STAT5010. Adv. Statistical Inference

Term 1, 2022/23

Department of Statistics

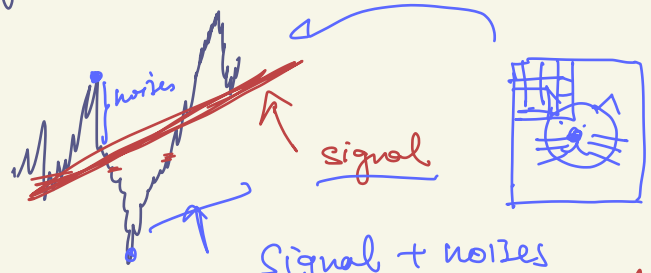
CUHK

Lecture 1

Different disciplines:

- genetic, medical, earth sciences
  |
  Cox PH ...
- Eco, finance
- Networks, images, recordings

} Data → statistics → inference / conclusion



Questions concerned:

* capture the uncertainty.
* methodology → point estimator. $\hat{\beta}\,(= (X^TX)^{-1}X^TY)$ e.g. — random.
  → testing ~ confidence region. $se(\hat{p}) = 0.0001$, $\beta = 0.01$
  $H_0: \theta = \theta_0$ vs $H_1: \theta = \theta_1$ $\quad \beta_2 = 0$
* optimality / optimal inference.

$\leq$ iid $(X_1, \dots, X_n)$ (obs: $x_1, \dots, x_n$)

→ finite-sample optimality

→ Asymptotic properties ($n \to \infty$) ✓

1. **Decision Theory** (Wald, 1939) $\quad E(X) = \int x\, dF(x)$

$(X_1, X_2, \dots, X_n) \overset{iid}{\sim} F$ (distribution) $\quad$ (density: f)

A statistical model is a family of distributions $\mathbb{P}$ indexed by a parameter θ. We denote

$$\mathbb{P} = \{P_\theta : \theta \in \Omega\}$$

$\Omega \in \mathbb{R}^k$: parameter space. $\quad \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$

Example.

obs: a sequence of coin flips $(X_1, X_2, \dots, X_n)$ (data) $\xrightarrow{\text{estimator}} \hat{\theta}(X)$ $\quad h(\theta)$

one can write $\mathbb{P} = \{\text{Bernoulli}(\theta) : \theta \in [0,1] \triangleq \Omega\}$

$P_\theta(X_i = 1) = \theta$.

estimation + testing.

A Decision Procedure

Define δ (estimator) as a map from the sample space $\mathcal{X}$ to the decision space $\mathbb{D}$.

Example. Take $\mathbb{P} = \{\text{Bernoulli}(\theta)\}$ as shown earlier,

(a) Estimating θ:

the decision space is $\mathbb{D} = [0,1]$ and one possible decision procedure can be $\delta(\underset{\sim}{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i \triangleq \bar{X}_n$.

(b) Hypothesis testing:

Accept / Reject the null hypothesis (H_0) that $\theta = 1/2$.

Correspondingly, the decision space is $\mathbb{D} = \{\text{Accept}, \text{Reject}\}$. one possible procedure is

$\delta(\underset{\sim}{X}) =$ "Reject H_0 if $\bar{x}_n > 0.5$ and accept it otherwise".

(c) A loss function is mapping $L : \Omega \times \mathbb{D} \to \mathbb{R}^+ \cup \{0\}$

$L(\theta, d)$ represents the penalty for making the decision d when θ is in fact the true parameter for the distribution generating the data.

e.g. Squared-error loss function: a) $L(\theta,d) = (\theta - d)^2$. ...

b) $L(\theta,d) = |\theta - d|$, c) ~~$L(\theta,d) = \int w(\theta - d)^k d\Lambda(\theta)$~~ ...

- Frequentist: θ is fixed yet unknown.
- Bayesian: $\Theta \sim$ distribution....

Risk function: $R(\theta, \delta) = E_\theta\left(L(\theta, \delta(X))\right)$

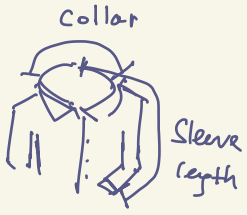
overall weighted avg of $L(\theta,\delta)$

Admissiability:

δ is inadmissible if there exists δ' such that $R(\theta,\delta') \le R(\theta,\delta)\ \forall \theta \in \Omega$ and $R(\theta',\delta') < R(\theta',\delta)$ for some $\theta' \in \Omega$.

1) Parametric models: Bernoulli (θ), $N(\mu, \sigma^2)$, $t(k)$ ...

structure.

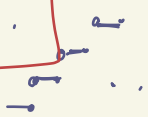


2) Nonparametric models:

$(X_1, \dots, X_n) \overset{iid}{\sim} F$ $\quad h(F)$

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x).$$

empirical cdf.



$$\sup_x \left| \hat{F}_n(x) - F_0(x) \right| \to 0, \quad \text{as } n \to \infty.$$

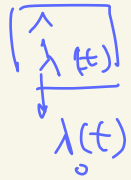
3) Semiparametric models:

Cox proportional hazard model:

baseline (not specified) — parametric

$$\lambda(t \mid Z(t)) = \lambda_0(t) \exp(\beta^T Z(t)), \qquad e^{\beta^T Z}$$

$$\Lambda(t) = \int_0^t \lambda(s)$$



where $\lambda(t \mid z) = \dfrac{f(t \mid z)\,dt}{1 - F(t \mid z)}$.

$T \sim F(f)$

$P(T \ge t \mid Z)$.

Large-Sample Theory

Convergence in probability:

$$\bar{Y}_n(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Definition: A sequence of random variables Y_n converges in probability to a random variable Y as $n \to \infty$, written as $Y_n \xrightarrow{P} Y$ if $\forall \varepsilon > 0$,

$$P(|Y_n - Y| \ge \varepsilon) \xrightarrow{n \to \infty} 0.$$

$\longrightarrow (\hat{\theta}_n \xrightarrow{P} \theta_0) \longleftarrow$

Chebyshev's inequality ...

Example: Weak Law of Large Numbers: (WLLNS)

$X_1, \dots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$, $\sigma^2 < \infty$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$E(\bar{X}_n - \mu)^2 = \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \to 0 \text{ as } n \to \infty,$$

in which case $\bar{X}_n \xrightarrow{P} \mu$ as $n \to \infty$. ||

[Proposition] If $f(\cdot)$ is continuous at c, and if $Y_n \xrightarrow{P} c$, then $f(Y_n) \xrightarrow{P} f(c)$.

· Proof · ...

Lecture 3

19 Sept 2022.

Large-Sample Theory

· Proof. Since f is continuous at c, given any $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that $|f(y) - f(c)| < \varepsilon$ whenever $|y - c| < \delta_\varepsilon$. Thus

$$P(|Y_n - c| < \delta_\varepsilon) \le P(|f(Y_n) - f(c)| < \varepsilon)$$

which implies

$$P(|f(Y_n) - f(c)| \ge \varepsilon) \le P(|Y_n - c| > \delta_\varepsilon) \to 0$$

as $Y_n \xrightarrow{P} c$ as $n \to \infty$. □

Con. in prob.

$d \to Z$

$Y_n = (-1)^n Z$

$Z \sim N(0,1)$

Convergence in distribution. (≈CLT) ← Stein's method...

Definition. A sequence of r.v.'s Y_n $(n \ge 1)$ with cdfs H_n converges in distribution (in law) to a random variable Y with cdf H if

$$H_n(y) \to H(y)$$

as $n \to \infty$ whenever H is continuous at y. For notation, we write $Y_n \Rightarrow Y$ or $Y_n \xrightarrow{\mathcal{D}/\mathcal{L}} Y$ ...

(Example). Suppose $Y_n = \frac{1}{n}$, a degenerate r.v. and that Y is always 0. Then

$$H_n(y) = P(Y_n \le y) = I\left(\frac{1}{n} \le y\right).$$

If $y > 0$, then $H_n(y) = I(\frac{1}{n} \le y) \to 1$ as $n \to \infty$ for eventually $1/n$ will be less than y. If $y \le 0$, then $H_n(y) = I(1/n \le y) = 0$ for all n and so $H_n(y) \to 0$ as $n \to \infty$.

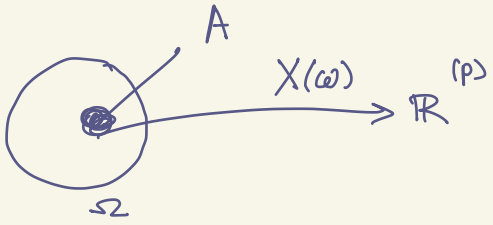
Note that $H(y) = P(Y \le y) = I(0 \le y)$. Comparisons with the limits just obtained show that $H_n(y) \to H(y)$ if $y \ne 0$ But $H_n(0) = 0 \to 0 \ne 1 = H(0)$. So, in this example, $Y_n \xrightarrow{d} Y$ but the cumulative distribution function do not converge to $H(y)$ as $y = 0$.

10 ways of looking at a r.v.

1) Prob. space $(\Omega, \mathcal{F}, \mathbb{P})$.
- Ω: sample space (ω)
- $\mathcal{F}$: σ-algebra / σ-field
- $\mathbb{P}$: prob. measure.

(Measure Theory)
↳ Billingsley (1995) Prob. and Measure.

2) Random variable.



