments. The two most common definitions under squared error loss are as follows.

(a) Risk under **sum of squared errors loss**:

$$R(\boldsymbol{\theta}, \boldsymbol{T}) = E_\theta[\boldsymbol{T} - \boldsymbol{g}(\boldsymbol{\theta})]^T[\boldsymbol{T} - \boldsymbol{g}(\boldsymbol{\theta})] = \sum_{i=1}^{r}[T_i - g_i(\boldsymbol{\theta})]^2 = \text{ scalar.}$$

This gives a complete ordering of estimators, i.e., for $\boldsymbol{T}$ and $\boldsymbol{T}'$, exactly one of $R(\boldsymbol{\theta}, \boldsymbol{T}) < R(\boldsymbol{\theta}, \boldsymbol{T}')$, $R(\boldsymbol{\theta}, \boldsymbol{T}) = R(\boldsymbol{\theta}, \boldsymbol{T}')$, or $R(\boldsymbol{\theta}, \boldsymbol{T}) > R(\boldsymbol{\theta}, \boldsymbol{T}')$, holds. (It will be our default risk function.)

(b) Risk under **concentration matrix loss**:

$$\begin{aligned} \mathcal{R}(\boldsymbol{\theta}, \boldsymbol{T}) &= E_\theta[\boldsymbol{T} - \boldsymbol{g}(\boldsymbol{\theta})][\boldsymbol{T} - \boldsymbol{g}(\boldsymbol{\theta})]^T = \text{ matrix } (r \times r) \\ &= \mathrm{Var}_\theta \boldsymbol{T}, \qquad\qquad \text{if } \boldsymbol{T} \text{ is unbiased.} \end{aligned}$$

We say that $\boldsymbol{T}$ is more concentrated about $\boldsymbol{g}(\boldsymbol{\theta})$ than $\boldsymbol{T}'$ if

$$\mathcal{R}(\boldsymbol{\theta}, \boldsymbol{T}') - \mathcal{R}(\boldsymbol{\theta}, \boldsymbol{T}) \geq 0, \qquad \text{(p.s.d.)} \tag{4.4.1}$$

This gives only a partial ordering of estimators, because the matrix in (4.4.1) may be neither psd nor nsd (e.g., when it has both negative and positve eigenvalues).

NOTE 4.4.1. It can be shown that if $R(\boldsymbol{\theta}, \boldsymbol{T}) \leq R(\boldsymbol{\theta}, \boldsymbol{T}')$ for every convex loss function $L(\boldsymbol{\theta}, \boldsymbol{d})$, then (4.4.1) holds (TPE Lemma 5.4.1).

With the (obvious) definition that $\boldsymbol{T}$ is unbiased for $\boldsymbol{g}(\boldsymbol{\theta})$ if and only if $E_{\boldsymbol{\theta}}\boldsymbol{T} = \boldsymbol{g}(\boldsymbol{\theta})$, we have the following extensions of earlier results.

(1) Rao-Blackwell Theorem (Ch. 2). The multivariate version is essentially the same; if $\boldsymbol{T}_0$ is unbiased for $\boldsymbol{g}(\boldsymbol{\theta})$ and $\boldsymbol{S}$ is complete & sufficient, then $E(\boldsymbol{T}_0|\boldsymbol{S})$, has uniformly minimum risk among all unbiased estimators (is UMVU), and is thus more concentrated about $\boldsymbol{g}(\boldsymbol{\theta})$ than any other unbiased estimator.
(2) Equivariant Estimation (Ch. 3). All definitions and results apply without change.
(3) Bayes, Minimaxity, Admissibility (Ch. 4).
   - The definition of Bayes estimator remains unchanged, but one can often compute these componentwise by marginalizing both the likelihood and the prior (e.g., TPE Problem 5.4.3). This marginalization trick always holds true under sq. error loss (TPE Lemma 5.4.3).
   - The definition of minimaxity remains unchanged, but results have to be derived individually for each situation.
   - For admissibility the story is quite different, as we will see in the remainder of this section.

James-Stein Estimator. Let $X_i \sim N(\theta_i, 1), i = 1, \cdots, s$ be independent random variables and let

$$L(\boldsymbol{\theta}, \boldsymbol{d}) = \|\boldsymbol{d} - \boldsymbol{\theta}\|^2 = \sum (d_i - \theta_i)^2 .$$

The usual (MLE) estimator of $\boldsymbol{\theta}$

$$T(\boldsymbol{X}) = (X_1, \cdots, X_s)$$

will be shown to be inadmissible if $s > 2$.

Let

$$\delta_{i,c} = \left(1 - c\frac{s-2}{S^2}\right) X_i,\ i = 1, \cdots, s,$$

where $S^2 = \sum_{i=1}^{s} X_i^2$. Here

$$\boldsymbol{\delta}_c := \begin{pmatrix} \delta_{1,c} \\ \vdots \\ \delta_{s,c} \end{pmatrix}.$$

THEOREM 4.4.2.

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}_c) = s - (s-2)^2 E_\theta \left(\frac{2c - c^2}{S^2}\right).$$

To prove the theorem, we need the following two lemmas.

LEMMA 4.4.3. *If $X \sim N(0,1)$ and $g : \mathbb{R} \to \mathbb{R}$ is absolutely continuous with derivative g' then*

$$E|g'(X)| < \infty \Rightarrow Eg'(X) = E(Xg(X)).$$

PROOF. Let $\phi(x) = (2\pi)^{-\frac{1}{2}} \exp(-x^2/2)$. Then

$$x\phi(x) = -\phi'(x) \qquad (*).$$

Then

$$\begin{aligned}
Eg'(X) &= \int g'(x)\,\phi(x)\,dx \\
&= \int_0^\infty + \int_{-\infty}^0 g'(x)\,\phi(x)\,dx \\
&= \int_0^\infty g'(x) \left(\int_x^\infty z\phi(z)\,dz\right) dx - \int_{-\infty}^0 g'(x) \left(\int_{-\infty}^x z\phi(z)\,dz\right) dx \text{ by } (*) \\
&= \int_0^\infty \left(\int_0^z g'(x)\,dx\right) z\phi(z)\,dz - \int_{-\infty}^0 \left(\int_z^0 g'(x)\,dx\right) z\phi(z)\,dz \text{ by Fubini} \\
&= \int_0^\infty (g(z) - g(0))\, z\phi(z)\,dz - \int_{-\infty}^0 (g(0) - g(z))\, z\phi(z)\,dz \\
&= E(Xg(X)).
\end{aligned}$$

□

LEMMA 4.4.4. *Suppose* $\boldsymbol{X} = (X_1, \cdots, X_s)$ *where* $X_1, \cdots, X_s$ *are independent and*

$$X_i \sim N(\mu_i, v_i).$$

Suppose $f : \mathbb{R}^s \to \mathbb{R}$ *is absolutely continuous in* x_i *for almost all* $(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_s)$. *Then if*

$$E\left|\frac{\partial}{\partial x_i} f(\boldsymbol{X})\right| < \infty,$$

$$v_i E \frac{\partial}{\partial x_i} f(\boldsymbol{X}) = E(X_i - \mu_i) f(\boldsymbol{X}).$$

PROOF. Let

$$Z = \frac{X_i - \mu_i}{\sqrt{v_i}}.$$

then from lemma (4.4.3),

$$\begin{aligned} & E\left(\tfrac{\partial}{\partial z} f\left(x_1, \cdots, x_{i-1}, \mu_i + \sqrt{v_i}\zeta, x_{i+1}, \cdots, x_s\right)|_{\zeta = Z}\right) \\ = \quad & E\left(Z f\left(x_1, \cdots, x_{i-1}, \mu_i + \sqrt{v_i} Z, x_{i+1}, \cdots, x_s\right)\right). \end{aligned}$$

Thus,

$$\begin{aligned} & \sqrt{v_i} E\left(\tfrac{\partial f}{\partial x_i}(\boldsymbol{X}) | X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_s\right) \\ = \quad & E\left(\tfrac{X_i - \mu_i}{\sqrt{v_i}} f(\boldsymbol{X}) | X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_s\right). \end{aligned}$$

Taking expectation of each side, we obtain the desired result. □

Now, we are ready to prove the theorem.

PROOF. (of theorem) Let

$$f_i(\boldsymbol{X}) = \frac{c(s-2)}{S^2} X_i \text{ and } \boldsymbol{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_s \end{pmatrix}.$$

Then

$$\boldsymbol{\delta}_c = \boldsymbol{X} - \boldsymbol{f}$$

and

$$\begin{aligned} R(\boldsymbol{\theta}, \boldsymbol{\delta}_c) &= E\left(\sum (X_i - f_i - \theta_i)^2\right) \\ &= E\left(\sum (X_i - \theta_i)^2 - 2\sum (X_i - \theta_i) f_i + \sum f_i^2\right) \\ &= s - 2\sum_1^s E \frac{\partial}{\partial x_i} f_i(\boldsymbol{X}) + \sum E\left(\frac{X_i^2}{S^4}\right) c^2 (s-2)^2. \end{aligned}$$

Since

$$\frac{\partial}{\partial x_i} f_i(\boldsymbol{X}) = c(s-2)\frac{S^2 - X_i \cdot 2X_i}{S^4} = c(s-2)\left(\frac{1}{S^2} - \frac{2X_i^2}{S^4}\right),$$

$$\begin{aligned} R(\boldsymbol{\theta}, \boldsymbol{\delta}_c) &= s - 2c(s-2) E\left(\frac{s-2}{S^2}\right) + c^2 (s-2)^2 E\frac{1}{S^2} \\ &= s - (s-2)^2 E\left(\frac{2c-c^2}{S^2}\right). \end{aligned}$$

□

COROLLARY 4.4.5. *For $0 < c < 2$ and $s > 2$,*

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}_c) < s \text{ for all } \theta$$

and $\boldsymbol{\delta}_1$ dominates all the other $\boldsymbol{\delta}_c$'s.

PROOF. $2c - c^2 > 0$ for all $c \in (0,2)$ and has a maximum value at $c = 1$. □

REMARK 4.4.6. The *James-Stein estimator* $\left(1 - \frac{s-2}{S^2}\right)\boldsymbol{X}$ has risk $R(\boldsymbol{\theta}, \boldsymbol{\delta}_1) = s - (s-2)^2 E\left(\frac{1}{S^2}\right)$. The risk of $\boldsymbol{X}$ is

$$R(\boldsymbol{\theta}, \boldsymbol{X}) = E\sum (X_i - \theta_i)^2 = s.$$

Therefore, $\boldsymbol{X}$ is not admissible for $\boldsymbol{\theta}$. In fact, the James-Stein estimator is not admissible either. A strictly risk-preferable estimator can be arrived at as follows.

Empirical Bayes Interpretation of the James-Stein Estimator. Suppose $\Theta_1, \cdots, \Theta_s$ are iid $N(0, \tau^2)$, i.e. this is the prior on each Θ_i. If τ^2 were known, the Bayes estimator with respect to squared error loss would be

$$\hat{\theta}_i = \frac{X_i}{1 + \tau^{-2}} = \left(1 - \frac{1}{1+\tau^2}\right) X_i, \qquad i = 1, \cdots, s.$$

However, if τ^2 is unknown it must be replaced by some estimate. Write

$$X_i = \Theta_i + Z_i, \{Z_i\} \sim \text{iid } N(0,1) \text{ with } \{Z_i\} \text{ independent of } \{\Theta_i\},$$

so that $X_i | \Theta_i = \theta_i \sim N(\theta_i, 1)$. Then,

$$\{X_i\} \sim \text{iid } N\left(0, \tau^2 + 1\right).$$

Since $S^2 = \sum_{i=1}^s X_i^2$ is complete and sufficient for τ^2, $\frac{s-2}{S^2}$ is UMVU for $\frac{1}{1+\tau^2}$. So a natural (empirical Bayes) estimator of θ_i is

$$\boldsymbol{\delta}_1 = \left(1 - \frac{s-2}{S^2}\right)\boldsymbol{X}.$$

Moreover since

$$\frac{1}{1+\tau^2} < 1,$$

a better estimate of $\frac{1}{1+\tau^2}$ is $\min\left(\frac{s-2}{S^2}, 1\right)$ which suggests using

$$\boldsymbol{\delta}_1^* = \left(1 - \min\left(\frac{s-2}{S^2}, 1\right)\right)\boldsymbol{X} = \max\left(1 - \frac{s-2}{S^2}, 1\right)\boldsymbol{X}.$$

$\boldsymbol{\delta}_1^*$ is strictly risk-preferable to $\boldsymbol{\delta}_1$ (so $\boldsymbol{\delta}_1$ is inadmissible). But $\boldsymbol{\delta}_1^*$ is also inadmissible. It was a difficult problem to find an estimator which is strictly risk-preferable to δ_1^*, in fact it took twenty years. It is now known that there are many admissible minimax estimators (TPE p. 357).

4.5. Discussion (Efron & Hastie, 2016)

- Although we gave a Bayesian interpretation of the JSE (James-Stein Estimator), it does not rely on any Bayesian assumptions!
- It's hard to construct JSE-like estimators; have to be done case-by-case (like UMVUEs). MLE on the other hand provides automatically asymptotically UMVU estimates (Ch. 6).
- The "shrinkage" in the above examples was toward zero, but in general it is toward a common central value like a mean (usually representing a null of no difference).
- Classical vs. Bigdata (to shrink or not to shrink): let n denote the sample size and p the number of parameters to estimate.

	Classical Data ($n \gg p$)	Big Data ($n \approx p$ or $n \ll p$)
Shrink?	no (generally)	yes (generally)
Methods	MLE	penalized/regularized likelihood
	least-squares	penalized/regularized least-squares
		max. posterior probability (MAP)

- Shrinkage tends to produce better results in general (on average), but this comes at the expense of extreme cases (outliers). E.g., if most of the $\theta_i \approx 0$ in JSE, but there are a very few large $|\theta_i|$, the result will be heavy shrinkage of the latter (toward overall mean ≈ 0). This situation is not uncommon in contemporary bigdata where the outliers are precisely the "interesting" cases swimming in a sea of uninterestingness... (see TPE p. 364–365).

CHAPTER 5

Large Sample Theory

This chapter introduces definitions, tools, and techniques for establishing asymptotic results. Before detailing the different modes of convergence, we recall the following basic result which is often used in proofs (along with the Triangle Inequality).

PROPOSITION 5.0.1 (Chebychev's Inequality). *If $Eg(X) < \infty$, where $g(\cdot)$ is a nonnegative function and $\varepsilon > 0$, then*

$$P(g(X) \geq \varepsilon) \leq Eg(X)/\varepsilon.$$

PROOF. TPE Problem 1.8.1. □

5.1. Convergence in Probability and Order in Probability

DEFINITION 5.1.1. A sequence of random variables X_n is said to converge to 0 in probability if for any $\varepsilon > 0$,

$$P(|X_n| > \varepsilon) \to 0 \quad \text{as } n \to \infty,$$

in which case we write

$$X_n \xrightarrow{p} 0$$

or equivalently

$$X_n = o_p(1).$$

DEFINITION 5.1.2. $\{X_n\}$ is *bounded in probability* (or *tight*) if for any $\varepsilon > 0$, there exists $M(\varepsilon) < \infty$ such that

$$P(|X_n| > M) < \varepsilon \text{ for all } n, \text{ (or equivalently for all } n \text{ sufficiently large)},$$

in which case we write

$$X_n = O_p(1).$$

DEFINITION 5.1.3.

$$\begin{aligned} X_n \xrightarrow{p} X &\iff X_n - X \xrightarrow{p} 0 \iff X_n - X = o_p(1) \\ X_n = o_p(a_n) &\iff \frac{X_n}{a_n} = o_p(1) \\ X = O_p(a_n) &\iff \frac{X_n}{a_n} = O_p(1). \end{aligned}$$

PROPOSITION 5.1.4. *Let $\{X_n\}$ and $\{Y_n\}$ be sequences of r.v's and suppose $a_n > 0$ and $b_n > 0$. Then the following results hold:*

(1) *If* $X_n = o_p(a_n)$ *and* $Y_n = o_p(b_n)$, *then*

$$\begin{aligned} X_n Y_n &= o_p(a_n b_n) \\ X_n + Y_n &= o_p(\max(a_n, b_n)) \\ |X_n|^r &= o_p(a_n^r),\ r > 0. \end{aligned}$$

(2) *If* $X_n = o_p(a_n)$ *and* $Y_n = O_p(b_n)$, *then*

$$X_n Y_n = o_p(a_n b_n).$$

(3) *If* $X_n = O_p(a_n)$ *and* $Y_n = O_p(b_n)$, *then*

$$\begin{aligned} X_n Y_n &= O_p(a_n b_n) \\ X_n + Y_n &= O_p(\max(a_n, b_n)) \\ |X_n|^r &= O_p(a_n^r),\ r > 0. \end{aligned}$$

PROOF. We only prove the first part and leave the remaining parts as exercises.

If $\left|\frac{X_n Y_n}{a_n b_n}\right| > \varepsilon$, then

$$\begin{aligned} &\text{either } \left|\frac{Y_n}{b_n}\right| \leq 1 \text{ and } \left|\frac{X_n}{a_n}\right| > \varepsilon \\ &\text{or } \left|\frac{Y_n}{b_n}\right| > 1 \text{ and } \left|\frac{X_n}{a_n}\frac{Y_n}{b_n}\right| > \varepsilon. \end{aligned}$$

Thus, if $X_n = o_p(a_n)$ and $Y_n = o_p(b_n)$, then

$$P\left(\frac{|X_n Y_n|}{a_n b_n} > \varepsilon\right) \leq P\left(\left|\frac{X_n}{a_n}\right| > \varepsilon\right) + P\left(\left|\frac{Y_n}{b_n}\right| > 1\right) \to 0 \text{ as } n \to \infty.$$

If $\frac{|X_n + Y_n|}{\max(a_n, b_n)} > \varepsilon$, then since $|X_n + Y_n| \leq |X_n| + |Y_n|$,

$$\frac{|X_n|}{a_n} > \frac{\varepsilon}{2} \text{ or } \frac{|Y_n|}{b_n} > \frac{\varepsilon}{2}.$$

Thus, as in the previous part,

$$P\left(\left|\frac{X_n + Y_n}{\max(a_n, b_n)}\right| > \varepsilon\right) \to 0$$

If $\frac{|X_n|^r}{a_n^r} > \varepsilon$, then $\frac{|X_n|}{a_n} > \varepsilon^{\frac{1}{r}}$. Thus,

$$P\left(\frac{|X_n|^r}{a_n^r} > \varepsilon\right) \to 0.$$

□

DEFINITION 5.1.5. For a sequence of random vectors $\mathbf{X}_n = (X_{n1}, \cdots, X_{nm})$, we define O_p and o_p as follows:

$$\begin{aligned}
\mathbf{X}_n = o_p(a_n) &\iff X_{nj} = o_p(a_n),\ j = 1, \cdots, m. \\
\mathbf{X}_n = O_p(a_n) &\iff X_{nj} = O_p(a_n),\ j = 1, \cdots m. \\
\mathbf{X}_n \xrightarrow{p} \mathbf{X} &\iff \mathbf{X}_n - \mathbf{X} = o_p(1) \iff X_{nj} \xrightarrow{p} X_j,\ j = 1, \cdots, m.
\end{aligned}$$

DEFINITION 5.1.6.

$$\|\mathbf{X}_n - \mathbf{X}\|^2 := \sum_{j=1}^{m} |X_{nj} - X_j|^2.$$

PROPOSITION 5.1.7.

$$\mathbf{X}_n - \mathbf{X} = o_p(1) \iff \|\mathbf{X}_n - \mathbf{X}\| = o_p(1).$$

PROOF. $\Rightarrow$)

$$\begin{aligned}
P\left(\|\mathbf{X}_n - \mathbf{X}\|^2 > \varepsilon\right) &= P\left(\sum_{1}^{m} |X_{nj} - X_j|^2 > \epsilon\right) \\
&\leq P\left(\bigcup_{j=1}^{m} \left\{|X_{nj} - X_j|^2 > \frac{\varepsilon}{m}\right\}\right) \\
&\leq \sum_{j=1}^{m} P\left(|X_{nj} - X_j|^2 > \frac{\varepsilon}{m}\right) \to 0.
\end{aligned}$$

$\Leftarrow$)

$$|X_{ni} - X_i|^2 \leq \|\mathbf{X}_n - \mathbf{X}\|^2 \Rightarrow X_{ni} - X_i = o_p(1).$$

□

PROPOSITION 5.1.8. *If* $\mathbf{X}_n - \mathbf{Y}_n \xrightarrow{p} 0$ *and* $\mathbf{Y}_n - \mathbf{Y} \xrightarrow{p} 0$, *then*

$$\mathbf{X}_n - \mathbf{Y} \xrightarrow{p} 0.$$

PROOF. By the triangle inequality:

$$\|\mathbf{X}_n - \mathbf{Y}\| \leq \|\mathbf{X}_n - \mathbf{Y}_n\| + \|\mathbf{Y}_n - \mathbf{Y}\| = o_p(1).$$

□

PROPOSITION 5.1.9 (Continuous Mapping). *If* $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ *and* $g : \mathbb{R}^m \to \mathbb{R}^s$ *is continuous, then*

$$g(\mathbf{X}_n) \xrightarrow{p} g(\mathbf{X}).$$

PROOF. We note that $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ if and only if every subsequence $\{\mathbf{X}_{n_j}\}$ has a subsequence $\left\{\mathbf{X}_{n_{j_k}}\right\}$ such that $\mathbf{X}_{n_{j_k}} \to \mathbf{X}$ a.s. as $k \to \infty$. Hence if $\{\mathbf{X}_{n_j}\}$ is any subsequence

of $\{\mathbf{X}_n\}$, then there exists a subsequence $\mathbf{X}_{n_{j_k}} \xrightarrow{\text{a.s.}} \mathbf{X}$ whence $g\left(\mathbf{X}_{n_{j_k}}\right) \xrightarrow{\text{a.s.}} g(\mathbf{X})$. But this implies that $g(\mathbf{X}_n) \xrightarrow{p} g(\mathbf{X})$. □

EXAMPLE 5.1.10. If $X_n \overset{d}{=} X$, then $X_n = O_p(1)$ and $X_n = o_p(a_n)$ for any sequence $\{a_n\}$ such that $a_n \to \infty$.

EXAMPLE 5.1.11. Suppose $X_1, X_2, \cdots$ iid. Then $X_n = O_p(1)$ and $X_n = o_p(a_n)$ if $a_n \to \infty$. Also, by WLLN and CLT,

$$\sum_1^n X_i \equiv S_n = \begin{cases} o_p(n) & \text{if } EX_1 = 0 \\ O_p(n) & \text{if } EX_1 \neq 0 \\ O_p(\sqrt{n}) & \text{if } EX_1 = 0 \text{ and } \operatorname{Var}(X_1) < \infty. \end{cases}$$

Taylor Expansions in Probability

PROPOSITION 5.1.12. *Suppose $X_n = a + O_p(r_n)$ with $r_n \to 0$ and $r_n > 0$. If g has s derivatives at a then*

$$g(X_n) = \sum_{j=0}^{s} \frac{g^{(j)}(a)}{j!} (X_n - a)^j + o_p(r_n^s).$$

PROOF. Let

$$h(x) := \begin{cases} \frac{g(x) - \sum_{j=0}^{s} \frac{g^{(j)}(a)}{j!}(x-a)^j}{\frac{(x-a)^s}{s!}} & \text{if } x \neq a \\ 0 & \text{if } x = a. \end{cases}$$

Then h is continuous since g has s derivatives at a. Since $\frac{X_n - a}{r_n} = O_p(1)$,

$$X_n - a = o_p(1).$$

Thus,

$$h(X_n) \xrightarrow{p} h(a) = 0$$

, i.e.

$$h(X_n) = o_p(1).$$

Thus,

$$h(X_n) \frac{(X_n - a)^s}{s!} = o_p(r_n^s).$$

□

EXAMPLE 5.1.13. Suppose $\{X_n\} \sim$ iid (μ, σ^2), $\mu > 0$. By Chebychev,

$$\overline{X}_n = \mu + O_p\left(n^{-\frac{1}{2}}\right).$$

Thus,

$$\log \overline{X}_n = \log \mu + \frac{1}{\mu}\left(\overline{X} - \mu\right) + o_p\left(n^{-\frac{1}{2}}\right)$$

and

$$\sqrt{n}\left(\log \overline{X}_n - \log \mu\right) = \frac{\sqrt{n}}{\mu}\left(\overline{X}_n - \mu\right) + o_p(1).$$

We end this section with a multivariate version of Proposition 5.1.12.

PROPOSITION 5.1.14. *Suppose* $\mathbf{X}_n = \mathbf{a} + O_p(r_n)$ *with* $\mathbf{a} \in \mathbb{R}^m$ *and* $r_n \to 0$. *If* $g : \mathbb{R}^m \mapsto \mathbb{R}$ *with continuous derivatives* $\partial g/\partial x_j$ *in a neighborhood of* $\mathbf{a}$, *then:*

$$g(\mathbf{X}_n) = g(\mathbf{a}) + \sum_{j=1}^{m} \frac{\partial g}{\partial x_j}(\mathbf{a})(X_{nj} - a_j) + o_p(r_n).$$

PROOF. Brockwell and Davis (1991, Proposition 6.1.6). □

5.2. Convergence in Distribution

DEFINITION 5.2.1. We say that a sequence of random vector $\mathbf{X}_n$ converges to $\mathbf{X}$ and denote

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X}$$

if $F_{\mathbf{X}_n}(\mathbf{x}) \to F_{\mathbf{X}}(\mathbf{x})$ for all $\mathbf{x} \in C = \{\mathbf{x} : F_{\mathbf{X}} \text{ is continuous at } \mathbf{x}\}$.

REMARK 5.2.2. Convergence for all $\mathbf{x}$ is too stringent a requirement as illustrated by $X_n \equiv \frac{1}{n}$. We would like to say $X_n \xrightarrow{d} X \equiv 0$ even though $F_{X_n}(0) = 0 \not\to F_X(0) = 1$.

THEOREM 5.2.3. *Suppose* $\mathbf{X}_n \sim F_n$ *and* $\mathbf{X} \sim F_0$. *Then the following are equivalent:*

(1) $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$.
(2) $\int g(\mathbf{x})\, dF_n(\mathbf{x}) \to \int g(\mathbf{x})\, dF_0(\mathbf{x})$ *for all bounded continuous function* g.
(3) $\int e^{i\mathbf{t}^T\mathbf{x}} dF_n(\mathbf{x}) \to \int e^{i\mathbf{t}^T\mathbf{x}} dF_0(\mathbf{x})$ *for all* $\mathbf{t} \in \mathbb{R}^m$
(i.e. $\phi_n(\mathbf{t}) := E\left(e^{i\mathbf{t}^T\mathbf{X}_n}\right) \to E\left(e^{i\mathbf{t}^T\mathbf{X}}\right) =: \phi_0(\mathbf{t})$ *for all* $\mathbf{t} \in \mathbb{R}^m$.*)*

PROOF. See Billingsley pp.378-383. □

Note: This theorem enables us to prove the **Cramer-Wold device**:

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \iff \boldsymbol{\lambda}^T\mathbf{X}_n \xrightarrow{d} \boldsymbol{\lambda}^T\mathbf{X}, \quad \text{for all } \boldsymbol{\lambda} \in \mathbb{R}^m.$$

PROOF. ($\Rightarrow$) Apply Theorem 5.2.3 (#2) with $g(\mathbf{x}) = e^{it\boldsymbol{\lambda}^T\mathbf{x}}$ to get

$$\phi_{\boldsymbol{\lambda}^T\mathbf{X}_n}(t) \to \phi_{\boldsymbol{\lambda}^T\mathbf{X}}(t) \implies \boldsymbol{\lambda}^T\mathbf{X}_n \xrightarrow{d} \boldsymbol{\lambda}^T\mathbf{X}.$$

($\Leftarrow$) Apply Theorem 5.2.3 (#3) to get

$$\phi_{\mathbf{X}_n}(\boldsymbol{\lambda}) = \phi_{\boldsymbol{\lambda}^T\mathbf{X}_n}(1) \to \phi_{\boldsymbol{\lambda}^T\mathbf{X}}(1) = \phi_{\mathbf{X}}(\boldsymbol{\lambda}).$$

□

PROPOSITION 5.2.4. *If* $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$, *then*

(1) $E\left|e^{i\mathbf{t}^T\mathbf{X}_n} - e^{i\mathbf{t}^T\mathbf{X}}\right| \to 0$ *for all* $\mathbf{t} \in \mathbb{R}^m$ *and*
(2) $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$.

PROOF. (1) Since

$$E\left|1 - e^{i\mathbf{t}^T(\mathbf{X}_n - \mathbf{X})}\right| \leq E\left|1 - e^{i\mathbf{t}^T(\mathbf{X}_n - \mathbf{X})}\right| I_{\|\mathbf{X}_n - \mathbf{X}\| \leq \delta} + 2P\left(\|\mathbf{X}_n - \mathbf{X}\| > \delta\right),$$

given any $\varepsilon > 0$, we can choose δ to make the first term less than $\varepsilon/2$. Then choose n to make the second term less than $\varepsilon/2$. Hence the left hand side converges to 0 as $n \to \infty$.

