Inference Problem.

(i) You are given a collection of prob. measures
$\{P_\theta : \theta \in \Theta\}$ on a sample space $(\mathcal{X}, \mathcal{F})$
where $\mathcal{X}$ is a set and $\mathcal{F}$ is a σ-field on $\mathcal{X}$.

(ii) Observe $X \sim P_\theta$ for some $\theta \in \Theta$

(iii) Infer θ from X.

Let $L(\theta, \boxed{\delta(X)})$ be the loss in estimating θ by $\delta(X)$, an estimator.

Define $\boxed{R(\theta,\delta) = E_{X\sim P_\theta}\{L(\theta, \boxed{\delta(X)})\}}$ to be the risk function of the estimator δ.

e.g. $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$, $\theta \in \mathbb{R}$, $X \sim \mathbb{R}^n$

$\boxed{X = (X_1, \dots, X_n)}$ $\quad \underset{P_\theta(X\in A)}{\overset{}{P_\theta(A)}} = \frac{1}{(\sqrt{2\pi})^n} \int_A e^{-\sum_{i=1}^n \frac{(x_i-\theta)^2}{2}} \, dx_1, \dots dx_n.$

$L(\theta, \delta(X)) = \{\theta - \delta(X)\}^2$

Two proposed estimator $\begin{cases} \delta_1(X) = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \\ \delta_2(X) = 0 \end{cases}$

Correspondingly,

$\boxed{R(\theta, \delta_1) = E_\theta (\bar{X}-\theta)^2 = \frac{1}{n}}$ and $R(\theta, \delta_2) = E_\theta(\theta^2) = \theta^2$

Strategies

$\boxed{\text{Strategy 1}}$

Def We say $\delta(X)$ is unbiased for θ if $E_{X\sim P_\theta}\delta(X) = \theta$, $\forall\, \theta \in \Theta$

For the previous example,

$E_\theta\{\delta_1(X)\} = E_\theta\{n^{-1}\sum_{i=1}^n X_i\} = \theta \qquad E_\theta\{\delta_2(X)\} = 0$

Later on, we shall show that δ_1 is the "best" amongst the class of all unbiased estimators in the problem.

$\boxed{\text{Strategy 2}}$ (Minimax)

We shall look at $\sup_{\theta\in\Theta} R(\theta, \delta)$ for compassion.

In our example, $\sup_{\theta\in\Theta} R(\theta,\delta_1) = \frac{1}{n}$, $\sup_{\theta\in\Theta} R(\theta, \delta_2) = +\infty$

$\boxed{\text{Strategy 3}}$ (Bayes)

Assume θ is random and has a distribution. π.

We may compare the Bayes risk, which is $\boxed{E_{\theta\sim\pi}\{R(\theta,\delta)\}}$

In our case, let $\pi \sim N(\mu, \tau)$

Bayes risk of δ_1, is $E_{\theta\sim\pi}\{R(\theta,\delta_1)\} = E_{\theta\sim\pi}\{\frac{1}{n}\} = \frac{1}{n}$

δ_2, ... δ_2 $= E_{\theta\sim\pi}\{\theta^2\} = \mu^2 + \tau$

$\boxed{\text{Strategy 4}}$ What happens when n is large?

In this case, by WLLN, $\delta_1(X) = \bar{X}_n \xrightarrow{P} \theta$.

$\delta_2(X) \xrightarrow{P} 0$.

Asymptotic optimality (to be learnt)

Chap 1 Probability & Measure

C&B.

Def The set S of all possible outcomes of a particular experiment is called the sample space for the experiment.

e.g. $S = \{H, T\}$ (coin flip) ; $S = [0, \infty)$ (stock price)

Def Event : any subset of S including S itself.

equality : $A \subset B$, $B \subset A \Leftrightarrow A = B$

union : $\cup$ intersection : $\cap$

De Morgan's Laws : $(A \cup B)^c = A^c \cap B^c$

$(A \cap B)^c = A^c \cup B^c$

Def Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \phi$.

For the disjoint events A_i $(i=1,\dots,n)$, $A_i \cap A_j = \phi$ for $i \neq j$.

Furthermore, if $\bigcup_{i=1}^{\infty} A_i = S$, the ~~ca~~ collection of A_i $(i \in \mathbb{Z}^+)$ forms ~~a~~ partition of S

Prob. Theory

Remark $(\bigcup_{i=1}^{\infty} A_i^c)^c = \bigcap_{i=1}^{n} A_i \in \mathcal{B}$

Def A measure μ ~~and~~ on a set S assigns a non-negative value $\mu(A)$ to a subset of S

Def A collection of subsets of S is called a σ-algebra, denote by by $\mathcal{B}$, or $\sigma(S)$, if

a. $\phi \in \mathcal{B}$

b. $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$ (closure under complementation)

c. If $A_i \in \mathcal{B}$ for $i=1,2,\dots$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$. (closure under countable unions)

Def A function μ on a σ-field of $\mathcal{A}$ is a measure if

a. For every $A \in \mathcal{A}$, $0 \le \mu(A) \le \infty$, $\mu: \mathcal{A} \to [0, \infty]$

b. If A_i are disjont, $A_i \in \mathcal{A}$, then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

c. $\mu(\mathcal{S}) = 1$ (prob. measure)

<u>10 ways looking at a random variable</u>

1. A function $X: \Omega \to \mathbb{R}$ such that images $X^{-1}(B)$ for any Borel set are elements of $\mathcal{A}$ is called a <u>random variable</u>. A p-tuple of r.v.'s is called a random vector.

2. Associated with a random vector of X on $(\Omega, \mathcal{A}, P)$ is a distribution function d.f.

$$F_{\underset{\sim}{X}}(X) = F_{X_1, X_2, \dots X_p}(x_1, x_2, \dots x_p) = P(w: X_1(w) \le x_1, \dots X_p(w) \le x_p)$$

Remark: F is right continuous. càdlàg.

3. For any scalar r.v. X with d.f. F, the quantity $Q(u) = F^{-1}(u) = \inf\{x : F(x) \ge u\}$ is called the u^{th} quantile of X (as of F), $u \in (0,1)$

$Q(1/2)$ = median. $Q(1/4)$, $Q(3/4)$: Q1. lower quantile. Q3. upper quantile.

4. If the df F is absolutely continuous with respect to the measure μ, then F has a density f w.r.t. μ. $F(x) = \int_{-\infty}^{x} f(u)\,du$. $f(x) = F'(x)$

5. The expectation of a r.v. X is $E(X) = \int_{\Omega} X(w)\,dP(w) = \int_{-\infty}^{\infty} x\,dF(x) = \int_{-\infty}^{\infty} x\, \overset{\text{R.N. dirvative}}{f(x)}\,dx$

Likewise, define the expectations of functions of X

$$E\{g(X)\} = \int_{\Omega} g(X(w))\,dP(w) = \int g(x)\,dF(x)$$

e.g. $g(x) = I(x \in B) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{ow.} \end{cases}$ $\quad E\{I(X \in B)\} = \int_B dP(w) = P(B)$.

6. Moments. Moments of higher powers of $X-\mu$ are often used to describe the basic charecteristics of the distributions of r.v.'s. In particular, we denote

$$\mu_k = E(X - EX)^k, \quad k = 1, 2, \dots \quad \text{as the } k\text{th central moment of } X.$$

e.g. $k=2$, μ_2: $Var(X) = E(X-EX)^2 = \sigma^2$ [variance].

$k=3$, $\frac{\mu_3}{\sigma^3}$: skewness (X) [asymmetry]

$k=4$ $\frac{\mu_4}{\sigma^4}$: kurtosis (X)

7. Moment Generating Function mgf.

To compute moments, it is often convenient to use the mgf, which is defined as.

$$m_X(t) = E\{\exp(tX)\} = \int e^{tx}\, dF(x)$$ Laplace transform

where $m_X(t)$ exists and its derivative exists in some neighborhood of 0.

Essentially, we have $\quad V_k = m_X^{(k)}(0) = E(X^k) \quad , k = 0, 1, 2, \ldots$

The property hold :

(a) For constant (μ, σ) $\quad : \; m_{\mu+\sigma X}(t) = \exp(\mu t)\, m_X(\sigma t)$

(b) For indep. X, Y $\quad : \; m_{X+Y}(t) = m_X(t)\, m_Y(t)$.

Example Suppose we have a discrete r.v. on $\{0, 1, 2, \ldots\}$ with

$$P(X = j) = a_j$$

We define the "generating function" of X as $\quad g(z) = \sum_{j\geq 0} a_j z^j = \sum_{j\geq 0} P(X=j)\, z^j$

Since $\sum_{j\geq 0} a_j = 1$, it is clear that

$$|g(z)| \leq \sum_j |a_j||z|^j \leq \sum_j a_j = 1 \quad \text{for } |z| \leq 1.$$

Consider the derivatives.

$$g'(z) = a_1 + 2a_2 z + 3a_3 z^2 + \cdots = \sum_{n=1}^{\infty} n a_n z^{n-1}$$

$$g''(z) = 2a_2 + 6a_3 z + \cdots = \sum_{n=1}^{\infty} n(n-1)\, a_n\, z^{n-2}$$

$$\vdots$$

$$g^{(j)}(z) = \sum_{n=j}^{\infty} (n(n-1)\cdots(n-j+1))\, a_n^{\,n-j} = \sum_{n-j}^{\infty} \binom{n}{j}(j!)\, a_n\, z^{n-j}$$

Thus, $\quad g^{(j)}(0) = j!\, a_j \quad$ or $\quad a_j = (j!)^{-1} g^{(j)}(0)$.

So all the information about the a_j's are contained within the function g and is made accessible, by diffrentiating and evaluating $g^{(k)}$ at 0.

Suppose the moments exist, then

$$g'(1) = \sum_{n=0}^{\infty} n a_n = \sum_{n=0}^{\infty} n P(X=n) = EX.$$

$$g''(1) = \sum_{n=0}^{\infty} n^2 a_n - \sum_{n=0}^{\infty} n a_n = E X^2 - (EX)^2$$

The distribution of a non-negative integer valued r.v. is uniquely determined by its generating function. $\quad a_j = (j!)^{-1} g^{(j)}(0)$.

We can write in a slightly fancy. notation:

$$g(X) = E(z^X) = E(e^{-\lambda X})$$

if X takes arbitrary real values, and consider $0 < j \le 1$, any such

z can be written as $e^{-\lambda}$, for $0 \le \lambda < \infty$

$$E(e^{-\lambda X}) = \sum_{j=0}^{\infty} p_j e^{-j\lambda}$$

$\downarrow$ $\quad$ $p_j \longrightarrow$ prob. that X takes the value of x_j.

$$E(e^{-\lambda X}) = \int e^{-\lambda u} f(u)\, du.$$

8. chracteristic function: $\phi_X(\lambda) \triangleq E[e^{i\lambda X}] = \int e^{i\lambda u} f(u)\, du$ $\leftarrow$ characteristic function.

$$|E(e^{itX})| \le E|e^{itX}| = E|\cos tX + i \sin tX|$$

$$= E(\cos^2 tX + \sin^2 tX)$$

$$= 1.$$

see e.g. 2.3.10 <u>Nonunique moments</u>

9. Cumulants.

$K_X(t) = \log M_X(t)$

Ch. 2 and 3. of C&B

Ch. 4. (bivariate transformation)

10. Conditional prob.

The conditional prob. of an event B given that an event A has occurred

is $P(B|A) = \dfrac{P(A \cap B)}{P(A)}$

If (X,Y) has a jonit density of $f_{X,Y}(x,y)$ and X with marginal $f_X(x)$

then the conditional density

$$f_{Y|X}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

Let X and Y be random k-vectors

(a) If $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}^k$, then $F_X = F_Y$. $(\phi_X(t) = E(\exp(i\vec{t}X)))$

(b) If $m_X(t) = \underset{E\{\exp(t^T Y)\}}{\overset{\|}{m_Y(t)}} < \infty$ for all t in the neighborhood of 0, then $F_X = F_Y$.

Pf: (a) See Billingsley (1968. P.395) [Inversion formula].

For any $a = (a_1, \dots a_k) \in \mathbb{R}^k$, $b = (b_1, \dots b_k) \in \mathbb{R}^k$ and $(a,b] = (a_1, b_1] \times (a_2, b_2] \times \cdots (a_k, b_k]$

satisfying $F_X = 0$, $F_X((a,b]) = \lim_{c\to\infty} \int_{-c}^{c} \cdots \int_{-c}^{c} \frac{\phi_X(t_1, \dots t_k)}{(2\pi)^{k}} \prod_{j=1}^{k} \frac{e^{-it_j a_j} - e^{it_j b_j}}{it_j}\, dt_j.$

Lecture 2

continuous

Characteristic Function $\phi_X(t)$ & Moment Generating Function $M_X(t)$

[Thm] Let X and Y be random k-vector

(a) If $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}^k$, then $F_X(x) = P(X \le x) = F_Y(x)$

(b) If $M_X(t) = M_Y(t) < \infty$ for all t in the neighborhood of 0, then $F_X = F_Y$.

Pf for (b): First consider the case $k=1$.

From $e^{s|x|} \le e^{sx} + e^{-sx}$, we can conclude that $|X|$ has an mgf that is finite in the neighborhood $(-c, c)$ for some $c>0$ and a constant s.

Observe that $$\left| e^{itx} \left\{ e^{iax} - \sum_{j=0}^{n} \frac{(iax)^j}{j!} \right\} \right| \le \frac{|ax|^{n+1}}{(n+1)!} \qquad \text{(exercise)}$$

we obtain
$$\left| \phi_X(t+a) - \sum_{j=0}^{n} \frac{a^j}{j!} E\{(iX)^j e^{itX}\} \right| \le \frac{|a| E|X|^{n+1}}{(n+1)!} \qquad (*)$$

Since
$$M_X(t) = \sum_{(r_1,\dots r_k)\in \mathbb{Z}} \frac{E(X_1^{r_1} \cdots X_k^{r_k})\, t_1^{r_1} \cdots t_k^{r_k}}{r_1!\, r_2! \cdots r_k!} \qquad (i)$$

tuto

and
$$\frac{\partial^r \phi_X(t)}{\partial t_1^{r_1} \cdots \partial t_k^{r_k}} = (-1)^{\frac{r}{2}} E\left(X_1^{r_1} X_2^{r_2} \cdots X_k^{r_k} e^{it^T X}\right), \quad r = \sum_{i=1}^{k} r_i \qquad (ii)$$

We can write $$\phi_k(t+a) = \sum_{j=0}^{\infty} \frac{\phi_k^{(j)}(t)}{j!} a^j, \quad |a| < c \qquad (+),$$
which ~~can~~ also holds when ϕ_X is replaced by ϕ_Y.

Under the assumption that $m_X = m_Y < \infty$ in the neighborhood of 0, X and Y has the same moments of all orders. By (ii), $\phi_X^{(j)}(0) = \phi_Y^{(j)}(0)$ for all $j = 1, 2, \dots$ which, and (+) with $t=0$ imply that ϕ_X and ϕ_Y are the same on the integral $(-c, c)$ and hence have identical derivatives there.

Choose $t = c-\varepsilon$ and $-c+\varepsilon$ and an arbitrary ~~some~~ small $\varepsilon > 0$ in (+), we can show that ϕ_X and ϕ_Y also agree on $(-2c+\varepsilon, 2c+\varepsilon)$ and hence $[-2c, 2c]$. Likewise, by the same arguement, ϕ_X and ϕ_Y are the same on $(-3c, 3c)$. Hence $\phi_X(t)$ and $\phi_Y(t)$ for all (t) and by (a) $F_X = F_Y$.

e.g. ~~tuto~~

For $k \geq 2$, suppose $F_X \neq F_Y$, then by (a), there exists $t \in \mathbb{R}^k$ such that $\phi_X(t) \neq \phi_Y(t)$. Then $\phi_{t^\top X}(1) \neq \phi_{t^\top Y}(1)$, which implies that $F_{t^\top X} \neq F_{t^\top Y}$. But $m_X = m_Y < \infty$ in a neighborhood of $0 \in \mathbb{R}^k$. This implies that $m_{t^\top X} = m_{t^\top Y} < \infty$ in a nbh of $0 \in \mathbb{R}$, which leads to the conclusion that $F_{t^\top X} = F_{t^\top Y}$. contradiction.

Some useful inequalities
→ Shorack and Wellner (1986) CB Chapter 3.5 ineq; convergence. Keener

[Thm] Let Z be a real r.v. and g a non-negative, non-decreasing function on the support of Z, i.e. a set B such that $P(Z \in B) = 1$, then

$$P(Z \geq a) \leq \frac{Eg(Z)}{g(a)}$$

Pf: observe that $g(a) I(Z \geq a) \leq g(Z) I(Z \geq a) \leq g(Z)$

Taking expectation. done.

Examples:

(a) Markov: $Z = |X|$, $g(t) = \overset{\max(0,t)}{\overset{\parallel}{t^+}} \Rightarrow P(|X| \geq a) \leq \frac{E|X|}{a}$

(b) Chebyshev: $Z = |X|$, $g(t) = t^2 \Rightarrow P(|X| \geq a) \leq \frac{E(X^2)}{a^2}$

(c) Bernstein: $Z = X$, $g(t) = e^{st} \Rightarrow P(X \geq a) \leq \frac{E(e^{sX})}{e^{sa}}$

One can construct examples for which they are actually sharp. For example, in the case of Markov inequality, suppose

$$X = \begin{cases} a & \mu/a \\ 0 & 1-\mu/a, \end{cases}$$

then $E(X) = \mu$ and obviously $P(X \geq a) = \frac{E(X)}{a}$.

[Thm] (Cauchy-Schwarts)

Let $X = (X_1, \dots, X_p)$ be a p-vector of real r.v.'s and $U = E(XX^\top)$. The matrix U is symmetric, non-negative definite with singularity ($|U| = 0$) iff there exists a p-vector $\alpha \neq 0$ such that $\underline{E(\alpha^\top X)^2 = 0 \ (*)}$

Proof: Since Expectation (E) is applied componentwise (and multiplication commutes) symmetry is immediate. Non-negative definiteness follows from

$$\alpha^\top U \alpha = E(\alpha^\top X)^2 \geq 0.$$

If equality holds, then clearly $(*)$ holds and U is singular since $U\alpha = 0$. On the other hand, if U is singular, there must exist $\alpha \neq 0$ s.t. $U\alpha = 0$ as the equality holds

[Corollary] For r.v.'s X_1 and X_2

$$\{E(X_1X_2)\}^2 \leq E(X_1^2)E(X_2^2)$$

$\tilde{X}_j = X_j - E(X_j) \quad j=1,2,\ldots$

and centering X_1 and X_2 at there respective measure yields.

$$\{cov(X_1, X_2)\}^2 \leq Var(X_1)Var(X_2). \Rightarrow corr(X_1, X_2) = \frac{cov(X_1,X_2)}{\sqrt{Var(X_1)Var(X_2)}} \in [-1,1]$$

Proof: Consider the previous thm with $p=2$, and recalling that $|u|$ may be expressed as the product of its eigenvalue which are non-negative, we have

$$0 \leq |u| = \frac{E(X_1^2)E(X_2^2)}{-\{E(X_1X_2)\}^2}$$

[Thm] (Jensen) If X and $g(X)$ are integrable r.v.'s and $g(\cdot)$ is convex, then

$$g(E(X)) \leq E(g(X)).$$

Proof: Convexity of g implies that for any ξ, there exists a line L through the point $(\xi, g(\xi))$ such that the graph g is above the line, i.e.

$$g(x) \geq g(\xi) + \lambda(x-\xi)$$

In particular, we let $\xi = E(X)$, then for all x,

$$g(x) \geq g(E(X)) + \lambda\{x - E(X)\}.$$

Note that λ depends on ξ but not on x. Now, let $x = X$ and

$$g(X) \geq g(E(X)) + \lambda\{X - EX\}.$$

Taking expectation, done.

[Corollary] (Liapounov) $\{E|X|^r\}^{1/r}$ is $\nearrow$ in r for $r \geq 0$

↳ Hölder. Minkowski..