3) Distribution ...

... ↙ ↑ Mgf.

8) Characteristic functions. ⟷ dist. fn.

$$\phi_X(t) = E(e^{itX}) = \int e^{itx}\, dF(x),$$

where $i = \sqrt{-1}$ $(i^2 = -1)$ and $e^{itx} = \cos(tx) + i\sin(tx)$.

(Example). The characteristic function of $Z \sim N(0,1)$:

$$\boxed{\phi_Z(t) = \exp\{-\tfrac{1}{2}t^2\}.}$$

$$|E(e^{itX})| \le E|e^{itX}| = E|\cos(tX) + i\sin(tX)|$$
$$= E\left(\sqrt{\cos^2(tX) + \sin^2(tX)}\right) = 1.$$

Fourier transform:

Inversion formula: $$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_X(t)\,dt.$$

$$F_X(x) - F_X(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itx} - e^{-ity}}{it}\phi_X(t)\,dt,$$

for the cont. points of $F_X(x)$ and $F_X(y)$.

[THEOREM] Uniqueness.

Let X and Y be random k-vectors. Then

(1) If $\phi_X(t) = \phi_Y(t)$ $\forall t \in \mathbb{R}^k$, then $F_X = F_Y$. ✓

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

$E(e^{tY})$ mgf. → Non-unique moments... (C&B, ch.2.x).

· Proof.

(1) For any $\underset{\sim}{a} = (a_1, \dots, a_k)^T \in \mathbb{R}^k$

and $\underset{\sim}{b} = (b_1, \dots, b_k)^T \in \mathbb{R}^k$,

and $(a,b] = (a_1, b_1] \times (a_2, b_2] \times \dots \times (a_k, b_k]$

satisfying $P_X(\text{the boundary of } (a,b]) = 0$,

$f_1(x) = \frac{1}{\sqrt{2\pi}x} \exp\{-\frac{1}{2}(\log x)^2\}$ (log-normal) $0 \le x < \infty$

$f_2(x) = f_1(x)\{1 + \sin(2\pi \log x)\}$ $0 \le x < \infty$

cdf ~ df.

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

(inversion formula for the k-dim version; Billingsley (95)).

(2) First, consider the case where $k = 1$.

Since $e^{s|x|} \le e^{sx} + e^{-sx}$,

we conclude that $|X|$ has an mgf that is finite in the neighbourhood $(-c, c)$ for some $c > 0$, and that $|X|$ has finite moments of all orders.

Observe that (Lagrange error bound: $|R_n| \le \frac{f^{(n+1)}(z)|x-a|^{n+1}}{(n+1)!}$)

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

$e^{itx}\{e^{iax}\} = \sum_{j=0}^{\infty}$

We can write ✓

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

↑

Facts:

1) If $M_X(t)$ is finite in a neighbourhood of 0, then $E(X_1^{r_1} X_2^{r_2} \dots X_k^{r_k})$ is finite for any integers $r_1, r_2, \dots, r_k \geq 0$ and $M_X(t)$ has the power series expansion

$$M_X(t) = \sum_{(r_1, r_2, \dots, r_k) \in \mathbb{Z}^+ \cup \{0\}} \frac{E(X_1^{r_1} \dots X_k^{r_k}) t_1^{r_1} \dots t_k^{r_k}}{r_1! r_2! \dots r_k!}$$

for any k-dim random vector X.

2) $$\frac{\partial^r \phi_X(t)}{\partial t_1^{r_1} \dots \partial t_k^{r_k}} = (-1)^{r/2} E(X_1^{r_1} \dots X_k^{r_k} e^{it^T X})$$

with ! $$\left. \frac{\partial^r \phi_X(t)}{\partial t_1^{r_1} \dots \partial t_k^{r_k}} \right|_{t=0} = (-1)^{r/2} E(X_1^{r_1} \dots X_k^{r_k})$$

We can write, for any $t \in \mathbb{R}$, (jth der. of ϕ wrt t.)

$$\phi_X(t+a) = \sum_{j=0}^{\infty} \frac{\phi_X^{(j)}(t)}{j!} a^j \qquad |a| < c \qquad (*)$$

(b/c: $E(e^{i(t+a)X}) = E\left(e^{itX} \sum_{j=0}^{\infty} \frac{a^j}{j!} (iX)^j\right)$

$= \sum_{j=0}^{\infty} \frac{a^j}{j!} \underbrace{E(e^{itX} \cdot (iX)^j)}_{\text{by (2)} = \phi_X^{(j)}(t)}$

The result/expansion also holds for Y. Under the assumption that $M_X = M_Y < \infty$ in a neighbourhood of 0, X and Y have the same moments of all orders. By (F2), $\phi_X^{(j)}(0) = \phi_Y^{(j)}(0)$ for all $j = 1, 2, \dots$, which and (*) with $t = 0$ imply that ϕ_X and ϕ_Y are the same on the interval $(-c, c)$.

Considering $t = c - \varepsilon$, and $-c + \varepsilon$ for an arbitrarily small $\varepsilon > 0$ in (*) shows that ϕ_X and ϕ_Y also agree on $(-2c + \varepsilon, 2c - \varepsilon)$ and hence $(-2c, 2c)$. Likewise, the same argument ϕ_X and ϕ_Y are the same on $(-3c, 3c)$ and so on. Hence, $\phi_X(t) = \phi_Y(t)$ for all t and by (i) $F_X = F_Y$.

For the case $k \geq 2$, if $F_X \neq F_Y$, then by part (i), there exists $t \in \mathbb{R}^k$ such that $\phi_X(t) \neq \phi_Y(t)$. Then $\phi_{t^TX}(1) \neq \phi_{t^TY}(1)$, which implies that $F_{t^TX} \neq F_{t^TY}$.

But $M_X = M_Y < \infty$ in a neighbourhood of $0 \in \mathbb{R}^k$ implies that $M_{t^TX} = M_{t^TY} < \infty$ in a neighbourhood of $0 \in \mathbb{R}$ and by the proved result for $k=1$, $M_{t^TX} = M_{t^TY}$. The contradiction shows that $F_X = F_Y$. □

9) Conditional Probability → ch.6 of Keener.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{when } P(B) \neq 0,$$

where by convention $P(B|A) = 0$ when $P(A) = 0$.

Suppose $E|f(X,Y)| < \infty$,

$$E(f(X,Y)) = E\Big(\underbrace{E(f(X,Y)|X)}\Big), \quad \text{(**)}$$ ← tower expectation. c.f. $E(Y|X) = \beta_0^T X$

with $E(f(X,Y)|X) \triangleq h(X)$, and

$$\rightarrow H(x) = E(f(X,Y)|X=x) = \int f(x,y)\, d\underline{Q_x(y)},$$

where $Q_x(y)$ denotes the conditional distribution of Y given $X=x$.

(**) becomes

$$E|f(X,Y)| = \int \underbrace{H(x)\, dP_X(x)}_{\text{dist. of } X} = \iint f(x,y)\, dQ_x(y)\, dP_X(x).$$

★ NFFC

Definition

The function Q is a conditional distribution of Y given X, written as $Y|X=x \sim Q_x$ if

(i) $Q_x(\cdot)$ is a prob. measure for all x,

(ii) $Q_x(B)$ is a measurable function of x for any Borel set B,

(iii) for any Borel sets A and B,

$$P(X \in A, Y \in B) = \int_A \underbrace{Q_x(B)}_{\uparrow A}\, dP_X(x)$$

10) Tail behaviour:

For a scalar random variable X with pdf f, we say X has

i) an exponential tail if

$$\lim_{a\to\infty} \frac{-\log\left(1-F(a)\right)}{ca^r} = 1 \quad \text{for some } c>0,\ r>0;$$

ii) an algebraic tail if

$$\lim_{a\to\infty} \frac{-\log\left(1-F(a)\right)}{m\log a} = 1 \quad \text{for some } m>0.$$

Examples:

Exponential $F(a) = 1-e^{-\lambda a} \Rightarrow c=\lambda,\ r=1$

Gaussian. $F(a) = \ldots \Rightarrow c=2,\ r=2.$

Student t. ← heavy-tail dist. for small d.f., $m=\nu$.

Sufficiency (data reduction / compression)

Def. A statistic $T: \mathcal{X} \mapsto T$ is a function of data

e.g. $T(X) = \frac{1}{n}\sum_{i=1}^n X_i \triangleq \bar{X}_n$.

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

(Example) Let $X_1, \ldots X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Is $\sum_{i=1}^n X_i \sim \text{Bin.}(n,\theta)$ sufficient?

$$P_\theta\left(X=x \mid T(x)=t\right) = \frac{P_\theta\left(X=x,\ T(X)=t\right)}{P_\theta\left(T(X)=t\right)}$$

$$= \frac{I\left(t=\sum_{i=1}^n x_i\right)\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}$$

$$= \frac{I\left(\sum_{i=1}^n x_i = t\right)}{\binom{n}{t}} = \frac{1}{\binom{n}{t}}$$

(Example). Let $X_1, \dots, X_n \overset{iid}{\sim} U(0,\theta)$. Show that $T(X) = \max_{1\le i\le n} X_i \triangleq X_{(n)}$ is sufficient.