(2) Since $|\cdot|$ is convex, by Jensen's inequality,

$$\left|Ee^{i\mathbf{t}^T\mathbf{X}_n} - Ee^{i\mathbf{t}^T\mathbf{X}}\right| \leq E\left|e^{i\mathbf{t}^T\mathbf{X}_n} - e^{i\mathbf{t}^T\mathbf{X}}\right| \to 0.$$

Thus, by theorem (5.2.3)

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X}.$$

□

PROPOSITION 5.2.5. (Slutzky's theorem) *If* $\mathbf{X}_n - \mathbf{Y}_n = o_p(1)$ *and* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$, *then*

$$\mathbf{Y}_n \xrightarrow{d} \mathbf{X}.$$

PROOF. For a random vector $\mathbf{Z}$, define $\phi_{\mathbf{Z}}(\underline{\theta})$ by

$$\phi_{\mathbf{Z}}(\underline{\theta}) := \int e^{i\underline{\theta}^T\mathbf{z}} dF_{\mathbf{Z}}(\mathbf{z})$$

Then, we have

$$|\phi_{\mathbf{Y}_n}(\mathbf{t}) - \phi_{\mathbf{X}}(\mathbf{t})| \leq |\phi_{\mathbf{Y}_n}(\mathbf{t}) - \phi_{\mathbf{X}_n}(\mathbf{t})| + |\phi_{\mathbf{X}_n}(\mathbf{t}) - \phi_{\mathbf{X}}(\mathbf{t})|$$

By Jensen's inequality and the same argument as in proposition (5.2.4)-(1),

$$|\phi_{\mathbf{Y}_n}(\mathbf{t}) - \phi_{\mathbf{X}_n}(\mathbf{t})| \leq E\left|1 - e^{-i\mathbf{t}^T(\mathbf{X}_n - \mathbf{Y}_n)}\right| \to 0.$$

Also, since $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$, the second term converges to 0. □

PROPOSITION 5.2.6 (Continuous Mapping). *If* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *and* $h : \mathbb{R}^m \to \mathbb{R}^s$ *is continuous, then*

$$h(\mathbf{X}_n) \xrightarrow{d} h(\mathbf{X}).$$

PROOF. Since $\exp\left\{i\mathbf{t}^T h(\cdot)\right\}$ is a bounded continuous function,

$$Ee^{i\mathbf{t}^T h(\mathbf{X}_n)} \to Ee^{i\mathbf{t}^T h(\mathbf{X})}.$$

□

PROPOSITION 5.2.7. (Generalized Slutzky's theorem) *If* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *and* $\mathbf{Y}_n \xrightarrow{p} \mathbf{b}$, *with* $\dim(\mathbf{Y}_n) = m_1$ *and* $\dim(\mathbf{X}_n) = m_2$, *then*

$$\begin{bmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{bmatrix} \xrightarrow{d} \begin{pmatrix} \mathbf{X} \\ \mathbf{b} \end{pmatrix}.$$

In particular, we have the following special cases:

(1) *If* $m_1 = m_2$, *then* $\mathbf{X}_n + \mathbf{Y}_n \xrightarrow{d} \mathbf{X} + \mathbf{b}$ *and* $\mathbf{Y}_n^T \mathbf{X}_n \xrightarrow{d} \mathbf{b}^T \mathbf{X}$.
(2) *If* $g : \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \to \mathbb{R}^s$ *is continuous at every point where the vector* $(\mathbf{b}, \mathbf{X})$ *takes its values, then* $g(\mathbf{Y}_n, \mathbf{X}_n) \xrightarrow{d} g(\mathbf{b}, \mathbf{X})$.

PROOF. Let $\mathbf{Z}_n = \begin{bmatrix} \mathbf{X}_n \\ \mathbf{b} \end{bmatrix}$. Then, $\mathbf{Z}_n - \begin{bmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{bmatrix} \xrightarrow{p} 0$, since each component converges to 0 in probability. Also $\mathbf{Z}_n \xrightarrow{d} \begin{bmatrix} \mathbf{X} \\ \mathbf{b} \end{bmatrix}$ since the sequence of characteristic functions converges. Hence by Slutzky,

$$\begin{bmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \mathbf{X} \\ \mathbf{b} \end{bmatrix}.$$

Because the mappings $(\mathbf{x}, \mathbf{y}) \to \mathbf{x} + \mathbf{y}$ and $(\mathbf{x}, \mathbf{y}) \to \mathbf{x}'\mathbf{y}$ are continuous from $\mathbb{R}^{2m} \to \mathbb{R}^m$ when $m_1 = m_2 = m$, we obtain result (1) from proposition (5.2.6). Result (2) follows similarly. □

Asymptotic Normality

DEFINITION 5.2.8. A sequence of random variables $\{X_n\}$ is said to be *asymptotically normal with mean* μ_n *and standard deviation* σ_n if $\sigma_n > 0$ for all sufficiently large n and if

$$\sigma_n^{-1} (X_n - \mu_n) \xrightarrow{d} Z$$

, where $Z \sim N(0, 1)$. We write

$$X_n \text{ is } AN(\underbrace{\mu_n}_{\substack{\text{asymptotic} \\ \text{mean}}} , \underbrace{\sigma_n^2}_{\substack{\text{asymptotic} \\ \text{variance}}}).$$

EXAMPLE 5.2.9. The classical CLT states that if $X_1, \cdots, X_n$ are iid with mean μ and variance σ^2, then $\overline{X}_n$ is $AN\left(\mu, \frac{\sigma^2}{n}\right)$ i.e.

$$\frac{\sqrt{n}}{\sigma} \left(\overline{X}_n - \mu\right) \xrightarrow{d} Z.$$

PROPOSITION 5.2.10 (Delta Method). *If* X_n *is* $AN(\mu, \sigma_n^2)$ *with* $\sigma_n \to 0$ *and* g *is differentiable at* μ, *then* $g(X_n)$ *is* $AN\left(g(\mu), g'(\mu)^2 \sigma_n^2\right)$.

PROOF. Because

$$Z_n := \frac{X_n - \mu}{\sigma_n} \xrightarrow{d} Z \sim N(0,1),$$
$$Z_n = O_p(1).$$

(Choose a pair of continuity points of F_z such that $F(z_1) < \varepsilon/4$ and $F(z_2) > 1 - \varepsilon/4$. Then for all $n > N(\varepsilon)$,

$$F_{Z_n}(z_1) < \frac{\varepsilon}{3} \text{ and } F_{Z_n}(z_2) > 1 - \frac{\varepsilon}{3}.$$

For $n = 1, \cdots, N(\varepsilon)$, there exists $v(\varepsilon)$ such that

$$P(|Z_n| > v(\varepsilon)) < \varepsilon \quad n = 1, \cdots, N(\varepsilon).$$

Choose $M(\varepsilon) \geq \max(v(\varepsilon), |z_1|, |z_2|)$. Then

$$P(|Z_n| > M(\varepsilon)) < \varepsilon \text{ for all } n = 1, 2, \cdots.)$$

Thus,

$$Z_n = O_p(1) \quad \Rightarrow \quad X_n = \mu + O_p(\sigma_n).$$

By proposition (5.1.12) ($g(X) = g(\mu) + g'(\mu)(X - \mu) + o_p(\sigma_n)$),

$$\frac{g(X_n) - g(\mu)}{\sigma_n} = \frac{g'(\mu)(X_n - \mu)}{\sigma_n} + o_p(1) \xrightarrow{d} N\left(0, g'(\mu)^2\right).$$

□

EXAMPLE 5.2.11. Let $\{X_n\} \sim$iid (μ, σ^2). Then $\overline{X}_n := \frac{1}{n}(X_1 + \cdots + X_n)$ is $AN\left(\mu, \frac{\sigma^2}{n}\right)$. Suppose $\mu \neq 0$. Then $g(x) := \frac{1}{x}$ has derivative at μ so $1/\overline{X}_n$ is $AN\left(\frac{1}{\mu}, \left(-\frac{1}{\mu^2}\right)^2 \frac{\sigma^2}{n}\right)$, i.e.

$$\frac{\sqrt{n}\mu^2}{\sigma}\left(\frac{1}{\overline{X}_n} - \frac{1}{\mu}\right) \xrightarrow{d} N(0,1).$$

Moreover since $\overline{X}_n \xrightarrow{p} \mu$ (by the WLLN), proposition (5.2.10) implies

$$\frac{\sqrt{n}\overline{X}_n^2}{\sigma}\left(\frac{1}{\overline{X}_n} - \frac{1}{\mu}\right) \xrightarrow{d} N(0,1).$$

NOTE 5.2.12. Although $\frac{1}{\mu}$ is the "asymptotic mean" of $1/\overline{X}_n$ in the above example, it is *not* the limit as $n \to \infty$ of $E\left(1/\overline{X}_n\right)$. In fact, $E\left|1/\overline{X}_n\right| = \infty$ if $X_1 \sim N(\mu, \sigma^2)$.

EXAMPLE 5.2.13. Suppose $X_1, \cdots, X_n \sim$ iid $(0, \sigma^2)$. Then

$$\frac{\sqrt{n}\overline{X}_n}{\sigma} \xrightarrow{d} Z \quad \sim N(0,1) \text{ by CLT}$$
$$\Rightarrow \quad n\overline{X}_n^2 \xrightarrow{d} \sigma^2 Z^2,$$

where $Z^2 \sim \chi^2(1)$ since $g(x) = x^2$ is continuous.

Multivariate Asymptotic Normality

DEFINITION 5.2.14. $\mathbf{X}_n$ is $AN(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ if

(1) $\boldsymbol{\Sigma}_n$ has no zero diagonal elements for all large enough n.
(2) $\boldsymbol{\lambda}^T \mathbf{X}_n$ is $AN\left(\boldsymbol{\lambda}^T \boldsymbol{\mu}_n, \boldsymbol{\lambda}^T \boldsymbol{\Sigma}_n \boldsymbol{\lambda}\right)$ for all $\boldsymbol{\lambda} \in \mathbb{R}^m$ such that $\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_n \boldsymbol{\lambda} > 0$ for all large enough n.

Recalling the Cramer-Wold device, i.e.

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \iff \boldsymbol{\lambda}^T \mathbf{X}_n \xrightarrow{d} \boldsymbol{\lambda}^T \mathbf{X} \text{ for all } \boldsymbol{\lambda} \in \mathbb{R}^m,$$

we see that $\mathbf{X}_n$ is $AN(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$where $\boldsymbol{\Sigma}_n$ satisfies (1) if and only if

$$\frac{\boldsymbol{\lambda}^T (\mathbf{X}_n - \boldsymbol{\mu}_n)}{\sqrt{\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_n \boldsymbol{\lambda}}} \xrightarrow{d} N(0,1),$$

for all $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_n \boldsymbol{\lambda} > 0$ for large enough n.

PROPOSITION 5.2.15 (Multivariate Delta Method). *Suppose* $\mathbf{X}_n$ *is* $AN(\boldsymbol{\mu}, c_n^2 \boldsymbol{\Sigma})$ *with* $c_n \to 0$. *If* $g : \mathbb{R}^m \to \mathbb{R}^k$ *is continuously differentiable in a neighborhood of* $\boldsymbol{\mu}$ *and* $\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T$ *has all diagonal elements greater than* 0 *where* $\mathbf{D} = [\partial g_i / \partial x_j]_{\mathbf{X}=\boldsymbol{\mu}}$, *then* $g(\mathbf{X}_n)$ *is* $AN\left(g(\boldsymbol{\mu}), c_n^2 \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T\right)$.

DEFINITION 5.2.16. A sequence of estimates $T_n = T_n(X_1, \cdots, X_n)$ of $g(\theta)$ is said to be *(weakly) consistent* if

$$T_n \xrightarrow{p} g(\theta) \text{ for all } \theta \text{ (where } p \text{ is under the measure } P_\theta),$$

and *strongly consistent* if

$$T_n \to g(\theta) \text{ a.s. } P_\theta \text{ for all } \theta.$$

EXAMPLE 5.2.17. (Moment Estimation) Let $\{X_N\} \sim$iid P_θ such that $E_\theta |X|^r < \infty$ for all θ. Suppose $m_j(\theta) = E_\theta X_1^j$ for $1 \le j \le r$ and $g(\theta) = \phi(m_1(\theta), \cdots, m_r(\theta))$ where ϕ is continuous. Then

$$T_n(X_1, \cdots, X_n) = \phi(\hat{m}_1, \cdots, \hat{m}_r) \to g(\theta) \text{ a.s. } P_\theta,$$

where

$$\hat{m}_j = \frac{1}{n} \sum_{k=1}^{n} X_k^j.$$

DEFINITION 5.2.18. A sequence of estimators is said to be *asymptotically normal* if there exists $\mu_n(\theta)$ and $\sigma_n(\theta) > 0$ such that T_n is $AN(\mu_n(\theta), \sigma_n^2(\theta))$ for all θ. i.e.

$$P_\theta\left(\frac{T_n - \mu_n(\theta)}{\sigma_n(\theta)} \le x\right) \to \Phi(x) \text{ for all } \theta.$$

REMARK 5.2.19. Suppose T_n is $AN(\mu_n'(\theta), \sigma_n'^2(\theta))$. If

$$\frac{\sigma_n(\theta)}{\sigma_n'(\theta)} \to 1 \text{ and } \frac{\mu_n(\theta) - \mu_n'}{\sigma_n} \to 0,$$

then T_n is $AN\left(\mu_n(\theta), \sigma_n^2(\theta)\right)$ since

$$\underset{\substack{d\downarrow \\ N(0,1)}}{\frac{T_n - \mu_n}{\sigma_n}} = \underset{\substack{\downarrow \\ 1}}{\frac{\sigma_n'}{\sigma_n}} \cdot \underset{\substack{d\downarrow \\ N(0,1)}}{\frac{T_n - \mu_n'}{\sigma_n'}} + \underset{\substack{d\downarrow \\ 0}}{\frac{\mu_n' - \mu_n}{\sigma_n}}$$

DEFINITION 5.2.20. A sequence of asymptotically normal estimators $\{T_n\}$ is said to be *asymptotically unbiased* for $g(\theta)$ if

$$\frac{\mu_n(\theta) - g(\theta)}{\sigma_n(\theta)} \to 0,$$

in which case

$$\frac{T_n - g(\theta)}{\sigma_n(\theta)} \xrightarrow{d} N(0,1)$$

and $\frac{\mu_n(\theta) - g(\theta)}{\sigma_n(\theta)}$ is called the *(standardized) asymptotic bias.*

It follows that if both the bias and variance of T_n go to zero, then T_n is consistent for its mean.

PROPOSITION 5.2.21. *If $ET_n \to \mu$ and $Var(T_n) \to 0$, then $T_n \xrightarrow{p} \mu$.*

PROOF. Chebychev's Inequality and Triangle Inequality. □

THEOREM 5.2.22 (Multivariate CLT). *If $\{\mathbf{X}_n\} \sim$ iid $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then*

$$\overline{\boldsymbol{X}}_n \sim AN\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right), \qquad \text{whence } \overline{\boldsymbol{X}}_n = \boldsymbol{\mu} + O_p(1/\sqrt{n}).$$

EXAMPLE 5.2.23. Let $\{X_n\} \sim$ iid P_θ such that $E_\theta |X|^{2r} < \infty$ for all θ. Suppose $m_j(\theta) = E_\theta X_1^j$ for $1 \le j \le r$, $g(\theta) = \phi(m_1(\theta), \cdots, m_r(\theta))$ for some continuously differentiable function ϕ, and define $T_n = \phi(\hat{m}_1, \cdots, \hat{m}_r)$, where $\hat{m}_j := \frac{1}{n}\sum_{j=1}^n X_i^j$. From the CLT,

$$\sqrt{n}\begin{bmatrix} \hat{m}_1 - m_1(\theta) \\ \vdots \\ \hat{m}_r - m_r(\theta) \end{bmatrix} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}), \qquad \text{where } \boldsymbol{\Sigma} = (\Sigma_{ij}) = \left(\text{Cov}\left(X_1^i, X_1^j\right)\right)_{i,j=1}^r.$$

Then, using Proposition 5.1.14, we have

$$T_n = \phi(m_1(\theta), \cdots, m_r(\theta)) + \sum_j \frac{\partial \phi}{\partial x_j}(\mathbf{m}(\theta))(\hat{m}_j - m_j(\theta)) + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Finally, applying Proposition 5.2.15 gives:

$$\frac{T_n - \phi(m_1(\theta), \cdots, m_r(\theta))}{\sigma_n(\theta)} \xrightarrow{d} N(0,1),$$

where (assuming $\boldsymbol{\Sigma} > 0$)

$$\sigma_n^2(\theta) = \frac{1}{n}\left(\frac{\partial \phi}{\partial m_1}, \cdots, \frac{\partial \phi}{\partial m_r}\right)\Sigma\begin{pmatrix} \partial\phi/\partial m_1 \\ \vdots \\ \partial\phi/\partial m_r \end{pmatrix}.$$

EXAMPLE 5.2.24. Let $X_1, \cdots, X_n \sim \Gamma(\alpha, \beta)$. Then $EX_1 = \alpha\beta$, $EX_1^2 = \alpha\beta^2 + \alpha^2\beta^2 = \beta^2(\alpha^2 + \alpha)$, and $\mathrm{Var}X_1 = \alpha\beta^2$. $\phi(m_1(\theta), m_2(\theta)) = \beta = \frac{m_2(\theta) - m_1^2(\theta)}{m_1(\theta)}$. $T_n = \frac{\frac{1}{n}\sum X_i^2 - \bar{X}^2}{\bar{X}} = \frac{\hat{m}_2 - \hat{m}_1^2}{\hat{m}_1}$. Then T_n is $AN(\beta, \sigma_n^2(\theta))$,where

$$\sigma_n^2(\theta) = \frac{1}{n}\left(\frac{-2m_1^2 - (m_2 - m_1^2)}{m_1^2}, \frac{1}{m_1}\right)\sum\begin{pmatrix} -1 - \frac{m_2}{m_1^2} \\ \frac{1}{m_1} \end{pmatrix}$$

and

$$\begin{aligned}\Sigma &= \begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_1^2) \\ Cov(X_1, X_1^2) & Cov(X_1^2, X_2^2) \end{bmatrix} \\ &= \begin{bmatrix} \alpha\beta^2 & \alpha(\alpha+1)(\alpha+2)\beta^3 - \alpha\beta\alpha\beta^2 \\ \alpha(\alpha+1)(\alpha+2)\beta^3 - \alpha\beta\alpha\beta^2 & \alpha(\alpha+1)(\alpha+2)(\alpha+3)\beta^4 \end{bmatrix}\end{aligned}$$

5.3. Asymptotic Comparisons (Pitman Efficiency)

DEFINITION 5.3.1. If $\sqrt{n}(T_n - g(\theta)) \xrightarrow{d} N(0, \sigma_0^2(\theta))$ and $\sqrt{n}\left(T'_{n'(n)}(X_1, \ldots, X_{n'}) - g(\theta)\right) \xrightarrow{d} N(0, \sigma_0^2(\theta))$, then the *Pitman asymptotic relative efficiency (ARE)* of $\{T_n\}$ relative to $\{T'_n\}$ is

$$e_{T,\,T'}(\theta) = \lim_{n\to\infty} \frac{n'(n)}{n}$$

provided the limit exists and is independent of the sequence $n'(n)$ chosen to satisfy

$$\sqrt{n}\left(T'_{n'(n)} - g(\theta)\right) \xrightarrow{d} N(0, \sigma_0^2(\theta))$$

Roughly speaking, $e_{T,\,T'}$ is the ratio of the number of observations required for the two estimators to achieve the same precision.

THEOREM 5.3.2. *if* $T_n \sim AN\left(g(\theta), \frac{\sigma_0^2(\theta)}{n}\right)$ *and* $T'_n \sim AN\left(g(\theta), \frac{\sigma_1^2(\theta)}{n}\right)$, *then* $e_{T,T'}(\theta) = \frac{\sigma_1^2(\theta)}{\sigma_0^2(\theta)}$.

PROOF. Let $n' = \left[n\frac{\sigma_1^2(\theta)}{\sigma_0^2(\theta)}\right]$, where $[x]$ is the integer part of x. Then

$$\begin{aligned}\sqrt{n'}(T'_{n'} - g(\theta)) &= \sqrt{\left[n\frac{\sigma_1^2(\theta)}{\sigma_0^2(\theta)}\right]}(T'_{n'} - g(\theta)) \\ &= \frac{\sigma_1(\theta)}{\sigma_0(\theta)}\sqrt{n}(T'_{n'} - g(\theta)) + O_p\left(\frac{1}{n}\right)\end{aligned}$$

and since LHS $\xrightarrow{d} N(0, \sigma_1^2(\theta))$, therefore

$$\sqrt{n}(T'_{n'} - g(\theta)) \to N\left(0, \sigma_0^2(\theta)\right)$$

Hence

$$e_{T,\,T'}\,(\theta) = \lim_{n\to\infty} \frac{n'(n)}{n} = \frac{\sigma_1^2(\theta)}{\sigma_0^2(\theta)}$$

provided the limit is the same for all $n'(n)$ such that $\lim n'(n)/n$ exists and

$$\sqrt{n}\,(T'_{n'} - g(\theta)) \to N\left(0, \sigma_0^2(\theta)\right)$$

Suppose $n'(n)$ is any such sequence, then the LHS of the previous line is equal to

$$\frac{\sqrt{n}}{\sqrt{n'}}\sqrt{n'}\,(T'_{n'} - g(\theta)) = \sqrt{\frac{n}{n'}}\frac{\sigma_1}{\sigma_0}\Bigg(\frac{\sigma_0}{\sigma_1}\sqrt{n'}\,(T'_{n'} - g(\theta))\Bigg)$$

$$\Bigg(\overset{d\downarrow}{N\left(0,\,\sigma_0^2(\theta)\right)}\Bigg)$$

therefore $\frac{n}{n'}\frac{\sigma_1^2}{\sigma_0^2} \to 1$, i.e. $\frac{n'}{n} \to \frac{\sigma_1^2}{\sigma_0^2}$. □

5.4. M-Estimation Theory

M-estimation is a very general method to derive consistency and asymptotic normality results for a lot of classical estimators that are obtained by solving a system of equation(s). This includes e.g., method of moments estimators and maximum likelihood estimators. The main idea is to write the system as an empirical average, to which the CLT is applied after a Taylor series expansion. See Serfling (1980, Ch 7) for a classical treatment, and Van der Vaart (1998, Ch 5) for a more recent coverage.

Background

- $X_1, \ldots, X_n \sim iid\, F_\theta(x)$, $f_\theta(x)$, with $\theta = (\theta_1, \ldots, \theta_d) \in \Omega \subseteq \mathbb{R}^d$, let θ_0 denote its true value, and $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ its estimator.
- Define $\hat{\theta}_n$ as an *M-estimator*:

  $$\hat{\theta}_n = \arg\max_{\theta\in\Omega} M_n(\theta), \qquad M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(x_i, \theta),$$

  where $m(x, \theta)$ is a scalar concave function of θ (which ensures the maximum exists, but can be relaxed).
- Often $\hat{\theta}_n$ is found as a root of the equation that results from differentiating $M_n(\theta)$:

  $$\boldsymbol{\Psi}_n(\theta) = \frac{\partial}{\partial\theta} M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(x_i, \theta) = 0,$$

  where $\boldsymbol{\psi}(x, \theta)$ is the vector-valued map

  $$\boldsymbol{\psi}(x, \theta) = \begin{bmatrix} \psi_1(x, \theta) \\ \vdots \\ \psi_d(x, \theta) \end{bmatrix} = \frac{\partial m(x, \theta)}{\partial\theta} := \dot{m}(x, \theta).$$

- In both cases, the true value θ_0 satisfies:

$$\theta_0 = \arg\max_{\theta\in\Omega} E_{\theta_0} m(X,\theta), \qquad \text{and} \qquad E_{\theta_0}\boldsymbol{\psi}(X,\theta) = 0.$$

(Note: finding the appropriate $m(.)$ and/or $\boldsymbol{\psi}(.)$ functions may be challenging.)

EXAMPLE 5.4.1 (Estimation of a location parameter). For estimating a measure of location like the mean (μ), median ($\xi_{0.5}$), and a quantile in general (ξ_p), proceed as follows:

- For μ use $m(x,\theta) = -(x-\theta)^2$ and $\psi(x,\theta) = (x-\theta)$. Then note that $\hat{\mu} = \bar{x}$ solves:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \theta) = 0.$$

- For $\xi_{0.5}$ use $m(x,\theta) = -|x-\theta|$ and $\psi(x,\theta) = \mathbf{1}(x > \theta) - \mathbf{1}(x < \theta)$. Then note that $\hat{\xi}_{0.5}$ solves:

$$\frac{1}{n}\sum_{i=1}^{n}[\mathbf{1}(x_i > \theta) - \mathbf{1}(x_i < \theta)] = 0.$$

EXAMPLE 5.4.2 (Maximum likelihood estimation). Using $m(x,\theta) = \log f_\theta(x)$ leads to the **log-likelihood** and **score** functions:

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log f(x_i,\theta), \qquad \boldsymbol{\Psi}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i,\theta) = 0.$$

Consistency. Consistency of $\hat{\theta}_n$ follows if the functions $\{m(x,\theta) : \theta \in \Omega\}$ or $\{\psi_j(x,\theta) : \theta \in \Omega,\ j = 1,\ldots,d\}$ are *Glivenko-Cantelli* (Van der Vaart, 1998, Ch 19). A simple set of sufficient conditions for this is:

(i) Ω is a compact set; and
(ii) the maps $\theta \mapsto m(x,\theta)$ or $\theta \mapsto \psi_j(x,\theta)$ are continuous $\forall\, x$ and $\forall\, j$, and are dominated by an integrable function $H(x)$ in the vicinity of θ_0; i.e.,

$$m(x,\theta) \le H(x), \qquad \psi_j(x,\theta) \le H(x), \qquad \text{with } EH(X) < \infty.$$

(Note: $H(x)$ can depend on θ_0 but not θ.)

Asymptotic Normality (AN). Under mild regularity assumptions $\hat{\theta}_n$ is AN:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, V_{\theta_0}^{-1} W_{\theta_0} V_{\theta_0}^{-T}\right), \tag{5.4.1}$$

where the $d\times d$ constituent matrices in the asymptotic covariance can (under appropriate conditions) be expressed either in terms of derivatives of the $m(.)$ or $\boldsymbol{\psi}(.)$ functions evaluated at θ_0, as follows:

$$W_{\theta_0} = E_{\theta_0}\boldsymbol{\psi}(X,\theta)\boldsymbol{\psi}(X,\theta)^T = E_{\theta_0}\left(\frac{\partial m(X,\theta)}{\partial\theta}\right)\left(\frac{\partial m(X,\theta)}{\partial\theta}\right)^T,$$

and

$$V_{\theta_0} = E_{\theta_0}\left(\frac{\partial \boldsymbol{\psi}(X,\theta)}{\partial \theta^T}\right) = \left(\frac{\partial^2 E_{\theta_0} m(X,\theta)}{\partial \theta \partial \theta^T}\right).$$

The regularity conditions needed here are not easy to state; most importantly they require continuity (in probability) of the map $\theta \mapsto \boldsymbol{\psi}(\cdot,\theta)$, that V_{θ_0} be nonsingular, and that the collection of maps $x \mapsto \boldsymbol{\psi}(x,\theta)$ form a *Donsker Class* (Van der Vaart, 1998, Ch 19).

Note: If changing the order of differentiation (w.r.t. θ) and integration (w.r.t. x) are permissible, then the two definitions of V_{θ_0} above coincide. (This fails e.g., in the case of quantiles, see below.)