Proof: By Jensen's ineq, since $|X|^r$ is convex in $|X|$ for $r \geq 1$,

Remark: we have $(E|X|)^r \leq E|X|^r$ in which case $E|X| \leq (E|X|^r)^{1/r}$

Moment inequality. Now, replace $|X|$ by $|X|^q$ for $0<q<r$,

If X has a rth moment, it also have the qth moment for $q<r$.

as $(E|X|^q)^{1/q} \leq (E|X|^{rq})^{1/(rq)} \triangleq (E|X|^s)^{1/s}$ where $s=rq$.

for $0<q<s<\infty$, since $r \geq 1$, so $q \leq rq = s$.

Convergence Results.

We are interested in sequences $X_1, X_2, \ldots$ of r.v.'s on a p-space $(\Omega, \mathcal{A}, P)$

I. Convergence in probability

Let $\{X_i\}_{i\geq 1}$ and X be real-valued r.v.'s on $(\Omega, \mathcal{A}, P)$.

We say that X_n converges in prob. to X if

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1, \quad \text{for any } \varepsilon > 0.$$

and we write usually $X_n \xrightarrow{P} X$.

Remark: Often X will be a degenerate r.v. e.g. $X_n = \bar{X}_n = \frac{\sum_{i=1}^{n} Z_i}{n}$.

where $Z_i \overset{iid}{\sim} N(\mu, \sigma^2)$, then $X_n \xrightarrow{P} \mu$, we can think of μ as the degenerate r.v. X, which takes the value μ with prob 1.

II Almost sure convergence / convergence with prob. 1

We say X_n converges almost surely (a.s.), or converges with prob. 1

if $\quad P(\lim_{n\to\infty} X_n = X) = 1$.

or equivalently, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_m - X| < \varepsilon \text{ for all } m > n) = 1$$ } tuto

One can show that a.s. convergence $\Rightarrow$ convergence in prob.

(and we have counter examples to show that the converse is wrong)

III. Convergence in the qth mean

X_n Converges in the qth mean to X if $\boxed{\lim_{n\to\infty} E|X_n - X|^q = 0}$

By the moment inequality introduced earlier,

$X_n \xrightarrow{q\text{th}} X \Rightarrow X_n \xrightarrow{p\text{th}} X$ for any $p < q$. Often, $q = 2$ in practice.

As a example of extreme behavior, suppose that $\boxed{X_n = \begin{cases} 0 & 1 - n^{-3} \\ n & n^{-3} \end{cases}}$,

then taking $X = 0$, we have $\lim_{n\to\infty} E|X_n - X|^q = 0$ for $q = 1, 2$, but $E|X_n - X|^3 = 1$.

IV. Convergence in distribution (in law)

X_n converges in distribution to X if for their respective distribution functions $\lim_{n\to\infty} F_n(x) = F(x)$ at each point of contunity of F.

We write $X_n \xrightarrow{d/D/L} X$ or as $F_n \Rightarrow F$ (F_n converges weakly to F)

Often, we write $X_n \rightsquigarrow X$ e.g. $X_n \rightsquigarrow N(0,1)$

e.g. tuto

$\text{II} \Rightarrow \text{I} \Rightarrow \text{IV} \Rightarrow \text{III}$

Big / Small O notation

For positive deterministic sequences $\{a_n\}, \{b_n\}$

a) If there is a $\Delta < \infty$ s.t $a_n/b_n \leq \Delta$ for sufficiently large n, we say $a_n = O(b_n)$

b) if $a_n/b_n \to 0$, we say $a_n = o(b_n)$

Clearly, if $a_n = O(n^r)$ and $b_n = O(n^s)$ then $a_n b_n = O(n^{r+s})$

and $a_n + b_n = O(n^{\max\{r,s\}})$

For sequences $\{X_n\}$ and $\{Y_n\}$ of r.v.'s on $(\Omega, \mathcal{A}, P)$ and any $\varepsilon > 0$.

a*) If there exists $\Delta < \infty$ s.t. $P(|X_n| \geq \Delta |Y_n|) < \varepsilon$ for sufficently large n,

we write $X_n = O_p(Y_n)$

b*) If $P(|X_n| > \varepsilon |Y_n|) \xrightarrow{n\to\infty} 0$, then we write $X_n = o_p(Y_n)$

In many cases, Y_n will be deterministic, we write correspondingly

$X_n = O_p(1)$: "bounded in prob."

$X_n = o_p(1)$: "tending 0 in prob."

[Thm] (Slusky) ... δ-method.

Let $\boxed{X_n \rightsquigarrow X}$ and $\boxed{Y_n \xrightarrow{P} y}$, a real constant. Then.

$X_n \xrightarrow{d} X$

(a) $X_n + Y_n \rightsquigarrow X + y$ (exercise).

(b) $X_n Y_n \rightsquigarrow yX$

Proof for (b)

Suppose $y=0$, and let $B>0$ be a real constant, and denote

$$X_n^B = X_n I(|X_n| \le B)$$

Then $\{|Y_n X_n| \ge \varepsilon\} = \{|Y_n||X_n^B| \ge \varepsilon\} \cup \{|Y_n||X_n - X_n^B| \ge \varepsilon\}$ (*)

$$\{|Y_n||X_n^B| \ge \varepsilon\} \subseteq \{|Y_n| > \frac{\varepsilon}{B}\}$$

and $P\{|Y_n||X_n^B| \ge \varepsilon\} \le P\{|Y_n| \ge \frac{\varepsilon}{B}\} \to 0$.

By the hypothesis that $X_n = O_p(1)$, there exists $\delta > 0$. and $B_j < \infty$ s.t. for n sufficiently large, $P(|X_n - X_n^{B_\delta}| > 0) < \delta$.

Since $\{|Y_n||X_n - X_n^B| \ge \varepsilon\} \subseteq \{|X_n - X_n^B| > 0\}$

(*) and addivity implies that $\lim_{n\to\infty} P\{\underline{|X_n||Y_n|} \ge \varepsilon\} < \delta$. $|X_nY_n - 0|$

Since ε and δ are arbitrary, we have shown that $X_n Y_n \xrightarrow{P} 0$.

The result follows by noticing that Y_n can be replaced by $Y_n - y$.

[Thm] (Continuous mapping)

If $X_n \rightsquigarrow X$ and g is continuous, $g(X_n) \rightsquigarrow g(X)$. Proof skipped.

[Thm] (δ-method)

Suppose $a_n(X_n - b) \rightsquigarrow X$, where a_n is a sequence of constants tending to ∞ and b is a fixed number. Let $g: \mathbb{R} \to \mathbb{R}$ be differentiable with continuous derivative g' at b. Then $a_n\{\underset{\uparrow\, x}{g(X_n)} - g(b)\} \rightsquigarrow \underset{\text{slope}}{g'(b)}\, \underset{\downarrow\, \mathrm{Var}(X)}{X}$.

Proof: By Slusky's thm,

$$X_n - b = a^{-1}\{a_n(X_n - b)\} \longrightarrow 0$$

and therefore $X_n \to b$. Now apply mean value thm to $g(X_n) - g(b)$, we have $g(X_n) - g(b) = g'(X_n^*)(X_n - b)$

where $|X_n^* - b| \le |X_n - b|$ where $X_n^* \to b$, so by the conuity of g' and cont. mapping thm (ACMT). $g'(X_n^*) \to g'(b)$. Multiplying a_n and again applying Slusky, we have the result. The above argument generalizes to $X_n, X \in \mathbb{R}^q$.

LLN, CLT (dependent)

Lecture 3

References: 1) C&B. Chapter 5&6 (data reduction) 2) Keener Chapter 8. 2-3. 3) Ferguson (convergence results)

Asmptotic behavior of sample mean.

* (i) Law of Large Numbers (LLNs) : $\hat{\mu} \to \mu$ $\quad \hat{\mu} = \bar{X} = n^{-1} \sum_{i=1}^{n} X_i$

* (ii) CLT : $\sqrt{n}(\hat{\mu}-\mu) \rightsquigarrow N(0, \sigma^2)$

(i) weak LLN : Let $Z_1 \ldots$ be indep. r.v.'s with means $\mu_1 \ldots$ and variance $\sigma_1^2, \ldots$

Z_n $\quad n^{-2}\sum_{i=1}^{n} \sigma_i^2 \to 0$ as $n \to \infty$, then $\bar{Z} = \hat{\mu} \to \bar{\mu}$

· Pf: $P(|\bar{Z}-\bar{\mu}| \geq \varepsilon) \leq \dfrac{E(\bar{Z}-\bar{\mu})^2}{\varepsilon^2} = \dfrac{\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2}{\varepsilon^2} \to 0$

Remark: If $\sigma_i^2 = \sigma^2$, then $\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 = \sigma^2/n$. $\quad \bar{Z}-\mu = o_p(\frac{1}{n})$
(for iid)

SLLN (Kolmogrov) For $\{Z_i, \mu_i, \sigma_i^2\}$ as above, if $\sum_{i=1}^{n} \frac{\sigma_i^2}{i^2} < \infty$, then $\bar{Z} \to \bar{\mu}$ a.s.

In particular, if $\{Z_i\}_{i=1}$'s are iid, with $E(Z_1) = \mu$, then $\bar{Z} - \mu$ a.s.

An important special case if the SLLN involves taking

$$Z_i = I_{[X_i,\infty)}(x) = I(X_i \leq x)$$

for iid r.v.'s $X_i \sim F$ and fixed x. Observe that

$$E(Z_i) = P(X_i \leq x) = F(x)$$

We can infer that $\quad F_n(x) \triangleq n^{-1}\sum_{i=1}^{n} I(X_i \leq x) \xrightarrow{a.s.} F(x)$ as $n \to \infty$.

This can be strengthened.

[Thm] (Glivenko-Cantelli)

$$P\left(\sup_x |F_n(x) - F(x)| \to 0\right) = 1.$$

· Pf: Let $\varepsilon > 0$ and find an integer $k > 1/\varepsilon$ and numbers

$-\infty = x_0 < x_1 \leq \cdots \leq x_{k-1} < x_k = \infty$ such that. $F(x_j^-) \leq \frac{j}{k} \leq F(x_j)$, $j = 1, \ldots k-1$.

where $F(x^-) = P(X < x)$.

Note that if $x_{j-1} < x_j$, then $F(x_j^-) - F(x_{j-1}) \leq \varepsilon$.

From SLLN, $F_n(x_j) \xrightarrow{a.s} F(x_j)$ and $F_n(x_j^-) \xrightarrow{a.s} F(x_j^-)$ for $j = 1, 2, \ldots k-1$.

Hence $\quad \Delta_n = \max\{|F_n(x_j) - F(x_j)|, |F_n(x_j^-) - F(x_j^-)|, j = 1, \ldots, k-1\} \to 0$.

Now, let x be arbitrary and find j s.t. $x_{j-1} < x \leq x_j$.

Then $F_n(x) - F(x) \leq F_n(x_j^-) - F(x_{j-1}) \leq F_n(x_j^-) - F(x_j^-) + \varepsilon$

and $F_n(x) - F(x) \geq F_n(x_{j-1}) - F(x_j) \geq F_n(x_{j-1}) - F(x_{j-1}) - \varepsilon$

Thus $\sup_x |F_n(x) - F(x)| \leq \Delta_n + \varepsilon \xrightarrow{a.s} \varepsilon$ as $n \to \infty$ (Van dei Vaant, 98. Chap 79)

Since this holds for all $\varepsilon > 0$ and the results follow □. Vanpnik (99).

(ii) CLT.

Suppose $Z_1, \ldots, Z_n$ are iid $N(0,1)$ or $Z \sim N(0, I_n)$. we know that $\alpha^T Z \sim N(0, \alpha^T \alpha)$

So, for instance, if we take $\alpha = n^{-1/2} I_n$, we have $n^{-1/2} \sum_{i=1}^{n} Z_i \sim N(0,1)$.

or equivalently $\sqrt{n}\bar{Z} \sim N(0,1)$.

[Thm] Suppose $X_1, \ldots$ are iid with $E(X_i) = \mu$, $Var(X_1) = \sigma^2$, then $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow N(0,1)$.

Pf: Existence of μ and σ^2 implies that the moment expension of the cf of X_1 can be written as $\phi_{X_1}(t) = \exp\{i\mu t - \frac{1}{2}\sigma^2 t^2 + o(t^3)\}$.

Define $S_n = X_1 + \cdots + X_n$, which has $\phi_{S_n}(t) = \phi^n_{X_1}(t)$

and let $u_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma^2} = \frac{\sqrt{n}(S_n/n - \mu)}{\sigma}$

$$\phi_{u_n}(t) = E(e^{+itu_n}) = \phi^n_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right) \exp\left(-\frac{i\mu t\sqrt{n}}{\sigma}\right)$$

$$= \left[\exp\left\{\frac{i\mu t}{\sigma\sqrt{n}} - \frac{\frac{1}{2}\sigma^2 t^2}{\sigma^2 n} + o\left(\frac{t^2}{\sigma^2 n}\right)\right\}\right]^n \exp\left(-\frac{i\mu t\sqrt{n}}{\sigma}\right)$$

$$= \exp\left\{-\frac{1}{2}t^2 + n\, o\left(\frac{t^2}{n}\right)\right\} \longrightarrow \exp\left(-\frac{1}{2}t^2\right) \text{ as } n \to \infty$$

which is the cf of $N(0,1)$.

<u>Why normal?</u>

If $\sqrt{n}(\bar{X}_n - \mu) \rightsquigarrow X$, what does X look like?

only normal distr. fufills the setup. (Stable laws).

Consider $Z_{2n} = \frac{X_1 + \cdots + X_n + X_{n+1} + \cdots + X_{2n}}{\sqrt{2n}}$

Clearly, $Z_{2n} \rightsquigarrow X$ but $Z_{2n} = \frac{X_1 + \cdots + X_n}{\sqrt{2n}} + \frac{X_{n+1} + \cdots + X_{2n}}{\sqrt{2n}} \triangleq Z_{an} + Z_{bn}$

Chow and Teicher (3rd)

where Z_{an} and $Z_{bn} \rightsquigarrow \frac{X}{\sqrt{2}}$

[Thm] (Lyapunov) Let $X_1, \ldots$ be indep. with $EX_i = 0$ and $EX_i^2 < \sigma_i^2 < \infty$. $E|X_i|^3 < \infty$ and $S_n^2 = \sum_{i=1}^{n} \sigma_i^2$. If $\overline{\lim_{n\to\infty}} \sum_{i=1}^{n} E|X_i|^3 = 0$, then $S_n^{-1} \sum_{i=1}^{n} X_i \rightsquigarrow N(0,1)$.

Lindeberg Feller cond. Martingale CLT ...

CLT for dependent cases / sequences.

One solution : α-mixing

Given a sequence $X_1, X_2, \dots$ and sets $A \in \sigma(X_1, \dots X_k)$.

and $B \in \sigma(X_{k+n}, X_{k+n+1}, \dots)$ for $k \geq 1$ and $n \geq 1$, then if there exists

a sequence of real numbers $\alpha_n \to 0$ s.t.

$$|P(A \cap B) - P(A)P(B)| \leq \alpha_n$$

then $\{X_n\}$ is α-mixing.

Special case If $\alpha_n = 0$ for $n > m$, then the sequence is said to be m-dependent.

CLT for α-mixing sequences

Suppose $X_1, X_2, \dots$ is stationary and α-mixing with $\alpha_n = O(n^{-5})$. $E(X_n) = 0$ and $E(X_n^{12}) < \infty$. Set $S_n = X_1 + \dots + X_n$. If $n^{-1} Var(S_n) \to \sigma^2 = E(X_1^2) + \sum_{k=1}^{\infty} E(X_1 X_{k+1})$ converges absolutely with $\sigma^2 > 0$, then $S_n / \sigma\sqrt{n} \rightsquigarrow N(0,1)$.

Extra examples With assumptions in CLT, if f is differentiable at μ, then

$$\sqrt{n}\{f(\bar{X}_n) - f(\mu)\} \rightsquigarrow N(0, [f'(\mu)]^2\sigma^2).$$

Pf : Write $f(\bar{X}_n) = f(\mu) + f'(\mu_n)(\bar{X}_n - \mu)$, where μ_n is an immediate point between $\bar{X}_n$ and μ. Since $|\mu_n - \mu| \leq |\bar{X}_n - \mu|$ and $\bar{X}_n \xrightarrow{P} \mu$ (LLNs) and since f' is continuous, $f'(\mu_n) \xrightarrow{P} f'(\mu)$. If $Z \sim N(0, \sigma^2)$, then $\sqrt{n}(\bar{X}_n - \mu) \rightsquigarrow Z \sim N(0, \sigma^2)$, by CLT. Thus by Slusky's thm, $\boxed{\sqrt{n}\{f(\bar{X}_n) - f(\mu)\}} = f'(\mu_n)\{\sqrt{n}(\bar{X}_n - \mu)\} \to f'(\mu)Z \sim N(0, [f'(\mu)]^2\sigma^2)$.

Asymptotics of medians and percentiles

For regularity, assume F has a unique median θ, so $F(\theta) = \frac{1}{2}$, and that $F'(\theta)$ exists, which is finite and positive. We want to study the asymptotic distribution of $\sqrt{n}(M_n - \theta)$, where M_n denotes the sample median of $(X_1, \dots X_n) \overset{iid}{\sim} f$. (with d.f. F)

$$P(\sqrt{n}(M_n - \theta) \leq a) = P(M_n \leq \theta + a/\sqrt{n})$$

Define $S_n = \#\{i \leq n : X_i \leq \theta + \frac{a}{\sqrt{n}}\} = \sum_{i=1}^{n} I(X_i \leq \theta + \frac{a}{\sqrt{n}})$

m: the middle integer.

Note that $\{M_n \leq \theta + \frac{a}{\sqrt{n}}\}$ if $\{S_n \geq m\}$. It is evident that, if we treat the observation i as a success if $X_i \leq \theta + \frac{a}{\sqrt{n}}$, then $S_n \sim \text{Binomial}(n, F(\theta + \frac{a}{\sqrt{n}}))$

Let $Y_n \sim \text{binomial}(n, p)$, then by CLT,

$$\sqrt{n}\left(\frac{Y_n}{n} - p\right) = \frac{Y_n - np}{\sqrt{n}} \rightsquigarrow N(0, p(1-p))$$

in which case $P\left(\frac{Y_n - np}{\sqrt{n}} > y\right) \to 1 - \Phi\left(\frac{y}{\sqrt{p(1-p)}}\right) = \Phi\left(\frac{-y}{\sqrt{p(1-p)}}\right)$ as $n \to \infty$.

where $\Phi(\cdot)$ denotes the cdf of $Z \sim N(0,1)$

Hence, the normal approximation for the binomial distribution gives

$$P(\sqrt{n}(M_n - \theta) \le a) = P(S_n > m-1)$$

$$= P\left\{ \frac{S_n - nF(\theta + a/\sqrt{n})}{\sqrt{n}} > \frac{m-1-nF(\theta+a/\sqrt{n})}{\sqrt{n}} \right\}$$

$$= \Phi\left(\frac{[nF(\theta + a/\sqrt{n}) - m + 1]/\sqrt{n}}{\sqrt{F(\theta + a/\sqrt{n})(1 - F(\theta + a/\sqrt{n}))}} \right) + o(1) \qquad (*)$$

Keener (p138)

Since F is continuous at θ.

$$\left[F(\theta + a/\sqrt{n}) \left\{ 1 - F(\theta + \tfrac{a}{\sqrt{n}}) \right\} \right]^{1/2} \to 1/2 \quad \text{as } n \to \infty.$$

Because F is differentiable at θ,

TPE, TSH > lehmann.

$$\frac{nF(\theta + \frac{a}{\sqrt{n}}) - m + 1}{\sqrt{n}} = \frac{aF(\theta + \frac{a}{\sqrt{n}}) - F(\theta)}{a/\sqrt{n}} - \frac{nF(\theta) - m + 1}{\sqrt{n}}$$

$$= \frac{aF(\theta + a/\sqrt{n}) - F(\theta)}{a/\sqrt{n}} + \frac{1}{2\sqrt{n}} \to aF'(\theta).$$

Since the numerator and the denominator of the argument of Φ in $(*)$ both converge, we can write

$$P(\sqrt{n}(M_n - \theta) \le a) \to \Phi(2aF'(\theta))$$

The limit here is the cdf of a normal r.v. with mean 0 and variance $[4\{F'(\theta)\}^2]^{-1}$ evaluates at a and so

$$\sqrt{n}(M_n - \theta) \to N\left(0, \frac{1}{4[F'(\theta)]^2}\right)$$

[Thm] Let $X_1, \dots, X_n$ be iid with common cdf F. and let $\tau \in (0,1)$ and let $\tilde{\theta}_n$ be the $\lfloor \tau n \rfloor$ th order statistic for $X_1, \dots, X_n$. where $\lfloor X \rfloor$, floor of X.
If $F(\theta) = \tau$ and if $F'(\theta)$ exists and is finite and positive.
then $\sqrt{n}(\tilde{\theta}_n - \theta) \rightsquigarrow N\left(0, \frac{\tau(1-\tau)}{\{F''(\theta)\}^2}\right)$

Data Reduction : Sufficiency ...