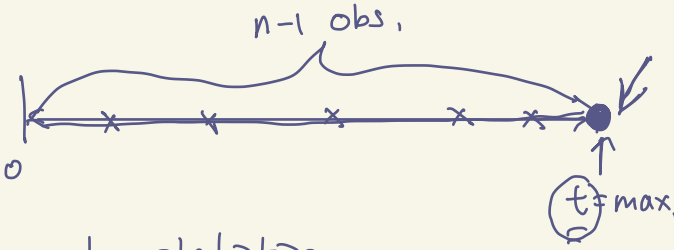
order statistic

$$F_T(t) = P\left(T \le t\right) = P\left(\max_{1\le i\le n} X_i \le t\right) = P\left(X_1 \le t, \dots, X_n \le t\right)$$

$$= \left\{P\left(X_1 \le t\right)\right\}^n = \left(\frac{t}{\theta}\right)^n \checkmark$$

$$f_T(t) = \frac{n t^{n-1}}{\theta^n} \quad \text{and} \quad P\left(X \mid T\right) \perp\!\!\!\perp \theta.$$

Intuition.



(Example) The order statistics $T(\underset{\sim}{X}) = \left(X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}\right)$ from a random sample of $X_i \overset{iid}{\sim} f$ are sufficient. Given $T(\underset{\sim}{X})$, the possible values of X are $n!$ permutations of T:

$$P\left(X_1 = X_{(1)}, \dots, X_n = X_{(n)}\right) = \frac{1}{n!}$$

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

Remark: Suppose we are given an independent source of randomness, say $U \perp\!\!\!\perp \theta$, we can generate a new dataset X' from the cond. dist. $P\left(X \mid T(X)\right)$ and define a randomized procedure:

$$\delta^*(X, U) \triangleq \delta\left(f(T(X)), U\right) = \delta(X') \overset{d}{=} \delta(X)$$

(Example) X and $Y \overset{iid}{\sim} f_\theta$, where $f_\theta(x) = \begin{cases} \theta e^{-\theta x} & , x \geq 0 \\ 0 & , \text{otherwise.} \end{cases}$

Let $U \sim \text{uniform}(0,1) \perp\!\!\!\perp (X, Y)$. Define $T = X + Y$. (T: suff.)

$\longrightarrow$ $\tilde{X} = UT$ and $\tilde{Y} = (1-U)T$. $\quad \tilde{X} + \tilde{Y} = X + Y$

$$\begin{aligned} P(T \leq t \mid Y = y) &= P(X + Y \leq t \mid Y = y) \\ &= E\left(I(X+Y \leq t) \mid Y = y\right) \\ &= \int I(x \leq t - y)\, dF_X(x) \\ &= F_X(t-y) \end{aligned}$$

We can write $P(T \leq t \mid Y) = F_X(t-Y)$, a r.v. of Y.

$$\begin{aligned} P_T(t) = P(T \leq t) &= E_Y\left(F_X(t-Y)\right) \\ &= \int_0^t \left(1 - e^{-\theta(t-y)}\right) \theta e^{-\theta y}\, dy \\ &= 1 - e^{-\theta t} - t\theta e^{-\theta t}, \end{aligned}$$

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

Note that (U, T) has the joint density

$$f_\theta(t,u) = \begin{cases} t\theta^2 e^{-\theta t} & , t \geq 0,\ u \in (0,1) \\ 0 & , \text{otherwise} \end{cases}$$

$$P_\theta\left(\begin{pmatrix} \tilde{X} \\ \tilde{Y} \end{pmatrix} \in B\right) = \iint_B \underset{\tilde{x}}{(tu}, \underset{\tilde{y}}{(1-u)t}) f_\theta(t,u)\, du\, dt$$

Hence, $\tilde{x}$ and $\tilde{Y}$ have the joint density of t

$$\frac{P_\theta\left(x+y, \frac{x}{x+y}\right)}{x+y} = \begin{cases} \theta^2 e^{-\theta(x+y)} & , x \geq 0, y \geq 0 \\ 0 & , \text{otherwise.} \end{cases}$$

which is the same joint density as (X, Y). //

[THEOREM]. Neyman-Fisher factorisation Criterion.

Suppose each $P_\theta \in \mathcal{P}$ has density $p(x;\theta)$ with respect to a common σ-finite measure μ, i.e. $dP_\theta/d\mu = p(x;\theta)$. Then, $T(X)$ is sufficient if and only if

(e.g. (μ, σ^2) for normal)

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

for some functions g_θ and h where $h(x)$ is free of θ.

·Proof. (Discrete: Refer to Casella & Berger, 2001)...

For the continuous case, T and X will not have a joint density with respect to any product measure. To begin, suppose P_θ, $\theta \in \Omega$ has density

$$\longrightarrow \quad p_\theta(x) = g_\theta(T(x))\, h(x) \quad \leftarrow \text{assumed}$$

(support of x.)

with respect to μ. Modifying h, we can assume WLOG that μ is a prob. measure equivalent to the family $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$.

$$\mu(N) = 0 \text{ if and only if } P_\theta(N) = 0 \quad \forall \theta \in \Omega.$$

Let E^* and P^* denote the expectation and probability when $X \sim \mu$; let G^* and $g(\theta)$ denote the marginal distributions for $T(X)$ when $X \sim \mu$ and $X \sim P_\theta$; and let Q be the conditional distribution for X given T when $X \sim \mu$.

To find the densities for T:

$$E_\theta(f(T)) = \int f(T(x))\, g_\theta(T(x))\, h(x)\, d\mu(x)$$

$$= E^*\left(f(T)\, g_\theta(T)\, h(X)\right)$$

(tower exp.) (cond. dist. of X given T.)

$$= \iint f(t)\, g_\theta(t)\, h(x)\, dQ_t(x)\, dG^*(t)$$

$$\triangleq \int f(t)\, g_\theta(t)\, w(t)\, dG^*(t)$$

(Leb. mea.) $\int g(x) f(x) dx$ (density)

where $w(t) = \int h(x)\, dQ_t(x)$

If $f(\cdot)$ is an indicator function, this shows that G_θ has density $g_\theta(t)w(t)$ with respect to G^*

Next, define $\tilde{Q}_t$ to have the density $h/w(t)$ with respect to Q_t so that

$$\tilde{Q}_t(B) = \int_B \frac{h(x)}{w(t)}\, dQ_t(x)$$

Then,

$$\begin{aligned} E_\theta\big(f(X,T)\big) &= E^*\Big(f(X,T)\, g_\theta(T)\, h(X)\Big) \\ &= \iint f(x,t)\, g_\theta(t)\, h(x)\, dQ_t(x)\, dG^*(t) \\ &= \iint f(x,t) \left[\frac{h(x)}{w(t)}\, dQ_t(x)\right] \left[g_\theta(t)\, w(t)\, dG^*(t)\right] \\ &= \iint f(x,t)\, d\tilde{Q}_t(x)\, dG_\theta(t) \end{aligned}$$

By the definition, this shows that $\tilde{Q}$ is a conditional distribution of X given T under P_θ. Because $\tilde{Q}$ does not depend on θ, T is sufficient.

Assume that T is suff ... (next lecture).

Lecture 4
26 Sept.

NFFC proof (cont'd)

• Proof • Mixture distributions: Given a marginal distribution G^* and a conditional distribution Q, we can define a mixture distribution $\hat{P}$ by

(cond. "given")

$$\hat{P}(B) = \int Q_t(B)\, dG^*(t) = \iint 1_B(x)\, dQ_t(x)\, dG^*(t).$$

Then for any integrable f,

$$\iint f\, d\hat{P} = \iint f(x)\, dQ_t(x)\, dG^*(t).$$

Suppose now that T is sufficient (with Q_T the conditional distribution for X given T). Let g_θ be the G^* density of T when $X \sim P_\theta$.

Ch. 1 of Keener

Existence: If $G^*(N) = 0$, $\mu(N_0) = 0$, where $N_0 = T^{-1}(N)$ ← null sets

$T(N_0) = N$

and so $G_\theta(N) = P_\theta(T \in N)$

$= P_\theta(X \in N_0) = \int_{N_0} p_\theta\, d\mu = 0$

Then $P_\theta(X \in B) = E_\theta\big(P_\theta(X \in B \mid T)\big)$

$= E_\theta\big(Q_T(B)\big)$

$= \int Q_t(B)\, g_\theta(t)\, dG^*(t)$

$= \iint 1_B(x)\, dQ_t(x)\, g_\theta(t)\, dG^*(t)$

$= \iint 1_B(x)\, g_\theta(T(x))\, dQ_t(x)\, dG^*(t)$

$= \int_B g_\theta(T(x))\, d\hat{P}(x),$

$P(X \in B) = \int_B f(x)\, dx.$

$h(x)$

which shows that P_θ has the density $g_\theta(T(\cdot))$ with respect to $\hat{P}$.

The mixture distribution $\hat{P}$ is absolutely continuous with respect to μ.

Keener (2010) Thm 1.10.

If a finite measure P is abs. cont. w.r.t. a σ-finite measure μ, then there exists a non-negative measurable function f such that

"

$$P(A) = \int_A dP = \int_A \frac{dP}{d\mu}\, d\mu \triangleq \int_A f\, d\mu$$

$$P(A) = \int_A f \, d\mu = \int f 1_A \, d\mu.$$

The function f is called the Radon-Nikodym derivative of P with respect to μ, or the density, denoted as $f = dP/d\mu$.