PROOF. A sketch of the AN proof in the $d = 1$ case is as follows. Taylor expand $\psi(x,\theta)$ about θ_0:

$$\sum \psi(x_i,\theta) = \sum \psi(x_i,\theta_0) + (\theta - \theta_0)\sum \dot{\psi}(x_i,\theta_0) + \text{remainder}.$$

Now substitute $\theta = \hat{\theta}_n$, whence the LHS is zero since $\hat{\theta}_n$ solves $\sum \psi(x_i,\hat{\theta}_n) = 0$, and since it can be shown that the remainder is $o_p(1)$, we obtain, after rearranging terms and multiplying by $\sqrt{n}$:

$$\sqrt{n}(\theta - \theta_0) = \frac{-\frac{1}{\sqrt{n}}\sum \psi(x_i,\theta_0)}{\frac{1}{n}\sum \dot{\psi}(x_i,\theta_0)} := \frac{A}{B}.$$

Now analyze the numerator and denominator terms separately. For the first we obtain that

$$A = \sqrt{n}\left(-\frac{1}{n}\sum \psi(X_i,\theta_0)\right) = \sqrt{n}\left(\bar{Y}_n - 0\right),$$

where $Y_i = -\psi(X_i,\theta_0) \sim$ iid (μ_Y, σ_Y^2) with $\mu_Y = -E_{\theta_0}\psi(X,\theta_0) = 0$ and $\sigma_Y^2 = Var_{\theta_0}\psi(X,\theta_0) = E_{\theta_0}\psi^2(X,\theta_0)$. Thus by applying the CLT we obtain: $A \xrightarrow{d} N(0,\sigma_Y^2)$. For the denominator, apply the WLLN to $Z_i = \dot{\psi}(X_i,\theta_0) \sim$ iid (μ_Z, σ_Z^2), where $\mu_Z = E_{\theta_0}\dot{\psi}(X,\theta_0)$, to see that: $B \xrightarrow{p} \mu_Z$. Putting it all together using Slutzky gives:

$$\sqrt{n}(\theta - \theta_0) \xrightarrow{d} \frac{N(0,\sigma_Y^2)}{\mu_Z} \sim N\left(0, \frac{\sigma_Y^2}{\mu_Z^2} = \frac{E_{\theta_0}\psi^2(X,\theta_0)}{[E_{\theta_0}\dot{\psi}(X_i,\theta_0)]^2}\right).$$

□

EXAMPLE 5.4.3 (Estimation of mean). For $X_1,\ldots,X_n \sim$ iid (μ_0,σ_0^2), use $\psi(x,\theta) = (x-\mu)$ for estimation of μ_0 via the sample mean $\bar{X}$, implying $\dot{\psi}(x,\mu) = -1$, so that $E_{\mu_0}\dot{\psi}(X,\mu) = -1$. Now, $E_{\mu_0}\psi^2(X,\mu) = E_{\mu_0}(X-\mu_0)^2 = \sigma_0^2$, and thus we obtain the classical CLT result:

$$\sqrt{n}(\bar{X} - \mu_0) \xrightarrow{d} N\left(0, \frac{E_{\mu_0}\psi^2(X,\mu_0)}{[E_{\mu_0}\dot{\psi}(X,\mu_0)]^2} = \sigma_0^2\right).$$

5.5. Example: AREs of Mean, Median, Trimmed Mean

PROPOSITION 5.5.1 (AN for a Quantile). *Let $F(x)$ be a cdf such that $F(x)$ is differentiable at $\xi_p = F^{-1}(p)$, the p-quantile of F, and that $f(\xi_p) = \frac{dF(x)}{dx}\big|_{x=\xi_p} > 0$. Let $X_1, \ldots, X_n$ be iid F and let $Y_1 \leq \cdots \leq Y_n$ be the order statistics. Then, if $[np]$ denotes any of the integers on either side of np, we have for any $0 < p < 1$:*

$$Y_{[np]} \sim AN\left(\xi_p, \frac{p(1-p)}{nf^2(\xi_p)}\right).$$

PROOF. The p-quantile $\theta_0 = \xi_p$ is an M-estimator with $m(x,\theta) = (1-p)(x-\theta)\mathbf{1}(x < \theta) - p(x-\theta)\mathbf{1}(x > \theta)$. To see why, consider:

$$E_\theta m(X,\theta) = (1-p)\int_{-\infty}^{\theta}(x-\theta)dF(x) - p\int_{\theta}^{\infty}(x-\theta)dF(x) := g(\theta).$$

Differentiating using Leibnitz's Rule:

$$\left[\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} h(x,\theta)dx = h(b,\theta)b'(\theta) - h(a,\theta)a'(\theta) + \int_{a(\theta)}^{b(\theta)}\frac{\partial}{\partial\theta}h(x,\theta)dx\right]$$

gives,

$$g'(\theta) = -(1-p)[F(\theta) - 0] + [1 - F(\theta)]p = p - F(\theta), \qquad g''(\theta) = -f(\theta) = V_\theta,$$

whence we note that $\theta_0 = \xi_p = \arg\max E_{\theta_0} m(X,\theta)$. To obtain W_θ:

$$\begin{aligned}
\frac{\partial m(x,\theta)}{\partial\theta} = \psi(x,\theta) &= p\mathbf{1}(x > \theta) - (1-p)\mathbf{1}(x < \theta) \\
\psi^2(x,\theta) &= p^2\mathbf{1}(x > \theta) + (1-p)^2\mathbf{1}(x < \theta) \\
W_{\theta_0} = E_{\theta_0}\psi^2(X,\theta) &= p^2\int_{\theta_0}^{\infty} dF(x) + (1-p)^2\int_{-\infty}^{\theta_0} dF(x) \\
&= p^2(1-p) + (1-p)^2 p = p(1-p).
\end{aligned}$$

Thus invoking the result in (5.4.1):

$$V_{\theta_0}^{-1} W_{\theta_0} V_{\theta_0}^{-T} = \frac{E_{\theta_0}\psi^2(X,\theta)}{[g''(\theta_0)]^2} = \frac{p(1-p)}{f^2(\xi_p)}.$$

□

PROPOSITION 5.5.2. *Suppose $F(x)$ is continuous at the quantiles ξ_{p_1} and ξ_{p_2}, and that $f(\xi_{p_1}) > 0$ and $f(\xi_{p_2}) > 0$. Then, for $0 < p_1 < p_2 < 1$, we have that*

$$\begin{bmatrix} Y_{[np_1]} \\ Y_{[np_2]} \end{bmatrix} \sim AN\left(\begin{bmatrix} \xi_{p_1} \\ \xi_{p_2} \end{bmatrix}, \frac{1}{n}\begin{bmatrix} p_1(1-p_1)/f^2(\xi_{p_1}) & p_1(1-p_2)/(f(\xi_{p_1})f(\xi_{p_2})) \\ p_1(1-p_2)/(f(\xi_{p_1})f(\xi_{p_2})) & p_2(1-p_2)/f^2(\xi_{p_2}) \end{bmatrix}\right).$$

PROOF. This is a $d = 2$ version of the previous proposition, $\theta = (\theta_1, \theta_2)$, where $\theta_0 = (\xi_{p_1}, \xi_{p_2})$, and we take

$$m(x, \theta) = \sum_{j=1}^{2} \left[(1 - p_j)(x - \theta_j)\mathbf{1}(x < \theta_j) - p_j(x - \theta_j)\mathbf{1}(x > \theta_j) \right].$$

Then, $E_\theta m(X, \theta) = g_1(\theta_1) + g_2(\theta_2) := g(\theta)$, with

$$g_j(\theta) = (1 - p_j) \int_{-\infty}^{\theta_j} (x - \theta_j) dF(x) - p_j \int_{\theta_j}^{\infty} (x - \theta_j) dF(x),$$

which implies

$$V_{\theta_0} = \left. \frac{\partial^2 g(\theta)}{\partial \theta \partial \theta^T} \right|_{\theta = \theta_0} = \begin{bmatrix} -f(\xi_{p_1}) & 0 \\ 0 & -f(\xi_{p_2}) \end{bmatrix}.$$

It's now easy to see that θ_0 is a root of the Jacobian of $g(\theta)$, and that its Hessian V_{θ_0} is negative definite (hence nonsingular), whence $E_{\theta_0} m(X, \theta)$ is maximized at θ_0. The elements of W_{θ_0} follow similarly to the $d = 1$ case:

$$\frac{\partial m(x, \theta)}{\partial \theta} = \begin{bmatrix} p_1 \mathbf{1}(x > \theta_1) - (1 - p_1)\mathbf{1}(x < \theta_1) \\ p_2 \mathbf{1}(x > \theta_2) - (1 - p_2)\mathbf{1}(x < \theta_2) \end{bmatrix} = \begin{bmatrix} \psi_1(x, \theta_1) \\ \psi_2(x, \theta_2) \end{bmatrix}.$$

so that, upon noting that $E_{\theta_0} \psi_j^2(X, \theta_j) = p_j(1 - p_j)$ and $E_{\theta_0} \psi_1(X, \theta_1)\psi_2(X, \theta_2) = p_1(1 - p_2)$, leads to

$$W_{\theta_0} = E_{\theta_0} \begin{bmatrix} \psi_1^2(X, \theta_1) & \psi_1(X, \theta_1)\psi_2(X, \theta_2) \\ \psi_1(X, \theta_1)\psi_2(X, \theta_2) & \psi_2^2(X, \theta_2) \end{bmatrix} = \begin{bmatrix} p_1(1 - p_1) & p_1(1 - p_2) \\ p_1(1 - p_2) & p_2(1 - p_2) \end{bmatrix}.$$

Computing $V_{\theta_0}^{-1} W_{\theta_0} V_{\theta_0}^{-T}$ then leads to the stated asymptotic covariance matrix. □

PROPOSITION 5.5.3 (AN for Trimmed Mean). *Let $F(x)$ be symmetric about 0 and suppose $\exists 0 < c \leq \infty$ such that $F(-c) = 0$, $F(c) = 1$, and that $f(x)$ is strictly positive and continuous on $(-c, c)$. If $X_1, \ldots, X_n$ are iid $F(x - \theta)$, then for any $0 < \alpha < 1/2$*

$$\bar{X}_\alpha \sim AN\left(0, \frac{\sigma_\alpha^2}{n}\right),$$

where

$$\bar{X}_\alpha = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} Y_i, \qquad \sigma_\alpha^2 = \frac{2}{(1 - 2\alpha)^2} \left[\int_0^{\xi_{1-\alpha}} t^2 f(t) dt + \alpha \xi_{1-\alpha}^2 \right],$$

$Y_1 \leq \cdots \leq Y_n$ are the order statistics, and $\xi_\alpha = F^{-1}(\alpha)$.

PROOF. Omitted, but can be proved similarly via M-estimation. □

These results can now be used to effect asymptotic relative efficiency calculations.

REMARK 5.5.4. Suppose $X_1, \ldots, X_n \sim iid\, F(x-\theta)$, $F(0) = 1/2$, and $f(0) > 0$. Then

$$\tilde{X} := \text{median}(X_1, \cdots, X_n) = X_{([n/2])} \sim AN\left(\theta, \frac{1}{4nf^2(0)}\right).$$

If $EX_1 = \theta$ and $\sigma^2 = Var_\theta X_1$, then from Example 5.4.3, $\bar{X} \sim AN(\theta, \sigma^2/n)$. Therefore, $e_{\tilde{X},\,\bar{X}} = 4\sigma^2 f^2(0)$. So if $2\sigma f(0) < 1$ then $\bar{X}$ is more efficient, whereas if $2\sigma f(0) > 1$ then $\tilde{X}$ is more efficient.

REMARK 5.5.5. Again assume $X_1, \ldots, X_n \sim iid\, F(x-\theta)$, $F(-c) = 0$, and f is symmetric continuous and positive. Then, since $\bar{X}_\alpha \to \tilde{X}$ as $\alpha \uparrow 1/2$ and $\bar{X}_\alpha \to \bar{X}$ as $\alpha \downarrow 0$:

$$\lim_{\alpha \uparrow \frac{1}{2}} \sigma_\alpha^2 = \frac{1}{4f^2(0)}, \qquad \text{and} \qquad \lim_{\alpha \downarrow 0} \sigma_\alpha^2 = \sigma^2,$$

so that the ARE's of $\tilde{X}$ and $\bar{X}_\alpha$ relative to $\bar{X}$ are:

$$e_{\tilde{X},\,\bar{X}}(f) = 4\sigma^2 f^2(0), \qquad \text{and} \qquad e_{\bar{X}_\alpha,\,\bar{X}}(f) = \sigma^2/\sigma_\alpha^2.$$

Defining the following mixture of two normals

$$T(\epsilon, \tau) = (1-\epsilon)\frac{1}{\sqrt{2\pi}} e^{-x^2/2} + \epsilon \frac{1}{\tau\sqrt{2\pi}} e^{-x^2/2\tau},$$

some numerical computations for different f's and α's then lead to the following AREs of $\bar{X}_\alpha$ relative to $\bar{X}$:

$f(x)\backslash\alpha$	.125	.25	.5
$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$	.94	.84	$\frac{2}{\pi}$=.64
$\frac{1}{\pi(1+x^2)}$	∞	∞	∞
$\frac{1}{2} e^{-\lvert\lambda\rvert/2}$	1.40	1.63	2
$T(.01, 3)$	.98	.89	.68
$T(.05, 3)$	1.19	1.09	.83
t_3	1.91	1.97	1.62
t_5	1.24	1.21	.96

REMARK 5.5.6. $\bar{X}$ is inefficient for heavy tails. This is because it is sensitive to one or two extreme observations.

REMARK 5.5.7. The optimal α depends on the distribution sampled. For large n the distribution can be estimated and α chosen accordingly as $\hat{\alpha}$.

CHAPTER 6

Maximum Likelihood Estimation

Assessing the performance of different types of estimators (UMVUE, MRE, Bayes, MLE) is usually difficult in finite samples. This task is made considerably simpler as $n \to \infty$, and especially for the Maximum Likelihood Estimator (MLE). This chapter discusses the various notions of asymptotic assessment and optimality, and establishes the relevant classical results for the MLE.

Efron & Hastie (2016) poignantly summarize the success of maximum likelihood estimation in the following quote:

> If Fisher had lived in the era of "apps", maximum likelihood estimation might have made him a billionarie. Arguably the 20th century's most influential piece of applied mathematics, maximum likelihood continues to be a prime method of choice in the statistician's toolkit. Roughly speaking, maximum likelihood provides nearly unbiased estimates of nearly minimum variance, and does so in an automatic way.

6.1. Consistency

Suppose that $X_1, X_2, \ldots$ are iid P_θ, $\theta = (\theta_1, \ldots, \theta_d) \in \Omega \subseteq \mathbb{R}^d$, and **make the following assumptions.**

- (A_0) $P_\theta \neq P_{\theta'}$ if $\theta \neq \theta'$.
- (A_1) $P_\theta,\ \theta \in \Omega$, have common support.
- (A_2) $\frac{d}{d\mu} P_\theta(x) = f(x, \theta)$.
- (A_3) The true parameter $\theta_0 \in int(\Omega)$.

DEFINITION 6.1.1. $L(\boldsymbol{x}, \theta) = \prod_1^n f(x_i, \theta)$ is called the **likelihood**, and $\ell(\boldsymbol{x}, \theta) = \sum_1^n \log f(x_i, \theta)$ is the **log likelihood**. An estimator $\hat{\theta}$ of θ is called a (global) **MLE** if

$$\ell(\boldsymbol{x}, \hat{\theta}(\boldsymbol{x})) = \sup_{\theta \in \Omega} \ell(\boldsymbol{x}, \theta).$$

The **MLE of** $g(\theta)$ is defined to be $g(\hat{\theta})$.

Likelihood equations. If $\theta \in \Omega$ and Ω is an open subsect of $\mathbb{R}^d$ and ℓ is differentiable on Ω, then $\hat{\theta}$ (if it exists) satisfies

$$\frac{\partial \ell}{\partial \theta_j}\left(\boldsymbol{x}, \hat{\theta}(\boldsymbol{x})\right) = 0, \qquad 1 \le j \le d.$$

In general

$$\bigtriangledown_\theta \ell(\boldsymbol{x}, \theta) = \mathbf{0}$$

may not have a unique solution. Sometimes the likelihood is unbounded and the MLE does not exist.

EXAMPLE 6.1.2.

$$\begin{aligned}
F(x,\theta) &= \begin{cases} 1 - e^{-ax} & 0 \le x < \tau \\ 1 - e^{-a\tau - b(x-\tau)} & x \ge \tau \end{cases} \\
\theta \in \Omega &= \left\{(a,b,\tau) \in (0,\infty)^3\right\} \\
f(x,\theta) &= ae^{-ax} I_{[0,\tau)}(x) + be^{-a\tau - b(x-\tau)} I_{[\tau,\infty)}(x) \\
\text{Hazard rate} &= \frac{f(x,\theta)}{1 - F(x,\theta)} = \begin{cases} a & x < \tau, \\ b & x \ge \tau. \end{cases} \\
\ell(\boldsymbol{x}, a, b, \tau) &= \sum_1^k \left[-ax_{(i)} + \log a\right] + \sum_{k+1}^n \left[-a\tau - b(x_{(i)} - \tau) + \log b\right],
\end{aligned}$$

where $x_{(i)}$ is the ith order statistic, and $k = \#\{x_i \colon x_i < \tau\}$.

For *any* fixed a, letting $b = \frac{1}{x_{(n)} - \tau}$ and $\tau \uparrow x_{(n)}$, we see that $\ell(\boldsymbol{x}, a, \frac{1}{x_{(n)}-\tau}, \tau) \to \infty$ as $\tau \uparrow x_{(n)}$, and hence a global MLE does not exist. However if we restrict $\tau < x_{(n-1)}$ the constrained MLE exists and is consistent.

LEMMA 6.1.3. *Let*

$$\begin{aligned}
I(\theta_j|\theta_i) &= -E_{\theta_i} \log \frac{f(X,\theta_j)}{f(X,\theta_i)} \\
&= -\int \log \frac{f(x,\theta_j)}{f(x,\theta_i)} f(x,\theta_i) d\mu(x) \\
&= \textbf{Kullback-Leibler discrepancy of } f(\cdot,\theta_j) \textbf{ relative to } f(\cdot,\theta_i)
\end{aligned}$$

Then $I(\theta_j|\theta_i) \ge 0$ with equality holding if and only if $\theta_j = \theta_i$.

PROOF. Jensen's inequality gives

$$\begin{aligned}
-E_{\theta_i} \log \frac{f(X,\theta_j)}{f(X,\theta_i)} &\ge -\log E_{\theta_i} \frac{f(X,\theta_j)}{f(X,\theta_i)} \\
&\ge -\log \int_{\{x : f(x,\theta_i) > 0\}} f(x,\theta_j) d\mu(x) \\
&\ge -\log 1 = 0
\end{aligned}$$

Equality holds if and only if $\frac{f(X,\theta_j)}{f(X,\theta_i)}$ is constant *a.s.* P_{θ_i}, and $\int_{\{x:f(x,\theta_i)>0\}} f(x,\theta_j)d\mu(x) = 1$. The latter equality implies $P_{\theta_j} \ll P_{\theta_i}$ and the former equality implies $P_{\theta_j} = P_{\theta_i}$, since $\int cf(x,\theta_i)d\mu = 1$ implies that $c = 1$. □

Note: All theorems in this section apply only to the one-dimensional case of $\Omega \subset \mathbb{R}$, but they are the easiest to check! For multidimensional versions, use the M-estimation method in §5.4.

THEOREM 6.1.4. *Suppose $\Omega = \{\theta_0, \ldots, \theta_k\}$ is composed of finitely many elements, and conditions $A_0 - A_2$ hold. Then the MLE $\hat{\theta}_n = \hat{\theta}(X_1, \cdots, X_n)$ is unique for sufficiently large n and $\hat{\theta}_n \xrightarrow{a.s.} \theta$.*

PROOF. Suppose that θ_0 is the true parameter value. Then lemma 6.1.3 and the SLLN imply that

$$-\frac{1}{n}\sum_1^n \log\frac{f(X_i,\theta_j)}{f(X_i,\theta_0)} \to I(\theta_j|\theta_0) \text{ a.s. } P_{\theta_0},\ j = 1, \ldots, k.$$

Hence for n sufficiently large

$$-\frac{1}{n}\sum_1^n \log\frac{f(X_i,\theta_j)}{f(X_i,\theta_0)} > \frac{1}{2}I(\theta_j|\theta_0)$$

$$\text{i.e. } \ell(\boldsymbol{X},\theta_0) - \ell(\boldsymbol{X},\theta_j) > \frac{n}{2}I(\theta_j|\theta_0) > 0 \text{ if } \theta_j \neq \theta_0. \tag{6.1.1}$$

So for all n sufficiently large $\ell(\boldsymbol{X},\theta_j)$ has a unique maximum at $\theta_j = \theta_0$. Hence $\hat{\theta}_{ML} \to \theta_0$ *a.s.* P_{θ_0}, i.e. $\hat{\theta}_{ML}$ is strongly consistent. □

REMARK 6.1.5. Theorem 6.1.4 may not hold if Ω is countably infinite (see example on page 410 of TPE).

THEOREM 6.1.6. *Suppose conditions $A_0 - A_3$ hold and for almost all x, $f(x,\theta)$ is differentiable with respect to $\theta \in N$ with continuous derivative $f'(x,\theta)$, where N is an open subset of Ω containing θ_0 and $\Omega \subseteq \mathbb{R}$. Then with P_{θ_0} probability 1, for n large*

$$\ell'(\theta, X) = \sum \frac{f'(X_i,\theta)}{f(X_i,\theta)} = 0$$

has a root $\hat{\theta}_n$ and $\hat{\theta}_n \to \theta_0$ a.s. P_{θ_0}.

PROOF. Choose $\epsilon > 0$ small so that $(\theta_0 - \epsilon, \theta_0 + \epsilon) \subseteq N$ and define

$$S_n = \{\boldsymbol{x}\colon \ell(\theta_0, \boldsymbol{x}) > \ell(\theta_0 - \epsilon, \boldsymbol{x}) \text{ and } \ell(\theta_0, \boldsymbol{x}) > \ell(\theta_0 + \epsilon, \boldsymbol{x})\}.$$

From eqtn (6.1.1), a.s. P_{θ_0}, $\ell(\theta_0, \boldsymbol{X}) - \ell(\theta_0 \pm \epsilon, \boldsymbol{X}) > 0$ for all n large. Hence there exists $\hat{\theta}_n(\boldsymbol{X}) \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$ such that $\ell'(\hat{\theta}_n) = 0$. If there exist more than one $\hat{\theta}_n \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$, choose the one closest to θ_0 (the set of roots is closed since $f'(x,\theta)$ is continuous in θ). Call this root $\hat{\theta}_n^*$. (Note that there could be 2 closest roots in which case choose the

larger.) Let $A_\epsilon = \{\boldsymbol{X} = (X_1, X_2, \ldots)\colon \exists\ \hat{\theta}_n^* \in (\theta_0 - \epsilon, \theta_0 + \epsilon)\ s.t.\ \ell'(\hat{\theta}_n^*) = 0,\ \forall$ sufficient large n and $\hat{\theta}_n^*$ is the closest to $\theta_0\}$. Then $P_{\theta_0}(A_\epsilon) = 1$. Define $A_0 = \lim_{k\to\infty} A_{1/k}$. Clearly $P_{\theta_0}(A_0) = 1$ and on A_0, $\hat{\theta}_n^* \to \theta_0$. □

REMARK 6.1.7. Theorem 6.1.6 says there exists a sequence of local maxima which converges a.s. P_{θ_0} to θ_0. However since we don't know θ_0, we can't determine the sequence unless ℓ has a unique local maximum for each n.

COROLLARY 6.1.8. *If $\ell'(\theta) = 0$ has a unique root $\hat{\theta}_n$ for all $\boldsymbol{X}$ and for all sufficiently large n, then*

$$\hat{\theta}_n \to \theta_0 \text{ a.s. } P_{\theta_0}$$

If in addition Ω is the open interval (θ_L, θ_U) and $\ell'(\theta)$ is continuous on Ω for all $\boldsymbol{X}$, then $\hat{\theta}_n$ maximizes the likelihood (globally), i.e. $\hat{\theta}_n$ is the MLE and hence the MLE is consistent.

PROOF. The first statement follows straight from theorem 6.1.6.

If $\hat{\theta}_n$ is not the MLE, then

$$\ell(\theta) \to \sup_\alpha \ell(\alpha) \text{ as } \theta \downarrow \theta_L \text{ or } \theta \uparrow \theta_U$$

But $\hat{\theta}_n$ is a local max by the proof of theorem 6.1.6 and hence ℓ must also have a local min and $\ell'(\theta) = 0$ for some $\theta \neq \hat{\theta}_n$, a contradiction. So $\hat{\theta}_n$ is the MLE. □

As we might expect, under mild conditions the MLE exists, is unique, and consistent for an **exponential family**.

THEOREM 6.1.9. *Consider a full-rank s-parameter exponential family in canonical form where the density can be written as*

$$p(x, \boldsymbol{\eta}) = \exp\left\{\sum_{i=1}^{s} \eta_i T_i(x) - A(\boldsymbol{\eta})\right\} h(x), \qquad \boldsymbol{\eta} \in \eta(\Omega),$$

and let the natural parameter space $\mathcal{N}$ be an open set. Let $\boldsymbol{x}$ is the observed data vector from a sample from this model, and $\boldsymbol{t}$ be the oberved value of the complete and sufficient statistic $\boldsymbol{T} = (T_1(\boldsymbol{x}), \ldots, T_s(\boldsymbol{x}))$. Then:

(i) *The MLE exists with probability tending to 1 as $n \to \infty$.*
(ii) *The MLE is consistent.*
(iii) *If the density function is continuous, then the MLE $\hat{\boldsymbol{\eta}}$ exists almost surely, and satisfies the equation:*

$$\left.\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}\right|_{\hat{\boldsymbol{\eta}}} = \boldsymbol{t}.$$

PROOF. This is a combination of Theorems 2.3.1, 2.3.2, and 5.2.2 in Bickell & Doksum (2015). □

EXAMPLE 6.1.10. $f(x,\theta) = \theta e^{-\theta x},\ 0 < x < \infty,\ 0 < \theta < \infty$. Obviously $\hat{\theta}_n = 1/\bar{x} \to \theta$ a.s. by SLLN (and continuous mapping), since $\mathbb{E}X = 1/\theta$. However let us show it by applying Theorem 6.1.9. For a random sample $\boldsymbol{x}$ the density is:

$$f(\boldsymbol{x}) = \exp\left\{-\theta n\bar{x} - (-n\log\theta)\right\},$$

from which we identify $\eta = -\theta$, $t = n\bar{x}$, and $A(\eta) = -n\log(-\eta)$, whence

$$\frac{\partial A(\eta)}{\partial \eta} = -\frac{n}{\eta} = n\bar{x}, \quad \Longrightarrow \quad \hat{\eta} = -1/\bar{x}, \quad \hat{\theta}_n = 1/\bar{x}.$$

Since $\Omega = (0,\infty)$ is open we have immediate consistency of the MLE.

6.2. Asymptotic Normality of the MLE

THEOREM 6.2.1. *Suppose that $X_1,\ldots,X_n$ are iid P_{θ_0}, $\hat{\theta}_n = \hat{\theta}_n(X_1,\ldots,X_n)$ is the MLE of the true parameter $\theta_0 \in \mathbb{R}^d$, and the following conditions hold:*

(i) *$\theta_0 \in int(\Omega) \subseteq \Omega$, and the model density $f(x,\theta)$ is 3 times differentiable w.r.t. θ in some open neighborhood of θ_0.*
(ii) $\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$.
(iii) $E_{\theta_0} \frac{\partial}{\partial \theta_j} \log f(X,\theta_0) = 0,\ 1 \le j \le d$.
(iv) *The **Fisher information matrix per observation**, defined as*

$$I(\theta_0) = E_{\theta_0}\left[\left(\frac{\partial}{\partial\theta}\log f(X,\theta_0)\right)\left(\frac{\partial}{\partial\theta}\log f(X,\theta_0)\right)^T\right] = -E_{\theta_0}\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f(X,\theta_0)\right],$$

is non-singular. (Note: this is the $I_1(\theta_0)$ defined in Ch. 2.)
(v) *There exists $\delta > 0$ such that $E_{\theta_0} W_\delta(X) < \infty$ where*

$$W_\delta(X) = \sup_{\|\theta-\theta_0\|<\delta} \sum_{1\le\alpha,\beta\le d} \left| \frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\log f(x,\theta) - \frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\log f(x,\theta_0)\right|$$

(vi) $\frac{\partial^2}{\partial\theta_\alpha\theta_\beta}\log f(x,\theta)$ *is continuous in θ for all x.*

Then

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(\underline{0}, I^{-1}(\theta_0)\right)$$

NOTE 6.2.2. If $\ell(\theta)$ is the log-likelihood for $X_1,\ldots,X_n$ under the above iid model, then

$$I(\theta_0) = -\frac{1}{n} E_{\theta_0}\ell''(\theta_0).$$

PROOF. From (vi), $W_\epsilon(X) \downarrow 0$ as $\epsilon \downarrow 0$. Hence from (v), $E_{\theta_0} W_\epsilon(X) \downarrow 0$ as $\epsilon \downarrow 0$.

Now for some ϕ between θ_0 and $\hat{\theta}_n$,

$$\frac{1}{n}\sum_{1}^{n}\frac{\partial}{\partial\theta_\alpha}\log f(x_i,\theta_0) = \frac{1}{n}\sum_{1}^{n}\frac{\partial}{\partial\theta_\alpha}\log f(x_i,\hat{\theta}_n) + \frac{1}{n}\sum_{i=1}^{n}\sum_{\beta=1}^{d}\frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\log f(x_i,\phi)(\theta_\beta^0-\hat{\theta}_{n\beta}) \tag{6.2.1}$$

(The first term equals 0 since $\hat{\theta}_n = MLE$.)
We will show that

$$\frac{1}{n}\sum_{i}\frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\log f(x_i,\phi)\to -I_{\alpha\beta}(\theta_0) \tag{6.2.2}$$

Write

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\log f(x_i,\phi) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\log f(x_i,\theta_0) + \frac{1}{n}\sum_{i}\left[\frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\log f(x_i,\phi)-\frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\log f(x_i,\theta_0)\right]$$

The first term on the right hand side goes to $-I_{\alpha\beta}(\theta_0)$ in probability by the WLLN. For any $c>0$,

$$\begin{aligned}P_{\theta_0}(|2nd\ term|>c) =& P_{\theta_0}(\|\phi-\theta_0\|>\epsilon, |2nd\ term|>c) + P_{\theta_0}(\|\phi-\theta_0\|\le\epsilon, |2nd\ term|>c)\\ \le& P_{\theta_0}(\|\phi-\theta_0\|>\epsilon)\\ &+ P_{\theta_0}\left(\frac{1}{n}\sum_i \sup_{\|\theta-\theta_0\|<\epsilon}\left|\frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\log f(x_i,\theta)-\frac{\partial^2}{\partial\theta_\alpha\partial\theta_\beta}\log f(x_i,\theta_0)\right|>c\right)\\ =& I+II\end{aligned}$$

where $I\le P_{\theta_0}\left(\|\hat{\theta}_n-\theta_0\|>\epsilon\right)\to 0$, and $II = P_{\theta_0}\left(\frac{1}{n}\sum_i W_\epsilon(x_i)>c\right)\le\frac{1}{c}E_{\theta_0}W_\epsilon(X)\to 0$ as $\epsilon\downarrow 0$. This proves (6.2.2).