Definition (Statistic) A statistic T is a function of the data. $= T(\underset{\sim}{X}) = T(X_1, \dots, X_n)$

Def (Sufficient statistic) A statistic is sufficient for a mode $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if for all t, the conditional distribution $X \mid T(X) = t$ does not depend on θ.

Example (Weighted coin flips)

Let $X_1, X_2, \dots, X_n$ be iid according to Bernoulli (θ). is the number of heads. i.e. $\sum_{i=1}^n X_i$ sufficient ? To check this, let's show the conditional distribution of $\underset{\sim}{X}$ given $\sum_{i=1}^n X_i$.

We have $\quad P_\theta(X) = \prod_{i=1}^n \theta^{X_i} (1-\theta)^{1-X_i} = \theta^{\sum_{i=1}^n X_i} (1-\theta)^{n-\sum_{i=1}^n X_i}$.

So the conditional distribution is

$$P_\theta(X=x \mid T(x)) = \frac{P_\theta(X=x, T(x)=t)}{P(T(x)=t)} = \frac{I(\sum_{i=1}^n X_i = t)\, \theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \frac{I(\sum_{i=1}^n X_i = t)}{\binom{n}{t}},$$

which does not depend on θ, so the sum of heads is a sufficient stat.

Example (Max of uniform)

Let $X_1, \dots, X_n$ be iid uniform $(0, \theta)$. Then $T(X) = \max(X_1, \dots, X_n)$ is sufficient.

To see the intuition, think of $X_1, \dots, X_n$ as n numbers ~~at~~ on the real line, then the remaining $n-1$ numbers, given the maximum is fixed at t, behave like $n-1$ iid random samples drawn from $\mathcal{U}(0,t)$.

Example (Order statistic)

Let $X_1, \dots, X_n$ be iid with any model. Then the order statistic $T = \{X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}\}$ are sufficient.

[Thm] (TPE 1.6. Thm 6.1) Point estimation. P33.

If $X \sim P_\theta \in P$ and T is sufficient for P. then for any decision procedure δ, there is a (possibly randomized) decision procedure of equal risk. that depend on X only through $T(X)$.

[Thm] Neyman-Fisher Factorization Criterion (NFFC) ← TSH. P.19.

Suppose each $P_\theta \in P$ has density $p(X;\theta)$ w.r.t a common σ-finite measure μ, i.e. $\frac{dP_\theta}{d\mu} = p(x;\theta)$. Then $T(X)$ is sufficient iff $p(x;\theta) = g_\theta(T(x))\, h(x)$ for some g_θ and h.

- Pf: (Discrete)

Suppose $p(x;\theta) = g_\theta(T(x))\, h(x)$. Since $P_\theta(X=x \mid T(X)=t) = 0$ whenever $T \ne T(x)$. so we may focus our attention to the case where $P_\theta(X=x \mid T(X) = T(x))$.

We can write $P_\theta(X=x \mid T(X)=T(x)) = \dfrac{P_\theta(X=x, T(X)=T(x))}{P_\theta(T(X)=T(x))} = \dfrac{P_\theta(X=x)}{P_\theta(T(X)=T(x))}$

$$= \frac{g_\theta(T(x))\, h(x)}{\sum_{x'} p(x';\theta) I(T(x')=T(x))} = \frac{g_\theta(T(x))\, h(x)}{\sum_{x'} g_\theta(T(x))\, h(x')\, I(T(x')=T(x))}$$

$$= \frac{h(x)}{\sum_{x'} h(x') I(T(x')=T(x))},$$ which is ind. of θ and hence T is sufficient.

Conversely, suppose $P_\theta(X=x \mid T(X)=T(x))$ is indep. of θ.

Then, defining $h(x) = P_\theta(X=x \mid T(X)=T(x))$, we have

$$p(x;\theta) = P_\theta(X=x) = P_\theta(X=x, T(X)=T(x))$$

$$= P_\theta(X=x \mid T(X)=T(x))\, P_\theta(T(X)=T(x)) = h(x)\, g_\theta(T(x))$$

which establishes the criterion.

Example (normal)

Let X_i be iid $N(\mu, \sigma^2)$ and $\theta=(\mu,\sigma^2)$. The joint density is ~~$p(x;\theta)$~~

$$p(x;\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{1}{2\sigma^2}\left(-\sum_{i=1}^n x_i^2 + 2\mu\sum_{i=1}^n x_i - n\mu^2\right)}$$

$$= g_\theta(T(x)) \qquad \text{where } T(X) = \left(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i\right)$$

Example Suppose X and Y are indep. with common Lebesgue density $f_\theta(x) = \theta e^{-\theta x} I(x \geq 0)$. Let U be indep. of X and Y and uniformly distributed on $(0,1)$. Take. $T = X+Y$ and define $\tilde{X} = UT$ and $\tilde{Y} = (1-U)T$.

To find $f_{\tilde{X},\tilde{Y}}$, observe that $P(T \leq t \mid Y=y) = P(X+Y \leq t \mid Y=y)$

$$= E\{I(X+Y \leq t) \mid Y=y\}$$

$$= F_X(t-y)$$

$$\Rightarrow F_T(t) = P(T \leq t) = E\{F_X(t-Y)\} = \int_0^\infty 1 - e^{-\theta(t-y)}\, \theta e^{-\theta y}\, dy$$

$$= 1 - e^{-\theta t} - t\theta e^{-\theta t}$$

Hence $f_T(t) = \dfrac{\partial F(t)}{\partial t} = t\theta^2 e^{-\theta t}$, $t \geq 0$.

Also, $f_{T,U}(t,u;\theta) = t\theta^2 e^{-\theta t} I(t \geq 0, u \in [0,1])$.

From which, we have $P\left(\begin{pmatrix}\tilde{X}\\ \tilde{Y}\end{pmatrix} \in B\right) = \iint I_B(tu, t(1-u))\, f_{T,U}(t,u;\theta)\, du\, dt$

$$\overset{\text{simplication}}{=} \iint I_B(x,y)(x+y)^{-1} f_{T,U}\left(x+y, \frac{x}{x+y}\right) dy\, dx.$$

Thus $(\tilde{X}, \tilde{Y})$ has the density

$$\frac{f_{\bar{X},\bar{Y}}\left(x+y, \frac{x}{x+y}\right)}{x+y} = \begin{cases} \theta^2 e^{-\theta(x+y)} & , x \geq 0, y \geq 0 \\ 0 & , ow \end{cases}$$

Lecture 4

Keener: Ch2&3 C&B: Ch6.

References: 1. C&B ☆ 2. keener ☆☆ 3. TPE ☆☆☆ 4. TSH ☆☆☆ 5. Bickel & Doksum ☆ 6. Serrich (? CMU) ☆☆ 7. Shao jun ☆☆ 8. Van der Vaart ☆☆☆ (asymptotic). 9. Ferguson (?).

Recall: X, Y. U: unif$(0,1)$, UT, $(1-U)T$, $T = X+Y$.

NFFC: $p(x;\theta) = g_\theta(T(x))\,h(x)$

Def Exponential families: A dominated family $\{P_\theta : \theta \in \Theta\}$ is said to form a k-dimensional exponential family if the corresponding density function $\{p_\theta(x)\}_{\theta\in\Theta}$ are of the form $p_\theta(x) = \exp\left\{\sum_{i=1}^k \eta_i(\theta) T_i(x) - \underbrace{B(\theta)}_{\text{normalizing constant}}\right\} h(x)$, where $h, T_1, \dots, T_k : \mathcal{X} \to \mathbb{R}$. $B, \eta_1, \dots \eta_k : \Theta \longrightarrow \mathbb{R}$.

By NFFC, we can see that $(T_1, \dots, T_k)$ is sufficient.

E.g. $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$, $\Theta = \mathbb{R} \times (0,\infty)$

$$p_\theta(X) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\sum_{i=1}^n \frac{X_i^2}{\sigma^2} + \frac{\mu\sum_{i=1}^n X_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right\}$$

Hence. $T_1(X) = \sum_{i=1}^n X_i^2$, $\eta_1(\theta, \sigma^2) = -\frac{1}{2\sigma^2}$ natural parameters.

$T_2(X) = \sum_{i=1}^n X_i$, $\eta_2(\theta, \sigma^2) = \frac{\mu}{\sigma^2}$

$B(\mu, \sigma^2) = \frac{n\mu^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)$

$h(X) = I(x \in (-\infty, \infty))$

* A measure ν is dominated by the measure μ (say Lebesgue) if $\nu \ll \mu$, which means that for ~~some~~ all measurable A, $\mu(A) = 0$ implies $\nu(A) = 0$. A family of prob. measure $(P_\theta)_{\theta\in\Theta}$ is dominated by μ iff for each $\theta \in \Theta$, the measure P_θ is dominated by μ.

E.g. $X_1, \dots, X_n \overset{iid}{\sim}$ Cauchy, i.e. $p_\theta(X) = \frac{1}{\pi}\frac{1}{1+(X-\theta)^2}$ is the density of P_θ w.r.t Lebesgue measure. In this case, $T(X) = (X_{(1)}, \dots, X_{(n)})$ is sufficient, where $(X_{(1)} \le \dots \le X_{(n)})$ are the order statistics. In deed $T(X)$ is minimal sufficient.

[Thm] (Pitman-Koopman-Darmois) (1936).

Suppose $(X_1, \dots, X_n)$ are iid with density $\{p_\theta : \theta \in \Theta\}$ w.r.t Lebesgue measure, which are continuous in x for θ fixed and support on an interval $I \subseteq \mathbb{R}$. Suppose there exists a sufficient statistic $(T_1, \dots, T_k)$ which are continuous:

(i) If $k=1$, then $p_\theta(X) = e^{\eta(\theta)T(X) - B(\theta)}\,h(X)$.

(ii) If $n > k > 1$, and the function $x \mapsto p_\theta(x)$ are continuous differentiable. then $p_\theta(x) = e^{\sum_{i=1}^k \eta_i(\theta) T(X) - B(\theta)}\,h(X)$.

Def An exponential family is in <u>canonical form</u> when the density has the form

$$p_\eta(x) = \exp\left\{ \sum_{i=1}^{k} \eta_i T_i(x) - A(\eta) \right\} h(x)$$

This parametrizes the density in terms of the <u>natural parameters</u> η instead of θ.

Def The set of all valid natural parameters Θ is called the <u>natural parameter space</u>:

for each $\eta \in \Theta$, there exists a normalizing constant $A(\eta)$ s.t. $\int p_\eta(x)\,dx = 1$.

Equivalently, $\Theta = \left\{ \eta : 0 < \int \exp\left(\sum_{i=1}^{k} \eta_i T_i(x)\, h(x) \right) d\mu(x) < \infty \right\}$

Thus, for any canonical exponential family, $P = \{ p_\eta, \eta \in H \}$, we have $\eta \in \Theta$.

<u>Reducing the dimension</u>

There are two cases when the superficial dimension of a k-dim exponential family $P = \{ p_\eta : \eta \in H \}$ can be reduced.

<u>Case I</u> The $T_i(x)$'s satisfy an affine equality constraint $\forall x \in X$.

$\exp\{-(\eta_1+\eta_2)x + \log(\eta_1+\eta_2)\} I(x>0)$

E.g. $X \sim Exp(\eta_1, \eta_2)$ i.e. $p(x_i, \eta_1, \eta_2) = \exp\{ -\eta_1 x - \eta_2 x + \log(\eta_1+\eta_2) \}$.

Hence $T_1(x) = T_2(x) = x$. i.e. they are linearly dependent. $\Rightarrow$ unidentifiable

Def If $P = \{ p_\theta, \theta \in \Theta \}$, then θ is unidentifiable if for two parameters $\theta_1 \neq \theta_2$, $p_{\theta_1} = p_{\theta_2}$.

In the ~~piecewise~~ previous example, $p(x; \eta_1 + a, \eta_2 - a) = p(x; \eta_1, \eta_2)$ for any $a < \eta_2$.

<u>Case 2</u> The η_i's satisfy an affine equality constraint for all $\eta \in H$.

E.g. $p(x;\eta) \propto \exp(\eta_1 x + \eta_2 x^2)$ for all (η_1, η_2) satisfying $\eta_1 + \eta_2 = 1$.

$$= \exp\{ \eta_1 (x - x^2) + x^2 \}.$$

Def . . .

A canonical exponential family $P = \{ p_\eta : \eta \in H \}$ is <u>minimal</u> if

no linear combination constraints.

- $\sum_{i=1}^{k} \lambda_i T_i(x) = \lambda_0 \ \forall x \in X \Rightarrow \lambda_i = 0 \ \forall i \in \{0, \dots k\}$. (no affine T_i equality).
- $\sum_{i=1}^{k} \lambda_i \eta_i = \lambda_0 \ \forall \eta \in H \Rightarrow \lambda_i = 0 \ \forall i \in \{0, \dots, k\}$ (no affine η_i equality).

Keener Ch.5.

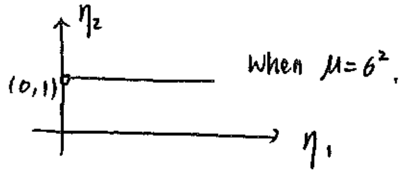
Def

Suppose $P = \{ p_\eta : \eta \in H \}$ is a k-dimensional minimal exponential family. If H contains an open k-dim rectangle, then P is called <u>full rank</u>, otherwise P is <u>curved</u>.

We illustrate three types of exponential families via normal dis. $N(\mu, \sigma^2)$, where $\eta_1 = \frac{1}{2\sigma^2}$, $\eta_2 = \frac{\mu}{\sigma^2}$, $T_1(x) = x^2$ and $T_2(x) = x$.

I) Non-minimal (so the dimension can be reduced)

when $\mu = \sigma^2$, $\eta_1 = \frac{1}{2\sigma^2}$, $\eta_2 = 1$.

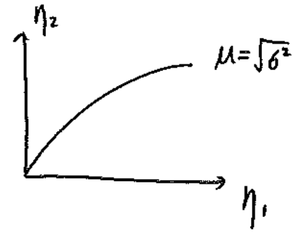


II) Minimal & Curved

e.g. $\mu = \sqrt{\sigma^2}$, so $\eta_1 = \frac{1}{2\sigma^2}$, $\eta_2 = \frac{1}{\sqrt{\sigma^2}}$

$\Downarrow$

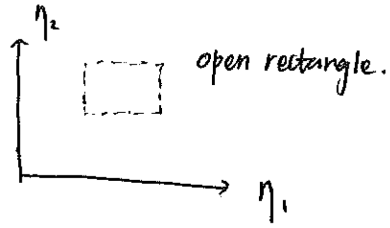
$\eta_2^2 = \eta_1$



III) Minimal & Full rank

e.g. no extra constraint on (μ, σ^2)

where the natural parameter space is $(0, +\infty) \times \mathbb{R}$.



Properties of exponential family

Property 1: If $X_1, \dots, X_n \overset{iid}{\sim} p(x;\theta) = \exp\{\sum_{i=1}^{k} \eta_i^{(\theta)} T_i(x) - B(\theta)\} h(x)$,

then $p(x_1, \dots x_n; \theta) = \exp\{\sum_{i=1}^{k} \eta_i(\theta) \sum_{j=1}^{n} T_i(x_j) - nB(\theta)\} \prod_{j=1}^{n} h(x_j)$.

By NFFC, $(\sum_{j=1}^{n} T_1(x_j), \dots, \sum_{j=1}^{n} T_k(x_j))$ is therefore a sufficient statistic.

Hence, exponential family data is highly compussible. (Pitman- koopman- Darmors).

Property 2: If f is integrable and $\eta \in \Theta$, then $G(f, \eta) = \int f(x) \exp\{\sum_{i=1}^{k} \eta_i T_i(x)\} d\mu(x)$ is infinitely differentiable w.r.t. η and the derivatives can be obtained by differentiating under the integral sign. (see. TSH, 2.7.1).

Property 3: Moments of T_i's.

Take, in particular, $f(x) = 1$, then

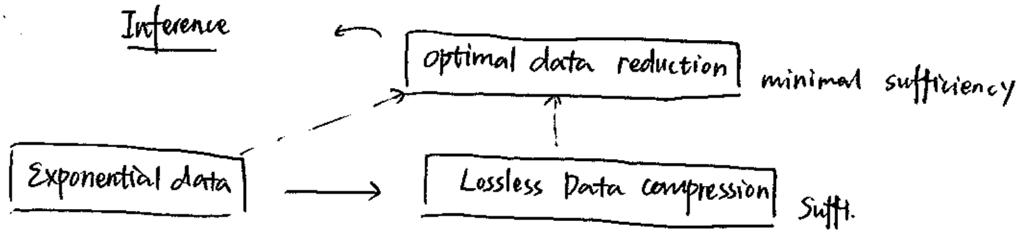
$$G(f, \eta) \triangleq \int \exp\{\sum_{i=1}^{k} \eta_i T_i(x)\} h(x) d\mu(x) = \exp\{A(\eta)\}.$$

$$\frac{\partial G(f,\eta)}{\partial \eta_i} = \int T_i(x) \exp\{\sum_{i=1}^{k} \eta_i T_i(x)\} h(x) d\mu(x) = \frac{\partial A(\eta)}{\partial \eta_i} \exp\{A(\eta)\}.$$

$$\frac{\partial A(\eta)}{\partial \eta_i} = \int T_i(x) \exp\{\sum_{i=1}^{k} \eta_i T_i(x) - A(\eta)\} h(x) d\mu(x) = E_\eta\{T_i(x)\}.$$

Minimal Sufficiency

Def A sufficient statistic T is minimal ~~sufficien~~ if for every statistic T'. T is a function of T'.
Equivalently, T is minimal if for every sufficient statistic T, $T(x)=T(y)$ whenever $T'(x)=T'(y)$.



[Thm] Let $\{p_\theta(x)\}_{\theta\in\Theta}$ be a family of densities w.r.t some measure μ (usually Lebesgue). Suppose that there exists a ~~stochastic~~ statistic s.t. for every $x, y \in \mathcal{X}$.

$$p_\theta(x) = C_{x,y}\, p_\theta(y) \iff T(x) = T(y)$$

for every $\theta \in \Theta$ and some $C_{x,y} \in \mathbb{R}$. Then T is a minimal sufficient statistic.

Pf: [T is sufficient].

Start with $T(\mathcal{X}) = \{t : t = T(x) \text{ for some } x \in \mathcal{X}\}$.

For each $t \in T(\mathcal{X})$, consider the preimage $A_t = \{x : T(x) = t\}$.

and select an arbitrary representative x_t from each A_t.

Then for any $y \in \mathcal{X}$, we have $y \in A_{T(y)}$ and $x_{T(y)} \in A_{T(y)}$.

By the definition of A_t, this implies that $T(y) = T(x_{T(y)})$

From the assumption of the thm,

$$\begin{aligned} p_\theta(y) &= C_{y, x_{T(y)}}\, p_\theta(x_{T(y)}) \\ &= h(y)\, g_\theta(T(y)) \end{aligned}$$

which yields sufficiency of T by NFFC.

[T is minimal]

Consider another sufficient statistic T'. By NFFC, $p_\theta(x) = \tilde{g}_\theta(T'(x))\,\tilde{h}(x)$

Take any x, y s.t. $T'(x) = T'(y)$. Then

$$\begin{aligned} p_\theta(x) &= \tilde{g}_\theta(T'(x))\,\tilde{h}(x) \\ &= \tilde{g}_\theta(T'(y))\,\tilde{h}(y) \cdot \frac{\tilde{h}(x)}{\tilde{h}(y)} \\ &= p_\theta(y) \cdot C_{x,y}. \end{aligned}$$

Hence, $T(x)=T(y)$ by the assumption of the thm. So, $T'(x)=T'(y)$ implies that $T(x)=T(y)$ for any sufficient statistic T' and x,y. As a result, T is a minimal sufficient statistic.