Suppose $\mu(N) = 0$, then $P_\theta(N) = \int Q_t(N) \, dG_\theta(t) = 0$, which implies that $G_\theta(\tilde{N}) = 0$, where $\tilde{N} = \{t : Q_t(N) > 0\}$. Because μ is equivalent to P and $G_\theta(\tilde{N}) = P_\theta(T \in \tilde{N}) = 0 \ \forall \theta \in \Omega$, $P^*(T \in \tilde{N}) = G^*(\tilde{N}) = 0$. Thus, $Q_t(N) = 0$ (a.e. G^*) and so $\hat{P}(N) = \int Q_t(N) \, dG^*(t) = 0$. Taking $\boxed{f_1 = d\hat{P}/dP^*}$, P_θ has density $g_\theta(T(x)) f_1(x)$ with respect to P^*.

μ G_θ □

(Example) Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$.

The joint distribution of $(X_1, \dots, X_n)$ is

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

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ + \frac{1}{2\sigma^2} \left(- \sum_{i=1}^n x_i^2 + 2\mu \sum_{i=1}^n x_i - n\mu^2 \right) \right\} \cdot \prod_{i=1}^n I(-\infty < x_i < \infty)$$

"$g_\theta(T)$" $\qquad$ $f_1(x)$

We have $T(X) = \left(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i \right)$, $\hat{\sigma}^2$, $\hat{\mu}$

By NFFC, we claim that T is sufficient for θ (or P_θ). //

Exponential families

The model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ forms an s-dimensional exponential family if each P_θ has the density of the form:

$$p(x; \theta) = \exp\left\{ \sum_{i=1}^s \eta_i(\theta) T_i(x) - B(\theta) \right\} h(x)$$

($T_i(x)$: suff. stat. by NFFC; $h(x)$: base measure, e.g. $I(x \in \mathbb{R})$; $\eta_i(\theta)$: natural parameter; $B(\theta)$: "standardizer")

$$B(\theta) = \log \left(\int \exp\left\{ \textstyle\sum_{i=1}^s \eta_i(\theta) T_i(x) \right\} h_1(x) \, d\mu(x) \right) \in \mathbb{R},$$

Examples include: Normal, Binomial, Poisson...

(Example). Exponential distribution $\mathcal{P} = \{\text{Exp}(\theta) : \theta > 0\}$

The densities take the form $p(x;\theta) = \theta e^{-\theta x} I(x \geq 0)$

$$= \exp\left(-\theta x + \log\theta\right) I(x \geq 0)$$

We have: $\eta_1(\theta) = -\theta$; $T_1(x) = x$, $B(\theta) = -\log\theta$, $h(x) = I(x \geq 0)$

$\widetilde{\eta}_1(\theta) = \theta$; $\widetilde{T}_1(x) = -x$, $B(\theta) = -\log\theta$, $h(x) = I(x \geq 0)$

(Example) Beta distribution $\mathcal{P} = \{\text{Beta}(\alpha,\beta) : \alpha, \beta > 0\}$; $\theta = (\alpha, \beta)$

$$p(x;\theta) = x^{\alpha-1}(1-x)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} I(0 < x \leq 1)$$

$$= \exp\left\{\underbrace{(\alpha-1)\log x}_{\eta_1(\theta)} + \underbrace{(\beta-1)\log(1-x)}_{\eta_2(\theta)} + \underbrace{\log\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{-B(\theta)}\right\}\underbrace{I(0<x<1)}_{h(x)}$$

$T = (T_1, T_2)^T$

$$= \exp\left\{\alpha\log x + \beta\log(1-x) + \log\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right\}\frac{I(0 \leq x \leq 1)}{x(1-x)}$$

Definition An exponential family is in canonical form when density has the form:

$$p(x;\eta) = \exp\left\{\sum_{i=1}^{s}\eta_i T_i(x) - A(\eta)\right\}h(x).$$

Definition The set of all valid natural parameters Ω is called the natural parameter space for each $\eta \in \Omega$, there exists a normalising constant $A(\eta)$ such that $\int p(x;\eta)dx = 1$.

Equivalently,

$$\Omega = \left\{\eta : 0 < \int \exp\left\{\sum_{i=1}^{s}\widehat{\eta}_i\widehat{T}_i(x)\right\}h(x)\,d\mu(x) < \infty\right\}.$$

For any canonical exponential family, $\mathcal{P} = \{f_\eta : \eta \in H\}$, we have $H \subseteq \Omega$, Ω is convex (see TSH, §2.7.1).

Definition If $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, then θ is unidentifiable if for two parameters $\theta_1 \neq \theta_2$, $f_{\theta_1} = f_{\theta_2}$.

Remark Two cases when the superficial dimension of an s-exponential family $\mathcal{P}$ can be reduced:

1) $X \sim \text{Exp}(\eta_1, \eta_2)$ with density

$$p(x;\eta_1,\eta_2) = \exp\{-\eta_1 x - \eta_2 x + \log(\eta_1+\eta_2)\} \mathbb{I}(x \geq 0), \quad (*)$$

where $T_1(X) = T_2(X) = X$ (are linearly dependant). We can actually combine (η_1, η_2) into $\eta_1+\eta_2$ and rewrite $(*)$ as

$$p(x;\eta_1,\eta_2) = \exp\{-\underbrace{(\eta_1+\eta_2)}_{\eta} x + \log \underbrace{(\eta_1+\eta_2)}_{\eta}\} \mathbb{I}(x \geq 0)$$

The $T_i(X)$'s satisfy an affine equality (linearly dep.) constraint for all $x \in \mathcal{X}$ which will result in unidentifiability, i.e.

$$p(x;\eta_1+c, \eta_2-c) = p(x;\eta_1,\eta_2) \text{ for all } c < \eta_2$$

2. The η_i's satisfy an affine equality constraint for all $\eta \in H$

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

$$p(x;\eta) \propto \exp(\eta_1 x + \eta_2 x^2) = \exp\left(\eta_1 (x - x^2) + x^2\right).$$

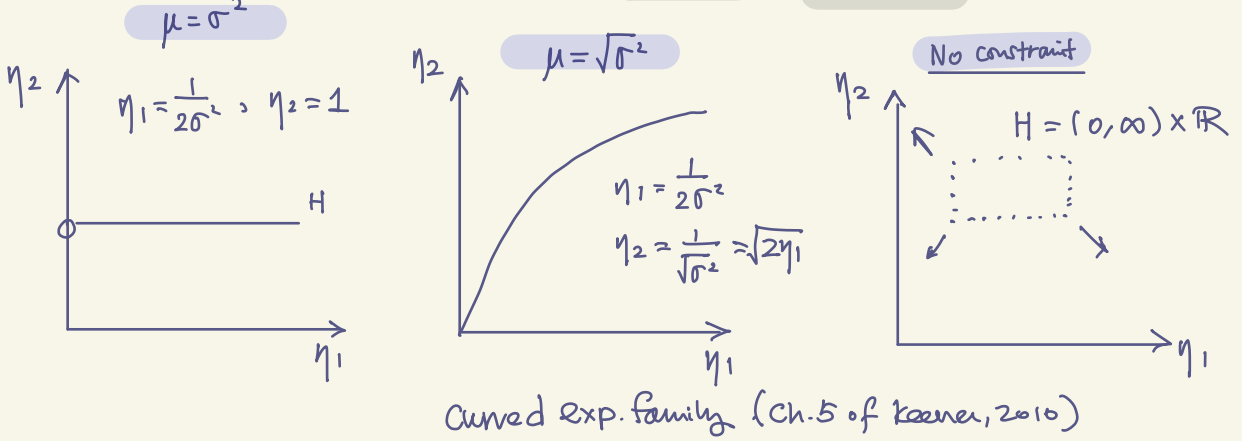
Definition A canonical exponential family $\mathcal{P} = \{P_\eta : \eta \in H\}$ is **minimal** if

(a) No affine T_i's: $\sum_{i=1}^s \lambda_i T_i(x) = \lambda_0 \ \forall x \in \mathcal{X} \Rightarrow \lambda_i = 0 \ \forall i \in \{0, \dots s\}$

(b) No affine η_i's: $\sum_{i=1}^s \lambda_i \eta_i = \lambda_0 \ \forall \eta \in H \Rightarrow \lambda_i = 0 \ \forall i \in \{0, \dots s\}$

Definition Suppose $\mathcal{P} = \{P_\eta : \eta \in H\}$ is an s-dimensional minimal exponential family. If H contains an open rectangle (s-dim), then $\mathcal{P}$ is called full-rank. otherwise, $\mathcal{P}$ is curved.

(Example) Consider $N(\mu, \sigma^2)$, where $\eta_1 = \frac{1}{2\sigma^2}$, $\eta_2 = \frac{\mu}{\sigma^2}$, $T_1(X) = -X^2$, $T_2(X) = X$



Curved exp. family (Ch.5 of Keener, 2010)

Remarks: Properties of Exponential Families.

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

then, by NFFC, $T = \left(\sum_{j=1}^{n} T_1(X_j), \dots, \sum_{j=1}^{n} T_s(X_j)\right)$ is sufficient. The exponential family data is highly compressible.

2) If f is integrable, and $\eta \in \Omega$, then

$$G(f,\eta) = \int f(x) \exp\left\{\sum_{i=1}^{s} \eta_i T_i(x)\right\} h(x)\, d\mu(x)$$

is infinitely differentiable with respect to η and the derivatives can be obtained by differentiating it under the integral sign.

3) Moments of T_i's...