Now write (6.2.1) in matrix form

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta_\alpha}\log f(X_i,\theta_0)\right]_{\alpha=1,\dots,d} = (-I(\theta_0)+o_p(1))\left(\theta_0-\hat{\theta}_n\right)$$

By CLT,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta_\alpha}\log f(X_i,\theta_0)\right]_{\alpha=1,\dots,d}\xrightarrow{d} N\left(\underline{0}, I(\theta_0)\right)$$

It follows that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \left(I(\theta_0) + o_p(1)\right)^{-1} \frac{1}{\sqrt{n}} \sum_1^n \frac{\partial}{\partial\theta} \log f(x_i, \theta_0)$$
$$\xrightarrow{d} N\left(\underline{0}, I^{-1}(\theta_0) I(\theta_0) I^{-1}(\theta_0)\right).$$

□

NOTE 6.2.3. Condition (v) can be replaced by: $\exists\ \delta > 0$ s.t. $\forall x$ in the support of $f(x, \theta)$

$$\left| \frac{\partial^3}{\partial\theta_\alpha \theta_\beta \theta_\gamma} \log f(x, \theta) \right| \leq M_{\alpha,\beta,\gamma}(x),$$

for any θ with $\parallel \theta - \theta_0 \parallel < \delta$, and where $E_{\theta_0} M_{\alpha,\beta,\gamma}(X) < \infty$. Condition (vi) can be replaced by:

$$E_{\theta_0} \sum_{j=1}^d \left(\frac{\partial \log f(x, \theta)}{\partial \theta_j} \right)^2 < \infty.$$

(See Van der Vaart, 1998, Theom 5.41.)

EXAMPLE 6.2.4. One parameter exponential family

$$f(x_i, \eta) = e^{\eta T(x_i) - A(\eta)} h(x_i)$$

The likelihood equation $\ell'(\eta) = 0$ implies $\frac{1}{n}\sum T(x_i) = A'(\eta) = E_\eta T(X_1)$. Since $A''(\eta) = I(\eta) = Var_\eta T > 0$, $A'(\eta)$ is strictly increasing so that $\ell'(\eta) = 0$ has at most one solution. By Theorem 6.1.6 and its Corollary, $\hat{\eta} \to \eta$ a.s. Also $\frac{d^3}{d\eta^3} \log f(x, \eta) = A'''(\eta)$ is independent of x and continuous. Hence by Theorem 6.2.1,

$$\sqrt{n}\left(\hat{\eta} - \eta\right) \xrightarrow{d} N\left(0, I^{-1}(\eta)\right) = N\left(0, \frac{1}{Var_\eta T}\right).$$

In fact, this is a special case of a more general result, which complements Theorem 6.1.9.

THEOREM 6.2.5. *Consider a full-rank s-parameter exponential family in canonical form:*

$$p(x, \boldsymbol{\eta}) = \exp\left\{ \sum_{i=1}^s \eta_i T_i(x) - A(\boldsymbol{\eta}) \right\} h(x), \qquad \boldsymbol{\eta} \in \eta(\Omega),$$

and let the natural parameter space $\mathcal{N}$ be an open set, with $\boldsymbol{T} = (T_1(\boldsymbol{x}), \ldots, T_s(\boldsymbol{x}))$ the complete and sufficient statistic. Then, if $\hat{\boldsymbol{\eta}}$ is defined to be the MLE, if it exists, and some fixed value otheriwse, we have that

$$\sqrt{n}\left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) \xrightarrow{d} N\left(\mathbf{0}, I^{-1}(\boldsymbol{\eta})\right), \quad \textit{where} \quad I(\boldsymbol{\eta}) = Var(\boldsymbol{T}) = \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}'}.$$

PROOF. This is Theorem 5.3.5 in Bickell & Doksum (2015). □

EXAMPLE 6.2.6. **(Censoring)**. Suppose $X_1, \cdots, X_n \sim iid\, E(\frac{1}{\theta})$ $(i.e.\, EX_i = \theta)$, and suppose we observe the censored data $Y_i = min(X_i, T)$ with T fixed. Let μ be the measure on $[0, T]$ defined by

$$\mu(A) = \int_A dx + I_A(T)$$

(Lebesgue measure plus unit mass at T).

Then the density of Y with respect to μ is

$$\begin{aligned}
p(y,\theta) &= \begin{cases} \frac{1}{\theta} e^{-y/\theta} & 0 \le y < T \\ e^{-T/\theta} & y = T \qquad (= P_\theta(X_i \ge T)) \end{cases} \\
\therefore -\frac{\partial^2}{\partial\theta^2} \log p(y,\theta) &= \begin{cases} \frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta}\left(\log\theta + \frac{y}{\theta}\right) & 0 \le y < T \\ \frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta}\left(\frac{T}{\theta}\right) & y = T \end{cases} \\
&= \begin{cases} -\frac{1}{\theta^2} + \frac{2y}{\theta^3} & 0 \le y < T \\ \frac{2T}{\theta^3} & y = T \end{cases} \\
\therefore I(\theta) &= -E\frac{\partial^2}{\partial\theta^2} \log p(y,\theta) \\
&= \int_0^T \left(-\frac{1}{\theta^2} + \frac{2y}{\theta^3}\right) \frac{1}{\theta} e^{-y/\theta} dy + \frac{2T}{\theta^3} e^{-T/\theta} \\
&= \frac{1}{\theta^2}\left(1 - e^{-T/\theta}\right) \\
\frac{\partial^3}{\partial\theta^3} \log p(y,\theta) &= \begin{cases} -\frac{2}{\theta^3} + \frac{6y}{\theta^4} & 0 \le y < T \\ \frac{6T}{\theta^4} & y = T \end{cases}
\end{aligned}$$

Since $\theta_0 \in (0, \infty)$, $\left|\frac{\partial^3}{\partial\theta^3} \log f(y,\theta)\right| \le Ay + B$ for $\frac{\theta_0}{2} < \theta < \frac{3\theta_0}{2}$, and $E_{\theta_0}(Ay + B) < \infty$. Check that $\hat{\theta}_n \xrightarrow{P} \theta_0$ under P_{θ_0}. Then by theorem (6.2.1)

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \frac{\theta^2}{1 - e^{-T/\theta}}\right)$$

6.3. Asymptotic Optimality of the MLE

Under the conditions of Theorem 6.2.1, the MLE satisfies

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, I^{-1}(\theta_0)\right)$$

We will show now that $I^{-1}(\theta_0)$ is the minimal attainable covariance matrix for a class of asymptotically normal estimates. (Note. As in section 6.2, $I(\theta)$ denotes the Fisher information matrix *per observation*.)

THEOREM 6.3.1. *Suppose $\{T_n\}$ is a sequence of asymptotically normal estimators of θ with $Var_\theta(T_n) < \infty$ for all n, and define*

$$\triangle_n(\theta) = \frac{\partial}{\partial\theta} E_\theta(T_n) = \left[\frac{\partial}{\partial\theta_i} E_\theta(T_{nj})\right]_{i,j=1,\ldots,d}.$$

If all the following conditions hold:

(i) $\sqrt{n}\,(T_n - \theta) \xrightarrow{d} N\,(0, \Sigma(\theta))$,
(ii) $n\,Cov_\theta(T_n) \geq \triangle_n^T I^{-1}(\theta)\triangle_n$, $\quad (A \geq B \iff A - B$ *psd, as defined earlier*),
(iii) $\exists\delta(\theta)$ *such that* $\sup_n E_\theta \parallel \sqrt{n}\,(T_n - \theta) \parallel^{2+\delta} < \infty$,
(iv) $\triangle_n \to I_{d\times d}$,

then

$$\Sigma(\theta) \geq I^{-1}(\theta).$$

PROOF. (iii) implies that $\{\sqrt{n}(T_n - \theta)\}$ and $\{n(T_{n,\alpha} - \theta_\alpha)(T_{n,\beta} - \theta_\beta)\}$ are uniformly integrable (Billingsley, p. 338). Hence from (i), if $Z \sim N(0, \Sigma(\theta))$,

$$E_\theta\sqrt{n}(T_n - \theta) \to EZ = 0$$

and

$$n\,Cov(T_n) \to Cov(Z) = \Sigma(\theta)$$

(Billingsley, p.338). As $n \to \infty$, the LHS of (ii) $\to \Sigma(\theta)$, and the RHS of (ii) $\to I^{-1}(\theta)$ by (iv). Therefore, $\Sigma(\theta) \geq I^{-1}(\theta)$. □

NOTE 6.3.2. Assumption (ii) will be satisfied under the conditions of corollary (2.4.4)(in the UMVU section).

DEFINITION 6.3.3. If $\{T_n\}$ satisfies (i) with $\Sigma(\theta) = I^{-1}(\theta)$, then it is said to be *asymptotically efficient.*

COROLLARY 6.3.4. *Let $g = (g_1, \cdots, g_r)^T \colon \Omega \to \mathbb{R}^r$ have continuous derivatives $\frac{\partial g_i}{\partial\theta_\alpha}$, $1 \leq \alpha \leq d,\ 1 \leq i \leq r$. Define*

$$\triangle(\theta) = \frac{\partial g}{\partial\theta} = \left[\frac{\partial g_j}{\partial\theta_i}\right]_{i=1,\cdots,d,\ j=1,\cdots,r}$$

and assume that the sequence $\{T_n\}$ of estimators satisfies

(i) $\sqrt{n}\,(T_n - g(\theta)) \xrightarrow{d} N\,(0, \Sigma(\theta))$,
(ii) $n\,Cov_\theta(T_n) \geq \triangle_n^T(\theta) I^{-1}(\theta)\triangle_n(\theta)$, *where* $\triangle_n(\theta) = \frac{\partial}{\partial\theta} E_\theta T_n$,
(iii) $\exists\delta(\theta) > 0$ *such that* $\sup_n E_\theta \parallel \sqrt{n}\,(T_n - g(\theta)) \parallel^{2+\delta} < \infty$,
(iv) $\triangle_n \to \triangle$,

then

$$\Sigma(\theta) \geq \triangle^T I^{-1}(\theta)\triangle$$

PROOF. The same as the proof of Theorem 6.3.1. ((ii) again holds under the conditions of Corollary 2.4.4.) □

Next we show that $g(\hat{\theta}_n)$, where $\hat{\theta}_n = MLE(\theta)$, achieves the lower bound in Corollary (6.3.4) provided the conditions of Theorem 6.2.1 hold.

COROLLARY 6.3.5. *Suppose* $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, I^{-1}(\theta)\right)$ *and* $\partial g_j/\partial\theta_i$ *is continuous for* $1 \le j \le r$, $1 \le i \le d$, *then*

$$g(\hat{\theta}_n) \text{ is } AN\left(g(\theta), \frac{1}{n}\triangle^T I^{-1}(\theta)\triangle\right).$$

PROOF. Since

$$\hat{\theta}_n - \theta = O_p\left(\frac{1}{\sqrt{n}}\right)$$

$$g(\hat{\theta}_n) - g(\theta) = \triangle^T\left(\hat{\theta}_n - \theta\right) + o_p\left(\frac{1}{\sqrt{n}}\right)$$

$$\therefore \sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow{d} N\left(0, \triangle^T I^{-1}(\theta)\triangle\right) \text{ by Slutzky.}$$

□

EXAMPLE 6.3.6. Suppose $X_1, \cdots, X_n$ are iid lognormal(μ, σ^2), (i.e. $Y_i = \log X_i \sim N(\mu, \sigma^2)$) with $\sigma^2 = 1$. The MLE of μ is $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \log X_i$. Suppose $g(\mu) = EX_1 = e^{\mu+1/2}$ (Set $\lambda = 1$ in $Ee^{\lambda Y_1} = e^{\lambda\mu+\lambda^2\sigma^2/2} \Rightarrow EX_1 = e^{\mu+\sigma^2/2}$).

Consider the two estimators of $\theta := g(\mu)$

$$\hat{\theta}_n = g(\hat{\mu}_n) = e^{\hat{\mu}_n+1/2} \quad \text{(MLE)}$$

and

$$\tilde{\theta}_n = \frac{1}{n}\sum_1^n X_i \quad \text{(Moment estimator).}$$

The Fisher information for μ in X_i = the Fisher information for μ in Y_i since the transformations $Y_i = \log X_i$ is one to one. So it equals $E_\mu\left(\frac{\partial}{\partial\mu}\log f(Y_i, \mu)\right)^2 = E_\mu(Y_i - \mu)^2 = 1$ (notice $\log f(Y, \mu) = -\frac{1}{2}\ln 2\pi - \frac{(Y-\mu)^2}{2}$). Therefore

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, \triangle^T I^{-1}(\mu)\triangle\right)$$

where $\triangle = \frac{d}{d\mu}g(\mu) = e^{\mu+1/2}$, i.e.

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, e^{2\mu+1}\right)$$

On the other hand, for the moment estimator $\tilde{\theta}_n = \frac{1}{n}\sum X_i$, we have

$$\sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow{d} N\left(0, Var(X_i)\right) = N\left(0, e^{2\mu+1}(e-1)\right)$$

($EX_i^2 = Ee^{2Y} =$ (MGF of $N(\mu,1)$ evaluated at $\lambda = 2) = e^{2\mu+2}$, therefore $Var(X_i) = e^{2\mu+2} - \left(e^{\mu+1/2}\right)^2 = e^{2\mu+1}(e-1)$.)

The ARE of $\tilde{\theta}_n$ relative to $\hat{\theta}_n$ is thus

$$ARE(\tilde{\theta}_n, \hat{\theta}_n) = 1/(e-1) = .582$$

The sample mean thus has poor ARE as an estimator of the mean of a log normal distribution. (The log normal distribution has heavy tails so this is not too surprising.)

EXAMPLE 6.3.7. (Hodges, TPE, p.440) $X_1, \cdots, X_n \sim iid\ N(\theta, 1)$, $I(\theta) = 1$. Define

$$T_n = \begin{cases} \bar{X}_n & if \quad |\bar{X}_n| > n^{-1/4} \\ 0 & if \quad |\bar{X}_n| < n^{-1/4} \end{cases}$$

Then, from a homework problem, T_n is $AN(\theta, \frac{v(\theta)}{n})$, where $v(\theta) = 1$ if $\theta \neq 0$; and 0 if $\theta = 0$. So $v(\theta) < I(\theta)$ at $\theta = 0$ and the parameter value 0 is called a point of **superefficiency**. Note the following:

- Theorem 6.3.1 does not apply to this example, since condition (ii) fails.
- T_n is not uniformly better than $\bar{X}_n$ for finite n, for example $\theta_n = n^{-1/4} \Rightarrow E_{\theta_n} n(T_n - \theta_n)^2 \to \infty > 1 = \lim_{n\to\infty} E_{\theta_n} n(\bar{X}_n - \theta_n)^2$.

REMARK 6.3.8. LeCam (1953) showed that for any estimator satisfying

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, v(\theta)\right)$$

the set of θ's where $v(\theta) < I(\theta)^{-1}$ has Lebesgue measure zero. (See Theorem 2.6 in TPE.)

Iterative Methods Suppose we have a sequence of estimators $\tilde{\theta}_n$ such that

$$\tilde{\theta}_n = \theta_0 + O_p(\frac{1}{\sqrt{n}})$$

Set

$$T_n = \tilde{\theta}_n - \frac{\ell'(\tilde{\theta}_n)}{\ell''(\tilde{\theta}_n)}$$

Then

$$\ell'(\tilde{\theta}_n) = \ell'(\theta_0) + (\tilde{\theta}_n - \theta_0)\ell''(\theta_n^*)$$

where θ_n^* is between θ_0 and $\tilde{\theta}_n$. Thus

$$\begin{aligned}\sqrt{n}\,(T_n-\theta_0) &= \sqrt{n}\left(\tilde{\theta}_n-\theta_0-\frac{\ell'(\tilde{\theta}_n)}{\ell''(\tilde{\theta}_n)}\right)\\ &= \sqrt{n}\left(\tilde{\theta}_n-\theta_0-\frac{\ell'(\theta_0)}{\ell''(\tilde{\theta}_n)}-(\tilde{\theta}_n-\theta_0)\frac{\ell''(\theta_n^*)}{\ell''(\tilde{\theta}_n)}\right)\\ &= \frac{-\frac{1}{\sqrt{n}}\ell'(\theta_0)}{\frac{1}{n}\ell''(\tilde{\theta}_n)}+\sqrt{n}(\tilde{\theta}_n-\theta_0)\left[1-\frac{\ell''(\theta_n^*)}{\ell''(\tilde{\theta}_n)}\right]\end{aligned}$$

Under the conditions of Theorem 6.2.1

$$\begin{aligned}&\frac{1}{\sqrt{n}}\ell'(\theta_0)\xrightarrow{d}N\left(0,I(\theta_0)\right),\\ &\frac{1}{n}\ell''(\tilde{\theta}_n)\xrightarrow{P}-I(\theta_0)\text{ and}\\ &\frac{1}{n}\ell''(\theta_n^*)\xrightarrow{P}-I(\theta_0).\end{aligned}$$

Hence the term in square brackets is $o_p(1)$.

Thus

$$RHS=\frac{-\frac{1}{\sqrt{n}}\ell'(\theta_0)}{\frac{1}{n}\ell''(\tilde{\theta}_n)}+o_p(1)\xrightarrow{d}N\left(0,I^{-1}(\theta_0)\right)$$

and so we have proved the following theorem.

THEOREM 6.3.9. *Suppose that (A_0)–(A_2) and all conditions of Theorem 6.2.1 hold, with the possible exception of (ii), and that $\tilde{\theta}_n$ is $\sqrt{n}$-consistent estimator of θ, i.e.*

$$\tilde{\theta}_n=\theta+O_p(\frac{1}{\sqrt{n}}).$$

Then,

$$T_n:=\tilde{\theta}_n-\frac{\ell'(\tilde{\theta}_n)}{\ell''(\tilde{\theta}_n)}$$

is asymptotically efficient.

COROLLARY 6.3.10. *If $I(\cdot)$ is continuous then the estimator*

$$T_n'=\tilde{\theta}_n+\frac{\ell'(\tilde{\theta}_n)}{nI(\tilde{\theta}_n)}$$

is asymptotically efficient.

PROOF.

$$\begin{aligned}\sqrt{n}\,(T_n' - T_n) &= \sqrt{n}\left(\frac{\ell'(\tilde{\theta}_n)}{\ell''(\tilde{\theta}_n)} + \frac{\ell'(\tilde{\theta}_n)}{nI(\tilde{\theta}_n)}\right)\\ &= \ell'(\tilde{\theta}_n)\left(\frac{\sqrt{n}}{\ell''(\tilde{\theta}_n)} + \frac{1}{nI(\tilde{\theta}_n)}\right)\\ &= \frac{\ell'(\tilde{\theta}_n)}{\sqrt{n}}\left(\frac{I(\tilde{\theta}_n) + \frac{\ell''(\tilde{\theta}_n)}{n}}{I(\tilde{\theta}_n)\frac{\ell''(\tilde{\theta}_n)}{n}}\right)\\ &= o_p(1).\end{aligned}$$

(The first factor is $O_p(1)$, the numerator of the second factor is $o_p(1)$ and the denominator $\xrightarrow{P}\ I(\theta_0)^2$.) □

EXAMPLE 6.3.11. **Location family**. Suppose that $X_1, X_2, \cdots$ are *iid* $f(x-\theta)$ where f is differentiable and symmetric, $f(x) > 0$ for all x and f' is continuous. Then the conditions of Theorem 6.1.6 hold

A_0 : $P_\theta \neq P_{\theta_0},\ \theta \neq \theta_0$
A_1 : common support
A_2 : $\frac{dP_\theta}{d\lambda}(x) = f(x-\theta)$
A_3 : $\theta \in int(\Omega) = (-\infty, \infty)$

Hence

$$\ell'(\theta, X) = \sum_1^n \frac{f'(x_i - \theta)}{f(x_i - \theta)} = 0 \tag{6.3.1}$$

has a sequence of roots $\hat{\theta}_n$ for n large such that $\hat{\theta}_n \to \theta_0$ a.s. P_{θ_0}. Since $\ell(\theta, X) \to 0$ as $\theta \to \pm\infty$, $\ell(\theta, X)$ must have a max, however there may be several solutions of (6.3.1).

Provided $f(x-\theta)$ satisfies conditions (v) and (vi) of Theorem 6.2.1, i.e.

$$E_{\theta_0}\left[\sup_{|\theta-\theta_0|<\delta}\left|\frac{\partial^2}{\partial\theta^2}\log f(x-\theta) - \frac{\partial^2}{\partial\theta^2}\log f(x-\theta_0)\right|\right] < \infty$$

and $\frac{\partial^2}{\partial\theta^2}\log f(x-\theta)$ is continuous in θ for all x, then all the conditions of theorem (6.2.1) (apart from (ii)) are satisfied. Also $\bar{X}$ is $AN(\theta, \frac{\sigma^2}{n})$, and hence

$$\bar{X} = \theta + O_p(\frac{1}{\sqrt{n}}).$$

The corollary of Theorem 6.3.9 therefore implies the asymptotic efficiency of

$$T_n = \bar{X}_n + \frac{\sum_{i=1}^n \frac{f'(X_i - \bar{X}_n)}{f(X_i - \bar{X}_n)}}{nI(\bar{X}_n)}$$

where

$$I(\theta) = \int \left(\frac{f'(y)}{f(y)}\right)^2 f(y)dy.$$

NOTE 6.3.12. The significance of this result is that, e.g. in Example 6.3.6 (lognormal), if the likelihood has multiple roots (so one doesn't know which to take), Theorem 6.3.9 and Corollary 6.3.10 say that one gets just as good an estimator (asymptotically), by starting with a $\sqrt{n}$-consistent one (e.g., the MOME $\bar{X}$), and using T_n obtained from one iteration of Newton-Raphson.

6.4. Asymptotic Efficiency of Bayes Estimators

EXAMPLE 6.4.1. Suppose $X \sim \text{Bin}(n, p)$, with $p \sim \text{B}(a, b)$, then from Example 4.1.4, the Bayes estimator of p is $T_n = (a + x)/(a + b + n)$, and hence:

$$\sqrt{n}(T_n - p) = \sqrt{n}\left(\frac{X}{n} - p\right) + \frac{\sqrt{n}}{a+b+n}\left[a - (a+b)\frac{X}{n}\right] \equiv S_1 + S_2.$$

Note that $S_1 \xrightarrow{d} \text{N}(0, p(1-p))$ by CLT, and since $X/n \xrightarrow{p} p$ by WLLN, then we have that $S_2 \to 0$ as $n \to \infty$. Thus both the Bayes estimator T_n and the MLE X/n have the same limiting asymptotic distribution.

Questions:

- Does this limiting result also hold for an arbitrary prior?
- And does it extend to more general models (not just the Binomial)?

The answer to both questions is YES, but requires some regularity conditions. (In the ensuing, let $\theta \in \Omega$ denote the d-dimensional parameter vector, and θ_0 its true value.)

Regularity Conditions:

(B1): The log-likelihood function $\ell(\theta)$ satisfies all the statements and assumptions of Theorem 6.2.1 (asymptotic normality of MLE).

(B2): Given $\epsilon > 0$, $\exists \delta > 0$ such that

$$P\left(\sup\{|R_n(\theta)/n| : |\theta - \theta_0| \le \delta\} \ge \epsilon\right) \longrightarrow 0, \quad \text{as } n \to \infty,$$

where R_n is the remainder term in a Taylor-series expansion of $\ell(\theta)$ about θ_0

$$\ell(\theta) = \ell(\theta_0) + (\theta - \theta_0)\ell'(\theta_0) - \frac{1}{2}(\theta - \theta_0)^2[nI(\theta_0) + R_n(\theta)],$$

which satisfies $R_n(\theta)/n \xrightarrow{p} 0$ as $n \to \infty$.

(B3): Given $\delta > 0$, $\exists \epsilon > 0$ such that

$$P\left(\sup\left\{\frac{\ell(\theta) - \ell(\theta_0)}{n} : |\theta - \theta_0| \ge \delta\right\} \le -\epsilon\right) \longrightarrow 1, \quad \text{as } n \to \infty.$$

(Controls the behavior of $\ell(\theta)$ at a distance from θ_0; important since Bayes estimators involve integration over entire range of θ values.)

(B4): The prior density $\pi(\theta)$ on θ is continuous and positive for all $\theta \in \Omega$.

(B5): $\mathbb{E}_\pi|\Theta| = \int |\theta|\pi(\theta)d\theta < \infty$.

The following lemma establishes that under these conditions the posterior is AN:

$$\pi(\theta|\boldsymbol{x}) \sim \text{AN}\left(\mu_n = \theta_0 + \frac{\ell'(\theta_0)}{nI(\theta_0)},\ \sigma_n^2 = \frac{1}{nI(\theta_0)}\right).$$

LEMMA 6.4.2. *If $\pi^*(t|\boldsymbol{x})$ is the posterior density of $t \equiv \sqrt{n}(\theta - T_n)$, where $T_n = \theta_0 + \frac{\ell'(\theta_0)}{nI(\theta_0)}$, we have the following two results, where $\phi(\cdot)$ is the pdf of a $N(0,1)$.*

(i) *If (B1)–(B4) hold:*

$$\int \left|\pi^*(t|\boldsymbol{x}) - \sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)})\right| dt \xrightarrow{p} 0.$$

(ii) *If (B1)–(B5) hold:*

$$\int (1+|t|)\left|\pi^*(t|\boldsymbol{x}) - \sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)})\right| dt \xrightarrow{p} 0.$$

NOTE 6.4.3. (i) and (ii) imply $\int |t|\left|\pi^*(t|\boldsymbol{x}) - \sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)})\right| dt \xrightarrow{p} 0$.

PROOF. TPE Theorem 8.2. □

THEOREM 6.4.4 (Asymptotic Efficiency of Bayes Estimators). *If (B1)–(B5) hold, and if $\tilde{\theta}_n$ is the Bayes estimator under squared error loss with prior pdf $\pi(\theta)$, then:*

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} N\left(0,\ I^{-1}(\theta_0)\right),$$

so that $\tilde{\theta}_n$ is consistent and asymptotically efficient.

PROOF. Note the following relation, with T_n as defined in Lemma 6.4.2:

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) = \sqrt{n}(\tilde{\theta}_n - T_n) + \sqrt{n}(T_n - \theta_0).$$

Now, from the proof of Theorem 6.2.1:

$$\frac{\ell'(\theta_0)}{\sqrt{nI(\theta_0)}} \xrightarrow{d} \text{N}(0,1), \qquad \Longrightarrow\ \sqrt{n}(T_n - \theta_0) \xrightarrow{d} \frac{1}{\sqrt{I(\theta_0)}}\text{N}(0,1) \sim \text{N}\left(0, I^{-1}(\theta_0)\right),$$

whence we only need to show that $\sqrt{n}(\tilde{\theta}_n - T_n) \xrightarrow{p} 0$. Now, under squared error loss:

$$\begin{aligned}
\tilde{\theta}_n &= \int \theta\pi(\theta|\boldsymbol{x})d\theta \\
&= \int (T_n + t/\sqrt{n})\pi^*(t|\boldsymbol{x})dt, \qquad \text{(transforming } \theta \mapsto t = \sqrt{n}\theta - \sqrt{n}T_n\text{)}, \\
&= T_n + \frac{1}{\sqrt{n}}\int t\pi^*(t|\boldsymbol{x})dt,
\end{aligned}$$

which implies $\sqrt{n}(\tilde{\theta}_n - T_n) = \int t\pi^*(t|\boldsymbol{x})dt$. Finally, and noting that $\int y\phi(y)dy = 0$ (integral of an odd function), we have:

$$\begin{aligned}
\sqrt{n}|\tilde{\theta}_n - T_n| &= \left| \int t\pi^*(t|\boldsymbol{x})dt - \int t\sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)})dt \right| \\
&= \int \left| t\left\{ \pi^*(t|\boldsymbol{x}) - \sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)}) \right\} \right| dt \\
&\leq \int |t| \cdot \left| \left\{ \pi^*(t|\boldsymbol{x}) - \sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)}) \right\} \right| dt, \qquad (|xy| \leq |x|.|y|), \\
&\xrightarrow{p} 0, \qquad \text{(by Note 6.4.3).}
\end{aligned}$$

□

As we would expect, Bayes estimators in the context of exponential family models are asymptotically efficient.