Remark For any minimal k-dim exponential family, the statistic $\left(\sum_{j=1}^n T_1(x_j), \dots \sum_{j=1}^n T_k(x_j)\right)$ is a minimal sufficient statistic. (Keener Ex 3.12).

Remark The support of X $\{x \in \mathcal{X} : p_\theta(x) > 0\}$ should be indep. of θ. e.g. $U(0,\theta)$, Binomial (n,θ).

$U(\theta,\theta+1)$ qualifying may appear.

Ancillarity and completeness

E.g.* Consider $X_1, \dots, X_n \overset{iid}{\sim}$ Cauchy (θ) where distribution is given by

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

then $(X_{(1)}, \dots X_{(n)})$ is minimal sufficient (TPE § 1.5)

Def A statistic A is ancillary for $X \sim P_\theta \in \mathcal{P}$ if the distribution of $A(X)$ does not dep. on θ.

e.g. Consider the previous e.g*, then $A(X) = X_{(n)} - X_{(1)}$ is ancillary even though $(X_{(1)}, \dots X_{(n)})$ is minimal sufficient. To see this, note that $X_i = Z_i + \theta$ for $Z_i \overset{iid}{\sim}$ Cauchy (0), $X_{(i)} = Z_{(i)} + \theta$ and $A(X) = A(Z)$, which does not dep. on θ.

Def A statistic A is first order ancillary for $X \sim P_\theta \in \mathcal{P}$ if $E_\theta\{A(X)\}$ does not dep. on θ.

Def A statistic T is complete for $X \sim P_\theta \in \mathcal{P}$ if no non-constant function of T is first order ancillary. In other words, if $E_\theta\{f(T(X))\}=0$ for all $\theta \in \Theta$, then $f(T(X))=0$ with prob. 1 for all $\theta \in \Theta$

Remark: For many important situations, completeness is the needed condition for the minimal sufficient and ancillary statistics to be independent.

Remark Complete, sufficient statistics give "optimal" unbiased estimator.

Lecture 5 UMVUE, Cramér Rao Lower Bound.

Def A statistic T is complete for $\uparrow$ see above.

Remark (i) If T is complete sufficient, then T is minimal sufficient (Bahader's Thm) (TPE)

(ii) Complete sufficient statistics yield optimal unbiased estimators.

Example Let $X_1, \dots, X_n \overset{iid}{\sim}$ Bernoulli (θ), $\theta \in (0,1)$, Then $T(X) = \sum_{i=1}^n X_i$ is sufficient.

Suppose $E_\theta\{f(T(X))\}=0$ for all $\theta \in (0,1)$, this means $\sum_{j=0}^n f(j)\binom{n}{j}\theta^j(1-\theta)^{n-j} = 0$ $\forall \theta \in (0,1)$

Dividing both sides by θ^n, and using $\beta = \frac{\theta}{1-\theta}$, we can write

$$\sum_{j=1}^n f(j)\binom{n}{j}\beta^j = 0 \quad \forall \beta > 0.$$

If f are non-zero, then the polynomial on the left is a ploynomial of degree at most n which can only have n roots ~~at~~ at most. Hence, it is impossible to have LHS equal 0 for writing β unless $f=0$, in which case T is complete.

Example

Let $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ with an unkown $\theta \in \mathbb{R}$ and a known $\sigma^2 > 0$. We'd like to examine if $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ is complete. To simplify our calculation, we consider the case with $n=1$ and $\sigma=1$ so that $T(X) = X \sim N(\theta, 1)$.

Suppose $E_\theta\{f(X)\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-\frac{(x-\theta)^2}{2}} dx = 0 \quad \forall \theta \in \mathbb{R}$.

By multiplying $\sqrt{2\pi} e^{-\frac{\theta^2}{2}}$ on both sides, we obtain

$$\int_{-\infty}^{\infty} f(x) e^{-\frac{x^2}{2}} e^{\theta x} dx = 0 \qquad \forall \theta \in \mathbb{R}. \qquad (*)$$

Now, we decompose f into its positive part and negative part as $f(x) = f_+(x) - f_-(x)$, where $f_+(x) = \max\{f(x), 0\}$, $f_-(x) = \max\{-f(x), 0\}$. Then $f_+(x) \geq 0$ and $f_-(x) \geq 0$ for all $x \in \mathbb{R}$. and $f_+(x) = f_-(x)$ iff $f_+(x) = f_-(x) = 0$.

If $f(x) \geq 0$ a.s. or $f(x) \leq 0$ a.s., then $(*)$ implies that $f(x)=0$ a.s. because setting $\theta=0$ gives us and integral of a non-negative (or non-positive) function of being 0. This is completeness.

In other words, if f_+ and f_- have non-zero components, and we may write

$$\frac{\int_{-\infty}^{\infty} f_+(x) e^{-\frac{x^2}{2}} e^{\theta x} dx}{\int_{-\infty}^{\infty} f_+(x) e^{-\frac{x^2}{2}} dx} = \frac{\int_{-\infty}^{\infty} f_-(x) e^{-\frac{x^2}{2}} e^{\theta x} dx}{\int_{-\infty}^{\infty} f_-(x) e^{-\frac{x^2}{2}} dx} \qquad (**)$$

with the equality of the numerators follow from $(*)$ and equality of the denominator follows from $(*)$ with setting $\theta=0$. The quantity $\frac{f_+(x) e^{-\frac{x^2}{2}}}{\int_{-\infty}^{\infty} f_+(x) e^{-\frac{x^2}{2}} dx}$ defines a prob. density and the LHS of $(**)$ is the MGF of this density.

Likewise, the RHS of $(**)$ is the MGF of $\frac{f_-(x) e^{-\frac{x^2}{2}}}{\int_{-\infty}^{\infty} f_-(x) e^{-\frac{x^2}{2}} dx}$

$\Rightarrow f_+(x) = f_-(x)$ a.s. $\Rightarrow f_+(x) = f_-(x) = 0$ a.s.

$\Rightarrow f(x) = 0 \Rightarrow T$ is complete

[Thm] (Basu's Thm)

If T is complete and sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and A is ancillary, then $T(X) \perp\!\!\!\perp A(X)$.

Pf: N.T.S. $P_\theta(A \in \mathcal{A} \mid T) = P_\theta(A \in \mathcal{A})$ a.s. P_θ.

Observations:

$$\left.\begin{array}{l} \cdot \text{ LHS is free of } \theta \text{ by sufficiency of } T \\ \cdot \text{ RHS is free of } \theta \text{ since } A \text{ is ancillary.} \end{array}\right\} \text{free of } \theta.$$

$E_\theta(\text{LHS}) = E_\theta(\text{RHS})$ by tower property.

$\Rightarrow$ LHS = RHS with prob. 1 by completeness of T. //

<u>Example</u> $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ (μ, σ^2 both unkown)

Cliam: $\bar{X}_n \perp\!\!\!\perp \underbrace{(n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$ sample variance.

Pf: Fix any $\sigma > 0$ and consider a submodel $P_\sigma = \{ N(\mu, \sigma^2),\ \mu \in \mathbb{R} \}$

In each submodel, $\bar{X}_n$ is complete and sufficient, and $n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is ancillary.

By Basu's thm, $X_n \perp\!\!\!\perp n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ under $N(\mu, \sigma^2)$ for any μ.

Since σ ~~is an~~ arbitrary, we have $\bar{X}_n \perp\!\!\!\perp n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ for the full model. //

<u>From Data compression to Risk Reduction / "optimal" estimation</u>

[Thm] (Rao-Blackwell Thm) Keener 3.28.

Suppose that T is sufficient for $\mathcal{P} = \{ P_\theta : \theta \in \Theta \}$, that $\delta(X)$ is an estimator for $g(\theta)$ for which $E\{\delta(X)\}$ exists, and that $R(\theta, \delta) = E_\theta\{ L(\theta, \delta(X)) \} < \infty$.

If $L(\theta, \cdot)$ is convex, then

$$R(\theta, \eta) \leq R(\theta, \delta)$$

Jensen's ineq: $E\{\phi(X)\} \geq \phi\{E(X)\}$ $\phi(\cdot)$ convex

where $\eta(T(X)) = \underline{E\{ \delta(X) \mid T(X) \}}$.

If $L(\theta, \cdot)$ is strictly convex, then $R(\theta, \eta) < R(\theta, \delta)$ for any θ unless $\eta(T'(X)) = \delta$ with prob. 1.

Pf: By Jensen's inequality,

$$E_\theta\{ L(g(\theta), \delta(X)) \mid T \} \geq L[g(\theta, E_\theta\{\delta(X) \mid T\})]$$
$$= L(g(\theta), \eta(T))$$

Taking another expectation, we have

$$E_\theta\{ L(g(\theta), \delta(X)) \} \geq E_\theta\{ L(g(\theta), \eta(T)) \}$$

$$\Rightarrow R(g(\theta), \delta) \geq R(g(\theta), \eta)$$ //.

<u>Example</u> Let $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim}$ Bernoulli(θ), $\theta \in (0,1)$

Consider the loss function $L(\theta, d) = (\theta - d)^2$ [squared loss function].

Suppose we consider first an unreasonable estimator $\delta(X) = X_1$.

We have shown that $T(X) = \bar{X}_n$ is sufficient, so we may apply

Rao-Blackwell thm to improve δ. In particular,

$$\begin{aligned}\eta(T(X)) &= E\{\delta(X) \mid T(X)\} \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_1 \mid \bar{X}_n) \\ &\overset{iid}{=} \frac{1}{n}\sum_{i=1}^{n} E(X_i \mid \bar{X}_n) \\ &= E(\bar{X}_n \mid \bar{X}_n) \\ &= \bar{X}_n .\end{aligned}$$

Observe that $R(\theta,\eta) = \dfrac{\theta(1-\theta)}{n} < \theta(1-\theta) = R(\theta,\delta)$.

R.B. gives a strict improvement. //.

Remark : Rao-Blackwell thm, however, does not necessarily lead to a uniformly optimal estimator.

For example, consider $\delta_{naive}(X) = \frac{1}{2}$, then

$\eta(T(X)) = E\{\delta_{naive}(X) \mid \bar{X}\} = \frac{1}{2}$ as well.

Since $R(\theta,\eta) = (\frac{1}{2}-\theta)^2$, neither R.Bised outcome is uniformly better across all θ.

Unbiased Estimation

An estimator is unbiased if $E_\theta\{\delta(X)\} = g(\theta)$. We attempt to find an unbiased estimator with uniformly minimum risk. i.e. unbiased δ satisfying $R(\theta,\delta) \le R(\theta,\delta')$. for all $\theta \in \Theta$. and an unbiased estimator δ'. Such an estimator is called uniformly minimum risk unbiased estimator (UMRUE). If, in particular, $L(\theta,\delta) = (\theta-\delta)^2$ is the choosen loss function, then an UMRUE becomes UMVUE.

$$E_\theta\{g(\theta)-\delta(X)\}^2 = \underbrace{[E_\theta\{\delta(X)-g(\theta)\}]^2}_{\text{Bias}^2} + \underbrace{E_\theta[\delta(X)-E_\theta\{\delta(X)\}]^2}_{\text{Variance}}$$

$\longrightarrow 0$ for $\delta(X)$ is an unbiased est. for $g(\theta)$.

[Thm] (Lehmann-Scheffe Thm)

If T is a complete and sufficient statistics, and $E_\theta\{h(T(X))\} = g(\theta)$. i.e. $h(T(X))$ is unbiased for $g(\theta)$, then $h(T(X))$ is

(1) the only function of $T(X)$ that is unbiased for $g(\theta)$

(2) an UMRUE under any convex loss function

(3) the unique UMRUE (up to a P-null set) under any strictly convex loss function

(4) the unique UMVUE (up to a P-null set)

Pf: (1) Suppose $E_\theta\{\tilde{h}(T(X))\} = g(\theta)$, then $E_\theta\{\tilde{h}(T(X)) - h(T(X))\} = 0 \quad \forall \theta \in \Theta$.

thus $\tilde{h}(T(X)) = h(T(X))$ a.s. for each θ due to completeness.

(2) Consider an unbiased estimator $\delta(X)$, and let $\tilde{h}(T(X)) = E_\theta\{\delta(X) | T(X)\}$.

Then $E_\theta\{\tilde{h}(T(X))\} = E_\theta\{\delta(X)\} = g(\theta)$ by tower property of conditional expectation.

By (1). $\tilde{h}(T(X)) = h(T(X))$ and by the Rao-Blackwell thm,

$R(g(\theta), h(T(\cdot))) \leq R(g(\theta), \tilde{h}(T(\cdot)))$ for all θ if the loss function is convex.

Therefore, $h(T(X))$ is an UMRUE under any convex loss function.

(3) If the loss function is strictly convex, $R(g(\theta), h(T(\cdot))) < R(g(\theta), \delta)$ unless $\delta(X) = h(T(X))$ a.s. Thus, $h(T(X))$ is the unique UMRUE.

(4) Done by (3). //.

Strategies for obtaining UMVUEs (R)

A - Rao-Blackwellisation / Conditioning $E(\ |T)$

B - Solve for δ satisfying $E_\theta\{\delta(T(x))\} = g(\theta)$, $\forall \theta \in \Theta$.

Example

Suppose $X_1, \dots, X_n \overset{iid}{\sim}$ Bernoulli (θ)

$T(X) = \sum_{i=1}^{n} X_n$ is complete and suff. and

$E\{T(X)/n\} = \theta$. Therefore $\bar{X}_n$ is an UMRUE for θ under any convex function.

Suppose now we are interested in estimating $g(\theta) = \theta^2$.

If we choose $\delta(X) = I(X_1 = X_2 = 1) = X_1 X_2$, then $E_\theta\{\delta(X)\} = \theta^2$ is unbiased.

Apply strategy A to obtain:

$$
\begin{aligned}
E\{\delta(X) \mid T(X) = t\} &= P(X_1 = X_2 = 1 \mid T(X) = t) \\
&= \frac{P(X_1 = X_2 = 1, \sum_{i=3}^{n} X_i = t-2)}{P(T(X) = t)} \\
&= \frac{\theta^2 \binom{n-2}{t-2} \theta^{t-2} (1-\theta)^{n-t} I(t \geq 2)}{\binom{n}{2} \theta^t (1-\theta)^{n-t}} \\
&= \frac{t(t-1) I(t \geq 2)}{n(n-1)} = \frac{t(t-1)}{n(n-1)}
\end{aligned}
$$

Hence, $\dfrac{T(X)\{T(X)-1\}}{n(n-1)}$ is the UMVUE.

Example Suppose $X_1, \dots, X_n \overset{iid}{\sim}$ uniform $(0, \theta)$. In this case, $T(X) = X_{(n)} = \max_{1 \leq i \leq n} X_i$ is a complete and sufficient statistic and $\delta(X) = 2X_1$ is an unbiased estimator of θ.

Given the knowledge of $X_{(n)}$, X_1 is equal to $X_{(n)}$ with prob. $1/n$ and distributed according to uniform $(0, X_{(n)})$ with prob. $1 - 1/n$.

So $P(X_1 = x_1 \mid T(X)) = \frac{1}{n} I(T(X) = x_1) + \frac{(1-\frac{1}{n}) I(0 < x_1 < T(X))}{T(X)}$

Hence, our UMVUE,

$$\begin{aligned} E\{\delta(X) \mid T(X)\} &= 2E\{X_1 \mid T(X)\} \\ &= 2\left\{\frac{1}{n} T(X) + (1-\frac{1}{n}) \int_0^{T(X)} \frac{x_1 \, dx_1}{T(X)}\right\} \\ &= 2\left\{\frac{T(X)}{n} + (1-\frac{1}{n}) \frac{T(X)}{2}\right\} \\ &= \left(\frac{n+1}{n}\right) T(X) \quad // . \end{aligned}$$

Example

Let $X_1, X_2, \dots, X_n \overset{iid}{\sim}$ Poisson (θ). Since this is a one-dimensional full-rank exponential family, X is a complete sufficient statistic. X is furthermore unbiased and therefore UMV for θ. Suppose that our goal is to estimate $g(\theta) = e^{-a\theta}$ for some given $a \in \mathbb{R}$. We need to find an estimator δ such that $E\{\delta(X)\} = g(\theta)$ for all θ. Under our model, we may reexpress this system of equations as:

$$\sum_{x=0}^{\infty} \delta(x) \frac{e^{-\theta}\theta^x}{x!} = e^{-a\theta} \quad \text{for all } \theta$$

$$\Rightarrow \sum_{x=0}^{\infty} \frac{\delta(x)\theta^x}{x!} = e^{(1-a)\theta} = \sum_{x=0}^{\infty} \frac{(1-a)^x \theta^x}{x!}$$

$\Rightarrow \delta(x) = (1-a)^x$ is the UMVUE of $g(\theta)$.

Remark:

This estimator is not satisfying. If $a=2$, for example, it will change its sign according to X even though we realize that our estimand $e^{-\theta a}$ is non-negative. The estimator is in fact inadmissible when $a > 1$ and dominated by $\max[\delta(x), 0]$.

Ch5 of TPE.

Suppose we have δ_i UMVU for $g(\theta)$ for $i \in \{1,2\}$. Is $\delta_1 + \delta_2$ then UMVU for $\underline{g_1(\theta) + g_2(\theta)}$?

[Thm] (Characterization of UMVUEs, see TPE 2.1.7).

Let $\Delta = \{\delta : E_\theta(\delta^2) < \infty\}$. Then $\delta_0 \in \Delta$ is UMVUE for $g(\theta) = E(\delta_0)$ iff $E\{\delta_0(\theta) u\} = 0$ for every $u \in \mathcal{U}$, where $\mathcal{U} = \{\text{unbiased estimator of } 0\}$

$$= \{u : \mathcal{X} \to \mathbb{R} \text{ s.t. } E_\theta(u(X)) = 0,\ E_\theta(u(X)^2) = \infty\}.$$

Pf: If δ_0 is UMVUE, let us consider $\delta_\lambda = \delta_0 + \lambda u$ for $\lambda \in \mathbb{R}$, $u \in \mathcal{U}$.

Since δ_0 has the minimal variance,

$$Var(\delta_\lambda) = Var(\delta_0) + \lambda^2 Var(u) + 2\lambda \, cov(\delta_0, u) \geq Var(\delta_0) \quad (\text{UMVU}). \quad (\#)$$

Consider the quadratic form $q(\lambda) = \lambda^2 Var(u) + 2\lambda \, cov(\delta_0, u)$, then q has the roots 0 and $-2\,cov(\delta_0, u)/Var(u)$.

If the roots are distinct, then the form must be negative at some point, which would violet the inequality (#). Hence $-2\,cov(\delta_0, u)/Var(u) = 0$, and thus $E(u\delta_0) = cov(\delta_0, u) = 0$.

For the converse result, we assume $E(\delta_0 u) = 0 \ \forall u \in \mathcal{U}$. and consider any unbiased estimator δ for $g(\theta)$. Then $\delta - \delta_0 \in \mathcal{U}$ so $E\{\delta_0(\delta-\delta_0)\} = 0$.

This implies that $E(\delta_0\delta) = E(\delta_0^2)$ and substracting $E(\delta_0)E(\delta)$ on both sides, we have

$$Var(\delta_0) = cov(\delta_0, \delta) \overset{cs}{\leq} \sqrt{Var(\delta_0)Var(\delta)}$$

Hence $Var(\delta_0) \leq Var(\delta)$ for any arbitrary estimator δ and δ_0 is UMVUE.

$\forall u \in \mathcal{U}, \ E((\delta_1+\delta_2)u) = E(\delta_1 u) + E(\delta_2 u) = 0$

$\Rightarrow \delta_1 + \delta_2$ is UMVU for $g_1(\theta) + g_2(\theta)$

Lecture 6

- Rao-Blackwell, LS, Basu, ...

Cramer-Rao Lower Bound / TPE §2.5 & 2.6. Keener Ch.3? C&B Ch.7.

Assume the following:

(a) $\Theta \subseteq \mathbb{R}$ is an open interval

(b) $\{P_\theta : \theta \in \Theta\}$ have common support A

(c) $P_\theta'(X) = \frac{\partial P_\theta(X)}{\partial X}$ exists and is finite for all $x \in A$.