Take $f(x) \equiv 1$, then

$$G(1,\eta) = \int \exp\left\{\sum_{i=1}^{s} \eta_i T_i(X)\right\} h(x)\, d\mu(x) = \exp\{A(\eta)\},$$

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

$$\frac{\partial A(\eta)}{\partial \eta_i} = \int T_i(x) \exp\left\{\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right\} h(x)\, d\mu(x)$$

$$= E_\eta(T_i(X))$$

$$\frac{\partial^2 A(\eta)}{\partial \eta_i \partial \eta_j} = \int T_i(x)\left(T_j(x) - \frac{\partial A(\eta)}{\partial \eta_j}\right) \exp\left\{\sum_{\ell=1}^{s} \eta_\ell T_\ell(x) - A(\eta)\right\} h(x)\, \partial\mu(x)$$

$$= E_\eta\left(T_i(X) T_j(X)\right) - E_\eta(T_i(X)) E_\eta(T_j(X))$$

$$= \mathrm{Cov}_\eta\left(T_i(X), T_j(X)\right).$$

Minimal Sufficiency

Definition A sufficient statistic T is minimal if for every sufficient statistic T' and for every $x, y \in \mathcal{X}$, $T(x) = T(y)$ when $T'(x) = T'(y)$. In other words, T is a function of T', i.e. there exists a function f such that $T(x) = f(T'(x))$ for any $x \in \mathcal{X}$.

$X_1, \dots, X_n$ (raw obs.) $\xrightarrow{\text{sufficient}}$ $T_1(X), T_2(X)$ $\xrightarrow{\text{min. suff.}}$ $T(X)$

say $T(X) = T_1(X) + T_2(X)$,

[THEOREM]. Let $\{p(x;\theta) : \theta \in \Omega\}$ be a family of densities with respect to some measure μ (Lebesgue measure for continuous distribution dx; counting measure for discrete distribution $(\chi_i - \chi_{i-1})$). Suppose that there exists a statistic T such that, for every $x, y \in \mathcal{X}$,

$$\rightarrow \boxed{p(x;\theta) = C_{x,y}\, p(y;\theta)} \Leftrightarrow \underline{T(x) = T(y)}$$

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

· Proof. We first prove that T is sufficient and then T is minimal.

⇒) Sufficiency for T:

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

For each $t \in T(\mathcal{X})$, we consider the preimage $\underline{A_t} = \{x : T(x) = t\}$. (select a rep χ_t from A_t.)

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

By the definition of A_t, we can see that $\underline{T(y) = T(\chi_{T(y)})}$

Because of the assumption/condition imposed in the statement,

$$\Rightarrow \quad p(y;\theta) = \underbrace{C_{y,\chi_{T(y)}}}\, p(\underline{\chi_{T(y)}};\theta) \triangleq \underline{h(y)\, g_\theta(T(y))},$$

which implies that T is sufficient as a result of the NFFC.

⇐) Minimality of T:

Consider another sufficient statistic, say T' (\$T\prime\$). By NFFC,

$$p(x;\theta) = \tilde{g}_\theta(T'(x))\,\tilde{h}(x)$$

Take any x and y such that $\underline{T'(x) = T'(y)}$, then

$$p(x;\theta) = \tilde{g}_\theta(T'(x))\,\tilde{h}(x)$$

$$= \underbrace{\tilde{g}_\theta(T'(y))\,\tilde{h}(y)} \cdot \underbrace{\frac{\tilde{h}(x)}{\tilde{h}(y)}} \triangleq p(y;\theta)\, C_{x,y}$$

- Hence, $T(x) = T(y)$ by the assumption of the minimal sufficient statistic theorem. So $T'(x) = T'(y)$ implies $T(x) = T(y)$ for any sufficient statistic T' and any x and y. Hence, T is a minimal sufficient statistic. □

(Example) Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with both μ and σ^2 unknown. Let $\underset{\sim}{x}$ and $\underset{\sim}{y}$ be any two sample points, and let $(\bar{x}_n, s_x^2)$ and $(\bar{y}_n, s_y^2)$ be the sample means and variances corresponding to the $\underset{\sim}{x}$ and $\underset{\sim}{y}$ samples, respectively.

$$\frac{f(\underset{\sim}{x}; \mu, \sigma^2)}{f(\underset{\sim}{y}; \mu, \sigma^2)} = \frac{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{n(\bar{x}_n - \mu)^2 + (n-1)s_x^2}{2\sigma^2}\right\}}{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{n(\bar{y}_n - \mu)^2 + (n-1)s_y^2}{2\sigma^2}\right\}}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left\{-n(\bar{x}_n^2 - \bar{y}_n^2) + 2n\mu(\bar{x}_n - \bar{y}_n) - (n-1)(s_x^2 - s_y^2)\right\}\right\}$$

The ratio will be constant as a function of $\theta = (\mu, \sigma^2)$ iff $\bar{x}_n = \bar{y}_n$ and $s_x^2 = s_y^2$. Thus, by the "theorem" $(\bar{x}_n, s^2)$ is a minimum sufficient statistic for θ, where $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$.

(Example) (Curved exponential family)

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\sigma, \sigma^2)$, where $\sigma > 0$. Then $\theta = \sigma$

$$\frac{p(\underset{\sim}{x}; \theta)}{p(\underset{\sim}{y}; \theta)} = \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2\right) + \frac{1}{\sigma}\left(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i\right)\right\}.$$

This ratio will be constant as a function of $\theta = \sigma$ iff $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ and $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$. Thus, by the above theorem, $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is minimal sufficient.

Remark: If $p(x; \theta) = C_{x,y}\, p(y; \theta)$ and x and y must be supported by the sample θ. (Support of X: $\{x \in \mathcal{X} : p(x; \theta) > 0\}$. Otherwise, the "constant" $C_{x,y}$ will be θ-dependent.

(Example) Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$ and $T(X) = \max_{1 \le i \le n} X_i = X_{(n)}$

In that case, for $\underset{\sim}{x} = (x_1, \dots, x_n)$ such that $x_i > 0$, for $i = 1, \dots, n$,

$$p(\underset{\sim}{x}; \theta) = \prod_{i=1}^{n} \frac{1}{\theta} I(x_i \le \theta) = \frac{1}{\theta^n} I\Big(\underbrace{\max_{1 \le i \le n} x_i}_{x_{(n)}} < \theta\Big)$$

If $T(x) = T(y)$, then $p(\underset{\sim}{x}; \theta) = 1 \times p(\underset{\sim}{y}; \theta)$. The ratio between the two distributions does not depend on θ, so T is sufficient.

Conversely, if $\underset{\sim}{x}, \underset{\sim}{y} > 0$, i.e. $x_i, y_i > 0$ for $i = 1, \dots, n$, are supported by the same θ's, then

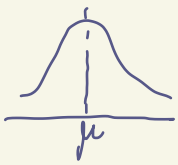
$$\{\theta \text{ supporting } x\} = (T(x), \infty) = (T(y), \infty) = \{\theta \text{ supporting } y\}$$

Therefore, $T(x) = T(y)$ and $I_{X_{(n)}}$ is a min. suffi statistic. //.

Mid term test

* 5-6 questions (shorter than long asg problems)
* Not open-book/-notes exam.; have 1 double-sided, A4 size info sheet. → final
* 2hr exam...

Lecture 5 30 OCTOBER 2022,



Ancillarity and Completeness.

(Example) Consider $X_1, \dots, X_n \overset{iid}{\sim} \text{Cauchy Loc}(\theta)$, whose distribution is given by

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

(see TPE. 1.5 : we can see that $(X_{(1)}, X_{(2)} \dots, X_{(n)})$ is minimal sufficient)

(This is also true for the double exponential location model: $p(x;\theta) \propto \exp(|x-\theta|)$. ↓ ↘ Kou (2002) MS. jump diffusion.

Question: What determines the level of compressibility?

[**Definition** A statistic A is ancillary for $X \sim P_\theta \in \mathcal{P}$ if the distribution of $A(X)$ does not depend on θ.]

(Example) Consider again $X_1 \dots X_n \overset{iid}{\sim} \text{Cauchy loc}(\theta)$.

Define $A(X) = \underbrace{X_{(n)}}_{(Z_{(n)}+\theta)} - \underbrace{X_{(1)}}_{(Z_{(1)}+\theta)}$, we can see that $A(X)$ is ancillary w.r.t. θ.

Let $X_i = Z_i + \theta$, for $Z_i \overset{iid}{\sim} \text{Cauchy loc}(0)$

So $X_{(i)} = Z_{(i)} + \theta$, $A(X) = A(Z)$

Hence $A(X)$ does not depend on θ. ||

↓

! The statistic that we make use of should include as little ancillary information as possible.

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

Definition 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 θ, then $f(T(X)) = 0$ with prob. 1 for all θ.

Completeness formalises our ideal concept of optimal data reduction. Minimal sufficiency is our achievable notion of optimal data reduction.

Properties:

1) If T is complete sufficient, then T is minimal sufficient. This is known as Bahadur's theorem.

2) Complete sufficient statistics yield optimal unbiased estimator. (More results later)... ↑ $E_\theta(\cdot)$

(Example)

Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, $\theta \in (0,1)$. Then we know that $T(X) = \sum_{i=1}^n X_i$ is sufficient. Suppose $E_\theta(f(T(X))) = 0$ for all $\theta \in (0,1)$.