EXAMPLE 6.4.5 (One parameter exponential family).

$$f(x_i, \theta) = e^{\theta T(x_i) - A(\theta)} h(x_i) = \text{canonical form}$$

From Theorems 2.4.7 and 1.3.4, $A'(\theta) = \mathbb{E}T(X)$, $A''(\theta) = \text{Var}T(X) = I(\theta)$. Now check that (B1)–(B5) hold:

- (B1) holds since this is exponential family.
- (B4)-(B5) are conditions on the prior of choice.
- For (B2), since $\ell(\theta) = \sum \theta T(x_i) - nA(\theta)$, we have

$$\begin{aligned}
\ell(\theta) - \ell(\theta_0) &= (\theta - \theta_0) \sum T(x_i) - n[A(\theta) - A(\theta_0)] \\
&= (\theta - \theta_0) \underbrace{\sum [T(x_i) - A'(\theta_0)]}_{=\ell'(\theta_0)} - n \underbrace{\{[A(\theta) - A(\theta_0)] - [(\theta - \theta_0)A'(\theta_0)]\}}_{=\frac{1}{2}(\theta-\theta_0)^2 A''(\theta^*)},
\end{aligned}$$

where the 2nd underbrace follows from the T-series expansion:

$$A(\theta) = A(\theta_0) + (\theta - \theta_0)A'(\theta_0)\frac{1}{2}(\theta - \theta_0)^2 A''(\theta^*),$$

which holds for some θ^* between θ and θ_0, with $\theta^* \to \theta_0$ as $n \to \infty$. Thus,

$$\ell(\theta) = \ell(\theta_0) + (\theta - \theta_0)\ell'(\theta_0) - \frac{1}{2}(\theta - \theta_0)^2 \frac{A''(\theta^*)}{n}.$$

Matching this up with the T-series expansion of $\ell(\theta)$ in the statement of Condition (B2) implies that

$$\frac{A''(\theta^*)}{n} = nI(\theta_0) + R_n(\theta) \qquad \Longrightarrow \qquad \frac{1}{n}R_n(\theta) = A''(\theta^*) - I(\theta_0).$$

Therefore we must show that given $\epsilon > 0$, $\exists \delta > 0$ such that

$$P\left(\sup\left\{|A''(\theta^*) - I(\theta_0)| : |\theta^* - \theta_0| \leq \delta\right\} \geq \epsilon\right) \longrightarrow 0,$$

which is satisfied because $A''(\theta) = I(\theta)$ is continuous (Assumption (vi) in Theorem 6.2.1) and $\theta^* \to \theta_0$.

Finally (B3). From the proof of (B2) we can write:

$$(\dagger) \quad \frac{1}{n}\left[\ell(\theta) - \ell(\theta_0)\right] = (\theta - \theta_0)\left\{\frac{1}{n}\sum\left[T(x_i) - A'(\theta_0)\right] - \left[\frac{A(\theta) - A(\theta_0)}{\theta - \theta_0} - A'(\theta_0)\right]\right\}.$$

Now, since $A''(\theta) = I(\theta) > 0$, we have that $A(\theta)$ is strictly convex, whence assuming w.l.o.g. that $\theta > \theta_0$, implies (from the def. of convexity) that

$$\frac{A(\theta) - A(\theta_0)}{\theta - \theta_0} > A'(\theta_0).$$

Since

$$\frac{1}{n}\sum\left[T(x_i) - A'(\theta_0)\right] = \frac{\ell'(\theta_0)}{n} = \underbrace{\frac{1}{\sqrt{n}}}_{\xrightarrow{p}0} \quad \underbrace{\frac{\ell'(\theta_0)}{\sqrt{n}}}_{\xrightarrow{d}N(0,I(\theta_0))} \quad \xrightarrow{p} 0,$$

where the distributional convergence of $\ell'(\theta_0)/\sqrt{n}$ follows from the Iterative Methods discussion of section 6.3, it therefore follows from (†) that

$$\frac{1}{n}\left[\ell(\theta) - \ell(\theta_0)\right] < 0, \qquad \text{w.p. 1 as } n \to \infty,$$

whence for given $\delta > 0$, let $\theta - \theta_0 \geq \delta$, so that

$$\begin{aligned} \sup\left\{\frac{\ell(\theta) - \ell(\theta_0)}{n}\right\} &\leq \delta\left\{\frac{1}{n}\sum\left[T(x_i) - A'(\theta_0)\right] - \inf\left[\frac{A(\theta) - A(\theta_0)}{\theta - \theta_0} - A'(\theta_0)\right]\right\} \\ &\leq -\epsilon, \qquad \text{w.p. 1 as } n \to \infty. \end{aligned}$$

6.5. Discussion: MLE vs. Shrinkage (Efron & Hastie, 2016)

- Although MLE and its accompanying asymptotic optimality theory is one of the crowning achievements of classical statistical inference, it has proved to be an inadequate and dangerous tool in many 21st century applications (bigdata). To quote Efron & Hastie (2016):

  > "Unbiasedness can be an unaffordable luxury when there are 100's or 1000's of parameters to estimate at the same time."

- As we saw in Ch. 4, deliberate introduction of bias via shrinkage in order to improve overall performance (at a possible danger to some individual estimates) is usually preferable in such (bigdata) cases.
- However, whereas MLE comes equipped with an elegant theory for optimal unbiased estimation, **at present there is no equivalent optimality theory for shrinkage estimation.**

CHAPTER 7

Optimal Testing Theory

Whereas we can view point estimation as a primary level type of inference, tests (or equivalently, confidence regions), are a second level type of inference; one usually first desires the former before embarking on a quest for the latter.

In this chapter we will see that the UMVU notion of optimal estimation translates into UMP and UMPU tests. The former are rather restrictive in that they typically do not exist for two-sided situations; the notion of *unbiasedness* helps to remedy this situation, so that one can derive UMPU two-sided tests for a large class of "nice" problems, including the s-parameter exponential family.

After battling in this ground of provably-optimal procedures, we end with feasible and practical guidance. In the failure of identifying an optimal procedure (almost always the case), one settles for the near-optimal likelihood ratio, Wald, or Score test. This general approach parallels our point estimation story, where in the failure of identifying an optimal UMVU, MRE, or minimax estimator, we settled for the near-optimal MLE (asymptotically UMVU), or the Bayes estimator (admissible).

The UMP procedures (§7.1–7.5) apply only to the one-dimensional parameter θ. In §7.6 we see how UMPU optimality accommodates the case when there is additionally a vector of *nuisance* parameters. Finally, §7.7 deals with the most general case when θ is partitioned into two vectors, only one of which is the parameter of interest.

7.1. Uniformly Most Powerful (UMP) Tests

Our basic decision problem is to either accept or reject a given hypothesis about θ based on an observation of a r.v. X when the underlying p.m.

$$\mathcal{P} = \{P_\theta,\ \theta \in \Omega\}.$$

Suppose that $\Omega = \Omega_K \cup \Omega_H$, where $\Omega_K \cap \Omega_H = \emptyset$.

$$\text{Hypotheses} \quad \begin{cases} H: & \theta \in \Omega_H \quad \text{(null)} \\ K: & \theta \in \Omega_K \quad \text{(alternative)} \end{cases}$$

Non-random test
Divide sample space S as: $S = S_0 \cup S_1$, where $S_0 \cap S_1 = \emptyset$.
Accept H if $X \in S_0$.

Reject H if $X \in S_1$.
S_1 is called the **critical region** (or the **rejection region**).
The **power** of the test is defined (for all θ) as:

$$\beta(\theta) = P_\theta(X \in S_1) = P_\theta(\text{reject } H).$$

The test is said to have significance **level** α if

$$\beta(\theta) \leq \alpha, \qquad \forall \theta \in \Omega_H.$$

In contrast to level, the test is said to have **size** α if this is the maximum power over the null space:

$$\sup_{\theta \in \Omega_H} \beta(\theta) = \alpha.$$

(In continuous settings "size" and "level" are synonymous – it's only in discrete situations that we make a distinction.)

Ideally we would like

$$\begin{aligned} P_\theta(X \in S_1) &= 0, \qquad \forall \theta \in \Omega_H \text{ (probability of Type I error)}, \\ P_\theta(X \in S_0) &= 0, \qquad \forall \theta \in \Omega_K \text{ (probability of Type II error)}. \end{aligned}$$

However, in general such an ideal test is impossible to construct, and so we search instead for a Uniformly Most Powerful (UMP) test.

Randomized test
If $X = x$ is observed, we toss a coin with $P(\text{Head}) = \phi(x) \in [0, 1]$. If the coin lands Head we reject H, otherwise we accept H. Note therefore that $\text{Head}|X \sim \text{Bern}(\phi(X))$, where:

$$\phi(x) = P_\theta(\text{Head} \mid X = x),$$

is called the **critical function**. If $\phi(x) \in \{0, 1\}$, then we are back in the non-random case with:

$$S_1 = \{x : \ \phi(x) = 1\}, \qquad \text{and} \qquad S_0 = \{x : \ \phi(x) = 0\}.$$

The probability of rejection (of H) by the randomized test is thus:

$$P_\theta(\text{Head}) = E_\theta(\text{Head}) = E_\theta\left[E_\theta(\text{Head} \mid X)\right] = E_\theta\phi(X) = \beta(\theta).$$

Problem: Choose $\phi(\cdot)$ to maximize the power

$$\beta_\phi(\theta) = E_\theta\phi(X), \qquad \forall \theta \in \Omega_K,$$

subject to the level α test constraint:

$$\beta_\phi(\theta) \leq \alpha, \qquad \forall \theta \in \Omega_H.$$

DEFINITION 7.1.1 (UMP test). A test ϕ is UMP of level α if the following two conditions are satisfied.

(i) $\beta_\phi(\theta) \leq \alpha, \ \forall \theta \in \Omega_H$. (The test has level α.)

(ii) $\beta_\phi(\theta) \geq \beta_{\phi'}(\theta)$, $\forall \theta \in \Omega_K$, and for every critical function ϕ' such that $\beta_{\phi'}(\theta) \leq \alpha$, $\forall \theta \in \Omega_H$. (The power of the test is at least as large as that of any other level α test.)

7.2. The Neyman-Pearson Lemma

A class of distributions is called *simple* if it contains a single distribution; otherwise it is said to be *composite*. The solution ϕ to the problem stated above of maximizing the power subject to being of level α, is given by the Neyman-Pearson (NP) Lemma if K is simple.

THEOREM 7.2.1 (Neyman-Pearson Lemma). *Suppose Ω_0 and Ω_1 are simple, consisting of the probabilty measures P_0 and P_1, respectively, with corresponding densities $p_0 = dP_0/d\mu$ and $p_1 = dP_1/d\mu$, with respect to dominating measure μ (e.g., take $\mu = P_0 + P_1$). Then, defining $A_\mu := \mu\{x : p_1(x) = kp_0(x)\}$, we have the following results.*

Existence & Sufficiency: *For $0 \leq \alpha \leq 1$, there exists a test ϕ and a constant k such that:*

(i) *$E_0\phi = \alpha$, (i.e., the test has size α).*

(ii) *The test is a <u>likelihood ratio test</u> given by*

$$\phi(x) = \begin{cases} 1, & p_1(x) > kp_0(x), \\ 0, & p_1(x) < kp_0(x), \\ \gamma, & p_1(x) = kp_0(x) \text{ and } A_\mu \neq 0, \end{cases}$$

where $0 \leq \gamma \leq 1$ is an arbitrary constant.

(iii) *$E_1\phi \geq E_1\phi'$, for every test ϕ' satisfying $E_0\phi' \leq \alpha$.*

Necessity: *If ϕ^* is a UMP level α test, then ϕ^* satisfies (ii) for some k, a.e. μ. It also satisfies (i) unless there is a test of size less than α with a power of 1.*

PROOF. *Existence & Sufficiency.* If $\alpha = \{0, 1\}$, choose $k = \{\infty, 0\}$, respectively. If P_0 and P_1 are mutually singular (the intersection of their supports has μ-measure zero), then taking $k = 0$ gives $\phi(x) = 1$ if $p_1(x) > 0$, and we can set $\phi(x) = \alpha$ where $p_0(x) > 0$. Thus the result of the Lemma follows since: (i) $E_0\phi = E_0\alpha = \alpha$; (ii) $\phi(x) = 1$ if $p_1(x) > 0$ and $\phi(x) = 0$ if $p_1(x) < 0$; (iii) $E_1\phi = 1 \geq E_1\phi'$ for any other ϕ'.

It remains to consider the case $0 < \alpha < 1$ and $\mu(x : p_0(x)p_1(x) > 0) > 0$. Let

$$\begin{aligned} k &= \inf\left\{k' : P_0\left(\frac{p_1(X)}{p_0(X)} \geq k'\right) \geq \alpha \geq P_0\left(\frac{p_1(X)}{p_0(X)} > k'\right)\right\} \\ &= \inf\left\{k' : P_0\left(\frac{p_1(X)}{p_0(X)} < k'\right) \leq 1 - \alpha \leq P_0\left(\frac{p_1(X)}{p_0(X)} \leq k'\right)\right\}, \end{aligned}$$

and note that $0 \leq k < \infty$. Analogously to A_μ, let $A_j := P_j(\{x : p_1(x) = kp_0(x)\})$ for $j = 0, 1$. Define

$$\phi(x) = \begin{cases} 1, & p_1(x) > kp_0(x), \\ 0, & p_1(x) < kp_0(x), \\ \frac{\alpha - P_0(p_1 > kp_0)}{A_0}, & p_1(x) = kp_0(x) \text{ and } A_0 \neq 0, \\ 0, & p_1(x) = kp_0(x) \text{ and } A_0 = 0. \end{cases}$$

Observe now that (i) follows because, from the above def. of k,

$$E_0\phi(X) = \begin{cases} P_0\left(\frac{p_1(X)}{p_0(X)} > k\right) = P_0\left(\frac{p_1(X)}{p_0(X)} \geq k\right) = \alpha, & A_0 = 0 \\ P_0\left(\frac{p_1(X)}{p_0(X)} > k\right) + \left[\frac{\alpha - P_0(p_1 > kp_0)}{A_0}\right] A_0 = \alpha, & A_0 \neq 0. \end{cases}$$

Since the test is clearly of the form given by (ii), it remains to show (iii). To this end, suppose ϕ' is such that $E_0\phi' \leq \alpha$. Since $0 \leq \phi' \leq 1$, we have

$$\begin{aligned} \phi - \phi' > 0 &\implies \phi > 0 \implies p_1 - kp_0 \geq 0, \quad \text{[by 1st and 3rd branches of } \phi\text{]}, \\ \phi - \phi' < 0 &\implies \phi \neq 1 \implies p_1 - kp_0 \leq 0, \quad \text{[by 2nd and 3rd branches of } \phi\text{]}, \end{aligned}$$

Thus $(\phi - \phi')(p_1 - kp_0) \geq 0$, whence

$$0 \leq \int (\phi - \phi')(p_1 - kp_0)d\mu = E_1\phi - E_1\phi' - k(E_0\phi - E_0\phi') \leq E_1\phi - E_1\phi', \tag{7.2.1}$$

since $E_0\phi = \alpha$ and $E_0\phi' \leq \alpha$ implies $k(E_0\phi - E_0\phi') \geq 0$. Therefore $E_1\phi \geq E_1\phi'$.

Necessity. If ϕ^* is a UMP level α test, then from (7.2.1)

$$0 \leq \int (\phi - \phi^*)(p_1 - kp_0)d\mu,$$

with equality holding only if $(\phi - \phi^*)(p_1 - kp_0) = 0$ μ-a.e., since the integrand is non-negative. This implies ϕ^* must satisfy (ii) μ-a.e. (Note that if ϕ^* has size smaller than α, then ϕ^* can be increased until either the size equals α or the power equals 1.) □

REMARK 7.2.2. UMP tests are determined uniquely up to sets of μ-measure 0 by (i) and (ii), provided $A_\mu = 0$. If $A_\mu = 0$, then the UMP test is non-random. If $A_\mu > 0$, then the UMP test can be randomized by choosing ϕ to be constant ($= \gamma$) on the boundary set A_μ. However, any ϕ will do provided the test has size α.

COROLLARY 7.2.3. *Let β be the power of a UMP level α test for testing P_0 vs. P_1, with $0 < \alpha < 1$. Then, $\alpha < \beta$ unless $P_0 = P_1$.*

PROOF. Take $\phi^*(x) \equiv \alpha$. Then $E_1\phi^* = \alpha \leq \beta$, by def. of UMP. If $\alpha = \beta < 1$, then ϕ^* is UMP and must satisfy (ii). Therefore $p_0(x) = kp_1(x)$ for every x, i.e., $P_0 = P_1$ (must have $k = 1$, otherwise p_0 will not integrate to 1). □

Geometric interpretation

For testing P_0 vs. P_1 via the NP Lemma, define $\mathcal{N} = (\alpha, \beta)$ such that $\exists$ a test ϕ with

$\alpha = E_0\phi(X)$ and $\beta = E_1\phi(X)$. Then, obviously $\mathcal{N} \subset [0,1] \times [0,1]$, and it can be shown that:

(i) $\mathcal{N}$ is convex,
(ii) both $(0,0)$ and $(1,1)$ are in $\mathcal{N}$,
(iii) $\mathcal{N}$ is symmetric about $(1/2, 1/2)$, so that $(\alpha, \beta) \in \mathcal{N} \implies (1-\alpha, 1-\beta) \in \mathcal{N}$,
(iv) $\mathcal{N}$ is a closed set.

Plotting α vs. β, we see that $\mathcal{N}$ describes a convex set extending from $(0,0)$ to $(1,1)$, centered at $(1/2, 1/2)$. For a given level α_0, the level α_0 tests are represented by the portion of $\mathcal{N}$ to the left of the vertical line $\alpha = \alpha_0$ (shaded region). The UMP test (tests) is (are) the single point (line) with largest β value in the shaded region at $\alpha = \alpha_0$.

Geometric proof of Corollary
Clearly $\beta \geq \alpha$ since $\phi(x) = \alpha$ for every x is an α level test with power α. If $\exists \alpha_0$ for which the level α_0 UMP test has power α_0, then by convexity and symmetry, $\mathcal{N}$ is the line segment joining $(0,0)$ and $(1,1)$. Therefore $\int \phi dP_0 = \int \phi dP_1$ for every test function ϕ, which implies $P_0 = P_1$.

The NP Lemma can usually be invoked to find a general one-sided UMP test for composite hypotheses, as the next (classical) example shows.

EXAMPLE 7.2.4 (UMP one-sided test for normal mean). Consider the single obs $X \sim \mathrm{N}(\mu, \sigma^2)$, where σ^2 is known. To find the UMP test of $H : \mu = 0$ vs. $K : \mu = \mu_2 > 0$, note that $A_\mu = 0$, so that the UMP is non-random, and is given by the NP Lemma as

$$\phi(x) = \begin{cases} 1, & \frac{p_2(x)}{p_0(x)} > k \\ 0, & \frac{p_2(x)}{p_0(x)} < k \end{cases} = \begin{cases} 1, & \exp\left\{\frac{\mu_2 x}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}\right\} > k \\ 0, & \exp\left\{\frac{\mu_2 x}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}\right\} < k \end{cases} = \begin{cases} 1, & x > k' \\ 0, & x < k' \end{cases}$$

where the last equality follows by the equivalence of events

$$\left\{\frac{p_2(x)}{p_0(x)} > k\right\} \Longleftrightarrow \{x > k'\}.$$

The cutoff k' is found by requiring the test to have size α:

$$\alpha = P_0(X > k') = P\left(Z > \frac{k'-0}{\sigma}\right), \qquad \Longrightarrow \quad k' = \sigma z_{1-\alpha},$$

where $Z \sim \mathrm{N}(0,1)$ and $z_{1-\alpha}$ is its $1-\alpha$ quantile, i.e., $\Phi(z_{1-\alpha}) = 1-\alpha$. Noting that the test did not require specific knowledge of μ_2, only that $\mu_2 > 0$, we can in fact conclude that the level α UMP test for $H : \mu = 0$ vs. $K_2 : \mu > 0$, rejects when $x > \sigma z_{1-\alpha}$. Its power function is

$$\beta_2(\mu) = P_2\left(X > \sigma z_{1-\alpha}\right) = 1 - \Phi(z_{1-\alpha} - \mu/\sigma).$$

Similarly, the test that rejects for $x < \sigma z_\alpha$ is UMP level α for $H : \mu = 0$ vs. $K_1 : \mu < 0$. Its power function is

$$\beta_1(\mu) = P_1\left(X < \sigma z_\alpha\right) = \Phi(z_\alpha - \mu/\sigma).$$

REMARK 7.2.5 (Nonexistence of two-sided UMP). UMP tests typically do not exist for two-sided alternatives. E.g., consider testing $H : \mu = 0$ vs. $K : \mu \neq 0$, a pair of hypotheses with a simple null, in the previous example. Sketching the power function $\beta_1(\mu)$ over all $\mu \in \mathbb{R}$, note that

$$\lim_{\mu \downarrow -\infty} \beta_1(\mu) = 1, \qquad \text{and} \qquad \lim_{\mu \uparrow \infty} \beta_1(\mu) = 0,$$

and is monotone decreasing betwen these two endpoints. Similarly, $\beta_2(\mu)$ is monotone increasing between the endpoints:

$$\lim_{\mu \downarrow -\infty} \beta_2(\mu) = 0, \qquad \text{and} \qquad \lim_{\mu \uparrow \infty} \beta_2(\mu) = 1.$$

By the necessity part of the NP Lemma, a UMP test for K would therefore have to coincide with $\beta_1(\mu)$ for $\mu < 0$ and $\beta_2(\mu)$ for $\mu > 0$, but neither of these two is UMP over all of $\mathbb{R}$. (The power function of the obvious test that rejects when either $x < \sigma z_{\alpha/2}$ or $x > \sigma z_{1-\alpha/2}$ is below each of these over their respective optimal regions.) Thus, no UMP test exists here.

7.3. P-Values

See §3.3 of TSH.

7.4. Monotone Likelihood Ratio

We saw in Example 7.2.4 that we can sometimes extend the NP simple hypotheses results to a composite one, which hold for all $\theta \in K$. This will now be seen to be an instance of a general result that holds whenever the family of measures $\{P_\theta\}$ has a *monotone likelihood ratio* (MLR).

DEFINITION 7.4.1 (MLR). The family $\mathcal{P} = \{p_\theta := dP_\theta/d\mu : \theta \in \Omega \subset \mathbb{R}\}$ has MLR in $T(\cdot)$ (usually a sufficient statistic) if $\forall \theta_1 < \theta_2$ there exists a non-decreasing function h_{θ_1,θ_2} of $T(\cdot)$ such that

$$\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} = h_{\theta_1,\theta_2}(T(x)), \qquad \text{on the set } A_+(\theta_1,\theta_2) = \{x : p_{\theta_2}(x)p_{\theta_1}(x) > 0\}.$$

(Families with non-increasing MLR may be treated by symmetry by reparametrizing, $\phi := -\theta$, which has the effect of reversing the inequalities in the next theorem; see Remark 7.4.5.)

EXAMPLE 7.4.2 (Uniform). $\boldsymbol{X} = (X_1, \ldots, X_n)$, where $X_1, \ldots, X_n \sim$ iid $\mathrm{U}(0, \theta)$. Then,

$$p_\theta(\boldsymbol{x}) = \frac{1}{\theta^n} I_{(0,\theta)}(x_{(1)}) I_{(0,\theta)}(x_{(n)}) = \frac{1}{\theta^n} I_{(0,x_{(n)})}(x_{(1)}) I_{(0,\theta)}(x_{(n)})$$

whence we see that $T(\boldsymbol{x}) = x_{(n)}$ is sufficient. Hence for $\theta_1 < \theta_2$,

$$\frac{p_{\theta_2}(\boldsymbol{x})}{p_{\theta_1}(\boldsymbol{x})} = \begin{cases} \left(\frac{\theta_1}{\theta_2}\right)^n, & 0 < x_{(n)} < \theta_1, \\ +\infty, & \theta_1 < x_{(n)} < \theta_2. \end{cases}$$

Since this ratio is constant over $A_+(\theta_1, \theta_2) = \{\boldsymbol{x} : 0 < x_{(n)} < \theta_1\}$, $\{p_\theta\}$ has MLR in $x_{(n)}$.

EXAMPLE 7.4.3 (One-parameter exponential family).

$$p_\theta(\boldsymbol{x}) = \exp\{\theta T(\boldsymbol{x}) - A(\theta)\}h(\boldsymbol{x}).$$

For $\theta_1 < \theta_2$, we have that

$$\frac{p_{\theta_2}(\boldsymbol{x})}{p_{\theta_1}(\boldsymbol{x})} = \exp\{(\theta_2 - \theta_1)T(\boldsymbol{x}) - [A(\theta_1) - A(\theta_2)]\},$$

which is increasing in $T(\boldsymbol{x})$, and thus the family has MLR in $T(\boldsymbol{x})$.

The most important result under MLR is the following theorem, which states that there is a one-sided UMP composite hypotheses test.

THEOREM 7.4.4 (One-sided UMP test under MLR). *Suppose $\{p_\theta\}$ has MLR in T. Then we have the following results.*

(i) *For testing $H : \theta \le \theta_0$ vs. $K : \theta > \theta_0$, there exists a UMP level α test, given by*

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

where $-\infty \le c \le \infty$ and $0 \le \gamma \le 1$ are determined by the level α constraint:

$$E_{\theta_0}\phi(\boldsymbol{X}) = \alpha.$$

(ii) *The power function $\beta(\theta) = E_\theta\phi(\boldsymbol{X})$ is strictly increasing on the set*

$$\{\theta : 0 < \beta(\theta) < 1\}.$$

(iii) *For any $\theta < \theta_0$, the test ϕ minimizes the Type I error, i.e., it minimizes $\beta(\theta)$ among all tests ϕ' satisfying*

$$E_{\theta_0}\phi'(\boldsymbol{X}) = \alpha. \qquad (\textit{i.e., } E_\theta\phi \le E_\theta\phi',\ \forall\theta < \theta_0.)$$

PROOF. We will consider only the case $0 < \alpha < 1$.

(i) Letting $c = \inf\{c' : P_{\theta_0}(T > c') < \alpha\}$, it is clear that

$$P_{\theta_0}(T > c) \le \alpha \le P_{\theta_0}(T \ge c).$$

Let

$$\gamma = \begin{cases} \frac{\alpha - P_{\theta_0}(T>c)}{P_{\theta_0}(T=c)}, & \text{if } P_{\theta_0}(T = c) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Using similar arguments to the proof of the NP Lemma, we have, for this (c, γ) pair, that $E_{\theta_0}\phi = \alpha$. Now consider $H : \theta = \theta_0$ vs. $K : \theta = \theta_1$, where $\theta_1 > \theta_0$. We

know from the NP Lemma that a UMP size α test is of the form

$$\phi^*(\boldsymbol{x}) = \begin{cases} 1, & p_{\theta_1}(\boldsymbol{x}) > kp_{\theta_0}(\boldsymbol{x}), \\ \gamma, & p_{\theta_1}(\boldsymbol{x}) = kp_{\theta_0}(\boldsymbol{x}) \\ 0, & p_{\theta_1}(\boldsymbol{x}) < kp_{\theta_0}(\boldsymbol{x}), \end{cases}$$

where k was defined in the proof. But ϕ is UMP for $H : \theta = \theta_0$ vs. $K : \theta = \theta_1$, since $E_{\theta_0}\phi(\boldsymbol{x}) = \alpha$ and,

$$\begin{aligned} \frac{p_{\theta_1}(\boldsymbol{x})}{p_{\theta_0}(\boldsymbol{x})} &> h_{\theta_1,\theta_2}(T(\boldsymbol{x})) \equiv h_{\theta_1,\theta_2}(c) \iff T(\boldsymbol{x}) > c, \\ \frac{p_{\theta_1}(\boldsymbol{x})}{p_{\theta_0}(\boldsymbol{x})} &< h_{\theta_1,\theta_2}(T(\boldsymbol{x})) \equiv h_{\theta_1,\theta_2}(c) \iff T(\boldsymbol{x}) < c, \end{aligned}$$

which follows from the fact that $h_{\theta_1,\theta_2}(\cdot)$ is monotone increasing in $T(\boldsymbol{x}) \equiv c$. Thus,

$$\phi(\boldsymbol{x}) = \begin{cases} 1, & \frac{p_{\theta_1}(\boldsymbol{x})}{p_{\theta_0}(\boldsymbol{x})} > h_{\theta_1,\theta_2}(c), \\ \gamma, & \frac{p_{\theta_1}(\boldsymbol{x})}{p_{\theta_0}(\boldsymbol{x})} = h_{\theta_1,\theta_2}(c), \\ 0, & \frac{p_{\theta_1}(\boldsymbol{x})}{p_{\theta_0}(\boldsymbol{x})} < h_{\theta_1,\theta_2}(c), \end{cases}$$

whence $\phi = \phi^*$ and $h_{\theta_1,\theta_2}(c) = k$. But, since c depends only on θ_0, ϕ is independent of θ_1 and is therefore UMP for $H : \theta = \theta_0$ vs. $K : \theta > \theta_0$. We next extend this to $K : \theta \leq \theta_0$. By the NP Lemma, note that ϕ is UMP for $H : \theta = \theta_1$ vs. $K : \theta = \theta_2$, for any $\theta_2 > \theta_1$, al level $\alpha' = E_{\theta_1}\phi = \beta(\theta_1)$. By Corollary 7.2.3, we then have that

$$\beta(\theta_2) > \alpha' = \beta(\theta_1), \qquad \text{provided } \beta(\theta_1) < 1, \tag{7.4.1}$$

whence $\beta(\theta) \leq \beta(\theta_0) = \alpha$, for every $\theta \leq \theta_0$. Finally, since ϕ maximizes $\beta(\theta)$ for each $\theta > \theta_0$ subject to $E_{\theta_0}\phi \leq \alpha$, it also does so subject to the more stringent condition: $E_\theta\phi \leq \alpha$ for every $\theta \leq \theta_0$. Hence ϕ is UMP level α for $H : \theta \leq \theta_0$ vs. $K : \theta > \theta_0$.