Define $I(\theta) = E_\theta\{\frac{\partial}{\partial\theta}\log P_\theta(x)\}^2 = \int_A \{\frac{P_\theta'(x)}{P_\theta(x)}\}^2 P_\theta(x)\,d\mu = \int_A \frac{\{P_\theta'(x)\}^2}{P_\theta(x)}\,d\mu$ to be the information function.

Lemma (i) Assume (a)-(c) hold, and $\frac{\partial}{\partial\theta}\int_A P_\theta(x)\,d\mu = \int_A \frac{\partial}{\partial\theta} P_\theta(x)\,d\mu$.

then $I(\theta) = Var_\theta\{\frac{\partial}{\partial\theta}\log P_\theta(x)\}$

(ii) In addition, $P_\theta''(x)$ exists $\forall \theta \in \Theta$, $x \in A$ and

(e) $\frac{\partial^2}{\partial\theta^2}\int_A P_\theta(x)\,d\mu = \int_A \frac{\partial^2}{\partial\theta^2} P_\theta(x)\,d\mu$, then $I(\theta) = -E_\theta\{\frac{\partial^2}{\partial\theta^2}\log P_\theta(x)\}$.

- Pf: (i) We need to show $I(\theta) = Var(\frac{\partial}{\partial\theta}\log P_\theta(X))$. Assume that $I_\theta = E_\theta(\frac{\partial}{\partial\theta}\log P_\theta(X))^2 < \infty$.

It suffices to show that $E_\theta(\frac{\partial}{\partial\theta}\log P_\theta(x)) = 0$. This is true because

$$E_\theta(\frac{\partial}{\partial\theta}\log P_\theta(x)) = \int_A \frac{P_\theta'(X)}{P_\theta(X)} P_\theta(X)\,d\mu = \int_A \frac{\partial}{\partial\theta} P_\theta(X)\,d\mu = \frac{\partial}{\partial\theta}\underbrace{\int_A P_\theta(x)\,d\mu}_{\to 1} = 0$$

(ii) Note that $\frac{\partial^2}{\partial\theta^2}\log P_\theta(x) = \frac{\partial}{\partial\theta}\left(\frac{P_\theta'(x)}{P_\theta(x)}\right) = \frac{P_\theta''(x)}{P_\theta(x)} - \left(\frac{P_\theta'(x)}{P_\theta(x)}\right)^2$

thus $E\{\frac{\partial^2}{\partial\theta^2}\log P_\theta(X)\} = \int_A P_\theta''(X)\,d\mu - E_\theta\{\frac{\partial}{\partial\theta}\log P_\theta(X)\}^2$

$$= \int_A \frac{\partial^2}{\partial\theta^2} P_\theta(x)\,d\mu - I(\theta) \overset{(a)}{=} \frac{\partial^2}{\partial\theta^2}\int_A P_\theta(x)\,d\mu - I(\theta) = -I(\theta)$$

$$\Rightarrow I(\theta) = -E_\theta\left\{\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)\right\}.$$

<u>Remark</u> Information depends on parametrization. For example, if $\eta = \tau(\theta)$, where $\tau \in C^2$, & $\tau'(\theta) \neq 0$, then $I(\tau(\theta)) = \dfrac{I(\theta)}{\{\tau'(\theta)\}^2}$ because

$$E_\tau\left\{\frac{\partial}{\partial\tau}\log p_{\theta(\tau)}(X)\right\}^2 = E_\theta\left\{\frac{\partial}{\partial\theta}\log p_\theta(X)\cdot\frac{\partial\theta}{\partial\tau}\right\}^2 = \frac{E_\theta\left\{\frac{\partial}{\partial\theta}\log p_\theta(X)\right\}^2}{[\tau'(\theta)]^2} = \frac{I(\theta)}{\{\tau'(\theta)\}^2}$$

<u>Example</u> Suppose $X \sim N(\theta, 1)$, $\theta > 0$

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\theta)^2}{2}} \Rightarrow \log p_\theta(x) = -\log\sqrt{2\pi} - \frac{(x-\theta)^2}{2} \quad \vee$$

$$\Rightarrow \frac{\partial}{\partial\theta}\log p_\theta(x) = x-\theta, \quad I(\theta) = E(X-\theta)^2 = Var(X) = 1$$

$$\left(\text{or } \frac{\partial^2}{\partial\theta^2}\log p_\theta(x) = -1 \Rightarrow I(\theta) = -(-1) = 1\right)$$

Let $\eta = \theta^2$, $X \sim N(\sqrt{\eta}, 1)$. $\log p_{\theta(\eta)}(x) = -\log(\sqrt{2\pi}) - \dfrac{(x-\sqrt{\eta})^2}{2}$

$$\Rightarrow \frac{\partial}{\partial\eta}\log p_{\theta(\eta)}(x) = \frac{x-\sqrt{\eta}}{2\sqrt{\eta}} \Rightarrow I(\eta) = \frac{E(X-\sqrt{\eta})^2}{4\eta} = \frac{1}{4\eta} = \frac{1}{4\theta^2} = \frac{1}{(\tau'(\theta))^2}$$

Multi-parameter Cramér-Rao Lower Bound

Suppose the following conditions hold:

(a) $\Theta \in \mathbb{R}^k$ is an open set

(b) $\{p_\theta(x) : \theta \in \Theta\}$ have common support I

(c) $\dfrac{\partial p_\theta(x)}{\partial \theta_i}$ exist $\forall i = 1, \dots, k$, $x \in I$ and is finite

(d) $\dfrac{\partial}{\partial \theta_i} \int_X p_\theta(x)\, d\mu = \int_X \dfrac{\partial}{\partial \theta_i} p_\theta(x)\, d\mu \quad \forall i = 1, \dots, k$

(e) $\dfrac{\partial}{\partial \theta_i} \int_X \delta(x) p_\theta(x)\, d\mu = \int_X \dfrac{\partial}{\partial \theta_i} \delta(x) p_\theta(x)\, d\mu \quad \forall i = 1, \dots, k.$

Define the $k \times k$ information matrix $I(\underset{\sim}{\theta})$ by $I_{ij}(\underset{\sim}{\theta}) = E_\theta\left\{ \left(\frac{\partial}{\partial \theta_i} \log p_\theta(x)\right)\left(\frac{\partial}{\partial \theta_j} \log p_\theta(x)\right)\right\}$

In particular, if $k=1$, $I(\theta) = E_\theta\left(\frac{\partial}{\partial \theta} \log p_\theta(x)\right)^2$.

Assume $I(\underset{\sim}{\theta})$ is finite and positive definite,

then $\operatorname{Var}_\theta(\delta(X)) \geq \alpha^T I(\underset{\sim}{\theta})^{-1} \alpha$, where $\alpha = (\alpha_1, \dots, \alpha_k)^T$

$$= \left(\frac{\partial}{\partial \theta_1} E_\theta(\delta(X)), \dots, \frac{\partial}{\partial \theta_k} E_\theta(\delta(X))\right)^T.$$

In particular, if $\delta(X)$ is unbiased for $g(\underset{\sim}{\theta})$,

then $\operatorname{Var}_\theta(\delta(X)) \geq \alpha^T I(\theta)^{-1} \alpha$, $\quad \alpha_i = \frac{\partial}{\partial \theta_i}\{g(\theta)\}$. $i = 1, 2, \dots, k$.

Pf: Let $\Psi_i(x) = \frac{\partial}{\partial \theta_i} \log p_\theta(x)$. then $E_\theta(\Psi_i(x)) = \int_X \left\{\frac{\partial}{\partial \theta_i} \log p_\theta(x)\right\} p_\theta(x_i)\, d\mu$

$$= \int_X \frac{\frac{\partial}{\partial \theta_i} p_\theta(x)}{p_\theta(x)} p_\theta(x)\, d\mu = \int_X \frac{\partial}{\partial \theta_i} p_\theta(x)\, d\mu = \frac{\partial}{\partial \theta_i} \int p_\theta(x)\, d\mu = 0.$$

Fix a non-zero vector $(a_1, \dots, a_k)$. Then $E_\theta\left\{\sum_{i=1}^k a_i \Psi_i(x)\right\} = 0$

Claim: $\operatorname{Var}\left(\sum_{i=1}^k a_i \Psi_i(x)\right) = a^T I(\underset{\sim}{\theta}) a$.

Observe that $\operatorname{Var}\left(\sum_{i=1}^k a_i \Psi_i(x)\right) = \sum_{i,j} a_i a_j \operatorname{cov}(\Psi_i(X), \Psi_j(X))$

$$= \sum_{i,j} a_i a_j E(\Psi_i(x) \Psi_j(x))$$

$$= \sum_{i,j} a_i a_j I_{ij}(\underset{\sim}{\theta}) = a^T I(\underset{\sim}{\theta}) a.$$

Finally, $\operatorname{cov}\left(\delta(x), \sum_{i=1}^k a_i \Psi_i(x)\right) = \sum_{i=1}^k a_i \operatorname{cov}(\delta(X), \Psi_i(X))$

$$= \sum_{i=1}^k a_i E(\delta(X)\Psi_i(X)) = \sum_{i=1}^k a_i \int_X \delta(x) \frac{\partial}{\partial \theta_i} \log p_\theta(x) \cdot p_\theta(x)\, d\mu.$$

$$= \sum_{i=1}^k a_i \int_X \delta(x) \frac{\partial}{\partial \theta} p_\theta(x)\, d\mu = \sum_{i=1}^k a_i \frac{\partial}{\partial \theta_i} \int_X \delta(x) p_\theta(x)\, d\mu = \sum_{i=1}^k a_i \alpha_i(\underset{\sim}{\theta})$$

By Cauchy - Schwarz inequality,

$$Var(\delta(X))\, Var\left(\sum_{i=1}^{n} a_i \psi_i(X)\right) \geq cov\left(\sum_{i=1}^{k} a_i \psi_i(X), \delta(X)\right)$$

$$\Rightarrow Var(\delta(X)) \geq \sup_{a \neq 0} \frac{\left(\sum_{i=1}^{k} a_i \alpha_i(\underset{\sim}{\theta})\right)^2}{a^T I(\underset{\sim}{\theta}) a} = \alpha^T I(\underset{\sim}{\theta})^{-1} \alpha.$$

E.g. $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\mu > 0$, $\sigma^2 > 0$.

Problem 1: We want to ~~simulate~~ estimate $g_1(\mu, \sigma^2) = \mu$

Problem 2: . . . . $g_2(\mu, \sigma^2) = \sigma^2$

Consider unbiased estimator only.

Claim 1. $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is UMVUE for σ^2.

why? ① $\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$ is complete sufficient. ② $\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right\}$.

③ $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1} \Rightarrow E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right\} = \frac{\sigma^2}{n-1}\cdot(n-1) = \sigma^2$ (unbiased).

Claim 2: $\bar{X}$ is UMVUE for μ. $E(\bar{X}-\mu)^2 = \frac{\sigma^2}{n}$

Note that $E\left\{\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n-1} \cdot \sigma^2\right\}^2 = Var\left(\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n-1}\right) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{(n-1)}$

$$\log p_{\mu,\sigma^2}(x) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^2} + C_n.$$

$\frac{\partial}{\partial\mu}\log p_{\mu,\sigma^2}(x) = \sum_{i=1}^{n}\frac{(X_i-\mu)}{\sigma^2}$ $\boxed{\frac{\partial^2}{\partial\mu^2}(\log p_{\mu,\sigma}(x)) = -\frac{\mu}{\sigma^2}}$ I_{11}

$$\frac{\partial}{\partial\sigma^2}\log p_{\mu,\sigma^2}(x) = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^4}$$

$$I = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

$$\frac{\partial^2}{\partial\mu\partial\sigma^2}\log p_{\mu,\sigma^2}(x) = -\sum_{i=1}^{n}\frac{(X_i-\mu)}{\sigma^4}$$

$$\frac{\partial^2}{\partial(\sigma^2)^2}\log p_{\mu,\sigma^2}(x) = \frac{n}{2\sigma^4} - \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^6}$$

$$I_{22} = -\frac{n}{2\sigma^4} + E\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^6} = -\frac{n}{2\sigma^4} + \frac{n\sigma^2}{\sigma^6} = \boxed{\frac{n}{2\sigma^4}}$$

$\Rightarrow$ CRLB for $\mu = [1\ 0]\, I^{-1} \begin{bmatrix}1\\0\end{bmatrix} = \frac{1}{I_{11}} = \frac{\sigma^2}{n}$

$\Rightarrow$ CRLB for $\sigma^2 = [0\ 1]\, I^{-1} \begin{bmatrix}0\\1\end{bmatrix} = \frac{1}{I_{22}} = \frac{2\sigma^4}{n}$

CRLB attained by $\bar{X}$ but not S^2.

Cramer-Rao lower bound (Information inequality)

Suppose (a) - (d) hold, let $\delta(X)$ be an estimator s.t. $\mathbb{E}_\theta\{\delta(X)\}^2 < \infty$. $I(\theta) \in (0, \infty)$

Assume further that $\int_A \delta(x) \frac{\partial}{\partial\theta} P_\theta(x)\, d\mu = \frac{\partial}{\partial\theta} E_\theta\{\delta(X)\}$. Then

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

Remark: If we want to estimate $g(\theta)$ using an unbiased estimator δ,

then $Var_\theta(\delta(X)) \geq \frac{\{g'(\theta)\}^2}{I(\theta)}$.

Pf: Let $V = \frac{\partial}{\partial\theta} \log P_\theta(x)$, so $E_\theta(V^2) = I(\theta) = Var_\theta(V)$. Also $E_\theta(V) = 0$.

By Cauchy Schwarz inequality, $Var_\theta(V)\, Var_\theta(\delta(X)) \geq \{cov(V, \delta(X))\}^2$

$$\Rightarrow \quad Var(\delta(X)) \geq \frac{\{cov(V, \delta(X))\}^2}{I(\theta)}.$$

It suffices to show that $cov(V, \delta(X)) = \frac{\partial}{\partial\theta} E_\theta(\delta(X))$

Observe that $cov(V, \delta(X)) = E_\theta(V(\delta(X))) = \int_A \frac{\partial}{\partial\theta} \log P_\theta(x)\, \delta(x)\, P_\theta(x)\, d\mu$

$$= \int_A \frac{\partial P_\theta(x)}{\partial\theta} \delta(x)\, d\mu = \frac{\partial}{\partial\theta} \int_A P_\theta(x)\,\delta(x)\, d\mu$$

$$= \frac{\partial}{\partial\theta} E_\theta(\delta(X)), \text{ where } A = \{x \in \mathcal{X} : P_\theta(x) > 0\}. \quad //.$$

Example Let $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$. We are interested in estimating $g(\theta) = \theta$.

Take $\delta(X) = \bar{X}_n$. Then $Var(\bar{X}_n) = \frac{1}{n}$.

For any unbiased estimator δ, we have

$$Var(\delta(X)) \geq \frac{1}{I_n(\theta)} = \frac{1}{n},$$

where $I_n(\theta) = E\left\{ \frac{\partial}{\partial\theta} \log P_\theta(x_1, \dots, x_n)\right\}^2 = n$,

~~then~~ Hence $\bar{X}_n$ is UMVUE.

$$I_n(\theta) = E\left\{\frac{\partial}{\partial\theta} \log P_\theta(X_1) + \cdots + \frac{\partial}{\partial\theta} \log P_\theta(X_n)\right\}^2$$

$$= \sum_{i=1}^n E\left\{\frac{\partial}{\partial\theta} \log P_\theta(X_i)\right\} + \sum_{i \neq j} \underbrace{E\left\{\frac{\partial}{\partial\theta} \log P_\theta(X_i) \frac{\partial}{\partial\theta} \log P_\theta(X_j)\right\}}_{0}$$

$$= \sum_{i=1}^n 1$$

$$= n.$$

Example $X_1, \dots, X_n \overset{iid}{\sim}$ Poisson (λ), $\lambda > 0$, We want to verify that $\bar{X}_n$ is UMVUE for λ.

Observe that $Var(\bar{X}_n) = \frac{Var(X_1)}{n} = \frac{\lambda}{n}$. $E(\bar{X}_n) = \lambda$.

Also, $Var_\lambda(\delta(X)) \geq \frac{1}{n I_1(\lambda)}$.

It suffices to show that $I_1(\lambda) = \frac{1}{\lambda}$.

Recall that $p_\lambda(x) = \frac{e^{-\lambda}\lambda^x}{x!}$

$$\log p_\lambda(x) = -\lambda + x\log\lambda - \log(x!)$$

$$\frac{\partial}{\partial\lambda}\log p_\lambda(x) = -1 + \frac{x}{\lambda} \Rightarrow I_1(\lambda) = E\left(\frac{x}{\lambda} - 1\right)^2 = \frac{Var(X)}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda} \quad //.$$

Remark: Assume (a)-(d) hold, then $Var_\theta(\delta(X)) \geq \frac{\left\{\frac{\partial}{\partial\theta}E_\theta(\delta(X))\right\}^2}{I(\theta)}$

If the equality holds, then $\frac{\partial}{\partial\theta}\log p_\theta(x) = a(\theta)\delta(x) + b(\theta)$ a.s. w.r.t. P_θ (measure)

Assume that $\frac{\partial}{\partial\theta}\log p_\theta(x)$ is continuous in θ, then $\theta \mapsto a(\theta)$ and $\theta \mapsto b(\theta)$ are also continuous, provided δ is not a degenerate r.v. It follows that for fixed $\theta_0 \in \Theta$,

$$p_\theta(x) = p_{\theta_0}(x)\, e^{\left\{\int_{\theta_0}^{\theta} a(t)\,dt\right\}\delta(x) + \left\{\int_{\theta_0}^{\theta} b(t)\,dt\right\}}$$

Thus, p_θ is a 1-parameter exponential family and $\delta(X)$ is the natural sufficient statistic.

Let $A = \left\{x : \frac{\partial}{\partial\theta}\log p_\theta(x) = a(\theta)\delta(x) + b(\theta)\right\}$. then $\exists\, x_1 \neq x_2$ s.t. $\delta(x_1) \neq \delta(x_2)$

For the x_1, x_2 equality holds in $\frac{\partial}{\partial\theta}\log p_\theta(x) = a(\theta)\delta(x) + b(\theta) = h(x,\theta)$

$$\Rightarrow h(x_1,\theta) = a(\theta)\delta(x_1) + b(\theta)$$
$$h(x_2,\theta) = a(\theta)\delta(x_2) + b(\theta)$$

$$\Rightarrow \frac{h(x_1,\theta) - h(x_2,\theta)}{\delta(x_1) - \delta(x_2)} = a(\theta)$$

$\Rightarrow$ a is cont. $\Rightarrow$ b is cont.

[Thm] Let $p_\theta(x) = e^{\eta(\theta)T(x) - B(\theta)}\tilde{h}(x)$, $\theta \in \Theta$ open interval. Let $\tau(\theta) = E_\theta(T)$

Assume T is not a constant r.v., then

(a) $\tau'(\theta) \neq 0$ and $I(\tau(\theta)) = \frac{1}{Var_\theta(T)}$

(b) $I(h(\theta)) = \left(\frac{\eta'(\theta)}{h'(\theta)}\right)^2 Var_\theta(T)$.

· Average Risk Optimality (Bayes Estimator) TPE Ch.4.

Suppose $\{P_\theta : \theta \in \Theta\}$ is a collection of prob. measures on X dominated by σ-finite measure μ. Assume that now θ is a random variable on Θ with dist. π, which is regarded as the prior dist.

Suppose we want to estimate $g(\theta)$, where $g: \Theta \to \mathbb{R}$. For an estimator $\delta(X)$, let the loss incurred be $L(g(\theta), \delta(X))$. Then the risk function, as defined before, is

$$R(g(\theta), \delta) = E_{X\sim P_\theta}\{L(g(\theta), \delta(X))\} = E\{L(g(\theta), \delta(X)) \mid \theta\}.$$

Define the Bayes risk of δ by $r(\pi, \delta) = E_{\theta\sim\pi}\{R(g(\theta), \delta(X))\}$. An estimator δ_0 is said to be a Bayes estimator if it minimizes the Bayes risk, i.e. for any other estimator, we have $r(\pi, \delta_0) \leq r(\pi, \delta)$.

The conditional distribution of $(\theta|X)$ is called the posterior distribution.