Observe:

$$E_\theta(f(T(X))) = \sum_{j=0}^{n} f(j) \binom{n}{j} \theta^j (1-\theta)^{n-j} = 0 \quad \forall \theta \in (0,1) \quad (*)$$

Dividing through by θ^n and let $\beta = \frac{\theta}{1-\theta}$, we have

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

If f are non-zero, then the LHS is a polynomial of degree at most n. However, an nth degree polynomial can have at most n roots. Hence, it is impossible for the LHS to equal 0 for EVERY $\beta > 0$ unless $f = 0$. We can conclude that T is complete.

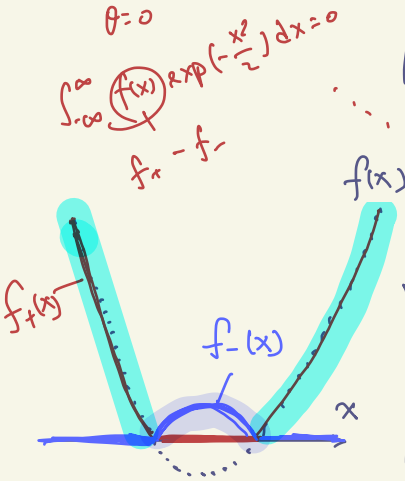
(Example) Let $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$, with $\theta \in \mathbb{R}$ unknown and a known $\sigma^2 > 0$. Is $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ complete for this model?

(min. suff.)

To simplify the notation, we consider a special case where $n = 1$ and $\sigma = 1$, so that $T(X) = X \sim N(\theta, 1)$.

Suppose (**)

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



Multiplying both sides by $\sqrt{2\pi}\, e^{\theta^2/2}$, we can rewrite (**) as

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

We decompose f into its positive and negative parts as

$$f(x) = f_+(x) - f_-(x),$$

where $f_+(x) = \max(f(x), 0)$ and $f_-(x) = \max(-f(x), 0)$.

Observe that $f_+(x) \geq 0$ and $f_-(x) \geq 0$ for all $x \in \mathbb{R}$ and $f_+(x) = f_-(x)$ if and only if $f_+(x) = f_-(x) = 0$.

If $f(x) \geq 0$ a.e. or $f(x) \leq 0$ a.e. then (***) implies that $f(x) = 0$ a.e. because setting $\theta = 0$ gives us an integral of a non-neg (or a non-pos) function keeping zero.

We may write

$M_X(t) = E(e^{tX}) = \int e^{tx} f(x) dx$

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

$M_{f_+}(t) = M_{f_-}(t)$

Notice that in (#),

$$\frac{f_+(x) \exp(-\frac{x^2}{2})}{\int_{-\infty}^{\infty} f_+(x) \exp(-\frac{x^2}{2}) dx}$$

defines a legitimate

probability density. Notice also that [redacted] is the mgf of this density. Similarly, the RHS is the mgf of the density $f_-(x)\exp(-x^2/2) \,/\, \int_{-\infty}^{\infty} f_-(x)\, e^{-x^2/2}\, dx$.

It implies that $f_+(x) = f_-(x)$ a.e. or, in other words, $f_+(x) = f_-(x) = 0$ a.e., i.e. $f(x) = 0$ a.s. Hence T is sufficient and complete. ||

(Example) / Exercise.

If $X \sim p(x;\theta) \propto h(x)\exp(\theta x)$, then the statistic $T(X) = X$ is complete.

Key steps:

1). Suppose $\int f(x)h(x)\exp(\theta x)\,dx = 0 \quad \forall \theta$.
2) Decompose $f(x) = f_+(x) - f_-(x)$ with $f_+ \cdot f_- \geq 0$
3) f_+ and f_- can be viewed as legitimate densities
4) The decomposition shows that the mgf's of f_+ and f_- coincide and hence $f(T) \equiv 0$.

[THEOREM]. $(T_1, \ldots, T_s)$ is complete for any θ-dimensional full rank exponential family. (TSH)

[THEOREM] (Basu's Theorem). If T is complete and sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ and V is ancillary, then $T(X) \perp\!\!\!\perp V(X)$.

· Proof. Define $g_A(t) = P_\theta(V \in A \mid T = t)$

or $g_A(T) = P_\theta(V \in A \mid T)$

and $p_A = P_\theta(V \in A)$.

$S^2 \perp\!\!\!\perp \theta$

$\bar{X}$

By sufficiency and ancillarity, neither p_A nor $g_A(t)$ depends on θ. By smoothing / tower expectation, we can write

$$p_A = P_\theta(V \in A) = E_\theta\left(P_\theta(V \in A \mid T)\right) = E_\theta\left(g_A(T)\right).$$

By completeness,

$$E_\theta\left(g_A(T) - p_A\right) \overset{\Delta}{=} E_\theta\left(f(T)\right) = 0$$

Hence $f(T) = g_A(T) - p_A = 0$, i.e. $g_A(T) = p_A$ a.s. P

Again, by smoothing / tower expectation,

$$P_\theta(T \in B, V \in A) = E_\theta\left(1_B(T)\, 1_A(V)\right) \qquad 1_B(T) = I(T \in B).$$

$$= E_\theta\left(E_\theta\left(1_B(T)\, 1_A(V) \mid T\right)\right)$$

$$= E_\theta\left(1_B(T)\, E_\theta\left(1_A(V) \mid T\right)\right)$$

$$= E_\theta\left(1_B(T)\, g_A(T)\right)$$

$$= E_\theta\left(1_B(T) \cdot p_A\right)$$

$$= p_A \cdot E_\theta(1_B(T))$$

$$= P_\theta(V \in A)\, P_\theta(T \in B).$$

Hence, T and V are independent as A and B are arbitrary Borel sets. ∎

(Example) Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Then $\bar{X}_n \perp\!\!\!\perp n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ and $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$.

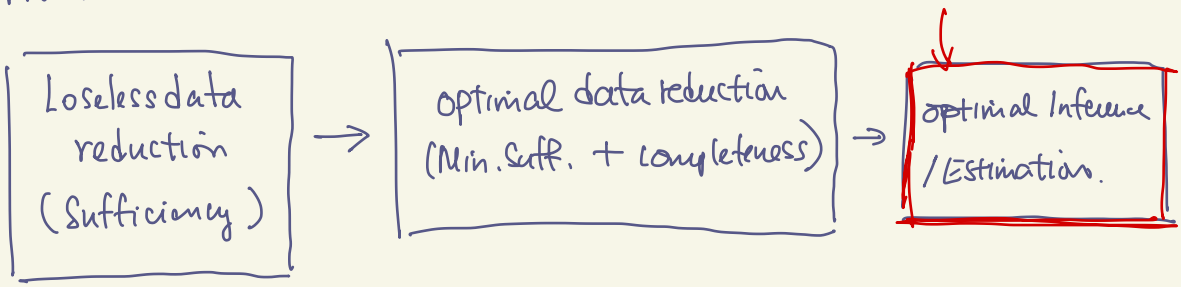
For any $\sigma > 0$, we consider the submodel $P_\sigma = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}\}$

In each submodel, $\bar{X}_n$ is complete and sufficient, and $n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is ancillary. (Let $X_i = Z_i + \mu$, then $X_i - \bar{X}_n = Z_i - \bar{Z}_n$, which does not depend on μ.

By Basu's theorem, $\bar{X}_n \perp\!\!\!\perp \sum_{i=1}^n (X_i - \bar{X}_n)^2$ under $N(\mu, \sigma^2)$ for any μ. Since σ is arbitrary, we can conclude that

$\bar{X}_n \perp\!\!\!\perp n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ holds for the full model $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R},\ \sigma^2 > 0\}$. ∎

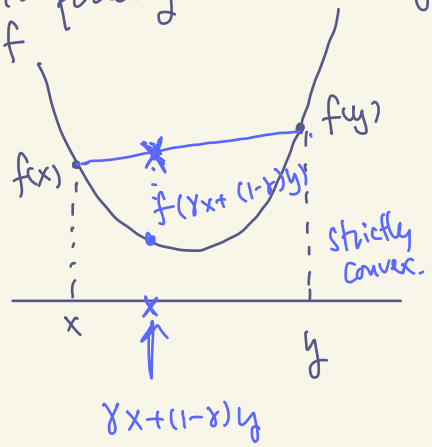
- From data reduction to optimal inference.

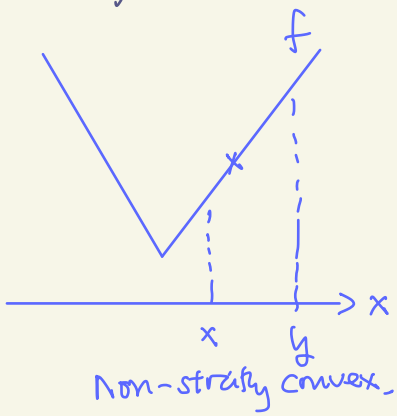


Definition A function f is a convex function if $x \neq y$ and $\gamma \in (0,1)$

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

The function f is said to be strictly convex if the above inequality holds strictly (i.e. '<')





(Examples) a) For any $\theta \in \Omega$, the function $f(d) = (d-\theta)^2$ is strictly convex on $\mathbb{R}$.

b) For any $\theta \in \Omega$, the function $\bar{f}(d) = |d-\theta|$ is convex but not strictly convex.