(ii) This was shown in (7.4.1) above.

(iii) For $\theta' < \theta_0$, $1 - \phi$ is a UMP test of $H : \theta = \theta_0$ vs. $K : \theta = \theta'$. Consequently, if $E_{\theta_0}\phi' = \alpha$, then $E_{\theta_0}(1 - \phi') = 1 - \alpha = E_{\theta_0}(1 - \phi)$, which implies, since $1 - \phi$ is UMP, that $E_{\theta'}(1 - \phi) \geq E_{\theta'}(1 - \phi')$, whence $E_{\theta'}\phi \leq E_{\theta'}\phi'$.

□

REMARK 7.4.5. We can make the following remarks concerning this theorem.

- To test $H : \theta \geq \theta_0$ vs. $K : \theta < \theta_0$, just reverse the inequalitites in the definition of $\phi(\boldsymbol{x})$ in (i).
- For testing $H : \theta \leq \theta_0$ vs. $K : \theta > \theta_0$ as in the theorem, analogous results hold if the likelihood ratio $p_{\theta_2}/p_{\theta_1}$ in the definition of MLR is instead non-increasing; again we just reverse the inequalitites in the definition of $\phi(\boldsymbol{x})$. Thus to have a one-sided UMP test, we just need the likelihood ratio to be **monotone**.

Under a one-parameter exponential family, we have the following consequence of this theorem.

COROLLARY 7.4.6. *Let $\theta \in \mathbb{R}$ and suppose $\boldsymbol{X}$ has density (w.r.t. a dominating measure) that is a one-parameter exponential family of the form:*

$$p_\theta(\boldsymbol{x}) = \exp\{Q(\theta)T(\boldsymbol{x}) - B(\theta)\}h(\boldsymbol{x}),$$

where $Q(\theta)$ is strictly monotone. Then, the UMP test of $H : \theta \leq \theta_0$ vs. $K : \theta > \theta_0$ is given by one of the following two cases.

Case $Q(\theta) \uparrow$ (increasing):

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

Case $Q(\theta) \downarrow$ (decreasing):

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

In each case, c and γ are determined by the level α constraint:

$$E_{\theta_0}\phi(\boldsymbol{X}) = \alpha.$$

For a UMP test of $H : \theta \geq \theta_0$ vs. $K : \theta < \theta_0$, just reverse all the above inequalitites.

EXAMPLE 7.4.7. $X_1, \ldots, X_n \sim$ iid Gamma(θ, λ), where $\theta > 0$ and $\lambda > 0$ are respectively, the shape and rate (inverse of scale) parameters, with density

$$f(x; \theta, \lambda) = \frac{\lambda^\theta}{\Gamma(\theta)} x^{\theta-1} e^{-\lambda x} I(x > 0), \qquad E(X) = \frac{\theta}{\lambda}.$$

If λ is known, the objective is to find the UMP test of $H : \theta \geq 1$ vs. $K : \theta < 1$. The density of the sample (likelihood) is seen to be a one-parameter exponential family:

$$L(\theta) = \exp\left\{\theta t(\boldsymbol{x}) - n[\log\Gamma(\theta) - \theta\log\lambda]\right\} h(\boldsymbol{x}), \qquad t(\boldsymbol{x}) = \sum \log x_i.$$

Since $Q(\theta) = \theta$ is increasing, the corollary identifies the UMP level α test as:

$$\phi(\boldsymbol{x}) = \begin{cases} 1, & t(\boldsymbol{x}) < c, \\ 0, & t(\boldsymbol{x}) > c, \end{cases}$$

where c solves

$$\alpha = E_{\theta_0}\phi(\boldsymbol{X}) = P_{\theta=1}(T < c) = P_{\theta=1}\left(-\sum \lambda \log(\lambda X_i) > c'\right).$$

Now, when $\theta = 1$, $f(x; \theta = 1) = \lambda e^{-\lambda x} I(x > 0) \sim \text{Exp}(\lambda)$, and it can be shown that

$$Y = -\lambda\log(\lambda X) \sim \text{Gumbel}(\underbrace{\mu = 0}_{\text{location}}, \underbrace{\sigma = \lambda}_{\text{scale}}),$$

so that the appropriate quantile c' can be found from the cdf of $\sum Y_i$ (which seems to be non-standard, but at least the mgf can be computed and quantiles obtained by inverting it via a saddlepoint approximation), or c can be found by Monte Carlo simulation directly from T.

A Decision-Theoretic Formulation

We can place the hypothesis testing problem on a decision-theoretic formulation, akin to the point estimation problem. For ϕ which tests

$$H : \theta \leq \theta_0 \quad \text{vs.} \quad K : \theta > \theta_0,$$

there are two possible decisions: $d_0 = \{\text{accept } H\}$, or $d_1 = \{\text{reject } H\}$. We can therefore define corresponding loss functions:

$$\begin{aligned} L_0(\theta) := L_0(\theta, d_0) &= \text{loss incurred when } \theta \text{ is the truth and we accept } H, \\ L_1(\theta) := L_1(\theta, d_0) &= \text{loss incurred when } \theta \text{ is the truth and we reject } H, \end{aligned}$$

so that the (total) loss is

$$L(\theta, \phi) = L_0(\theta)(1 - \phi) + L_1(\theta)\phi.$$

We can now define the risk in the usual way as expected loss:

$$R(\theta, \phi) = EL(\theta, \phi) = L_0(\theta)(1 - E_\theta\phi) + L_1(\theta)E_\theta\phi.$$

DEFINITION 7.4.8 (Inadmissible test). A test ϕ is *inadmissible* if $\exists\phi'$ such that

$$\begin{aligned} R(\theta, \phi') &\leq R(\theta, \phi), \qquad \forall\theta \\ R(\theta, \phi') &< R(\theta, \phi), \qquad \text{for some } \theta. \end{aligned}$$

That is, ϕ is inadmissible if $\exists\phi'$ which dominates ϕ. A test ϕ is *admissible* if its not inadmissible.

DEFINITION 7.4.9 (Complete classes). A class $\mathcal{C}$ of tests is *complete* if $\forall\phi \notin \mathcal{C}$, $\exists\phi' \in \mathcal{C}$ such that ϕ is dominated by ϕ'.

A complete class is *minimal* if it does not contain a proper complete subclass. (If a minimal complete class exists, it consists of precisely the admissible tests.)

A class $\mathcal{C}$ is *essentially complete* if $\forall\phi \notin \mathcal{C}$, $\exists\phi' \in \mathcal{C}$ which is "at least as good" as ϕ, i.e., $R(\theta, \phi') \leq R(\theta, \phi)$, $\forall\theta$. Such a class is *minimal* if it does not contain a proper essentially complete subclass.

The point is that if there is a (minimal) essentially complete class, then one need not bother with considering tests outside of this class.

THEOREM 7.4.10. *Under the setting and assumptions of Theorem 7.4.4, let $\mathcal{C}$ be the class consisting of all tests of the form given by (i) of that theorem. If*

$$\begin{aligned} L_1(\theta) - L_0(\theta) &> 0, \qquad \text{for } \theta < \theta_0, \\ L_1(\theta) - L_0(\theta) &< 0, \qquad \text{for } \theta > \theta_0, \end{aligned}$$

then we have the following results.

(i) *$\mathcal{C}$ is essentially complete.*
(ii) *If additionally the set $\{x : p_\theta(x) > 0\}$ is independent of θ, $\mathcal{C}$ is minimal essentially complete.*

PROOF. For any given test ϕ', let $\alpha = E_{\theta_0}\phi'$. Choose ϕ as in Theorem 7.4.4, i.e., $E_{\theta_0}\phi = \alpha$ with

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

Then, $E_\theta\phi \le E_\theta\phi'$, $\forall\theta < \theta_0$ (has smaller Type I error), and $E_\theta\phi \ge E_\theta\phi'$, $\forall\theta > \theta_0$ (is UMP). Hence

$$\begin{aligned} EL(\theta) := R(\theta,\phi) &= L_1(\theta)P(\text{reject } H) + L_0(\theta)P(\text{accept } H) \\ &= L_1(\theta)E_\theta\phi + L_0(\theta)(1 - E_\theta\phi) \\ &= L_0(\theta) + E_\theta\phi\,(L_1(\theta) - L_0(\theta)) \\ &\le L_0(\theta) + E_\theta\phi'\,(L_1(\theta) - L_0(\theta)) \\ &= R(\theta,\phi'), \end{aligned}$$

where the $\le$ part follows by the assumptions on L_1 and L_0 in the statement of the theorem. □

7.5. Confidence Bounds

UMP one-sided tests can be used to derive upper and lower confidence bounds (CBs). As we will see, inverting a UMP test leads to UMA confidence CBs (defined next). Since lower and upper bounds are analogous, it suffices to focus our attention, say, on lower bounds, $\underline{\theta}$.

DEFINITION 7.5.1 (UMA lower CB). We define the following based on sample data X.

(i) $\underline{\theta}(X)$ is a $(1-\alpha)$ *lower confidence bound* for θ if:

$$P_\theta(\underline{\theta}(X) \le \theta) \ge 1 - \alpha, \quad \forall\theta.$$

(The idea is that $\underline{\theta}$ falls below θ with a specified high probability of at least $1-\alpha$.)

(ii) The *confidence coefficient* or *confidence level* for $\underline{\theta}(X)$ is defined to be

$$\inf_{\theta} P_\theta(\underline{\theta}(X) \leq \theta).$$

(This usually turns out to be $1-\alpha$.)

(iii) $\underline{\theta}$ is a *uniformly most accurate* (UMA) lower CB for θ with confidence level $(1-\alpha)$, if, in addition to (i), we have

$$P_\theta(\underline{\theta}(X) \leq \theta') \leq P_\theta(\underline{\theta}^*(X) \leq \theta'), \quad \forall \theta' < \theta,$$

and for every other lower CB $\underline{\theta}^*(X)$ satisfying (i). (The idea is that we want to underestimate θ by as little as possible.)

Our aim is to find a lower CB for θ which falls below θ with high probability ($\geq 1-\alpha$), but not too far below. Excessive underestimation can be assessed via a loss function. Suppose the following conditions hold

$$\begin{aligned} L(\theta, \underline{\theta}) &= 0, \qquad \text{if } \underline{\theta} > \theta, \\ L(\theta, \underline{\theta}) &\geq 0, \qquad \forall\ \underline{\theta} \leq \theta, \\ L(\theta, \underline{\theta}) &\geq L(\theta, \underline{\theta}'), \quad \text{if } \underline{\theta} \leq \underline{\theta}' \leq \theta. \end{aligned} \tag{7.5.1}$$

Problem: Minimize the risk $E_\theta L(\theta, \underline{\theta})$, subject to

$$P_\theta(\underline{\theta}(X) \leq \theta) \geq 1-\alpha. \tag{7.5.2}$$

Solution: An UMA lower CB minimizes the risk subject to (7.5.2).(See Problem 3.44 in TSH.)

Thus the determination of an UMA lower CB also solves the more general problem formulated in terms of **any** loss function satisfying (7.5.1). Finding UMA CBs is faciltated by introducing the following concept.

DEFINITION 7.5.2 (Confidence Sets). A family of subsets $S(x)$ of Ω, where $x \in \mathcal{X}$, is said to be a family of *confidence sets* at confidence level $(1-\alpha)$, if

$$P_\theta(\theta \in S(X)) \geq 1-\alpha, \quad \forall \theta \in \Omega.$$

Thus, the random set $S(X)$ covers the true parameter with probability at least $(1-\alpha)$.

EXAMPLE 7.5.3. If $\theta(X)$ is defined as in (i) of Definition 7.5.1, then the sets $S(x) = [\underline{\theta}(x), \infty)$ constitute a family of $(1-\alpha)$-level confidence sets for θ.

The next theorem shows that inverting a UMP test leads to an UMA confidence set.

THEOREM 7.5.4. *For all $\theta_0 \in \Omega$, let $A(\theta_0)$ be the acceptance region of a (non-random) level α test of $H(\theta_0) : \theta = \theta_0$, and let*

$$S(x) = \{\theta :\ x \in A(\theta) \text{ and } \theta \in \Omega\}.$$

Then, we have the following results.

(i) *For $x \in \mathcal{X}$, $S(x)$ is a family of level $(1-\alpha)$ confidence sets for θ.*
(ii) *If, for all θ_0, $A(\theta_0)$ is the acceptance region of a level α UMP test of $H(\theta_0)$ vs. the alternative $K(\theta_0)$, then the corresponding confidence set $S(x)$, minimizes*

$$P_\theta(\theta_0 \in S(X)), \qquad \forall \theta \in K(\theta_0),$$

among all $(1-\alpha)$ level families of confidence sets for θ.

PROOF. (i) By def., $\theta \in S(x)$ if and only if $x \in A(\theta)$, and therefore

$$P_\theta(\theta \in S(X)) = P_\theta(X \in A(\theta)) \geq 1-\alpha.$$

(ii) If $S^*(x)$ is any other family of level $(1-\alpha)$ confidence sets, then $A^*(\theta) = \{x : \theta \in S^*(x)\}$ defines an α level test of $H(\theta_0)$ vs. $K(\theta_0)$, since

$$P_{\theta_0}(X \in A^*(\theta_0)) = P_{\theta_0}(\theta_0 \in S^*(X)) \geq 1-\alpha.$$

However, $A(\theta_0)$ is UMP, and hence

$$P_\theta(X \in A^*(\theta_0)) = P_\theta(\theta_0 \in S^*(X)) \geq P_\theta(X \in A(\theta_0)) = P_\theta(\theta_0 \in S(X)).$$

□

COROLLARY 7.5.5. *Suppose $\{p_\theta(x),\ \theta \in \Omega\}$ has MLR in $T(x)$, and that the cdf $F_\theta(t)$ of T is marginally continuous in each t and θ (when the other is fixed). Then, we have the following results.*

(i) *For each level $(1-\alpha)$, there exists a UMA lower confidence bound $\underline{\theta}$ for θ.*
(ii) *If $F_\theta(T(x)) = 1-\alpha$ has a solution $\theta = \hat{\theta}$ for each x, then the UMA (lower) bound is unique, and $\underline{\theta} = \hat{\theta}$.*

PROOF. (i) For each θ_0 there exists a $c(\theta_0)$ such that

$$P_{\theta_0}(T > c(\theta_0)) = \alpha. \tag{7.5.3}$$

Moreover, from Theom 7.4.4,

$$\phi(x) = \begin{cases} 1, & \text{if } T(x) > c(\theta_0) \\ 0, & \text{if } T(x) < c(\theta_0) \end{cases}$$

is a UMP level α test for $H(\theta_0)$ vs. $K : \theta > \theta_0$, with

$$\beta(\theta) = E_\theta \phi > \alpha, \qquad \forall \theta > \theta_0,$$

and consequently $P_\theta(T > c(\theta_0)) > \alpha$ for all $\theta > \theta_0$. By def., (7.5.3) holds also for every $\theta > \theta_0$, whence $c(\theta) > c(\theta_0)$, i.e., $c(\cdot)$ is strictly increasing (and continuous by the continuity of $F_\theta(t)$ in θ). Now, set $A(\theta) = \{x : T(x) \leq c(\theta)\}$, $S(x) = \{\theta : x \in A(\theta)\}$, and define $\underline{\theta}(x) = \inf\{\theta : T(x) \leq c(\theta)\}$. Then,

$$\theta \geq \underline{\theta}(x) \iff c(\theta) \geq T(x) \iff x \in A(\theta).$$

Consequently, it follows from (7.5.3) that for every θ, $P_\theta(\underline{\theta}(X) \leq \theta) = P_\theta(T(X) \leq c(\theta)) = 1-\alpha$. Since by Theom 7.5.4 $[\underline{\theta}(x), \infty)$ minimizes $P_\theta(\underline{\theta}(X) \leq \theta')$ for every $\theta' < \theta$, it follows that $\underline{\theta}$ is a UMA lower bound for θ.

(ii) Suppose $0 < F_{\theta_0}(t) < 1$. Then, setting

$$\phi = \begin{cases} 1, & \text{if } T > t \\ 0, & \text{if } T \le t \end{cases}$$

implies $E_{\theta_0}(1-\phi) = P_{\theta_0}(T \le t) = 1-\alpha$, which holds also for every $\theta > \theta_0$ (by the corollary to the Neyman-Pearson Lemma), and this means that $F_\theta(t)$ is a strictly decreasing function of θ at each θ such that $0 < F_\theta(t) < 1$. Consequently, $F_\theta(t) = 1-\alpha$ can have at most one solution: $\theta = \hat{\theta}$. Then $F_{\hat{\theta}}(t) = 1-\alpha$, and (by def.) $c(\hat{\theta}) = t$, so that

$$t \le c(\theta) \quad \Longleftrightarrow \quad c(\hat{\theta}) \le c(\theta) \quad \Longleftrightarrow \quad \hat{\theta} \le \theta.$$

(Follows by part (i) where it was shown $c(\cdot)$ is continuous and strictly increasing.) Setting $t = T(x)$ gives

$$\theta \ge \hat{\theta}(x) \quad \Longleftrightarrow \quad T(x) \le c(\theta) \quad \Longleftrightarrow \quad \theta \ge \underline{\theta}(x),$$

whence it follows that $\hat{\theta} = \underline{\theta}$.

□

EXAMPLE 7.5.6 (Exponential waiting times). If $X_1, \ldots, X_n \sim$ iid $\mathrm{E}(\lambda)$, we wish to derive UMA lower and upper CBs for λ. Since we have a one-parameter exponential family, this is most easily done by invoking first Corollary 7.4.6 to derive a corresponding UMP one-sided test, followed by Corollary 7.5.5 which guarantees an UMA lower/upper CB upon inversion of the (UMP) test. From the pdf of the sample

$$p_\lambda(\boldsymbol{x}) = \lambda^n \exp\{-\lambda T(\boldsymbol{x})\} \prod_i I(x_i > 0), \qquad T(\boldsymbol{x}) = \sum_i x_i \sim \Gamma(n, \lambda),$$

we note that $Q(\lambda) = -\lambda$ is monotone decreasing in λ, and the cdf of T is continuous in both t and λ. (For ease of quantile calculation, we also note that $2\lambda T \sim \chi^2(2n)$.) Thus the UMP test of $H : \lambda \ge \lambda_0$ accepts for $T \le c$, where

$$1-\alpha = P_{\lambda_0}(T \le c) = P(\chi^2(2n) \le 2\lambda_0 c), \quad \Longrightarrow \quad c = \frac{\chi^2_{1-\alpha}(2n)}{2\lambda_0}.$$

Thus,

$$\boldsymbol{x} \in A(\lambda_0) \quad \Longleftrightarrow \quad \sum_i x_i \le \frac{\chi^2_{1-\alpha}(2n)}{2\lambda_0} \quad \Longleftrightarrow \quad \lambda_0 \le \frac{\chi^2_{1-\alpha}(2n)}{2\sum x_i},$$

so that $\chi^2_{1-\alpha}(2n)/(2t)$ is a $(1-\alpha)$ UMA upper CB for λ.

Similarly, $\chi^2_\alpha(2n)/(2t)$ is a $(1-\alpha)$ UMA lower CB for λ, obtained by inverting the UMP test of $H : \lambda \le \lambda_0$, which accepts for $T \ge c$.

UMA Confidence Intervals (CIs)

Lower and upper CBs can be used to construct the more common CI, defined as follows.

DEFINITION 7.5.7 (Confidence Interval). Suppose all of the following hold:

- $\underline{\theta}$ is a lower CB with confidence level $1-\alpha_1$,
- $\overline{\theta}$ is a upper CB with confidence level $1-\alpha_2$,
- $\underline{\theta} < \overline{\theta}$ for every sample point x (occurs if, e.g., $\alpha_1+\alpha_2<1$).

Then, the interval $(\underline{\theta},\overline{\theta})$ is called a *confidence interval* for θ with level $(1-\alpha_1-\alpha_2)$, i.e.,

$$P_\theta(\underline{\theta}\le\theta\le\overline{\theta}) = 1-\alpha_1-\alpha_2, \quad \forall\theta\in\Omega.$$

If $\underline{\theta}$ and $\overline{\theta}$ are UMA, then they minimize the risks under their respective alternatives, $E_\theta L_1(\theta,\underline{\theta})$ and $E_\theta L_2(\theta,\underline{\theta})$, at their respective levels. This is so for any L_1 that is non-increasing in $\underline{\theta}$ for $\underline{\theta}<\theta$ and 0 for $\underline{\theta}\ge\theta$, and for any L_2 that is nondecreasing in $\overline{\theta}$ for $\overline{\theta}>\theta$ and 0 for $\overline{\theta}\le\theta$. Letting

$$L(\theta;\underline{\theta},\overline{\theta}) = L_1(\theta,\overline{\theta}) + L_2(\theta,\underline{\theta}),$$

the CI $(\underline{\theta},\overline{\theta})$ thus minimizes the risk under the alternative, $E_\theta L(\theta;\underline{\theta},\overline{\theta})$, subject to having confidence level $(1-\alpha_1-\alpha_2)$:

$$P_\theta(\underline{\theta}>\theta)\le\alpha_1, \quad \text{and} \quad P_\theta(\overline{\theta}<\theta)\le\alpha_2.$$

Examples of loss functions satisfying these propertis are as follows.

Natural measure:

$$L(\theta;\underline{\theta},\overline{\theta}) = \begin{cases} \overline{\theta}-\underline{\theta}, & \text{if } \underline{\theta}\le\theta\le\overline{\theta}, \\ \overline{\theta}-\theta, & \text{if } \theta<\underline{\theta}, \\ \theta-\underline{\theta}, & \text{if } \theta>\overline{\theta}. \end{cases}$$

Coverage: $L(\theta;\underline{\theta},\overline{\theta}) = \overline{\theta}-\underline{\theta}$.
Weighted distance from ends: $L(\theta;\underline{\theta},\overline{\theta}) = a(\theta-\underline{\theta})^2 + b(\overline{\theta}-\theta)^2$.

EXAMPLE 7.5.8 (Exponential waiting times (continued)). From the lower and upper $(1-\alpha)$ CBs we obtained in Example 7.5.6, it is easy to see that the interval

$$\left(\frac{\chi^2_\alpha(2n)}{2\sum x_i}, \frac{\chi^2_{1-\alpha}(2n)}{2\sum x_i}\right),$$

is a CI with confidence level $(1-2\alpha)$.

7.6. Uniformly Most Powerful Unbiased (UMPU) Tests

In Remark 7.2.5 the real reason there was no UMP two-sided test for the normal mean, is that the power function of the one-sided UMP tests dips below the size of the test in the null space. If this region is to be in the alternative space K for the two-sided test, then we should introduce a constraint that the power function over K must not dip below the size of the test. This is the concept of an **unbiased** test.

DEFINITION 7.6.1 (Unbiased test). A level α test ϕ of $H : \theta \in \Omega_H$ vs. $K : \theta \in \Omega_K$ is *unbiased* if

$$E_\theta \phi \geq \alpha, \quad \forall \theta \in \Omega_K. \tag{7.6.1}$$

(And since ϕ is level α we also have $E_\theta \phi \leq \alpha, \ \forall \theta \in \Omega_H$.)

Clearly any UMP level α test is unbiased, since $\phi'(x) = \alpha \ \forall x$ is a level α test, and so a UMP test ϕ (which has a power function at least as large), must satisfy (7.6.1).

Now, let ω denote the set of parameter points that are on the boundary of H and K, i.e., the set of points θ that are points or limit points of both Ω_H and Ω_K. If $\beta_\phi(\theta) = E_\theta \phi$ is a continuous function of θ, then for $\theta \in \omega$, we must have $\beta_\phi(\theta) = \alpha$. The reason for this is that if θ is a limit point of values $\theta_n \in \Omega_H$ and $\theta'_n \in \Omega_K$, then

$$\beta_\phi(\theta) = \lim_{n\to\infty} \beta_\phi(\theta_n) \leq \alpha, \qquad \text{and} \qquad \beta_\phi(\theta) = \lim_{n\to\infty} \beta_\phi(\theta'_n) \geq \alpha.$$

This embodies the concept of an α-similar test.

DEFINITION 7.6.2 (α-similar test). A test ϕ is *α-similar* on the parameter points ω that are on the boundary of H and K, if

$$\beta_\phi(\theta) = \alpha, \qquad \forall \theta \in \omega. \tag{7.6.2}$$

The importance behind this definition, is that it allows us to establish unbiasedness through the more tractable concept of α-similarity, as the next result shows.

LEMMA 7.6.3 (UMPU test). *Suppose $\Omega_H \cap \Omega_K \subset \mathbb{R}^k$ and $\beta_\phi(\theta)$ is continuous in θ for every test ϕ. If ϕ' is UMP α-similar of level α, then it is UMP unbiased (UMPU) of level α.*

PROOF. Since $\phi \equiv \alpha$ is α-similar, $E_\theta \phi' \geq E_\theta \phi = \alpha$, for every $\theta \in \Omega_K$, and hence ϕ' is unbiased. Now let ϕ be any unbiased level-α test. Since:

$$E_\theta \phi \geq \alpha \quad \forall \theta \in \Omega_K \quad \text{(unbiased)},$$

and

$$E_\theta \phi \leq \alpha \quad \forall \theta \in \Omega_H \quad \text{(level } \alpha\text{)},$$

we must have

$$E_\theta \phi = \alpha \quad \forall \theta \in (\partial\Omega_H \cap \partial\Omega_K)$$

(since for θ on the boundary $\partial\Omega_H$, there exists a sequence $\{\theta_n\} \in \Omega_H$ such that $\theta_n \to \theta$; likewise there exists a sequence $\{\theta'_n\} \in \Omega_K$ such that $\theta'_n \to \theta$; and $E_\theta \phi$ is continuous in θ). Hence ϕ is α-similar, and consequently

$$E_\theta \phi' \geq E_\theta \phi, \quad \forall \theta \in \Omega_K,$$

that is, ϕ' is UMPU of level α. □

One-parameter Exponential Families and Two-sided Tests

Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)$ has a density belonging to the one-parameter exponential family

$$p_\theta(\boldsymbol{x}) = \exp\{\theta T(\boldsymbol{x}) - A(\theta)\}h(\boldsymbol{x}).$$

Then, letting $\theta_2 > \theta_1$, we have the following results concerning the existence of UMP tests.

(i) A UMP test exists for $H : \theta \leq \theta_0$ vs. $K : \theta > \theta_0$, and the reverse situation, by Corollary 7.4.6.

(ii) A UMP test exists for $H : \theta \leq \theta_1$ or $\theta \geq \theta_2$ vs. $K : \theta \in (\theta_1, \theta_2)$, by TSH Theorem 3.7.1.

(iii) A UMP test does NOT exist for $H : \theta_1 \leq \theta \leq \theta_2$ vs. $K : \theta < \theta_1$ or $\theta > \theta_2$, by TSH Problem 3.54. In this case, consider the test

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

where $c_1, c_2 \in \mathbb{R}$ and $0 \leq \gamma \leq 1$ are determined by the level α constraint:

$$E_{\theta_1}\phi(\boldsymbol{X}) = E_{\theta_2}\phi(\boldsymbol{X}) = \alpha.$$

From the results on exponential families in Ch. 1 (Theorem 1.3.13), we know that $\beta_\phi(\theta) = E_\theta\phi$ is continuous in θ on $int(\mathcal{N})$, $\omega = \{\theta_1, \theta_2\}$, and $E_{\theta_1}\phi(\boldsymbol{X}) = E_{\theta_2}\phi(\boldsymbol{X}) = \alpha$, whence it follows that ϕ is α-similar. Now, by TSH Theorem 3.7.1, it follows that $1 - \phi$ is UMP level $(1 - \alpha)$ for $H' : \theta < \theta_1$ or $\theta > \theta_2$ vs. $K' : \theta_1 \leq \theta \leq \theta_2$, and also that $\forall \phi'$ such that $E_{\theta_1}\phi' = E_{\theta_2}\phi' = 1 - \alpha$,

$$E_\theta(1 - \phi) \leq E_\theta(1 - \phi'), \qquad \forall \theta < \theta_1 \text{ or } \theta > \theta_2,$$

whence

$$E_\theta\phi' \leq E_\theta\phi, \qquad \forall \notin [\theta_1, \theta_2].$$

Hence, the test $\phi(\boldsymbol{x})$ defined above is UMPU level α by Lemma 7.6.3.