Define the marginal distribution of X as M (which has the density m w.r.t. μ)

$$m(x) = \int_\Theta P_\theta(x)\,\pi(d\theta).$$

Example $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \underset{\text{known}}{\sigma^2})$, $\Theta = \mathbb{R}$. Assume that $\theta \sim N(\underset{\text{hyper parameter}}{\mu}, \tau^2)$ (prior dist.).

$$P_\theta(\underset{\sim}{X}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\sum_{i=1}^n \frac{(X_i-\theta)^2}{2\sigma^2}}, \quad \pi(\theta) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}$$

$$\text{Joint density} = P_\theta(x)\pi(\theta) \propto e^{-\sum_{i=1}^n \frac{(X_i-\theta)^2}{2\sigma^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}$$

$$\text{posterior density} \propto e^{-\sum_{i=1}^n (X_i-\theta)^2/2\sigma^2} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}$$

$$\propto \cdots = e^{-\frac{\theta^2}{2}\left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right] + \theta\left[\frac{\sum X_i}{\sigma^2} + \frac{\mu}{\tau^2}\right]}$$

Now if $(\theta|X) \sim N(a, b)$, this has density proportional to

$$e^{-\frac{(\theta-a)^2}{2b^2}} \propto e^{-\frac{\theta^2}{2b^2} + \frac{\theta a}{b^2}}$$

$$\frac{1}{b^2} = \frac{1}{\sigma^2} + \frac{1}{\tau^2} \Rightarrow b^2 = \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}$$

$$\frac{a}{b^2} = \frac{\sum X_i}{\sigma^2} + \frac{\mu}{\tau^2} \Rightarrow a = \frac{\sum_{i=1}^n X_i/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} = \frac{\mu\sigma^2 + n\bar{X}\tau^2}{\sigma^2 + n\tau^2}$$

$\Rightarrow$ Posterior dist. is $N\left(\frac{\mu\sigma^2 + n\bar{X}\tau^2}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$

Remark: If the prior has density π, the posterior has density $\pi(\theta|X)$ w.r.t. the same ~~dist.~~ dominated measure

$\to \pi(\theta|X)m(X) = p_\theta(X)\pi(\theta)$

$\Rightarrow \pi(\theta|X) \propto p_\theta(X)\pi(\theta)$

$\propto g_\theta(T(X))\pi(\theta)$ – posterior depends on X through sufficiency.

[Thm] If $L(g(\theta), \delta(X)) = \{g(\theta)-\delta(X)\}^2$, then $\delta_0(X) = E\{g(\theta)|X\}$ is a Bayes estimate.
with Bayes risk $E\{Var(g(\theta)|X)\}$
If $\delta(X)$ is another Bayes estimator, then $\delta_0(X)=\delta(X)$ w.p. 1.

Pf: Let δ be any estimator, then the risk of δ is

$$E\{(\delta(X)-g(\theta))^2\} = E\{\delta(X)-\delta_0(X)+\delta_0(X)-g(\theta)\}^2$$

$$= E\{(\delta(X)-\delta_0(X))^2\} + E\{(\delta_0(X)-g(\theta))^2\} + 2E\{(\delta(X)-\delta_0(X))(\delta_0(X)-g(\theta))\}.$$

$$\geq E\{(\delta(X)-\delta_0(X))\}^2 + E\{\delta_0(X)-g(\theta)\}^2 \text{ if } \underline{E\{(\delta(X)-\delta_0(X))(\delta_0(X)-g(\theta))\}=0}. \ (*)$$

δ is a Bayes estimator ✓

$\Rightarrow$ δ_0 is a Bayes estimator. Furthermore,

iff $\delta(X)=\delta_0(X)$ w.p. 1.

Finally, Bayes risk of $\delta_0 = E[g(\theta)-E\{g(\theta)|X\}]^2$

$$= E(E[g(\theta)-E\{g(\theta)|X\}]^2|X)$$

$$= E\{Var(g(\theta)|X)\}$$

To see $(*)$,

$$E\{(\delta(X)-\delta_0(X))(\delta_0(X)-g(\theta))\}$$

$$= E[E\{(\delta_0(X)-\delta(X))(\delta_0(X)-g(\theta))\}|X]$$

$$= E[(\delta_0(X)-\delta(X))E\{\delta_0(X)-g(\theta)|X\}] = 0.$$

Next: Least favorable prior.

Lecture 7 (Average risk optimality) Bayes. est. minimaxity, admissibility.

[THM] If $L(g(\theta), \delta(X)) = \{g(\theta) - \delta(X)\}^2$, then $\delta_0(X) = E(g(\theta)|X)$ is a Bayes estimate with Bayes risk $E(Var(g(\theta)|X))$. If $\delta(X)$ is another Bayes estimator, then $\delta_0(X) = \delta(X)$ w.p. 1. (Pf in L6)

Remarks

(a) Here $\delta_0(X) = \delta(X)$ w.p. 1 refers to the joint probability when X and θ are both random.

This also means that $\delta_0(X) = \delta(X)$ w.p. 1 under the marginal dist. of X.

(b) This does not imply $P(\delta(X) = \delta_0(X) \mid \theta) = 1 \quad \forall \theta$.

(c) If, however, the marginal dist. of X dominates P_θ, $\theta \in \Theta$, then we have $\delta_0(X)$ is the unique Bayes estimate in the sense that $P_\theta(\delta(X) = \delta_0(X)) = 1 \quad \forall \theta$.

Example

Suppose $X \sim$ Binomial (n, θ), $\theta \in [0,1]$. Case 1: $\pi_1(\theta) = U(0,1)$, Case 2: $\pi_2(0) = \pi_2(1) = \frac{1}{2}$

Case 1:
$$P(X=x) = \int_0^1 \underbrace{\binom{n}{x} \theta^x (1-\theta)^{n-x}}_{p(x|\theta)} d\theta = \binom{n}{x} \int_0^1 \theta^x (1-\theta)^{n-x} d\theta = \frac{\binom{n}{x}}{B(x+1, n-x+1)} = \frac{1}{n+1}$$

Marginal dominates conditional, i.e. $P(X \in A) = 0 \Rightarrow P(X \in A \mid \theta) = 0$.

Bayes estimate is unique.

$$\pi(\theta|x) \propto \theta^x (1-\theta)^{n-x} = Beta(x+1, n-x+1) \Rightarrow \text{Bayes estimate} = \frac{x+1}{n+2}$$

Case 2:
$$P(X=x) = \tfrac{1}{2} P(X=x \mid \theta=0) + \tfrac{1}{2} P(X=x \mid \theta=1)$$
$$= \tfrac{1}{2}\{I(x=0) + I(x=n)\}$$
$$\Rightarrow P(X=0) = P(X=n) = \tfrac{1}{2} \text{ and } P(X=x) = 0 \text{ for } x = \{1, 2, \ldots, n-1\}.$$

$\Rightarrow$ Marginal does not dominate the conditional.

$\Rightarrow$ Bayes estimate is not unique. Correspondingly, the Bayes estimate is $E(\theta|X)$

$$E(\theta|X=0) = P(\theta=1 \mid X=0) = \frac{P(X=0|\theta=1)P(\theta=1)}{P(X=0|\theta=1)P(\theta=1) + P(X=0|\theta=0)P(\theta=0)} = \frac{0}{0+\frac{1}{2}} = 0$$

$$E(\theta|X=n) = \cdots = 1$$

Then the class of all Bayes estimators is given by $\delta_0(0) = 0$, $\delta_0(n) = 1$, $\delta_0(x)$: any arbitrary values for $x \in \{1, \ldots, n-1\}$.

Lemma A Bayes estimator (w.r.t squared error) can never be unbiased, unless $\delta(X) = g(\theta)$ w.p. 1.

Pf: Let $\delta_0(X) = E\{g(\theta)|X\}$ be the Bayes estimator. Assume that $E\{\delta_0(X)|\theta\} = g(\theta)$ [is unbiased]

We claim that $I \triangleq E\{\delta_0(X) - g(\theta)\}^2 = 0$.

$$I = E\{\delta_0(X)\}^2 + E\{g(\theta)\}^2 - 2E\{\delta_0(X)g(\theta)\}$$

where $E\{\delta_0(X)g(\theta)\} = E\{E\{\delta_0(X)g(\theta)|X\}\} = E\{\delta_0(X)E\{g(\theta)|X\}\} = E\{\delta_0(X)\}^2$

or $E\{\delta_0(X)g(\theta)\} = E\{E\{\delta_0(X)g(\theta)|\theta\}\} = E\{g(\theta)\}^2$

$\Rightarrow I = E\{g(\theta)\}^2 - E\{\delta_0(X)\}^2 = E\{\delta_0(X)\}^2 - E\{g(\theta)\}^2$

$\Rightarrow I = 0$. //.

Conjugate

A class of prob. distributions F is said to be a conjugate family of priors for a model $\{P_\theta : \theta \in \Theta\}$

if the posterior distribution $\pi(\theta|X)$ also belongs to F.

Example ① $X_1, \dots, X_n \sim N(\theta, \sigma^2)$, $\theta \sim N(\mu, \tau^2)$, $\pi(\theta|X) \sim N(\cdot, \cdot)$

② $X_1, \dots, X_n \overset{iid}{\sim}$ Binomial$(1, p)$, $p \sim$ Beta(α, β).

This has density $\pi_{\alpha,\beta}(p) = \dfrac{p^{\alpha-1}(1-p)^{\beta-1}}{\text{Beta}(\alpha,\beta)}$, where $\text{Beta}(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp$.

Note $\text{Beta}(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, where $\Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}dx$, $\Gamma(h+1) = h!$ for integer h.

$$f_p(\underset{\sim}{x}) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \Rightarrow \pi(p|\underset{\sim}{x}) \propto p^{\sum_{i=1}^n x_i + \alpha - 1}(1-p)^{n-\sum_{i=1}^n x_i + \beta - 1}$$

$$= \text{Beta}\left(\sum_{i=1}^n x_i + \alpha,\ n - \sum_{i=1}^n x_i + \beta\right)$$

$$E(p|\underset{\sim}{x}) = \frac{\sum_{i=1}^n x_i + \alpha}{n+\alpha+\beta} \rightarrow \frac{\sum x_i}{n} = \bar{x}_n$$

× empirical Bayes

× Hierachical Bayes.

③ $X_1, \dots, X_n \overset{iid}{\sim} P_0(\lambda)$, $\lambda \sim \Gamma(\alpha, \gamma)$ distribution

$$\pi_{\alpha,\gamma}(\lambda) = \frac{e^{-\alpha\lambda}\lambda^{\gamma-1}\alpha^\gamma}{\Gamma(\gamma)}, \quad \lambda > 0.$$

④ $X_1, \dots, X_n \overset{iid}{\sim} U(0, \theta)$, $\theta \sim$ Pareto (a, c), $\pi_{a,c}(\theta) = \dfrac{ac^a}{\theta^{a+1}}$, $\theta > c$.

Minimaxity Ch5. TPE.

Def The minimax risk of an estimator $\delta(X)$ for estimating $g(\theta)$ is $\sup_{\theta\in\Theta} R(g(\theta), \delta)$. An estimator δ_0 is said to be minimax, if, for any other estimator δ, we have $\sup_{\theta\in\Theta} R(g(\theta), \delta_0) \le \sup_{\theta\in\Theta} R(g(\theta), \delta)$.

Def Given a prob. dist. π (prior) on Θ, define the Bayes risk of the prior π by $r(\pi) = r(\pi, \delta_\pi)$
where δ_π is Bayes estimate w.r.t. π.

Def A prior π is said to be <u>least favorable</u> if $r(\pi) \geq r(\pi')$ for all π' (other prior dist. on Θ)

[THM] Suppose π is a distribution on Θ s.t. $r(\pi) = r(\pi, \delta_\pi) = \sup_{\theta\in\Theta} R(g(\theta), \delta_\pi)$

Then (a) δ_π is minimax

(b) If δ_π is unique Bayes w.r.t. π, then δ_π is unique minimax.

(c) π is least favorable. (pf) ↓

<u>Corollary</u> A Bayes estimator with constant risk is minimax.

Pf: This means $R(g(\theta), \delta_\pi) = \alpha$ (free of θ).

$\Rightarrow \quad r(\pi, \delta_\pi) = E_{\theta\sim\pi}\{R(g(\theta), \delta_\pi)\} = \alpha$

and $\quad \sup_{\theta\in\Theta} R(g(\theta), \delta_\pi) = \alpha$,

THM Pf: (a) Let δ be arbitrary, then

bayes estimate.

$$\sup_{\theta\in\Theta} R(g(\theta), \delta) \geq \int_\Theta R(g(\theta), \delta)\, \pi(d\theta) = r(\pi, \delta) \overset{(*)}{\geq} r(\pi, \delta_\pi) = \sup_{\theta\in\Theta} R(g(\theta), \delta_\pi)$$

(b) Let $\delta \neq \delta_\pi$, i.e. $\exists\, \theta$ st $P_\theta(\delta(X) \neq \delta_\pi(X)) > 0$

$\Rightarrow$ (*) is a strictly inequality as δ_π is unique Bayes.

(c) Let π' be any distribution. N.T.S. $r(\pi') \leq r(\pi)$

But observe that $r(\pi') = \int_\theta R(g(\theta), \delta_{\pi'})\, \pi'(d\theta) \leq \int_\theta R(g(\theta), \delta_\pi)\, \pi'(d\theta) \leq \sup_{\theta\in\Theta} R(g(\theta), \delta_\pi) = r(\pi)$

$\Rightarrow$ π is least favorable. //.

<u>Example</u> Let $X_1, \ldots, X_n \overset{iid}{\sim} B(1, p)$. Find a minimax estimator for p.

Let the prior on p be Beta(α, β), i.e. $\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$

Then the Bayes estimator is $\delta_\pi(X) = \dfrac{\sum_{i=1}^n X_i + \alpha}{n+\alpha+\beta}$.

$$R(p, \delta_\pi) = E\left(\frac{\sum_{i=1}^n X_i + \alpha}{n+\alpha+\beta} - p\right)^2 = \frac{np - np^2 + \alpha^2 - 2\alpha(\alpha+\beta)p + (\alpha+\beta)^2 p^2}{(n+\alpha+\beta)^2}$$

To make this free of p, $\begin{cases} n = 2\alpha(\alpha+\beta) \\ n = (\alpha+\beta)^2 \end{cases} \Rightarrow \begin{cases} \alpha+\beta = \sqrt{n} \\ 2\alpha\sqrt{n} = n \end{cases} \Rightarrow \begin{cases} \alpha = \sqrt{n}/2 \\ \beta = \sqrt{n}/2 \end{cases}$

$\delta_\pi(\underset{\sim}{X}) = \dfrac{\sum_{i=1}^n X_i + \sqrt{n}/2}{n+\sqrt{n}}$ is the unique minimax estimator (marginal dominates conditional).

Def A sequence of priors $\{\pi_n\}_{n\geq 1}$ is least favorable if $\lim_{n\to\infty} r(\pi_n) = \sup_{\pi} r(\pi)$

[THM] Suppose $\{\pi_n\}_{n\geq 1}$ is a sequence of priors such that $\lim_{n\to\infty} r(\pi_n) = \sup_{\theta\in\Theta} R(g(\theta), \delta_0)$, then

(a) δ_0 is minimax

(b) $\{\pi_n\}$ is least favorable. (Pf ↓)

Example (motivate the above THM) $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ (σ^2 known). Find a minimax estimator for θ with the squared loss function.

Claim: $\bar{X}_n$ is minimax.

Let $\pi_{\mu,\tau^2}(\theta) = N(\mu, \tau^2)$. The Bayes estimator is $\delta_\pi = \dfrac{\frac{n\bar{X}_n}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$

Bayes risk: $r(\pi) = r(\pi, \delta_\pi) = \dfrac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$.

Here $\delta_0 = \bar{X}_n$, $R(\theta, \bar{X}_n) = E(\bar{X} - \theta)^2 = \frac{\sigma^2}{n} \Rightarrow \sup_{\theta\in\Theta} R(\theta, \bar{X}_n) = \frac{\sigma^2}{n}$

Also, $\lim_{\tau\to\infty} r(\pi_\tau) = \frac{\sigma^2}{n} = \sup_{\theta\in\Theta} R(\theta, \bar{X}_n)$

$\Rightarrow$ $\bar{X}_n$ is minimax. Also, $\{\pi_\tau\}_{\tau\in\mathbb{N}}$ is a least favorable distribution.

[THM] ↓

Pf: (a) Let δ be any other estimator. Then

$$\sup_{\theta\in\Theta} R(g(\theta), \delta) \geq \int_\Theta R(g(\theta), \delta)\, \pi_n(d\theta) = r(\pi_n, \delta) \geq r(\pi_n)$$

Take limit to get $\sup_{\theta\in\Theta} R(g(\theta), \delta) = \lim_{n\to\infty} r(\pi_n)$

(b) N.T.S $\sup_{\pi} r(\pi) = \lim_{n\to\infty} r(\pi_n)$.

Observe that $\sup_{\pi} r(\pi) \geq r(\pi_n) \Rightarrow \sup_{\pi} r(\pi) \geq \lim_{n\to\infty} r(\pi_n)$.

For any π, $r(\pi) = \inf_{\delta} r(\pi, \delta) \leq r(\pi, \delta_0) \leq \sup_{\theta\in\Theta} R(g(\theta), \delta_0) = \lim_{n\to\infty} r(\pi_n)$

Hence, $\sup_{\pi} r(\pi) \leq \lim_{n\to\infty} r(\pi_n)$ //.

[Lemma] Suppose $\delta(X)$ is minimax for $g(\theta)$ on the parameter set $\theta \in \Theta_0$, where $\Theta_0 \subseteq \Theta$.

If $\sup_{\theta\in\Theta_0} R(g(\theta), \delta) = \sup_{\theta\in\Theta} R(g(\theta), \delta)$, then δ is minimax for $\theta \in \Theta$.

Pf see TPE.

Example $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$ (both unknown)

$\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$

For any estimator δ, $\sup_{\theta\in\Theta} R(\mu, \delta) \geq \sup_{\theta\in\Theta;\, \sigma=\sigma_0} R(\mu, \delta) \geq \frac{\sigma_0^2}{n}$

$\Rightarrow \sup_{\theta\in\Theta} R(\mu, \delta) \geq \sup_{\sigma_0 > 0} \frac{\sigma_0^2}{n} = +\infty$

cf.

Example : Assume $\mu \in \mathbb{R}$, $0 < \sigma \le M$, $\Theta = \mathbb{R} \times [0, M]$.

In this case $\bar{X}$ is again minimax. This is because

Let $\Theta_0 = \mathbb{R} \times \{M\}$. In this case, we know that $\bar{X}_n$ is minimax and $\sup_{\theta \in \Theta_0} R(\mu, \bar{X}_n) = \frac{M^2}{n}$

Also, $R(\mu, \bar{X}_n) = \frac{\sigma^2}{n} \Rightarrow \sup_{\theta \in \Theta} R(\mu, \bar{X}) = \sup_{\sigma \in [0,M]} \frac{\sigma^2}{n} = \frac{M^2}{n} = \sup_{\theta \in \Theta} R(\mu, \bar{X}_n)$

$\Rightarrow \bar{X}_n$ is minimax on Θ.

Admissibility

An estimator δ is said to be inadmissible if $\exists \delta'$ s.t. $R(g(\theta), \delta') \le R(g(\theta), \delta)$ with strict inequality for some $\theta \in \Theta$.

An estimator δ is admissible if there is no such δ'.

Remark If loss function is strictly convex, any estimator which is not a function of the minimal sufficient statistic is inadmissible (Rao-Blackwell).

Lemma Any unique Bayes estimator is admissible.

TPE 5.2. Suppose δ is a unique Bayes estimator, which is not admissible,

$\Rightarrow \exists \delta'$ better than $\delta \Rightarrow \delta'$ is Bayes. Contradiction.

Lemma An admissible estimator with constant risk is minimax.

Lemma : If δ is unique minimax, then δ is admissible.

Lecture 8

Asymptotic Optimality ~ M-estimator $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma)$ C&B Ch 7. K. Ch9.

Let $\{X_1, X_2, \dots, X_n\}$ be iid from $\{P_\theta : \theta \in \Theta\}$ with pdf $p_\theta(\cdot)$ w.r.t. some σ-finite measure.