[THEOREM] (Jensen's inequality). For a convex f, and $E(X)$ exists, then

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

- If f is strictly convex, then the above inequality holds strictly unless $X = E(X)$ with probability 1.

[THEOREM] (Rao-Blackwell Theorem).

Suppose T is sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$, 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, in particular, $L(\theta, \cdot)$ is convex, then

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

for $\boxed{\eta = E(\delta(X) \mid T(X))}$.

If $L(\theta, \cdot)$ is strictly convex, then $R(\theta, \eta) < R(\theta, \delta)$ for any θ unless $\eta(T(X)) = \delta(X)$ with prob. 1.

(Example). If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, where $\theta \in (0,1)$. Consider $L(\theta, d) = (\theta - d)^2$. Suppose we have a naive estimator $\delta(X) = X_1$. We know that $T(X) = \bar{X}_n$ is sufficient. So, we can apply the Rao-Blackwell Theorem to improve our estimator δ:

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

Recall that

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

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

Next lecture: UMVUE (uniform min var unbiased estimator).

Lecture 6

10/ october /2022.

Unbiased estimation

Definition An estimator is said to be unbiased if $E_\theta(\delta(X)) = g(\theta)$ for all θ.

? uniformly best estimator (challenging)

But we can find an unbiased estimator that gives minimum risk uniformly, i.e. $R(\theta, \delta) \leq R(\theta, \delta')$ for all $\theta \in \Omega$ and any other unbiased estimator.

Such an estimator is called a uniformly minimum risk unbiased estimator (UMRUE).

In particular, if we adapt $L(\theta, d) = (\theta - d)^2$, an UMRUE will become a UMVUE. because

variance ↑ , Bias ↓

$$\underbrace{E_\theta\left(\{g(\theta) - \delta(X)\}^2\right)}_{\text{MSE.}} = \underbrace{\{E_\theta(\delta(X)) - g(\theta)\}^2}_{\text{Bias}^2} + \underbrace{E_\theta\left(\{\delta(X) - E_\theta(\delta(X))\}^2\right)}_{\text{Variance of } \delta(X)}$$

and if $\delta(X)$ is unbiased, then the MSE will reduced to

$$E_\theta\left(\{g(\theta) - \delta(X)\}^2\right) = E_\theta\left(\{\delta(X) - E_\theta(\delta(X))\}^2\right)$$

Definition If an unbiased estimator exists, then $g(\cdot)$ is called U-estimable.

(Example) Suppose $X \sim$ uniform $(0, \theta)$. Then $\delta(X)$ is unbiased if

$$\int_0^\theta \frac{\delta(x)}{\theta}\, dx = g(\theta), \quad \forall \theta > 0$$

or if $$\int_0^\theta \delta(x)\, dx = \theta g(\theta), \quad \forall \theta > 0. \qquad (*)$$

So, g cannot be U-estimable unless $\theta g(\theta) \to 0$ as $\theta \downarrow 0$. If g' exists, then differentiating $(*)$, and also by the fundamental theorem of calculus, we have

$$\delta(x) = \frac{\partial}{\partial x}\{x g(x)\} = g(x) + x g'(x).$$

For example, say $g(\theta) = \theta$, then $\delta(X) = X + X \cdot 1 = \underline{2X}$.

(Example) Suppose $X \sim \text{Binomial}(n, \theta)$. If $g(\theta) = \sin\theta$, then $\delta(X)$ will be unbiased if

$$\sum_{k=0}^{n} \delta(k) \binom{n}{k} \theta^k (1-\theta)^{n-k} = \sin\theta \qquad \forall\, \theta \in (0,1). \quad (**)$$

The LHS (**) is a polynomial in θ with degree at most n. The Sine function cannot be written as a polynomial of degree n. Therefore, $\sin\theta$ is not U-estimable. ||

Definition An unbiased estimator δ is uniformly minimum variance unbiased (UMVU) if $\text{Var}_\theta(\delta) \le \text{Var}_\theta(\delta')$ $\forall\, \theta \in \Omega$ and for any other competing unbiased estimator δ'.

[THEOREM]. (Lehmann-Scheffé Theorem). If T is a complete and sufficient statistic, and $E_\theta(h(T(X))) = g(\theta)$ [i.e. $h(T(X))$ is unbiased for $g(\theta)$], then $h(T(X))$ is

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

(b) an UMRUE under any convex loss function,

(c) the unique UMRUE (hence UMVUE), up to a-P_θ-null set, under any strictly convex loss function.

. Proof. (a) Suppose $E_\theta\left(\tilde{h}(T(X))\right) = g(\theta)$, then

$$E_\theta\left(\tilde{h}(T(X)) - h(T(X))\right) = 0 \qquad \forall\, \theta \in \Omega.$$

Thus, $\tilde{h}(T(X)) = h(T(X))$ almost surely for all $\theta \in \Omega$ by Completeness.

(b) Consider any unbiased estimator $\delta(X)$ and let $\tilde{h}(T(X)) = E_\theta(\delta(X) \mid T(X))$. Then

$E_\theta\left(\tilde{h}(T(X))\right) = E_\theta\left(E_\theta(\delta(X)\mid T(X))\right) = E_\theta(\delta(X)) = g(\theta)$ (by "smoothing"). By (a), $\tilde{h}(T(X)) = h(T(X))$, then $R(\theta, \tilde{h}(T(X))) = R(\theta, h(T(X)))$. By Rao-Blackwell Theorem, we have $R(\theta, \tilde{h}(T(X))) \le R(\theta, \delta)$ for all $\theta \in \Omega$ if the loss function is convex. It follows that

$$R(\theta, h(T(X))) \le R(\theta, \delta), \quad \forall\, \theta \in \Omega.$$

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

(c) If the loss function is strictly convex, $R(\theta, h(T(X))) < R(\theta, \delta)$ unless $\delta(X) \overset{a.s.}{=} h(T(X))$. Hence, $h(T(X))$ is the unique UMRUE (resp. UMVUE if the loss function is the squared error loss function).

Possible solution a: Rao-Blackwellization

(Example) (Rao-Blackwellization). Suppose $X_1, \dots, X_n \overset{iid}{\sim}$ Bernoulli(θ).

We know that $T(X) = \sum_{i=1}^n X_i$ is a complete, sufficient statistic. We also know that $n^{-1} T(X)$ is an unbiased estimator for θ, i.e.

$E_\theta(T(X)/n) = n^{-1} \sum_{i=1}^n E_\theta(X_i) = \theta$.

Therefore, $n^{-1} T(X)$ is an UMRUE for θ under any convex loss function

Suppose that, instead, we are interested in estimating θ^2. Let's observe

$$\delta(X) = I(X_1 = X_2 = 1) = X_1 X_2$$

$$E_\theta(\delta(X)) = E_\theta(X_1 \cdot X_2) = \{E_\theta(X_1)\}^2 = \theta^2$$

$$\longrightarrow E_\theta(\delta(X) \mid T(X) = t) = P_\theta(X_1 = X_2 = 1 \mid T(X) = t)$$

$$= \frac{P_\theta(X_1 = X_2 = 1, \sum_{i=3}^n X_i = t-2)}{P_\theta(T(X) = t)}$$

$$= \frac{\theta^2 \binom{n-2}{t-2} \theta^{t-2} (1-\theta)^{n-t} I(t \geq 2)}{\binom{n}{t} \theta^t (1-\theta)^{n-t}}$$

$$= \frac{t(t-1) I(t \geq 2)}{n(n-1)}.$$

Hence, we conclude that a UMRUE for θ^2 is $\left[\frac{T(X)(T(X)-1)}{n(n-1)}\right]$. //

(Example) Suppose $X_1, \dots, X_n \overset{iid}{\sim} U(0, \theta)$. In this case $T(X) = X_{(n)}$ is a complete and sufficient statistic, and $\delta(X) = 2X$ is an unbiased estimator for θ, i.e. $E_\theta(2X) = 2 \cdot (\frac{\theta}{2}) = \theta$.

Given $X_{(n)}$, X_1 is equal to $X_{(n)}$ ($= X_{(n)}$) with prob. $1/n$ and follows uniform $(0, X_{(n)})$ with prob. $1 - 1/n$. Hence

$$P_\theta(X_1 = x_1 \mid T(X)) = \frac{I(T(X) = x_1)}{n} + \underbrace{\frac{I(0 < x_1 < T(X))}{T(X)}}_{\text{uniform.}} (1 - \frac{1}{n}).$$

To find the UMVUE, we calculate

$$\begin{aligned} E_\theta(\delta(X)\,|\,T(X)) &= 2E_\theta(X_1\,|\,T(X)) \\ &= 2\left\{\frac{1}{n}T(X) + \left(1-\frac{1}{n}\right)\int_0^{T(X)} \frac{x_1\,dx_1}{T(X)}\right\} \\ &= 2\left\{\frac{T(X)}{n} + \left(1-\frac{1}{n}\right)\frac{T(X)}{2}\right\} \\ &= \left(\frac{n+1}{n}\right)T(X), \end{aligned}$$

which gives the desired result. ||

Possible soln B: Solve for δ directly

(Example) Let $X \sim \text{Poisson}\,|\,\theta)$. X is a complete and sufficient statistic for θ. X is also unbiased and therefore UMVUE for θ.