(iv) A UMP test does NOT exist for $H : \theta = \theta_0$ vs. $K : \theta \neq \theta_0$, by TSH Problem 3.54. In this case, consider the test

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

where $c_1, c_2 \in \mathbb{R}$ and $0 \leq \gamma \leq 1$ are determined by the level α constraint

$$E_{\theta_0}\phi(\boldsymbol{X}) = \alpha, \quad \text{and} \quad E_{\theta_0}\phi(\boldsymbol{X})T(\boldsymbol{X}) = \alpha E_{\theta_0}T(\boldsymbol{X}).$$

Then, $\phi(\boldsymbol{x})$ is UMPU by the argument on pp. 111-113 of TSH.

EXAMPLE 7.6.4 (UMPU two-sided test for normal mean). Let $X_1, \ldots, X_n \sim$ iid $\mathrm{N}(\theta, \sigma^2)$, where σ^2 is known. By the result from case (iv) above, the UMPU test of $H : \theta = \theta_0$ vs. $K : \theta \neq \theta_0$, is given by

$$\phi(\boldsymbol{X}) = \begin{cases} 1, & \bar{X}_n < c_1 \text{ or } \bar{X}_n > c_2 \\ 0, & \text{otherwise,} \end{cases} \qquad \Longleftrightarrow \qquad \phi(Z) = \begin{cases} 1, & Z < z_1 \text{ or } Z > z_2 \\ 0, & \text{otherwise,} \end{cases}$$

where $Z = (\bar{X}_n - \theta_0)/(\sigma/\sqrt{n}) \sim \mathrm{N}(0,1)$ with density $f(z)$ under H, and z_1, z_2 satisfy

$$E_{\theta_0}(\phi) = P\left(Z < z_1\right) + P(Z > z_2) = \alpha \qquad \Longleftrightarrow \qquad \int_{z_1}^{z_2} f(z)dz = 1 - \alpha,$$

and

$$E_{\theta_0}(\phi Z) = \alpha E_{\theta_0}(Z) \quad \Longleftrightarrow \quad E_{\theta_0}[(1-\phi)Z] = (1-\alpha)E_{\theta_0}(Z) \quad \Longleftrightarrow \quad \int_{z_1}^{z_2} z f(z)dz = 0.$$

The first condition states that the interval $[z_1, z_2]$ must enclose an area of $(1-\alpha)$, while the second stipulates that $z_1 < 0$, $z_2 > 0$, with $|z_1| = z_2$ (the integral of an odd function can only be zero if the limits are the same distance apart and on opposite sides of zero). The only values that satisfy these are $z_1 = -z_{1-\alpha/2}$ and $z_2 = z_{1-\alpha/2}$. Thus the UMPU level α test rejects for

$$\bar{X}_n < \theta_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \text{or} \qquad \bar{X}_n > \theta_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

EXAMPLE 7.6.5 (UMPU two-sided test for normal std. deviation). Let $X_1, \ldots, X_n \sim$ iid $\mathrm{N}(0, \sigma^2)$. The density of the sample is

$$p_\sigma(\boldsymbol{X}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{\theta T(\boldsymbol{X})\}, \qquad \theta = -\frac{1}{2\sigma^2},\ T = \sum X_i^2,$$

so that from case (iv) above, the UMPU test of $H : \sigma = \sigma_0$ vs. $K : \sigma \neq \sigma_0$, which is equivalent to $H : \theta = \theta_0$ vs. $K : \theta \neq \theta_0$, accepts for

$$c_1 \leq T \leq c_2 \qquad \Longleftrightarrow \qquad t_1 \leq \frac{T}{\sigma^2} \leq t_2$$

where $T/\sigma_0^2 \sim \chi^2(n)$ with density $f_n(t) \propto t^{n/2-1}e^{-t/2}$ under H, and t_1, t_2 satisfy

(a) $$E_{\sigma_0}(1-\phi) = \int_{t_1}^{t_2} f_n(t)dt = 1 - \alpha,$$

and

$$E_{\sigma_0}[(1-\phi)T] = (1-\alpha)E_{\sigma_0}(T) \qquad \Longleftrightarrow \qquad \int_{t_1}^{t_2} t f_n(t)dt = (1-\alpha)n.$$

This 2nd condition is equivalent to (TSH Problem 4.5)

(b) $$t_1^{n/2}e^{-t_1/2} = t_2^{n/2}e^{-t_2/2}.$$

The system (a) and (b) of two equations in two unknowns can now be solved numerically for t_1, t_2. Alternatively, the equal-tails test with $t_1 = \chi^2_{\alpha/2}(n)$ and $t_2 = \chi^2_{1-\alpha/2}(n)$ provides (by the CLT since T is an empirical sum) a good approximation for large n.

UMPU Tests for Multi-parameter Exponential Families

Here we generalize the above to the situation when only one parameter ($\theta \in \mathbb{R}$) is of interest in an exponential family, while the remaining parameters ($\boldsymbol{\xi} \in \mathbb{R}^k$) comprise a vector of nuisance parameters. We suppose $\boldsymbol{X} = (X_1, \ldots, X_n)$ has a density of the form

$$\frac{dP_{\theta,\boldsymbol{\xi}}}{d\mu}(\boldsymbol{x}) = C(\theta, \boldsymbol{\xi}) \exp\{\theta U(\boldsymbol{x}) + \boldsymbol{\xi} \cdot \boldsymbol{T}(\boldsymbol{x})\},$$

where $(\theta, \boldsymbol{\xi}) \in \Omega$, where the parameter space Ω is convex and contains an open set of $\mathbb{R}^{k+1}$. (The density is in canonical form, but we combine the $A(\cdot)$ and $h(\cdot)$ functions into $C(\cdot)$ for a compact representation.)

We will assume that Ω contains points for which $\theta < \theta_0, \theta_1, \theta_2$, and points for which $\theta > \theta_0, \theta_1, \theta_2$. We are interested in tests of the following types:

(1) $H_1 : \theta \leq \theta_0$ vs. $K_1 : \theta > \theta_0$.
(2) $H_2 : \theta \leq \theta_1$ or $\theta \geq \theta_2$ vs. $K_2 : \theta_1 < \theta < \theta_2$.
(3) $H_3 : \theta_1 \leq \theta \leq \theta_2$ vs. $K_3 : \theta < \theta_1$ or $\theta > \theta_2$.
(4) $H_4 : \theta = \theta_0$ vs. $K_4 : \theta \neq \theta_0$.

Since the sufficient statistics U and $\boldsymbol{T}$ contain all the information in the sample regarding $(\theta, \boldsymbol{\xi})$, we can restrict attention to tests based on them.

Now, by Theorem 1.3.11, note that $(U, \boldsymbol{T})$ have the joint density

$$\frac{dP^{U,\boldsymbol{T}}_{\theta,\boldsymbol{\xi}}}{d\nu}(u, \boldsymbol{t}) = C(\theta, \boldsymbol{\xi}) e^{\theta u + \boldsymbol{\xi} \cdot \boldsymbol{t}},$$

with respect to the measure

$$\nu(B) = \mu\left\{\boldsymbol{x} \ : \ (U(\boldsymbol{x}), \boldsymbol{T}(\boldsymbol{x})) \in B\right\}, \qquad \forall\ B \in \mathcal{B}(\mathbb{R}^{k+1}).$$

The next result shows that the conditional distributions of U given $\boldsymbol{T} = \boldsymbol{t}$ constitute a one-parameter exponential family.

LEMMA 7.6.6 (Distribution of $U|\boldsymbol{T}$). *For any fixed $(\theta_0, \boldsymbol{\xi}_0) \in \Omega$, define*

$$d\nu'_{\boldsymbol{t}}(u) = dP^{U|\boldsymbol{T}=\boldsymbol{t}}_{\theta_0,\boldsymbol{\xi}_0}(u), \qquad d\nu_{\boldsymbol{t}}(u) = e^{-\theta_0 u} d\nu'_{\boldsymbol{t}}(u), \qquad \frac{1}{C_{\boldsymbol{t}}(\theta)} = \int e^{(\theta-\theta_0)u} d\nu'_{\boldsymbol{t}}(u).$$

Then, the distribution of $U|\boldsymbol{T} = \boldsymbol{t}$ constitutes a one-parameter exponential family of the form

$$dP^{U|\boldsymbol{t}}_{\theta}(u) = C_{\boldsymbol{t}}(\theta) e^{\theta u} d\nu_{\boldsymbol{t}}(u),$$

which therefore does not depend on the nuisance parameter $\boldsymbol{\xi}$.

PROOF. Note that since we can write

$$dP^{U,\boldsymbol{T}}_{\theta,\boldsymbol{\xi}}(u, \boldsymbol{t}) = \frac{C(\theta, \boldsymbol{\xi})}{C(\theta_0, \boldsymbol{\xi}_0)} e^{(\theta-\theta_0)u + (\boldsymbol{\xi}-\boldsymbol{\xi}_0)\cdot\boldsymbol{t}} dP^{U,\boldsymbol{T}}_{\theta_0,\boldsymbol{\xi}_0}(u, \boldsymbol{t}),$$

we obtain the density of $U|\boldsymbol{t}$ as the joint divided by the marginal of $\boldsymbol{T}$, using the measures defined above:

$$\begin{aligned} dP_\theta^{U|\boldsymbol{t}}(u) &= \frac{dP_{\theta_0,\boldsymbol{\xi}_0}^{U,\boldsymbol{T}}(u,\boldsymbol{t})}{\int dP_{\theta_0,\boldsymbol{\xi}_0}^{U,\boldsymbol{T}}(du,\boldsymbol{t})} = \frac{e^{(\theta-\theta_0)u}dP_{\theta_0,\boldsymbol{\xi}_0}^{U,\boldsymbol{T}}(u,\boldsymbol{t})}{\int e^{(\theta-\theta_0)u}dP_{\theta_0,\boldsymbol{\xi}_0}^{U,\boldsymbol{T}}(du,\boldsymbol{t})} = \frac{e^{(\theta-\theta_0)u}d\nu'_{\boldsymbol{t}}(u)}{\int e^{(\theta-\theta_0)u}d\nu'_{\boldsymbol{t}}(du)} \\ &= C_{\boldsymbol{t}}(\theta)e^{\theta u}d\nu_{\boldsymbol{t}}(u). \end{aligned}$$

□

We now tackle each of the four cases described above.

Case (1): For every $\alpha \in (0,1)$, there exist constants $c(\boldsymbol{t})$ and $\gamma(\boldsymbol{t})$ such that the test

$$\phi_1(u,\boldsymbol{t}) = \begin{cases} 1, & u > c(\boldsymbol{t}), \\ \gamma(\boldsymbol{t}), & u = c(\boldsymbol{t}), \\ 0, & u < c(\boldsymbol{t}), \end{cases}$$

satisfies

$$(\star) \qquad E_{\theta_0}[\phi_1 \mid \boldsymbol{T}] = \alpha.$$

Moreover, ϕ_1 is UMP level α conditional on $\boldsymbol{T} = \boldsymbol{t}$, i.e., for every ϕ satisfying $(\star)$, we have

$$E_\theta[\phi_1 \mid \boldsymbol{T} = \boldsymbol{t}] \geq E_\theta[\phi \mid \boldsymbol{T} = \boldsymbol{t}], \qquad \forall \theta > \theta_0, \tag{7.6.3}$$

since the distribution of $U|\boldsymbol{T} = \boldsymbol{t}$ is a one-parameter exponential family.

NOTE 7.6.7. The following points should be noted.

- We use the notation $E_\theta[\phi|\boldsymbol{T}]$ instead of $E_{\theta,\boldsymbol{\xi}}[\phi|\boldsymbol{T}]$, since the distribution of $U|\boldsymbol{T}$ is independent of $\boldsymbol{\xi}$ (Lemma 7.6.6.)
- Tests satisfying $(\star)$ are said to have *Neyman structure* (see TSH §4.3).
- Reverse the inequalities in the definition of ϕ_1 if one wishes to test instead $H_1 : \theta \geq \theta_0$ vs. $K_1 : \theta < \theta_0$.

Case (2): Similarly, the test

$$\phi_2(u,\boldsymbol{t}) = \begin{cases} 1, & c_1(\boldsymbol{t}) \leq u \leq c_2(\boldsymbol{t}), \\ \gamma_i(\boldsymbol{t}), & u = c_i(\boldsymbol{t}),\ i = 1,2, \\ 0, & u < c_1(\boldsymbol{t}) \text{ or } u > c_2(\boldsymbol{t}), \end{cases}$$

satisfying

$$E_{\theta_i}[\phi_2 \mid \boldsymbol{T}] = \alpha, \qquad i = 1,2,$$

is UMP level α conditional on $\boldsymbol{T} = \boldsymbol{t}$.

Case (3): Similarly, the test

$$\phi_3(u,\boldsymbol{t}) = \begin{cases} 1, & u < c_1(\boldsymbol{t}) \text{ or } u > c_2(\boldsymbol{t}), \\ \gamma_i(\boldsymbol{t}), & u = c_i(\boldsymbol{t}),\ i = 1,2, \\ 0, & c_1(\boldsymbol{t}) \leq u \leq c_2(\boldsymbol{t}), \end{cases}$$

satisfying

$$E_{\theta_i}[\phi_3 \mid \boldsymbol{T}] = \alpha, \qquad i = 1, 2,$$

is UMP level α conditional on $\boldsymbol{T} = \boldsymbol{t}$.

Case (4): Similarly, the test

$$\phi_4(u, \boldsymbol{t}) = \begin{cases} 1, & u < c_1(\boldsymbol{t}) \text{ or } u > c_2(\boldsymbol{t}), \\ \gamma_i(\boldsymbol{t}), & u = c_i(\boldsymbol{t}),\ i = 1, 2, \\ 0, & c_1(\boldsymbol{t}) \le u \le c_2(\boldsymbol{t}), \end{cases}$$

satisfying

$$E_{\theta_0}[\phi_4 \mid \boldsymbol{T}] = \alpha, \qquad \text{and} \qquad E_{\theta_0}[\phi_4 U \mid \boldsymbol{T}] = \alpha E_{\theta_0}[U \mid \boldsymbol{T}],$$

is UMP level α conditional on $\boldsymbol{T} = \boldsymbol{t}$.

The conditional UMP property of these tests is of limited usefulness; in practice we almost never want the value of T to be fixed/given. What one wishes for instead is some sort of unconditional optimality, and this is established by the following theorem.

THEOREM 7.6.8 (Unconditional UMPU). *The tests $\phi_1, \ldots, \phi_4$ defined for Cases (1)–(4) above are unconditionally UMPU level α.*

PROOF. We prove only Case (1); the remainder being similar. Suppose ϕ is an unbiased level α test. Then, because it's α-similar, $E_{\theta_0,\boldsymbol{\xi}}\phi(U, \boldsymbol{T}) = \alpha$ for every $\boldsymbol{\xi}$. Now let $g(\boldsymbol{T}) = E_{\theta_0}(\phi|\boldsymbol{T}) - \alpha$, and note that

$$E_{\theta_0,\boldsymbol{\xi}} g(\boldsymbol{T}) = E_{\theta_0,\boldsymbol{\xi}}\phi(U, \boldsymbol{T}) - \alpha = \alpha - \alpha = 0. \tag{7.6.4}$$

For $\theta = \theta_0$ the dist. of $\boldsymbol{T}$ belongs to the k-parameter exp. family with parameter set $\Omega_0 = \{\boldsymbol{\xi} : (\theta_0, \boldsymbol{\xi}) \in \Omega\}$. By our assumptions on Ω, Ω_0 contains an open subset of $\mathbb{R}^k$, and therefore $\boldsymbol{T}$ is complete for $\{P_{\theta_0,\boldsymbol{\xi}} : \boldsymbol{\xi} \in \Omega_0\}$. Now, because of (7.6.4), we deduce (from the def. of completeness) that $g(\boldsymbol{T}) = 0$ a.s., whence it follows immediately that

$$E_{\theta_0}(\phi|\boldsymbol{T}) = \alpha \quad \text{a.s.} \tag{7.6.5}$$

(This shows that an unconditional unbiased level α test is also level α conditionally.) Finally, from (7.6.3) it follows that for $\theta > \theta_0$,

$$E_{\theta,\boldsymbol{\xi}}(\phi_1) = E_{\theta,\boldsymbol{\xi}}\left[E_\theta(\phi_1 \mid \boldsymbol{T})\right] \ge E_{\theta,\boldsymbol{\xi}}\left[E_\theta(\phi \mid \boldsymbol{T})\right] = E_{\theta,\boldsymbol{\xi}}(\phi), \tag{7.6.6}$$

whence ϕ_1 is UMPU level α (UMPU by (7.6.6), and level α by (7.6.5)). □

EXAMPLE 7.6.9 (Comparison of Poisson means). For $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$, with X and Y independent, we see that the joint density is the 2-parameter exp. family:

$$f_{X,Y}(x, y) \;=\; \exp\left\{ \underbrace{y}_{u}\,\underbrace{\log(\mu/\lambda)}_{\theta} + \underbrace{(x+y)}_{t}\,\underbrace{\log(\lambda)}_{\xi} \right\} \frac{e^{-(\lambda+\mu)}}{x!y!} = e^{\theta u + \xi t}\, C(\theta, \xi),$$

with $(\lambda, \xi) \in \Omega = \mathbb{R}^2$ clearly convex. Now note that tests about $\theta = \log(\mu/\lambda) \Rightarrow \mu/\lambda = e^\theta$, correspond to comparing μ and λ. E.g., suppose we wish to test $H : \theta \le 0$ ($\Leftrightarrow\ \mu \le \lambda$)

vs. $K : \theta > 0$ ($\Leftrightarrow \ \mu > \lambda$). Then, from Case (1), the conditional UMP level α test is given by

$$\phi_1(u,t) = \begin{cases} 1, & u > c(t), \\ \gamma(t), & u = c(t), \\ 0, & u < c(t), \end{cases}$$

with $c(t)$ and $\gamma(t)$ satisfying $E_{\theta=0}[\phi_1|T] = \alpha$. To compute the power function we need the distribution of $U|t$, which will be seen to be $\text{Bin}(n = t, p = e^\theta/(1+e^\theta))$. Noting that $\lambda = e^\xi$ and $\mu = e^{\xi+\theta}$, the joint of (U,T) is

$$dP^{U,T}_{\theta,\xi}(u,t) = e^{\theta u + \xi t} \frac{\exp\{-e^\xi(1+e^\theta)\}}{u!(t-u)!} I_{\{0,\ldots,t\}}(u) I_{\{0,\ldots,\infty\}}(t),$$

where the fact that the support is the lattice region on the upper portion of the first quadrant of the u vs. t plane separated by the line $u = t$, stems form the fact that $y = u \geq 0$ and $x = t - u \geq 0$. Summing over u yields the marginal of T:

$$dP^T_{\theta,\xi}(t) = \exp\{e^{\xi t} - e^\xi(1+e^\theta)\} I_{\{0,\ldots,\infty\}}(t) \underbrace{\sum_{u=0}^{t} \frac{e^{\theta u}}{u!(t-u)!}}_{t!/(1+e^\theta)^t}.$$

where in the summation we used the following identity for $Z \sim \text{Bin}(n,p)$:

$$\sum_{z=0}^{n} \frac{1}{z!(n-z)!} \left(\frac{p}{1-p}\right)^z = \frac{1}{n!(1-p)^n}.$$

The density of $U|t$ now follows straightforwardly by dividing the joint of (U,T) by the marginal of T. To find the cuttoff points $c(t) \in \mathbb{R}$ and $0 \leq \gamma(t) \leq 1$, we solve:

$$\alpha = E_{\theta=0}[\phi_1|t] = P(Z_t > c(t)) + \gamma(t) P(Z_t = c(t)), \qquad Z_t \sim \text{Bin}(n = t, p = 1/2).$$

This can be solved exactly via tables, or approximated via the CLT (normal approximation to the Binomial).

By Theorem 7.6.8, the (unconditional) UMPU level α test is identical. So what is the difference? In both situations we observe the value of $t = x + y$, so t is known. The distinction is that if we don't fix anything, then ϕ_1 is only UMPU, whereas if we want the optimal test *among all those with the same value of* t (x and y vary but their sum is fixed), then we obtain the stronger result that ϕ_1 is in fact UMP.

EXAMPLE 7.6.10 (Testing a normal std. deviation). For $X_1, \ldots, X_n \sim \text{iid } \text{N}(\mu, \sigma^2)$, TSH §5.2 investigates the following 4 tests:

- $H_1 : \sigma \leq \sigma_0$ vs. $K_1 : \sigma > \sigma_0$.
- $H_2 : \sigma \geq \sigma_0$ vs. $K_2 : \sigma < \sigma_0$.
- $H_3 : \mu \leq \mu_0$ vs. $K_3 : \mu > \mu_0$.
- $H_4 : \mu \geq \mu_0$ vs. $K_4 : \mu < \mu_0$.

The difficulty of these situations is that both parameters are unknown. TSH §3.9 shows that H_1 is the only one for which there exists a UMP test (which rejects for large $\sum(x_i - \bar{x})^2$). Here we will investigate optimal tests for H_2. Treating μ as a nuisance parameter, we have the 2-parameter exp. family:

$$f_{\sigma,\mu}(\boldsymbol{x}) \;=\; \exp\left\{ \underbrace{-\frac{1}{2\sigma^2}}_{\theta} \underbrace{\sum}_{u} x_i^2 + \underbrace{\frac{\mu}{\sigma^2}}_{\xi} \underbrace{\sum}_{t} x_i \right\} C(\theta,\xi).$$

Thus, testing $H_2 : \sigma \geq \sigma_0$ is equivalent to $H_2 : \theta \geq \theta_0$. By Theorem 7.6.8, the UMPU level α test ϕ_2 rejects for $u < c(t)$, and accepts if $u > c(t)$, which is equivalent to rejecting if $\sum x_i^2 < c(\bar{x})$, since $t = \sum x_i = n\bar{x}$. To find the cutoff $c(\bar{x})$, we solve

$$\begin{aligned}
\alpha &= P_{\sigma_0}\left(\sum X_i^2 < c(\bar{X}) \mid \bar{X} \right) \\
&= P_{\sigma_0}\left(\sum X_i^2 - n\bar{X}^2 < c'(\bar{X}) \mid \bar{X} \right) && \text{subtract a constant} \\
&= P_{\sigma_0}\left((n-1)S^2 < c'(\bar{X}) \mid \bar{X} \right), && (n-1)s^2 = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2, \\
&= P_{\sigma_0}\left((n-1)S^2 < c' \right), && \text{since } S^2 \text{ is independent of } \bar{X}, \\
&= P_{\sigma_0}\left(\frac{(n-1)S^2}{\sigma_0^2} < \frac{c'}{\sigma_0^2} \right), && \text{divide by a constant}, \\
&= P\left(\chi^2(n-1) < c_2 \right), && \implies c_2 = \chi^2_\alpha(n-1).
\end{aligned}$$

Thus the rejection rule is:

$$\sum (x_i - \bar{x})^2 < \sigma_0^2 \chi^2_\alpha(n-1) \qquad \Longleftrightarrow \qquad \sum x_i^2 < \frac{t^2}{n} + \sigma_0^2 \chi^2_\alpha(n-1) \equiv c(t).$$

The power function of this test is:

$$\beta_2(\sigma) = P\left(\chi^2(n-1) < \frac{\sigma_0^2}{\sigma^2} \chi^2_\alpha(n-1) \right) = \int_0^{\sigma_0^2 \chi^2_\alpha(n-1)/\sigma^2} f_{n-1}(y)dy,$$

where $f_k(y)$ is the density of a $\chi^2(k)$.

It is interesting to compare this test to the UMP test ϕ_2^* of H_2 discussed in TSH Example 3.9.1, for the (much) simpler situation when μ is known, and which rejects for $\sum(x_i - \mu)^2 < c$. Using similar arguments, we find c by solving

$$\begin{aligned}
\alpha &= P_{\sigma_0,\mu}\left(\sum (X_i - \mu)^2 < c \right) = P_{\sigma_0,\mu}\left(\sum \left(\frac{X_i - \mu}{\sigma_0} \right)^2 < \frac{c}{\sigma_0^2} \right) \\
&= P\left(\chi^2(n) < \frac{c}{\sigma_0^2} \right), && \implies c = \sigma_0^2 \chi^2_\alpha(n),
\end{aligned}$$

whence the power function is given by:

$$\beta_2^*(\sigma) = P_{\sigma,\mu}\left(\sum \left(\frac{X_i - \mu}{\sigma} \right)^2 < \frac{\sigma_0^2 \chi^2_\alpha(n)}{\sigma^2} \right) = P\left(\chi^2(n) < \frac{\sigma_0^2 \chi^2_\alpha(n)}{\sigma^2} \right).$$

Plotting these two power functions, we might expect their difference to be small, with perhaps $\beta_2^*(\sigma)$ slightly larger than $\beta_2(\sigma)$ over most values of σ, since ϕ_2^* uses more information (note that ϕ_2^* cannot be implemented without knowledge of μ).

7.7. Likelihood Ratio (LR), Wald, and Score Tests

The Neyman-Pearson Lemma naturally suggests the LR as a good test. In the absence of an optimal test (UMP, UMPU, etc.), we fall back on LR, Wald, and Score tests. A complete coverage of this subject can be found in Severini (2000), Chs 3 & 4, and we follow Severini's compact notation here.

We continue to let $\ell(\theta)$ denote the log likelihood based on a sample of size n, where $\theta = (\theta_1, \ldots, \theta_d) \in \Omega \subset \mathbb{R}^d$. When needed, we partition $\theta = (\psi, \lambda)$, where $\psi = (\psi_1, \ldots, \psi_q)$ denotes the parameter of interest, while $\lambda \in \mathbb{R}^{d-q}$ is a nuisance parameter. Derivatives of $\ell(\theta)$ w.r.t. θ are denoted as:

$$\begin{aligned} \ell_\theta(\theta) &= \frac{\partial \ell(\theta)}{\partial \theta}, \qquad \text{Jacobian vector (a tensor of dim=1)} \\ \ell_{\theta\theta}(\theta) &= \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T}, \qquad \text{Hessian matrix (a tensor of dim=2)} \end{aligned}$$

etc. We assume the following (loosely stated) regularity conditions, satisfied by all "nice" (henceforth called **regular**) models:

R1. $\ell(\theta)$ can be approximated by a 4th order polynomial in θ around the *true* value $\theta_0 \in \Omega$,

$$\ell(\theta) = \ell(\theta_0) + \ell_\theta(\theta_0)(\theta - \theta_0) + \cdots + \frac{1}{4!}\ell_{\theta\theta\theta\theta}(\theta_0)(\theta - \theta_0)^4 + R_n(\theta),$$

with the remainder term satisfying the following condition over some neighborhood N_0 of θ_0:

$$\frac{\sup_{\theta \in N_0} |R_n(\theta)|}{n||\theta - \theta_0||^5} = O_p(1).$$

R2. The first 4 derivatives of $\ell(\theta)$, $\{\ell_\theta, \ldots, \ell_{\theta\theta\theta}\}$, have joint cumulants which are $O(n)$, and the vector of sample averages $\ell_\theta/\sqrt{n}$ obeys the CLT:

$$\ell_\theta/\sqrt{n} \xrightarrow{d} \mathrm{N}(0, I(\theta_0)),$$

where now, and throughout this chapter, $I(\theta)$ is as defined in (7.7.4).

R3. For non-negative integers $\{i_1, \ldots, i_4\}$ with $i_1 + \cdots + i_4 \leq 4$, and for $\{j, k, l, m\} \in \{0, 1, \ldots, d\}$, one is able to interchange up to 4th order derivatives with integrals as follows:

$$\frac{\partial^{i_1+\cdots+i_4}}{\partial\theta_j^{i_1}\partial\theta_k^{i_2}\partial\theta_l^{i_3}\partial\theta_m^{i_4}} E_{\theta_0} \exp\{\ell(\theta) - \ell(\theta_0)\}|_{\theta=\theta_0} = E_{\theta_0}\left\{\frac{\partial^{i_1+\cdots+i_4}}{\partial\theta_j^{i_1}\partial\theta_k^{i_2}\partial\theta_l^{i_3}\partial\theta_m^{i_4}} \exp\{\ell(\theta) - \ell(\theta_0)\}|_{\theta=\theta_0}\right\}.$$

Properties R1–R3 hold in most models of practical interest (and most, if not all, models covered in this course). Examples include models where the observations are independent but not identical (e.g., regression and GLM), and models where the observations are dependent (e.g., some types of stochastic processes).