Suppose we want to estimate $g(\theta)$ and a candidate estimator is $\delta_n(X_1, \dots, X_n)$.

Def We say $\delta_n(\underset{\sim}{X})$ is consistent 弱 for $g(\theta)$ if

$$\delta_n(\underset{\sim}{X}) \xrightarrow[P_\theta]{P} g(\theta) \quad \forall \theta \in \Theta, \text{ i.e.}$$

$$\forall \theta \in \Theta,\ \forall \varepsilon > 0,\ P_\theta(|\delta_n(\underset{\sim}{x}) - g(\theta)| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

Example. $X_1, \dots, X_n \overset{iid}{\sim} \text{Bin}(1, \theta)$, UMVUE for $g(\theta) = \theta$ is $\bar{X}_n$.

$\bar{X}_n \xrightarrow[P_\theta]{P} \theta$ by WLLN $\Rightarrow \bar{X}_n$ is consistent for θ.

Remarks For $X_1, \dots, X_n \overset{iid}{\sim} F$

a) Assume $E_F|X_1| < \infty$, then $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} E_F X_1$ (WLLN)

b) Assume $E_F X_i^2 < \infty$, then $W_n \triangleq \dfrac{\sum_{i=1}^n X_i - n E_F X_1}{\sqrt{n \operatorname{Var}_F(X_1)}} \xrightarrow{d} N(0,1)$ (CLT)

i.e. $\lim_{n\to\infty} P(W_n \le t) \to \Phi(t) \quad \forall t \in \mathbb{R}$.

Def Let $L(\theta|X_1,\dots,X_n) = \prod_{i=1}^n p_\theta(X_i)$ be the likelihood function, and $\ell(\theta|X_1,\dots,X_n) = \log L(\theta|X_1,\dots,X_n)$ be the log-likelihood function. If there exists a unique $\hat\theta_n$, which is a global maximizer of $\theta \mapsto L(\theta|\underset{\sim}{X})$ or $\theta \mapsto \ell(\theta|\underset{\sim}{X})$. then define $\hat\theta_n$ as the MLE for θ.

Example Suppose $X_1,\dots,X_n \overset{iid}{\sim} \mathrm{Bin}(1,\theta)$. $p_\theta(x) = \theta^x (1-\theta)^{1-x}$, $\theta \in (0,1)$.

$$L_n(\theta|\underset{\sim}{X}) = \prod_{i=1}^n p_\theta(X_i) = \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}$$

$$\ell_n(\theta|\underset{\sim}{X}) = \sum_{i=1}^n X_i \log\theta + \left(n - \sum_{i=1}^n X_i\right)\log(1-\theta)$$

regularity conditions

$$\ell_n'(\theta|\underset{\sim}{X}) = \frac{\sum_{i=1}^n X_i}{\theta} - \frac{n-\sum_{i=1}^n X_i}{1-\theta}$$

$$\ell_n''(\theta|\underset{\sim}{X}) = -\frac{\sum_{i=1}^n X_i}{\theta^2} - \frac{n-\sum_{i=1}^n X_i}{(1-\theta)^2} < 0 \Rightarrow \ell_n(\cdot|\underset{\sim}{X}) \text{ is strictly concave.}$$

Observe that $\ell_n'(\theta|\underset{\sim}{X})\Big|_{\theta=\hat\theta_n} = 0 \Rightarrow \frac{\sum_{i=1}^n X_i}{\theta} = \frac{n-\sum_{i=1}^n X_i}{1-\theta}$

$$\Rightarrow \hat\theta_n = \frac{\sum_{i=1}^n X_i}{n} = \bar X_n$$

$\Rightarrow$ MLE exists and equals $\bar X_n$.

Also, $\bar X_n \xrightarrow[P_\theta]{P} \theta \quad \forall \theta \in (0,1)$ [CONSISTENCY]

and $\underset{\text{slow down the motion}}{\sqrt{n}}(\bar X_n - \theta) \xrightarrow[P_\theta]{d} N\left(0, \frac{1}{\theta(1-\theta)}\right)$ [Asy. normality via CLT]

[Thm] Suppose $X_1,\dots,X_n$ are iid from P_θ for some $\theta \in \Theta$, with pdf $p_\theta(\cdot)$.

A0. $P_{\theta_1} \neq P_{\theta_2}$ whenever $\theta_1 \neq \theta_2$ (identifiability)

A1. $\{P_\theta, \theta \in \Theta\}$ have common support.

Then, $P_{\theta_0}\left(\ell_n(\theta_0|\underset{\sim}{X}) > \ell_n(\theta|\underset{\sim}{X})\right) \xrightarrow{n\to\infty} 1 \quad \forall \theta \neq \theta_0$

Pf: Let $T_n = \frac{1}{n}\sum_{i=1}^n \log\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}$, then $T_n \xrightarrow[P_{\theta_0}]{P} E_{\theta_0}\left[\log\frac{p_\theta(X_1)}{p_{\theta_0}(X_1)}\right]$

entropy

Now $E_{\theta_0}\left\{\log\frac{p_\theta(X_1)}{p_{\theta_0}(X_1)}\right\} = \int \log\left\{\frac{p_\theta(x)}{p_{\theta_0}(x)}\right\} p_{\theta_0}(x)\, d\mu = -D(\theta_0\|\theta) < 0$ for $\theta \neq \theta_0$

$\Rightarrow P_{\theta_0}(T_n < 0) \xrightarrow[n\to\infty]{} 1$ But $T_n < 0 \Leftrightarrow \frac{1}{n}\sum_{i=1}^n \log\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} < 0$

$$\Leftrightarrow \log\prod_{i=1}^n p_\theta(X_i) < \log\prod_{i=1}^n p_{\theta_0}(X_i)$$

$$\Leftrightarrow \ell_n(\theta|\underset{\sim}{X}) < \ell_n(\theta_0|\underset{\sim}{X})$$

[Corollary] Suppose (A0) and (A1) hold. If Θ is finite, then the MLE $\hat\theta_n$ exists with high prob. (prob $\to 1$) and $P_{\theta_0}(\hat\theta_n = \theta_0) \xrightarrow{n\to\infty} 1$.

Pf: Let $\Theta = \{\theta_0, \theta_1, \dots, \theta_K\} \Rightarrow P_{\theta_0}\left(\ell_n(\theta_0|\underset{\sim}{X}) > \ell_n(\theta_j|\underset{\sim}{X})\right) \xrightarrow[n\to\infty]{} 1, \quad 1 \le j \le K.$

$$\Rightarrow P_{\theta_0}\left(\ell_n(\theta_0|\underset{\sim}{X}) > \max_{1\le j\le K} \ell_n(\theta_j|\underset{\sim}{X})\right) \xrightarrow{n\to\infty} 1$$

Let $A_n = \{ \underset{\sim}{X} : l_n(\theta_n \mid \underset{\sim}{X}) > \max_{1 \le j \le k} l_n(\theta_j \mid \underset{\sim}{X}) \}$

If $\underset{\sim}{X} \in A_n$, then $\hat{\theta}_n(X) = \theta_0$ and $P_{\theta_0}(A_n) \to 1$.

<u>MLE expansion</u> $\quad \sqrt{n}(\hat{\theta}_n - \theta_0) \to 0$

interest

$$0 = l_n'(\hat{\theta}_n) = l_n'(\theta_0) + \boxed{(\hat{\theta}_n - \theta_0)} l_n''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 l_n'''(\xi_n) \qquad \xi_n \in [\theta_0, \hat{\theta}_n]$$

$$\Rightarrow (\hat{\theta}_n - \theta_0)\{ l_n''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 l_n'''(\xi_n) \} = -l_n'(\theta_0)$$

$$\Rightarrow \quad \sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-l_n'(\theta_0)/\sqrt{n}}{-l_n''(\theta_0)/n - \frac{1}{2}(\hat{\theta}_n - \theta_0) l'''(\xi_n)/n}.$$

It suffices to show

$$\frac{1}{\sqrt{n}} l_n'(\theta_0) \xrightarrow{d} N(0, I(\theta_0))$$

$$-\frac{1}{n} l_n''(\theta_n) \xrightarrow{P} I(\theta_0)$$

and $\frac{1}{n}(\hat{\theta}_n - \theta_0) l'''(\xi_n) \xrightarrow{P} 0$,

then $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \dfrac{N(0, I(\theta_0))}{I(\theta_0) + 0} \overset{d}{=} N(0, I(\theta_0)^{-1})$.

Observe that $\quad u(\hat{\theta}_n) = 0 \Rightarrow u'(\theta_0) + u''(\theta_0)(\hat{\theta}_n - \theta_0) + \cdots$

$$\frac{1}{\sqrt{n}} l_n'(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p_\theta(X_i) \Big|_{\theta = \theta_0} \xrightarrow[\theta_0]{d} N(0, I(\theta_0)) \quad \text{by CLT}.$$

$$\frac{1}{n} l_n''(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log p_\theta(X_i) \Big|_{\theta = \theta_0} \xrightarrow[\theta_0]{P} -I(\theta_0) \quad \text{by WLLN}.$$

$$\left| \frac{1}{n}(\hat{\theta}_n - \theta_0) l_n''(\xi_n) \right| \le |\hat{\theta}_n - \theta_0| \cdot \frac{1}{n} \sum_{i=1}^{n} M(X_i) \xrightarrow[\theta_0]{P} 0 \cdot E_{\theta_0} M(X_i) = 0 \quad \text{by consistency and WLLN.}$$

estimating equations

semi-parametric

Vander Vaart (98) Ch.5. $\quad$ M. Z estimator

$E\{N(t) = Y(t)\,d\Lambda(t)\} = 0$

poisson $\quad$ mean.

Cox model: proportional hazards model $\quad \lambda(t|Z) = \underbrace{\lambda_0(t)}_{\text{non-para}} \underbrace{\exp\{\beta^T Z\}}_{\text{para}}$ $\quad Z = \begin{cases} 0 & \text{placebo} \\ 1 & \text{drug} \end{cases}$

counting process

$\lambda_0(t)$: baseline hazard (unspecified) $\quad$ $\beta^T Z$: covariate effect

---

Main references: TSH (3rd)

<u>Setup</u> Let $\{P_\theta : \theta \in \Theta\}$ be a collection of prob. measure on $\mathcal{X}$, dominated by a σ-finite measure μ. Let $p_\theta(\cdot) = \dfrac{dP_\theta}{d\mu}$. Let Θ_0 and Θ_1 be two disjoint subsets of Θ. Given $X \sim P_\theta$ for some $\theta \in \Theta$, we want to decide whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$.

$$\frac{f(t|z)}{S(t|z)} = \frac{f(t|z)}{1 - F(t|z)} \qquad F(t) = P(T \le t)$$

the semi-parametric model

$\sqrt{n}(\hat{\beta} - \beta_0) \to ?$

$\therefore$ partial likelihood $\rightsquigarrow$ profile likelihood (Murphy & vdV, 2000 JRSSB)

(Cox, 72, 75)

72: JRSSB $\quad$ 75: Biometrika

E.g. $X \in \mathbb{R}^n$, $X = (X_1, \dots, X_n) \overset{iid}{\sim} N(\theta, 1)$

$\Theta \in \mathbb{R}$, $\Theta_0 = \{0\}$, $\quad \Theta_1 = \{1\}$. $\quad$ simple vs simple / composite

$\Theta_1 = \{\theta : \theta > 1\}$.

Top1 cited paper

Kaplan-Meier 337.

Cox & Hinkley (74) Theoretical Statistics

strongly recommended by Ying Zhiliang.

Def A function $\phi : X \to \{0,1\}$ is called a non-randomized test function.

Types of error

decision	$\theta \in \Theta_1$	$\theta \in \Theta_0$
$\phi = 1$	✓	Type I
$\phi = 0$	Type II	✓

need to be controlled. more important.

$P_\theta(\phi = 1)$, $\theta \in \Theta_0$ Type I

$P_\theta(\phi = 0)$, $\theta \in \Theta_1$ Type II

* Power function of ϕ : 1 - prob. of type II error $= P_\theta(\phi=1)$, $\theta \in \Theta_1$ prob. of correctly reject H_0
* size of a test ϕ : $\sup_{\theta \in \Theta_0} P_\theta(\phi = 1)$ type I

Let $\alpha \in (0,1)$, a test ϕ is called <u>level α</u> if $\sup_{\theta\in\Theta_0} P_\theta(\phi=1) \le \alpha$.

Def A test ϕ is called <u>uniformly most powerful</u> level α test (UMP) if given any other level α test ψ, we have $P_\theta(\phi=1) \ge P_\theta(\psi = 1)$ $\forall \theta \in \Theta_1$.

Def A function $\phi : X \to [0,1]$ is called a randomized test function or just a test function. If $\phi(\underset{\sim}{x}) = p$, toss a coin with prob. of heads p. If heads choose Θ_1, if tails choose Θ_0. In all previous definitions, replace $P_\theta(\phi = 1)$ by $E_\theta \phi$.

[Thm] (Neyman - Pearson)

Suppose we want to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ at level α

(a) There exists a test ϕ satisfying

(i) $E_{\theta_0}\phi = \alpha$

(ii) $\exists\, k \in [0,\infty)$ s.t. $\phi = \begin{cases} 1 & \text{if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k \\ 0 & \text{o/w.} \end{cases}$

(b) If a test satisfies (i) and (ii) above, then ϕ is a Most Powerful test for testing $\theta = \theta_0$ vs $\theta = \theta_1$ at level α.

(c) If ϕ is Most Powerful at level α, it must satisfy (ii), for the same k as in (a). It also satisfies (i) unless $E_{\theta_1}(\phi) = 1$ (power = 1).

Pf: (a) If $\alpha = 0$, take $k = \infty$, $\phi = 0$. If $\alpha = 1$, take $k = 0$, $\phi = 1$.

For $\alpha \in (0,1)$, let $\alpha(c) = P_{\theta_0}(p_{\theta_1}(x) > c\, p_{\theta_0}(x))$, $c > 0$

$$= P_{\theta_0}\left(\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > c\right) = 1 - P_{\theta_0}\left(\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \le c\right)$$

$\Rightarrow \alpha(\cdot)$ is non-decreasing and right-continuous.

Also, $\alpha(c-) - \alpha(c) = P_{\theta_0}\left(\frac{P_{\theta_1}(x)}{P_{\theta_0}(x)} = c\right)$, $\alpha(\infty) = 0$, $\alpha(0-) = 1$.

$\Rightarrow \exists\, C_0$ s.t. $\alpha(C_0) \le \alpha \le \alpha(C_0-)$.

Let

$$\phi = \begin{cases} 1 & \\ \dfrac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} & \text{if} \\ 0 & \end{cases} \quad \begin{array}{l} P_{\theta_1}(x) > C_0 P_{\theta_0}(x) \\ P_{\theta_1}(x) = C_0 P_{\theta_0}(x) \\ P_{\theta_1}(x) < C_0 P_{\theta_0}(x) \end{array}$$

with if $\alpha(C_0-) = \alpha(C_0)$, set $\phi = 1$ on $P_{\theta_1}(x) = C_0 P_{\theta_0}(x)$

$$\Rightarrow E_{\theta_0}\phi = P_{\theta_0}(P_{\theta_1}(x) > C_0 P_{\theta_0}(x)) + P_{\theta_0}(P_{\theta_1}(x) = C_0 P_{\theta_0}(x)) \cdot \frac{\alpha - \alpha(C_0)}{\alpha(C_0-) - \alpha(C_0)}$$

$$= \alpha(C_0) + [\alpha(C_0-) - \alpha(C_0)] \times \frac{\alpha - \alpha(C_0)}{\alpha(C_0-) - \alpha(C_0)}$$

$$= \alpha.$$

(b) Let ϕ be of the MP form, i.e. $\exists\, k$ s.t. $E_{\theta_0}\phi = \alpha$, $\phi(x) = \begin{cases} 1 & \text{if } \frac{P_{\theta_1}(x)}{P_{\theta_0}(x)} \underset{<}{>} k \\ 0 & \text{o/w}. \end{cases}$

Let $\phi^*(x)$ be the test s.t. $E_{\theta_0}\{\phi^*(x)\} \le \alpha$,

we need to show that $E_{\theta_1}\{\phi(x)\} - E_{\theta_1}\{\phi^*(x)\} \ge 0$.

Consider the integral:

$$\int \{\phi(x) - \phi^*(x)\}\{P_{\theta_1}(x) - kP_{\theta_0}(x)\}\, d\mu$$

$$= \int_{\phi > \phi^*} \{\phi(x) - \phi^*(x)\}\{P_{\theta_1}(x) - kP_{\theta_0}(x)\}\, d\mu$$

$$+ \int_{\phi < \phi^*} \{\phi(x) - \phi^*(x)\}\{P_{\theta_1}(x) - kP_{\theta_0}(x)\}\, d\mu$$

Observe if $\phi > \phi^*$, $\phi > 0 \Rightarrow P_{\theta_1}(x) \ge k P_{\theta_0}(x)$

if $\phi < \phi^*$, $\phi < 1 \Rightarrow P_{\theta_1}(x) \le k P_{\theta_0}(x)$

therefore, $0 \le \int \{\phi(x) - \phi^*(x)\}\{P_{\theta_1}(x) - kP_{\theta_0}(x)\}\, d\mu$

$$= E_{\theta_1}\phi(x) - E_{\theta_1}\phi^*(x) - k\{E_{\theta_0}\phi(x) - E_{\theta_0}\phi^*(x)\}$$

$$\Rightarrow E_{\theta_1}\phi(x) - E_{\theta_1}\phi^*(x) \ge k\{E_{\theta_0}\phi(x) - E_{\theta_0}\phi^*(x)\} \ge k(\alpha - \alpha) = 0.$$

(c) Let ϕ^* be an MP test. Let ϕ be the test from (a), we have $E_{\theta_1}\phi(x) = E_{\theta_1}\phi^*(x) = \alpha$

$$\Rightarrow \int (\phi(x) - \phi^*(x))\{P_{\theta_1}(x) - kP_{\theta_0}(x)\}\, d\mu = 0$$

$$\Rightarrow \int_{\phi > \phi^*} \{\phi(x) - \phi^*(x)\}\{P_{\theta_1}(x) - kP_{\theta_0}(x)\}\, d\mu = \int_{\phi < \phi^*} \{\phi(x) - \phi^*(x)\}\{P_{\theta_1}(x) - kP_{\theta_0}(x)\}\, d\mu$$

$\Rightarrow \phi = \phi^*$ $\forall\, x$ s.t. $P_{\theta_1} \ne kP_{\theta_0}(x)$, where k is defined as in (a).

Also, we must have $E_{\theta_0}\{\phi^*(x)\} = \alpha$ unless $E_{\theta_1}\phi^*(x) = 1$

because $E_{\theta_0}(\phi^*(x)) = E_{\theta_0}(\phi(x)) = \alpha$ unless $k=0$. $B_n + k = 0 \iff E_{\theta_1}\phi^*(x) = 1$.

If $E_{\theta_0}(\phi^*(x)) < \alpha$, $E_{\theta_1}(\phi^*(x)) < 1$, then ϕ^* is not MP.

Remark: If $\{x : P_{\theta_1}(x) = k P_{\theta_0}(x)\}$ is of measure 0, MP is unique.

E.g. $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$

test $H_0: \theta = 0$ vs $H_1: \theta = 1$ at level α.

$$\frac{P_{\theta=1}(x_1, \dots, x_n)}{P_{\theta=0}(x_1, \dots, x_n)} = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i-1)^2}}{\left(\frac{1}{\sqrt{2\pi}}\right) e^{-\frac{1}{2}\sum_{i=1}^n (x_i^2)}} = e^{\sum_{i=1}^n x_i - \frac{n}{2}}$$

$$\Rightarrow \phi = 1 \text{ if } \frac{p_{\theta_1}(\underset{\sim}{x})}{p_{\theta_0}(\underset{\sim}{x})} > k \iff \sum_{i=1}^n x_i - \frac{n}{2} > \log k \iff \sum_{i=1}^n x_i > k' = \log k + \frac{n}{2}$$

$$\Rightarrow \phi(\underset{\sim}{x}) = \begin{cases} 1 & \text{if } \sum x_i > k' \\ 0 & \sum x_i < k' \end{cases}$$

where $\alpha = E_{\theta=0}\phi(x) = P_{\theta_0}\left(\sum_{i=1}^n x_i > k'\right) \Rightarrow k' = \sqrt{n}\, z_{1-\alpha}$. $P(Z \le z_{1-\alpha}) = 1-\alpha$.