Bartlett Identities

Property R3 leads in particular to the so-called *Bartlett identities*. The key to this is the following equation, which in the scalar θ case is:

$$\frac{\partial^j}{\partial\theta^j}E_{\theta_0}\exp\{\ell(\theta)-\ell(\theta_0)\}|_{\theta=\theta_0} = E_{\theta_0}\left\{\frac{\partial^j}{\partial\theta^j}\exp\{\ell(\theta)-\ell(\theta_0)\}|_{\theta=\theta_0}\right\}, \tag{7.7.1}$$

Now, since

$$E_{\theta_0}\exp\{\ell(\theta)-\ell(\theta_0)\} = \int\frac{L(\theta)}{L(\theta_0)}L(\theta_0)dx = \int L(\theta)dx = 1,$$

it implies that, in particular, by (7.7.1) with $j=1$,

$$E_{\theta_0}\left\{\frac{\partial}{\partial\theta}\exp\{\ell(\theta)-\ell(\theta_0)\}|_{\theta=\theta_0}\right\} = \frac{\partial}{\partial\theta}E_{\theta_0}\exp\{\ell(\theta)-\ell(\theta_0)\}|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}(1) = 0,$$

and for general j (and $\forall\theta_0$):

$$E_{\theta_0}\left\{\frac{\partial^j}{\partial\theta^j}\exp\{\ell(\theta)-\ell(\theta_0)\}|_{\theta=\theta_0}\right\} = 0 = \frac{\partial^j}{\partial\theta^j}E_{\theta_0}\exp\{\ell(\theta)-\ell(\theta_0)\}|_{\theta=\theta_0}.$$

Thus, in the $j=1$ case,

$$0 = E\left\{\frac{\partial}{\partial\theta}\,e^{\ell(\theta)-\ell(\theta_0)}\big|_{\theta=\theta_0}\right\} = E\left\{\ell_\theta(\theta_0)e^{\ell(\theta_0)-\ell(\theta_0)}\right\} = E\ell_\theta(\theta_0),$$

and since this holds for every θ_0, we obtain the 1st Bartlett Identity: $E\ell_\theta(\theta)=0$. Similarly, in the $j=2$ case,

$$0 = E\left\{\frac{\partial^2}{\partial\theta^2}\,e^{\ell(\theta)-\ell(\theta_0)}\big|_{\theta=\theta_0}\right\} = E\left\{\ell_{\theta\theta}e^{\ell(\theta_0)-\ell(\theta_0)}+\ell_\theta^2e^{\ell(\theta_0)-\ell(\theta_0)}\right\},$$

which leads to the 2nd Bartlett Identity: $E\ell_{\theta\theta}(\theta)+E\ell_\theta(\theta)^2=0$. This generalizes to the vector θ case, and for every integer j there is a corresponding identity. The **first two Bartlett identities** are:

$$E\ell_\theta(\theta) = 0, \tag{7.7.2}$$

$$E\ell_{\theta\theta}(\theta)+E\ell_\theta(\theta)\ell_\theta(\theta)^T = 0. \tag{7.7.3}$$

Different types of Information

For regular models we have the following types of information-related quantites and results.

Score Function: $\ell_\theta(\theta)$. The first two Bartlett identities imply that the score vector has mean zero and its variance is equal to the (expected) Information matrix:

$$\begin{aligned} E\ell_\theta(\theta) &= 0, \\ \mathrm{Var}[\ell_\theta(\theta)] &= E\ell_\theta(\theta)\ell_\theta(\theta)^T := \mathcal{I}(\theta), \qquad (\textbf{Expected Information}). \end{aligned}$$

Observed Information: $\mathcal{J}(\theta) := -\ell_{\theta\theta}(\theta)$. The 2nd Bartlett identity implies that:

$$E\mathcal{J}(\theta) = \mathcal{I}(\theta).$$

Partial Information: Invoking the $\theta = (\psi, \lambda)$ partition, partition the Information matrix accordingly as:

$$\mathcal{I}(\theta) = \begin{bmatrix} \mathcal{I}_{\psi\psi}(\theta) & \mathcal{I}_{\psi\lambda}(\theta) \\ \mathcal{I}_{\lambda\psi}(\theta) & \mathcal{I}_{\lambda\lambda}(\theta) \end{bmatrix},$$

where, using obvious notation, we have e.g.,

$$\mathcal{I}_{\psi\lambda}(\theta) = E\ell_\psi(\theta)\ell_\lambda(\theta)^T = -E\ell_{\psi\lambda}(\theta),$$

and

$$\ell_\psi(\theta) = \frac{\partial \ell(\theta)}{\partial \psi}, \qquad \ell_\lambda(\theta) = \frac{\partial \ell(\theta)}{\partial \lambda}, \qquad \ell_{\psi\lambda}(\theta) = \frac{\partial^2 \ell(\theta)}{\partial \psi \partial \lambda^T}.$$

DEFINITION 7.7.1 (Average Information per observation). For regular models, we define the *average (expected) Information per observation* as

$$I(\theta) := \lim_{n\to\infty} \frac{1}{n}\mathcal{I}(\theta). \tag{7.7.4}$$

The asymptotic normality of the MLE result for regular models is more general than those in Ch. 6, and allows us to break free from the iid assumption (e.g., regression). If $\hat{\theta}$ denotes the MLE of θ based on a sample of size n from a regular model, then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathrm{N}(0, I^{-1}(\theta_0)), \tag{7.7.5}$$

with $I(\theta)$ as defined in (7.7.4). In particular, if the MLE is based on an iid sample of size n, then $\mathcal{I}(\theta) = nI(\theta)$, whence $I(\theta)$ coincides with the (expected) Information per observation of Ch. 6.

DEFINITION 7.7.2 (Partial Information). The partial (expected) Information for ψ, defined as

$$\mathcal{I}_\psi(\theta) := \mathcal{I}_{\psi\psi}(\theta) - \mathcal{I}_{\psi\lambda}(\theta)\mathcal{I}_{\lambda\lambda}^{-1}(\theta)\mathcal{I}_{\lambda\psi}(\theta), \tag{7.7.6}$$

plays the same role for inference on ψ that $\mathcal{I}(\theta)$ plays for inference on the entire θ, and can be derived as the appropriate CRLB by generalizing the argument in Remark 2.4.6. If λ is known, then $\mathcal{I}_\psi(\theta) = \mathcal{I}_{\psi\psi}(\theta)$, so that

$$\mathcal{I}_\psi(\theta) - \mathcal{I}_{\psi\psi}(\theta)$$

represents the loss of information about ψ due to the fact that λ is unknown. (Similar results hold for the definition of partial observed Information, $\mathcal{J}_\psi(\theta)$.) Note that we can define partial Information per observation by replacing $\mathcal{I}(\theta) \mapsto I(\theta)$ everywhere in (7.7.6), whence for an iid sample, $\mathcal{I}_\psi(\theta) = nI_\psi(\theta)$.

Let $\hat{\theta}$ denote the MLE of θ based on a sample of size n from a regular model. We describe the three tests for testing the two-sided hypothesis

$$H: \ \theta = \theta_0 \quad \text{vs.} \quad K: \ \theta \neq \theta_0.$$

The tests reject H for large values of the corresponding statistic (W, W_w, W_s), and as we will show next, the asymptotic distribution of each of these under H is $\chi^2(d)$.

(i) **LR Test**. The test statistic is:

$$W = W(\theta_0) := 2[\ell(\hat{\theta}) - \ell(\theta_0)]. \tag{7.7.7}$$

To derive the asymptotic distribution of W under H, Taylor-series expand $\ell(\theta) - \ell(\theta_0)$ around $\hat{\theta} = \theta_0$, so that

$$\frac{1}{2}W = \ell_\theta(\theta_0)^T(\hat{\theta} - \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^T\ell_{\theta\theta}(\theta_0)(\hat{\theta} - \theta_0) + O_p(1/\sqrt{n}).$$

Now, from Ch 6 results, and with I_d denoting the identity matrix of rank d, we have:

$$\begin{aligned} \hat{\theta} - \theta_0 &= \mathcal{I}^{-1}(\theta_0)\ell_\theta(\theta_0)) + O_p(1/\sqrt{n}), \\ \mathcal{I}^{-1/2}(\theta_0)\ell_\theta(\theta_0) &\xrightarrow{d} \mathrm{N}(0, I_d), \\ \ell_{\theta\theta}(\theta_0) &= -\mathcal{I}(\theta_0) + O_p(\sqrt{n}). \end{aligned}$$

Substituting these results into the above expression for $W/2$, we obtain

$$W = \left[\mathcal{I}^{-1/2}(\theta_0)\ell_\theta(\theta_0)\right]^T \left[\mathcal{I}^{-1/2}(\theta_0)\ell_\theta(\theta_0)\right] + O_p(1/\sqrt{n}), \tag{7.7.8}$$

so that we have an asymptotic chi-square distribution for the LR statistic under the null hypothesis:

$$W \xrightarrow{d} \chi^2(d).$$

(ii) **Wald Test**. The test statistic is:

$$W_w = W_w(\theta_0) := (\hat{\theta} - \theta_0)^T\mathcal{I}(\hat{\theta})(\hat{\theta} - \theta_0). \tag{7.7.9}$$

To derive the asymptotic distribution of W_w under H, Taylor-series expand the LR test statistic W around $\theta_0 = \hat{\theta}$ (before we expanded around $\hat{\theta} = \theta_0$), so that

$$-\frac{1}{2}W = \ell_\theta(\hat{\theta})^T(\theta_0 - \hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^T\ell_{\theta\theta}(\hat{\theta})(\theta_0 - \hat{\theta}) + O_p(1/\sqrt{n}).$$

Now, since $\ell_\theta(\hat{\theta}) = 0$ and

$$-\ell_{\theta\theta}(\hat{\theta}) = \mathcal{I}(\theta_0) + O_p(\sqrt{n}) = \mathcal{I}(\hat{\theta}) + O_p(\sqrt{n}),$$

it follows that

$$W = W_w + \underbrace{(\theta_0 - \hat{\theta})^T}_{O_p(1/\sqrt{n})} O_p(\sqrt{n}) \underbrace{(\theta_0 - \hat{\theta})}_{O_p(1/\sqrt{n})} + O_p(1/\sqrt{n}) = W_w + O_p(1/\sqrt{n}),$$

so that W_w has the same limiting distribution as W.

(ii) **Score Test**. The test statistic is:

$$W_s = W_s(\theta_0) := \ell_\theta(\theta_0)^T \mathcal{I}^{-1}(\theta_0) \ell_\theta(\theta_0). \tag{7.7.10}$$

The fact that W_s has the same limiting distribution as W follows straightforwardly from (7.7.8). (The Score test is also known as the Rao Score Test, or Lagrange Multiplier Test.)

NOTE 7.7.3. One can use any of the four versions of Information (expected or observed evaluated at θ_0 or $\hat{\theta}$) in the definition of W_w and W_s, namely

$$\{\mathcal{I}(\theta_0), \mathcal{I}(\hat{\theta}), \mathcal{J}(\theta_0), \mathcal{J}(\hat{\theta})\},$$

without affecting the asymptotics.

Testing only a subset of parameters

Partition $\theta = (\psi, \lambda)$, where $\psi \in \mathbb{R}^q$ is the parameter of interest, and $\lambda \in \mathbb{R}^{d-q}$ is a nuisance parameter. The (unrestricted) MLE is $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$, and let $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$ denote the (profile or restricted) MLE of θ when ψ is held fixed, which just involves maximizing $\ell(\theta)$ over λ, i.e., $\hat{\lambda}_\psi = \arg\max_\lambda \ell(\psi, \lambda)$. We now wish to test

$$H: \ \psi = \psi_0 \quad \text{vs.} \quad K: \ \psi \neq \psi_0.$$

The analogous versions of the LR, Wald, and Score Tests are now as follows:

$$W = W(\psi_0) = 2[\ell(\hat{\theta}) - \ell(\hat{\theta}_{\psi_0})], \tag{7.7.11}$$

$$W_w = W_w(\psi_0) = (\hat{\psi} - \psi_0)^T \mathcal{I}_\psi(\hat{\theta})(\hat{\psi} - \psi_0), \tag{7.7.12}$$

$$W_s = W_s(\psi_0) = \ell_\psi(\hat{\theta}_{\psi_0})^T \mathcal{I}_\psi^{-1}(\hat{\theta}_{\psi_0}) \ell_\psi(\hat{\theta}_{\psi_0}). \tag{7.7.13}$$

Using similar arguments as before, it can be shown that now

$$W \ \xrightarrow{d} \ \chi^2(q),$$

which is also the limiting distribution of W_w and W_s.

One-sided tests

Partition $\theta = (\psi, \lambda)$ as above, but ψ is a scalar ($q = 1$). To test, e.g.,

$$H: \ \psi \leq \psi_0 \quad \text{vs.} \quad K: \ \psi > \psi_0,$$

use the signed square root of the statistics in (7.7.11)–(7.7.13):

$$\begin{aligned} R = R(\psi_0) &= \text{sgn}(\hat{\psi} - \psi_0)\sqrt{W(\psi_0)}, \\ R_w = R_w(\psi_0) &= \text{sgn}(\hat{\psi} - \psi_0)\sqrt{W_w(\psi_0)}, \\ R_s = R_s(\psi_0) &= \text{sgn}(\hat{\psi} - \psi_0)\sqrt{W_s(\psi_0)}. \end{aligned}$$

To derive the asymptotics, one can show that (under H)

$$R(\psi_0) = \sqrt{\mathcal{I}_\psi(\hat{\theta})}(\hat{\psi} - \psi_0) + O_p(1/\sqrt{n}),$$

and thus, since the first term in the above summand converges to a standard normal, we obtain

$$R \xrightarrow{d} \mathrm{N}(0,1),$$

with identical conclusions for R_w and R_s.

Confidence Regions

Construction of confidence regions by inverting each of the tests is immmediate. E.g., if $W(\theta_0)$ denotes any of the three test (7.7.7), (7.7.9), or (7.7.10), inversion of the two-sided test leads to the $(1-\alpha)$ acceptance region

$$A(\theta_0) = \left\{\theta_0 \mid W(\theta_0) \leq \chi^2_{1-\alpha}(d)\right\}.$$

Likewise, inversion of the two-sided subset case tests (7.7.11)–(7.7.13), leads to the $(1-\alpha)$ acceptance region

$$A(\psi_0) = \left\{\psi_0 \mid W(\psi_0) \leq \chi^2_{1-\alpha}(q)\right\}.$$

EXAMPLE 7.7.4 (Inference for Weibull shape). Consider the following shape-scale parametrization for the density of a Weibull distribution

$$f_\theta(x) = \psi\lambda(\lambda x)^{\psi-1}\exp\left\{-(\lambda x)^\psi\right\} I(x>0), \qquad \theta = (\psi,\lambda),$$

where the parameter of interest $\psi > 0$ controls the shape, while the inverse of the nuisance parameter $\lambda > 0$ controls the scale. Since this is not an exponential family, we have little hope of deducing any kind of optimal test. Tedious computations lead to the following expression for the Information matrix (per oservation):

$$I(\theta) = \begin{bmatrix} \frac{\pi^2/6+\gamma^2-2\gamma}{\psi^2} & \frac{1-\gamma}{\lambda} \\ \frac{1-\gamma}{\lambda} & \frac{\psi^2}{\lambda^2} \end{bmatrix} = \begin{bmatrix} I_{\psi\psi}(\theta) & I_{\psi\lambda}(\theta) \\ I_{\lambda\psi}(\theta) & I_{\lambda\lambda}(\theta) \end{bmatrix}, \qquad \gamma = 0.5772\ldots (\text{ Euler's constant}).$$

From this we obtain the partial Information

$$I_\psi(\theta) = I_{\psi\psi}(\theta) - I_{\psi\lambda}(\theta)I^{-1}_{\lambda\lambda}(\theta)I_{\lambda\psi}(\theta), \qquad \Longrightarrow \quad \mathcal{I}_\psi(\theta) = nI_\psi(\theta) = \frac{n}{\psi^2}\left(\frac{\pi^2}{6} - 1\right).$$

The log likelihood based on a random sample of size n is

$$\ell(\theta) = \ell(\psi,\lambda) = n\psi\log(\lambda) + n\log\psi + (\psi-1)t - \lambda^\psi s_\psi, \qquad t = \sum \log x_i,\ s_\psi = \sum x_i^\psi.$$

It is possible to obtain the profile MLE of λ as $\hat{\lambda}_\psi = (n/s_\psi)^{1/\psi}$, which upon substitution leads to the profile log-likelihood

$$\ell(\hat{\theta}_\psi) = \ell(\psi,\hat{\lambda}_\psi) = n\log\left(\frac{n}{s_\psi}\right) + n\log\psi + (\psi-1)t - n.$$

This can now be maximized (numerically) for the MLE of ψ which is then substituted into the above formulas, yielding the following cascade of results:

$$\hat{\psi} = \arg\max_{\psi} \ell(\hat{\theta}_\psi), \qquad s_{\hat{\psi}} = \sum x_i^{\hat{\psi}}, \qquad \hat{\lambda} = \hat{\lambda}_{\hat{\psi}} = \left(\frac{n}{s_{\hat{\psi}}}\right)^{1/\hat{\psi}}, \qquad \ell(\hat{\theta}) = \ell(\hat{\psi}, \hat{\lambda}).$$

Straightforward substitution into (7.7.11)–(7.7.13) then leads to:

$$\begin{aligned} W(\psi_0) &= 2n \log\left(\frac{s_{\psi_0}\hat{\psi}}{s_{\hat{\psi}}\psi_0}\right) + 2\left(\hat{\psi} - \psi_0\right) t. \\ W_w(\psi_0) &= \left(\frac{\psi_0}{\hat{\psi}} - 1\right)^2 n \left(\frac{\pi^2}{6} - 1\right). \\ W_s(\psi_0) &= \left(\frac{\psi_0 \sum x_i^{\psi_0} \log x_i}{s_{\psi_0}} - \frac{\psi_0 t}{n} - 1\right)^2 \left(\frac{n}{\psi_0}\right)^3 \left(\frac{\pi^2}{6} - 1\right). \end{aligned}$$

A two-sided level α test of $H : \psi = \psi_0$ then rejects for $W(\psi_0) > \chi^2_{1-\alpha}(1)$, whereas the one-sided test of $H : \psi \leq \psi_0$ rejects for $R(\psi_0) = \text{sgn}(\hat{\psi} - \psi_0)\sqrt{W(\psi_0)} > z_{1-\alpha}$, etc. To illustrate construction of confidence intervals, inversion of Wald leads to the $(1-\alpha)$ acceptance region

$$A(\psi_0) = \left\{\psi_0 \mid W_w(\psi_0) \leq \chi^2_{1-\alpha}(1)\right\}.$$

EXAMPLE 7.7.5 (Neyman Smooth Test). Suppose we have a family of densities from a full-rank k-parameter exponential family

$$f_\theta(x) = c(\theta) \exp\{\sum_{j=1}^k \theta_j t_j(x)\} = \exp\{\sum_{j=1}^k \theta_j t_j(x) - \log c(\theta)^{-1}\},$$

where $\theta \in \Omega \subset \mathbb{R}^k$ is the natural parameter set (of which $\theta = 0$ is an interior point), and where the $t_j(x)$ are a set of orthonormal functions satisfying:

$$E_0(t_j(X)) = 0, \qquad \text{Cov}_0(t_i(X), t_j(X)) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

(The notation E_0 and Cov_0 here means that expectations are taken with respect to the measure for the case $\theta = 0$, which is a Unif$(0,1)$.) By Theom 2.4.7 in canonical form, we identify $A(\theta) = -\log c(\theta)$, whence the first two moments for the vector $t = (t_1, \ldots, t_k)'$ are:

$$0 = E_0(t) = \left.\frac{\partial A(\theta)}{\partial \theta}\right|_{\theta=0} = -\left.\frac{1}{c(\theta)} \frac{\partial c(\theta)}{\partial \theta}\right|_{\theta=0}, \tag{7.7.14}$$

$$I_k = \text{Cov}_0(t) = \left.\frac{\partial^2 A(\theta)}{\partial \theta \partial \theta'}\right|_{\theta=0} = \left.\frac{\partial}{\partial \theta'}\left[-\frac{1}{c(\theta)} \frac{\partial c(\theta)}{\partial \theta}\right]\right|_{\theta=0}. \tag{7.7.15}$$

For a random sample $x_1, \ldots, x_n$ from $f_\theta(x)$, and in the context of goodness-of-fit, Neyman (1937) proposed (what is now known to be) a Score test for $H : \theta = 0$ vs. $K : \theta \neq 0$.

Note that now the CSS is $T = (\sum t_1(x_i), \ldots, \sum t_k(x_i))'$, and the new "A" function is $A_n(\theta) = nA(\theta)$, so that the log-likelihood and its derivative are:

$$\ell(\theta) = \theta' T - nA(\theta), \qquad \ell_\theta(\theta) = T - n\frac{\partial A(\theta)}{\partial \theta}.$$

Thus, under H, we have from (7.7.14) that $\ell_\theta(0) = T$, and (7.7.15) implies that

$$\mathcal{I}(0) = n \left.\frac{\partial^2 A(\theta)}{\partial\theta\partial\theta'}\right|_{\theta=0} = nI_k.$$

which leads to

$$W_S = \ell_\theta(0)'\mathcal{I}(0)^{-1}\ell_\theta(0) = \frac{1}{n}T'T \xrightarrow{d} \chi^2(k), \qquad \text{under } H.$$

EXAMPLE 7.7.6 (Poisson GLM). In this log-linear regression model for counts, we observe the pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where the y_i are independent Poisson with means $\mu_i = \exp\{\lambda + \psi x_i\}$, and the x_i are known covariates. The joint density of the y_i is therefore seen to be the 2-parameter exponential family,

$$f(\boldsymbol{y}) = \exp\left\{\psi t + \lambda s - e^\lambda \sum_i e^{\psi x_i}\right\} \prod_i \frac{1}{y_i!} I_{\{0,1,\ldots\}}(y_i), \qquad s = \sum_i y_i,\ t = \sum_i x_i y_i,$$

and the goal is to test if there is an effect from the covariates, i.e. $H : \psi = 0$ vs. $K : \psi \neq 0$. Since this is in canonical form for $\theta = (\psi, \lambda)$ with (s, t) the CSS, we make the identification $A(\theta) = e^\lambda u(\psi)$, where $u(\psi) = \sum e^{\psi x_i}$, whence the Information matrix for the model is obtained straightforwardly as:

$$\mathcal{I}(\theta) = \frac{\partial^2 A(\theta)}{\partial\theta\partial\theta^T} = e^\lambda \begin{bmatrix} u(\psi) & v(\psi) \\ v(\psi) & w(\psi) \end{bmatrix}, \qquad v(\psi) = \sum x_i e^{\psi x_i},\ w(\psi) = \sum x_i^2 e^{\psi x_i},$$

and since this is a regular model, we have the asymptotic distribution for the MLE of θ as in (7.7.5). The partial information is:

$$\mathcal{I}_\psi(\theta) = \mathcal{I}_{\psi\psi}(\theta) - \mathcal{I}^2_{\psi\lambda}(\theta)/\mathcal{I}_{\lambda\lambda}(\theta) = e^\lambda[w(\psi) - v^2(\psi)/u(\psi)].$$

Now, since $\ell(\theta) = \psi t + \lambda s - A(\theta) + \text{constant}$, we have from the score equations

$$\begin{aligned} \ell_\psi(\theta) &= t - e^\lambda v(\psi) = 0, \\ \ell_\lambda(\theta) &= s - e^\lambda u(\psi) = 0, \qquad \Longrightarrow \quad \hat\lambda_\psi = \log(s/u(\psi)), \end{aligned}$$

but we must then solve for the MLE of ψ numerically to obtain: $\hat\theta = (\hat\psi, \hat\lambda_{\hat\psi})$, leading to,

$$\begin{aligned} \ell(\hat\theta_\psi) &= \ell(\psi, \hat\lambda_\psi) = \psi t + s\log(s/u(\psi)) - s, \\ \ell(\hat\theta) &= \ell(\hat\psi, \hat\lambda_{\hat\psi}) = \hat\psi t + s\log(s/u(\hat\psi)) - s. \end{aligned}$$

and,

$$\mathcal{I}_\psi(\hat\theta_\psi) = \frac{s}{u(\psi)}\left[w(\psi) - \frac{v^2(\psi)}{u(\psi)}\right], \qquad \mathcal{I}_\psi(\hat\theta) = \frac{s}{\hat u}\left[\hat w - \frac{\hat v^2}{\hat u}\right],$$

where we use the shorthand $\hat{u} \equiv u(\hat{\psi})$, etc. In addition, we will need the MLEs at the null value of $\psi_0 = 0$: $\hat{\theta}_{\psi_0} \equiv \hat{\theta}_0 = (0, \hat{\lambda}_0)$, where $\hat{\lambda}_0 \equiv \hat{\lambda}_{\psi_0} = \log(s/n) = \log(\bar{y})$, which leads to:

$$\ell_\psi(\hat{\theta}_0) \;=\; t - \bar{y}\bar{x}, \qquad \text{and} \qquad \mathcal{I}_\psi(\hat{\theta}_0) = \bar{y}\left[\sum x_i^2 - n\bar{x}^2\right].$$

From the above results, we can now calculate the triad of statistics, all of which are $\chi^2(1)$ under H:

- LR:
$$W(0) = 2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)] = 2\left[t\hat{\psi} + s\log(n/\hat{u})\right].$$
- Wald:
$$W_W(0) = (\hat{\psi} - \psi_0)^2 \mathcal{I}_\psi(\hat{\theta}) = \frac{s}{\hat{u}}\left[\hat{w} - \frac{\hat{v}^2}{\hat{u}}\right]\hat{\psi}^2.$$
- Score:
$$W_S(0) = \ell_\psi(\hat{\theta}_0)^2 / \mathcal{I}_\psi(\hat{\theta}_0) = \frac{(t - \bar{y}\bar{x})^2}{\bar{y}[\sum x_i^2 - n\bar{x}^2]}.$$

It would be interesting to compare this triad of tests with the UMPU, Case (4) of the multiparameter EF, with critical function:

$$\phi = \begin{cases} 1, & t < c_1(s) \text{ or } t > c_2(s), \\ \gamma_1(s), & t = c_1(s), \\ \gamma_2(s), & t = c_2(s), \\ 0, & \text{otherwise}, \end{cases}$$

where the cutoff points are determined from $E_{\psi=0}(\phi|S) = \alpha$ and $E_{\psi=0}(\phi T|S) = \alpha E_{\psi=0}(T|S)$. However, the $\prod_i (y_i!)^{-1}$ term in the expression for $f(\boldsymbol{y})$ appears to make the calculation of the joint distribution of (T, S) intractable!

7.8. Discussion

- The LR, Wald, and Score statistics can be shown to have a non-central chi-square asymptotic distribution under (local) alternatives (Severini, 2000, Ch 4).
- The asymptotic normality result for the MLE in (7.7.5) holds quite generally beyond iid data. Two common instances that considerably extend our range of applications include: (i) regression models where the data are independent but not identically distributed, and (ii) stationary time series models where the data are not independent but are identically (marginally) distributed.
- Example 7.7.6 is at the threshold of tractability in terms of what can feasibly be analytically computed for the sub-optimal LR, Wald, and Score tests. In practice (implemented in software packages) the process is automated by computing the MLEs numerically, and substituting the expected by the observed Information throughout, $\mathcal{I}(\theta) \mapsto \mathcal{J}(\theta)$, which requires only numerical evaluation of the Hessian of $\ell(\theta)$.
- For small n it may be necessary to compute the null distribution of the sub-optimal LR, Wald, and Score test statistics via Monte Carlo simulation, since the asymptotic χ^2 may be unreliable.
- Hypothesis testing in the Big Data era (Efron, 2010). The 21st century has ushered in the era of high-dimensional testing, where $\dim(\theta) = d \gg n$. One successful way forward here has been the formulation of this problem into a large-scale testing framework, where one constructs many simultaneous tests and tries to control the error rate. E.g., Efron (2010) describes a typical microarray study on the effect of $d = 6,033$ genes on $n = 102$ subjects (52 with disease and 50 without, serving as controls). The effect of each gene is then investigated individually by carrying out d two-sample t-tests. The struggle here has been twofold: (i) adapting existing multiple comparison procedures (e.g., Tukey's MCP) to cope with a number of comparisons far in excess of what they were designed for, and (ii) devising new types or definitions of error rate. The current best recommendations from Efron (2010) are usage of: (i) adapted *FamilyWise Error Rate* (FEWR) control procedures, and (ii) the *False Discovery Rate* (FDR) paradigm proposed by Benjamini & Hochberg (1995).

Bibliography

[1] Y. Benjamini and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57, 289–300.

[2] P. Billingsley (1995). *Probability and Measure*, 3rd Edition. Wiley.

[3] P.J. Bickell and K.A. Doksum (2015). *Mathematical Statistics (Vol. 1)*, 2nd Edition. CRC Press.

[4] P.J. Brockwell and R.A. Davis (1991). *Time Series: Theory and Methods*, 2nd Edition. Springer.

[5] B. Efron (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing and Prediction*. Cambridge.

[6] B. Efron and T. Hastie (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge.

[7] F.A. Graybill and R.B. Deal (1959). Combining unbiased estimators. *Biometrics*, 15, 543–550.

[8] W. James and C. Stein (1961). Estimation with quadratic loss. *Proc. 4th Berkeley Symp. Math. Statist. Prob.*, 1. Berkeley: University of California Press.

[9] E.L. Lehmann and G. Casella (1998). *Theory of Point Estimation*, 2nd Edition. Springer.

[10] E.L. Lehmann and J.P. Romano (2005). *Testing Statistical Hypotheses*, 3rd Edition. Springer.

[11] J. Neyman (1937). Smooth test for goodness of fit. *Scandinavian Actuarial Journal*, 3, 149–199.

[12] M.J. Schervish (1995). *Theory of Statistics*. Springer.

[13] R.J. Serfling (1980). *Approximation Theorems of Mathematical Statistics*. Wiley.

[14] T.A. Severini (2000). *Likelihood Methods in Statistics*. Oxford.

[15] A. Van der Vaart (1998). *Asymptotic Statistics*. Cambridge.