Lecture 9 UMP, MLR, least favorable dist, ... TSH. Ch3 K. Ch12. C&B Ch8.

* Neyman-Pearson (Simple vs Simple)

Recap. the above example ↑

$$\phi(x) = \begin{cases} 1 & > \\ \gamma & = \quad * \text{(randomized test)} \\ 0 & < \end{cases}$$

E.g. $X_1, X_2 \overset{iid}{\sim}$ Bernoulli(θ)

Test $H_0: \theta = \frac{1}{2}$ versus $H_1: \theta = \frac{2}{3}$ at level $\alpha = \frac{1}{2}$

Sample	(0,0)	(0,1)	(1,0)	(1,1)
$P_{\theta_0}(x_1, x_2)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$P_{\theta_1}(x_1, x_2)$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{4}{9}$
$\frac{P_{\theta_1}}{P_{\theta_0}}$	$\frac{4}{9}$ not reject H_0	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{16}{9}$ reject H_0

$\Rightarrow k = \frac{8}{9}$

Let $\phi = (x_1, x_2) = \begin{cases} 1 & (x_1, x_2) = (1,1) \\ 0 & (x_1, x_2) = (0,0) \\ \text{randomized} & \end{cases}$

s.t. $E_{\theta_0}\phi(x_1, x_2) = \frac{1}{2}$

Test procedure: $\left.\begin{array}{l} \phi(1,0) = 1, \ \phi(0,1) = 0 \\ \phi(1,0) = 0, \ \phi(0,1) = 1 \end{array}\right\}$ non-randomized

$\phi(1,0) = \phi(0,1) = \frac{1}{2}$ randomized characterized by MP.

[Corollary] Let $\beta = \beta(\theta_1)$ denote the power of the MP test for testing $H_0: \theta = \theta_0$ vs $H_1: \theta = \theta_1$ at level $\alpha \in (0,1)$. Then $\beta \geq \alpha$. Further more, $\beta > \alpha$ unless $P_{\theta_1} = P_{\theta_0}$.

Pf: Let ϕ be the MP test from part (a) of NP lemma.

Let $\psi(\underset{\sim}{X}) \equiv \alpha \Rightarrow \beta = E_{\theta_1}\phi(X) \geq E_{\theta_1}\psi(\underset{\sim}{X}) = \alpha$.

Suppose $\beta = \alpha$, then ψ is a MP test.

$\Rightarrow P_{\theta_1}(\underset{\sim}{X}) = k P_{\theta_0}(\underset{\sim}{X})$ a.s. μ.

$\Rightarrow k = 1 \Rightarrow P_{\theta_1} = P_{\theta_0}$ //.

E.g. Suppose $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$ $H_1: \theta \neq 0$ (no UMP test).

Test $\underbrace{H_0: \theta = 0}_{\text{simple}}$ vs $\underbrace{H_1: \theta > 0}_{\text{composite}}$ ($H_1: \theta < 0$) at level α.

Fix $\theta_1 > 0$ ($\theta_1 < 0$) $(\theta_1 \in \Theta_1)$

Test: $H_0: \theta = 0$ vs $H_1': \theta = \theta_1$ at level α.

MP test for this problem is

$$\phi(\underset{\sim}{X}) = \begin{cases} 1 & \text{if } \sum X_i \overset{<}{>} \sqrt{n}\, z_{1-\alpha} \\ 0 & \text{o/w.} \end{cases}$$

$\Rightarrow \phi$ is uniformly MP for testing $H_0: \theta = 0$ vs $H_1: \theta \overset{<}{>} 0$.

Monotone Likelihood Ratio (MLR)

Def Suppose Θ is an interval. We say that $\{P_\theta : \theta \in \Theta\}$ has the monotone likelihood ratio (MLR) property in a statistic $T(X)$ if $\forall\ \theta_1 < \theta_2 \in \Theta$, $\frac{P_{\theta_2}(\underset{\sim}{X})}{P_{\theta_1}(\underset{\sim}{X})}$ is a non-decreasing function of $T(X)$.

E.g. $P_\theta(X) = e^{\eta(\theta)T(X) - B(\theta)} h(x)$, $\theta \in (a,b)$, $\underline{\eta \text{ non-decreasing}}$ **

$$\frac{P_{\theta_2}(X)}{P_{\theta_1}(X)} = e^{\{\eta(\theta_2) - \eta(\theta_1)\}T(X)} e^{-B(\theta_2) + B(\theta_1)} \triangleq g(T(X)),$$

where $g(t) = e^{\{\eta(\theta_2) - \eta(\theta_1)\}t} e^{-B(\theta_2) + B(\theta_1)}$

E.g. $X_1, \dots, X_n \overset{iid}{\sim}$ uniform$(0,\theta)$. $P_\theta(\underset{\sim}{X}) = \frac{1}{\theta^n} I(X_{(n)} < \theta)$

Let $T = X_{(n)}$, $\theta_1 < \theta_2$. If $0 < T < \theta_1$, $\frac{P_{\theta_2}(\underset{\sim}{X})}{P_{\theta_1}(\underset{\sim}{X})} = \left(\frac{\theta_1}{\theta_2}\right)^n$

if $\theta_1 \leq T < \theta_2$, $\frac{P_{\theta_2}(\underset{\sim}{X})}{P_{\theta_1}(\underset{\sim}{X})} = \infty$

if $\theta_2 \leq T$, $\frac{P_{\theta_2}(\underset{\sim}{X})}{P_{\theta_1}(\underset{\sim}{X})} = \frac{0}{0}$ (set to be ∞).

Thm 3.4.1

[Thm] Let $\{P_\theta(\cdot), \theta \in \Theta\}$ be MLR in $T(X)$ s.t. $P_{\theta_1} \neq P_{\theta_2}$ if $\theta_1 \neq \theta_2$ and Θ is an interval.

(a) For testing $H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$ at level $\alpha \in (0,1)$, there exists a UMP test ϕ of the form

$$\phi(X) = \begin{cases} 1 & \text{if } T(\underset{\sim}{X}) > c \\ \gamma & \phantom{\text{if }} T(\underset{\sim}{X}) = c \\ 0 & \phantom{\text{if }} T(\underset{\sim}{X}) < c \end{cases} \quad \text{and } E_{\theta_0}\phi(X) = \alpha.$$

(b) The power function $\beta(\theta) = E_\theta \phi$ is strictly increasing on the set $\{\theta : 0 < \beta(\theta) < 1\}$,

i.e. if $\theta_1 < \theta_2 \in \Theta$ s.t. $\beta(\theta_1), \beta(\theta_2) \in (0,1)$, then $\beta(\theta_1) < \beta(\theta_2)$.

(c) For all $\theta' \in \Theta$, the test of part (a) is UMP for testing $H_0: \theta \le \theta'$ vs $H_1: \theta > \theta'$ at level $\alpha' = \beta(\theta')$

(d) For any $\theta < \theta_0$, ϕ minimizes $\beta(\theta)$ amongst all tests satisfying $E_{\theta_0}\psi(X) = \alpha$.

Pf: Let $f(c) = P_{\theta_0}(T(X) > c)$, $f(\infty) = 0$, $f(-\infty) = 1$,

$\exists\ C_0 \in [-\infty, \infty]$ s.t. $f(C_0-) \ge \alpha \ge f(C_0)$

Let $\phi(X) = \begin{cases} 1 & T(X) > C_0 \\ \dfrac{\alpha - f(C_0)}{f(C_0-) - f(C_0)} & T(X) = C_0 \\ 0 & T(X) < C_0 \end{cases}$, then we can check $E_{\theta_0}\phi(X) = \alpha$.

Fix $\theta_1 > \theta_0$, we need to show ϕ is MP for $\theta = \theta_0$ vs $\theta = \theta_1$.

Let $\dfrac{P_{\theta_1}(X)}{P_{\theta_0}(X)} = g_{\theta_0,\theta_1}(T(X))$ where $g_{\theta_0,\theta_1}(\cdot)$ is non-decreasing

Set $K = g_{\theta_0,\theta_1}(C_0)$. If $\dfrac{P_{\theta_1}(X)}{P_{\theta_0}(X)} > k \Leftrightarrow g_{\theta_0,\theta_1}(T(X)) > g_{\theta_0,\theta_1}(C_0)$

$\Rightarrow T(X) > C_0 \Rightarrow \phi = 1$.

If $\dfrac{P_{\theta_1}(X)}{P_{\theta_0}(X)} < k \Rightarrow \phi = 0$

$\Rightarrow \phi$ is of NP form $\Rightarrow \phi$ is MP for $\theta = \theta_0$ vs $\theta = \theta_1$ at level α.

$\Rightarrow \phi$ is UMP for $\theta = \theta_0$ vs $\theta > \theta_0$ at level α

N.T.S $\sup_{\theta < \theta_0} E_\theta \phi \le \alpha$ s.t. ϕ is UMP for $\theta \le \theta_0$ vs $\theta > \theta_0$

Fix $\theta_0' \le \theta_0$, consider the test problem of $\theta = \theta_0'$ vs $\theta = \theta_0$ at level $\beta(\theta_0') = E_{\theta_0'}\phi$.

ϕ is MP for this problem.

$\Rightarrow \beta(\theta_0') = E_{\theta_0'}\phi \le \underbrace{\boxed{E_{\theta_0}\phi}}_{\text{power}} = \alpha$ (because size $\le$ power)

$\Rightarrow \phi$ is level α UMP test for $\theta \le \theta_0$ vs $\theta > \theta_0$

(b) Fix $\theta' \le \theta''$, assume $\beta(\theta'), \beta(\theta'') \in (0,1)$. N.T.S $\boxed{\beta(\theta') < \beta(\theta'')}$

Consider the problem of testing $\theta = \theta'$ vs $\theta = \theta''$ at level $\beta(\theta')$, ϕ is MP for this problem,

$\Rightarrow \beta(\theta') < \beta(\theta'')$ because of again " size $\le$ power "

(c) Repeat the proof [TSH 3.4.1].

(d) Fix $\theta' < \theta_0$, N.T.S ϕ minimizes $E_{\theta'}\tilde{\phi}$ for all tests with $E_{\theta_0}\tilde{\phi} \le \alpha$.

$\Leftrightarrow$ $\{1-\phi$ maximizes $1-E_{\theta'}\tilde{\phi}$ subject to $1-E_{\theta_0}\tilde{\phi} = 1-\alpha\}$

$\Leftrightarrow$ $\{\psi = 1-\phi$ maximizes $E_{\theta'}\tilde{\psi}$ subject to $E_{\theta_0}\tilde{\psi} = 1-\alpha\}$

$\theta \ne \theta' \Rightarrow P_\theta \ne P_{\theta'}$ (identifiability)

i.e. ψ is ~~a~~ MP for $\theta = \theta_0$ vs $\theta = \theta'$ at level $1-\alpha$

where $\psi = \begin{cases} 1 & T(X) < c_0 \\ \gamma & T(X) = c_0 \\ 0 & T(X) > c_0 \end{cases}$ and $E_{\theta_0}\psi = 1-\alpha$.

E.g. $X_1, \dots, X_n \overset{iid}{\sim} U(0,\theta)$. Test $H_0: \theta = 1$ vs $H_1: \theta > 1$ at level α.

By the thm, a UMP test is given by $\phi = \begin{cases} 1 \\ 0 \end{cases}$ if $\begin{matrix} X_{(n)} > k \\ X_{(n)} < k \end{matrix}$ and

$$\alpha = E_{\theta=1}\phi = P_{\theta=1}(X_{(n)} > k) = 1 - P_{\theta=1}(X_{(n)} \le k) = 1 - k^n \Rightarrow k = (1-\alpha)^{1/n}$$

E.g. (Cauchy location level)

Let X have the density $p_\theta(x) = \frac{1}{1+(x-\theta)^2}$. We find two points at which the MLR condition fails.

For any fixed $\theta > 0$,

$$\frac{p_\theta(x)}{p_0(x)} = \frac{1+x^2}{1+(x-\theta)^2} \to 1 \quad \text{as } x \to \infty \text{ or } x \to -\infty$$

but $p_\theta(0)/p_0(0) = \frac{1}{1+\theta^2}$, which is strictly less than 1.

Thus, the ratio must increase at some values of x and decrease at others.

Hence, $p_\theta(x)$ is not monotone in x, or in other words, the likelihood ratio

$T(X) = x$ is not MLR.

Strategies for finding UMPs $\to$ existence not guaranteed.

1) Reduce the composite alternative to a simple alternative. If H_1 is composite, fix $\theta_1 \in \Theta_1$ and test H_0 against $H_1: \theta = \theta_1$. (Hope that it does not depend on θ_1) ✓

2) Collapse the composite null to a simple null $(\cdots)^{☆}$.

3) Apply Neyman Pearson Lemma: Find the MP LRT test for simple vs simple case / use MLR trick. ✓

Least favorable distribution

Consider: $H_0: X \sim f_\theta \quad \theta \in \Theta$
$H_1: X \sim g$ (known).

We now impose a prior distribution π on Θ_0. So we consider a new set of Hypotheses:

$$H_\pi: X \sim h_\pi(x) = \int_{\Theta_0} f_\theta(x)\, d\pi(\theta)$$

vs $H_1: X \sim g$

Let β_π be the power of the MP level α test ϕ_π for testing H_π vs. g.

Def π is a least favorable distribution if $\beta_\pi \le \beta_{\pi'}$ for any prior π'. (smallest power).

[Thm] (TSH 3.8.1) Suppose ϕ_π is a MP level α test for testing H_π against g. If ϕ_π is level α for the original hypothesis H_0 (i.e. $E_{\theta_0}\phi_\pi(X) \le \alpha \ \forall \theta_0 \in \Theta_0$), then

(a) The test ϕ_π is MP for the original $H_0: \theta \in \Theta_0$ vs. g.

(b) The distribution π is least favorable.

Pf: (a) Let ϕ^* be any other level-α test of $H_0: \theta \in \Theta_0$ vs g. Then ϕ^* is also a level α test for H_π v.s. g because

$$E_\theta \phi^*(X) = \int \phi^*(x) f_\theta(x)\, d\mu(x) \le \alpha \quad \forall \theta \in \Theta_0.$$

which implies that

$$\int \phi^*(x) h_\pi(x)\, d\mu(x) = \iint \phi^*(x) f_\theta(x)\, d\mu(x)\, d\pi(\theta) \le \int \alpha\, d\pi(\theta) = \alpha$$

Since ϕ_π is MP for H_π vs g, we have

$$\int \phi^*(x) \underbrace{g(x)\, d\mu(x)}_{\text{power}} \le \int \phi_\pi(x) g(x)\, d\mu(x)$$

Hence ϕ_π is a MP test for H_0 vs g because ϕ_π is also level α.

(b) Let π' be any distribution on Θ_0. Since $E_\theta \phi_\pi(X) \le \alpha \quad \forall \theta \in \Theta_0$, we know that ϕ_π must be level α for $H_{\pi'}$ vs g. Thus $\beta_\pi \le \beta_{\pi'}$ so π is least favorable dist.

Example (Testing in the presence of nuisance parameters)

Let $X_1, \dots, X_n$ be iid $N(\theta, \sigma^2)$, where both (θ, σ^2) are unknown.

We consider the test $H_0: \sigma \le \sigma_0$ against $H_1: \sigma > \sigma_0$. $\Theta_0 = \{(\theta,\sigma), \theta \in \mathbb{R}, \sigma < \sigma_0\}$

1. Fix a simple alternative (θ_1, σ_1) for some arbitrary $\theta_1, \sigma_1 > \sigma_0$.

2.* Choose a prior π to "collapse over null hypothesis". $\sigma = \sigma_0$

Consider the boundary case between H_0 and H_1: $\{\sigma = \sigma_0\}$

π will be a prob. dist. over $\theta \in \mathbb{R}$ for the fixed $\sigma = \sigma_0$.

* observation: Given any test function $\phi(X)$ and a sufficient satistic T, there exists a test function η that less than some power as ϕ but depends on X only through T.

$$\eta(T(X)) = E\{\phi(X) \mid T(X)\}.$$

Hence, we restrict our attention to sufficient statistics.

(Y, U) where $Y = \bar{X}_n$ and $U = \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$. We know that $Y \sim N(\theta, \frac{\sigma^2}{n})$ and $U \sim \sigma^2 \chi_{n-1}$ and Y is ind. of U due to Basu's thm.

Thus, for π supported on $\sigma = \sigma_0$, we obtain the joint density of (Y, U) under H_π as

$$C_0 U^{\frac{n-3}{2}} \exp\left(-\frac{u}{2\sigma^2}\right) \int \exp\left(-\frac{u}{2\sigma_0^2}(y-\theta)^2\right) d\pi(\theta)$$

and the joint density under the alternative hypothesis.

$$c_1 u^{\frac{n-3}{2}} \exp\left(-\frac{u}{2\sigma_1^2}\right) \exp\left(-\frac{n}{2\sigma_1^2}(y-\theta_1)^2\right).$$

To achieve the minimal maximum power against the alternative (i.e. to be least favorable) we need to choose π s.t. the two distributions become as close as possible. Under H_1, $Y \sim N(\theta_1, \frac{\sigma_1^2}{n})$. Under H_π, the distribution of Y is in a convolution form, i.e. $Y = Z + \Theta$, for $Z \sim N(0, \frac{\sigma_0^2}{n})$, $\Theta \sim \pi$, where Z and Θ are indep.. Hence, if we choose $\Theta \sim N(\theta_1, \frac{\sigma_1^2 - \sigma_0^2}{n})$, Y will become the same distribution under both H_π and H_1, which is $N(\theta_1, \frac{\sigma_1^2}{n})$. Under this prior, the LRT rejects for large values of $\exp\left\{-\frac{u}{2\sigma_1^2} + \frac{u}{2\sigma_0^2}\right\}$, i.e. large values of u.

So, the MP test rejects H_π if $\sum_{i=1}^n (X_i - \bar{X})^2$ lies above the threshold determined by the size constraint. In particular, it rejects H_π if $\sum_{i=1}^n (X_i - \bar{X}_n)^2 > \sigma_0^2 C_{n-1, 1-\alpha}$, where $C_{n-1,1-\alpha}$ is the $(1-\alpha)$th quantile of χ^2_{n-1}.

3. Check if this MP test is of level α for the composite null. For any (θ, σ) with $\sigma \le \sigma_0$, the prob. of rejection is $P_{\theta,\sigma}\left(\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} > \frac{\sigma_0^2 C_{n-1,1-\alpha}}{\sigma^2}\right) = P\left(\chi^2_{n-1} > \frac{\sigma_0^2}{\sigma^2} C_{n-1,1-\alpha}\right) \le \alpha$.
with equality holds iff $\sigma = \sigma_0$. Hence, it follows ~~that~~ from thm (TSH 3.8.1) that our test is MP for testing the original null vs $N(\theta_1, \sigma_1)$

4. Finally, the MP level α test for testing the composite null H_0 vs an arbitrary choosen alternative (θ_1, σ_1) ~~does~~ does not depend on (θ_1, σ_1). Hence, it is UMP for testing H_0 against H_1.

KS : $\sup_t |\hat{F}_\theta(t) - F(t)|$ (Emperical cdf: $\hat{F}_\theta$)

CrM : $\int (\hat{F}_n(t) - F(t))^2 dF$.

3 tests

* Likelihood ratio test (Keener)

$X_1, \dots, X_n \overset{iid}{\sim} P_\theta(\cdot)$. We want to test $H_0: \theta \in \Theta_0$ vs $H_1: \theta \in \Theta_1$.

$$\text{LRT} : \quad \Lambda(X_1, \dots, X_n) = \frac{\sup_{\theta \in \Theta_0} P_\theta(X_1, \dots, X_n)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} P_\theta(X_1, \dots, X_n)}$$

$$-2 \log \Lambda(\underset{\sim}{X}) \xrightarrow{d} \chi^2_{d(\Theta_0 \cup \Theta_1) - d(\Theta_0)}$$

* Wald test

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$$

* Rao Score test

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n u_{\theta_0}(X_i) \xrightarrow[\theta_0]{d} N(0, I(\theta_0)) \qquad u_\theta(x_i) = \frac{\partial}{\partial \theta} \log P_\theta(x_i).$$