Suppose we are interested in estimating $g(\theta) = e^{-a\theta}$ for $a \in \mathbb{R}$. We need to find an estimator δ such that $E_\theta(\delta(X)) = g(\theta)$ for all $\theta > 0$. Under this model, we have

$$E_\theta(\delta(X)) = \sum_{x=0}^{\infty} \delta(x)\cdot\frac{e^{-\theta}\theta^x}{x!} = e^{-a\theta}, \quad \forall\,\theta > 0$$

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

$$\Rightarrow \quad \delta(X) = (1-a)^X \text{ is the UMVUE for } g(\theta).$$

Note that this estimator is not ideal, in the sense that if $a=2$, then the estimator $\delta(X) = (-1)^X$ will change signs according to the "evenness" of X.

The estimator is hence inadmissible when $a > 1$ and will be dominated by $\max(\delta(X), 0)$. ||

Possible soln C: Guess

(Example) Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Consider the case where $\theta = (\mu, \sigma^2)$ is unknown.

(a) The UMVUE for $\sigma^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = S^2$. [Trg].

(b) How about the UMVUE for σ?

(c) What is the UMVUE for μ^2?

(b) Observe:

$$X_i - \bar{X}_n \sim N\left(0, \frac{n-1}{n}\sigma^2\right) \quad \checkmark$$

$$E\left(|X_i - \bar{X}_n|\right) = \sigma\sqrt{\frac{2}{\pi}} \times \sqrt{\frac{n-1}{n}} \quad \text{(Verify)}$$

This implies: $\delta' = \dfrac{\sqrt{\pi n}}{\sqrt{2(n-1)}} |X_i - \bar{X}_n| \;\Big|\; S.$

is unbiased for σ. [Rao-Blackwellisation may not work so easily here..]

Instead, we can observe another fact that

$$S_*^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 = (n-1)S^2 \sim \sigma^2 \chi^2_{n-1}$$

Hence,

$$E(S_*) = \sigma E(\chi_{n-1}) \Rightarrow \sigma = \frac{E(S_*)}{E(\chi_{n-1})},$$

meaning that $\dfrac{S_*}{E(\chi_{n-1})}$ is unbiased for σ and hence UMVUE.

(c) Take the expectation of the UMVUE for μ and squaring it,

we obtain $E(\bar{X}_n^2) = \mu^2 + \frac{\sigma^2}{n}$

↖ and so, $\delta_n(X) = \bar{X}_n^2 - \dfrac{S_*^2}{n(n-1)}$ is UMVUE for μ^2.

negative ...

$\max(0, \delta_n(X))$ ‖

Fisher Information

Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Define $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$, $\dfrac{S^2}{n-1}$ is the UMVUE for σ^2 ↙ unbiased

MLE for σ^2 is $\dfrac{S^2}{n}$ ↑

} neither MLE nor UMVUE is admissible

Shrunk estimator: $\dfrac{S^2}{n+1}$. Lower MSE.

⋮

← [James-Stein estimator...]

Rabinowitz(?) 2000 Sinica.
Semiparametric efficiency

Suppose we have δ_1 and δ_2 as UMVUEs for $g_1(\theta)$ and $g_2(\theta)$, respectively. Is $\delta_1 + \delta_2$ an UMVUE for $g_1(\theta) + g_2(\theta)$?

[THEOREM] (TPE 2.1.7) (Characterization of UMVUEs).

Let $\Delta = \{\delta : E_\theta(\delta^2) < \infty\}$. Then $\delta_0 \in \Delta$ is UMVU for $g(\theta) = E(\delta_0)$ if and only if $E(\delta(\theta) U) = 0$ for every $U \in \mathcal{U} = \{E(U) = 0\}$, unbiased estimators of 0.

· Proof. If δ_0 is an UMVUE, let's consider $\delta_1 = \delta_0 + \lambda U$ for a fixed $\lambda \in \mathbb{R}$ and $U \in \mathcal{U}$. Since δ_0 has minimal variance,

$$Var_\theta(\delta_1) = Var_\theta(\delta_0) + \lambda^2 Var_\theta(U) + 2\lambda Cov_\theta(\delta_0, U) \geq Var_\theta(\delta_0),$$

or equivalently,

$$\lambda^2 Var_\theta(U) + 2\lambda Cov(\delta_0, U) \geq 0.$$

Consider the quadratic form $q(\lambda) = \lambda^2 Var_\theta(U) + 2\lambda Cov(\delta_0, U)$. The form q has the roots

$$\lambda = 0 \quad \text{and} \quad \lambda = \frac{-2Cov_\theta(\delta_0, U)}{Var_\theta(U)}.$$

If the roots are distinct, the form must be negative at some point, which violates the inequality above. Hence $-2Cov_\theta(\delta_0, U)/Var_\theta(U) = 0$ in which case $E_\theta(U\delta_0) = Cov(\delta_0 U) = 0$.

We assume that $E_\theta(U\delta_0) = 0$ for all $U \in \mathcal{U}$, and consider any δ unbiased for $g(\theta)$.

It follows that $\delta - \delta_0 \in \mathcal{U}$. So $E_\theta\left(\delta_0(\delta - \delta_0)\right) = 0$. This implies that $E_\theta(\delta_0 \delta) = E_\theta(\delta_0^2)$ and subtracting $\{E_\theta(\delta_0)\}^2 (= E_\theta(\delta_0) E_\theta(\delta))$ on both sides, we obtain

$$Var_\theta(\delta_0) = Cov_\theta(\delta_0, \delta) \overset{CSI}{\leq} \sqrt{Var_\theta(\delta_0) Var_\theta(\delta)}$$

Hence $Var_\theta(\delta_0) \leq Var_\theta(\delta)$ for any arbitrary unbiased estimator δ. Hence δ_0 is a UMVUE for $g(\theta)$.

$$E_\theta\{(\delta_1+\delta_2)U\} = E_\theta(\delta_1 U) + E(\delta_2 U) = 0 \quad \forall U \in \mathcal{U}$$

$\Rightarrow \delta_1 + \delta_2$ is a UMVUE of $g_1(\theta) + g_2(\theta)$.

Variance Bounds and Information

$\mathrm{Var}_\theta(\delta) \geq$ (LB).

Recall from the Cauchy-Schwarz inequality,

ψ: psi, ϕ: phi

$$\mathrm{Cov}(X,Y) \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)} \quad \text{or} \quad |\mathrm{Cov}(X,Y)| \leq \sigma_X \sigma_Y.$$

If δ is an unbiased estimator for $g(\theta)$ and ψ is an arbitrary random variable, then

$$\mathrm{Var}_\theta(\delta) \geq \frac{\mathrm{Cov}_\theta^2(\delta, \psi)}{\mathrm{Var}_\theta(\psi)}$$

Suitable ψ so that the $\mathrm{Cov}_\theta(\delta,\psi)$ is the same for all δ that are unbiased for $g(\theta)$.

(above)

Let $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ be a dominated family with densities $f_\theta : \theta \in \Omega = \mathbb{R}$.

To begin, $E_{\theta+\Delta}(\delta) - E_\theta(\delta)$ gives $g(\theta+\Delta) - g(\theta)$ for any unbiased δ.

Here Δ must be chosen such that $\Delta + \theta \in \Omega$. Next, we write $E_{\theta+\Delta}(\delta) - E_\theta(\delta)$ as a covariance under f_θ. This step includes the use of "likelihood ratio." We assume here that $f_{\theta+\Delta}(x) = 0$ whenever $f_\theta(x) = 0$. Define

$$L(x) = \begin{cases} \dfrac{f_{\theta+\Delta}(x)}{f_\theta(x)}, & f_\theta(x) > 0 \\ 0, & f_\theta(x) = 0 \end{cases}$$

Let $f_{\theta+\Delta}(x) = \dfrac{f_{\theta+\Delta}(x)}{f_\theta(x)} \times f_\theta(x)$ a.s. x.

L

and so for any function h integrable, under $P_{\theta+\Delta}$,

chg of measure / importance sampling.

$$E_{\theta+\Delta}(h(X)) = \int h f_{\theta+\Delta}\, d\mu = \int h L f_\theta\, d\mu = E_\theta\big(L(X)h(X)\big).$$

Take $h = 1$, then,

$$E_\theta(L(X)) = \int \frac{f_{\theta+\Delta}(x)}{f_\theta(x)} f_\theta(x)\,dx = \int f_{\theta+\Delta}(x)\,dx = 1$$

Take $h = \delta$, then

$$E_{\theta+\Delta}(\delta) = E_\theta(L(X)\delta)$$

So, if we define $\psi(X) = L(X) - 1$,

then we can see that $E_\theta(\psi(X)) = E_\theta(L(X) - 1) = 0$

and

$$E_{\theta+\Delta}(\delta) - E_\theta(\delta) \approx E_\theta(L\delta) - E_\theta(\delta) = E_\theta(\psi\delta) = \mathrm{Cov}_\theta(\delta,\psi).$$

($E(\psi)E(\delta)$)

As a result, $\mathrm{Cov}_\theta(\delta,\psi) = g(\theta+\Delta) - g(\theta)$

for any unbiased estimator δ. With this particular choice of ψ, the above inequality can be rewritten as

$$\mathrm{Var}_\theta(\delta) \geq \frac{\mathrm{Cov}_\theta^2(\delta,\psi)}{\mathrm{Var}_\theta(\psi)} = \frac{(g(\theta+\Delta)-g(\theta))^2}{\mathrm{Var}_\theta(\psi)} = \frac{(g(\theta+\Delta)-g(\theta))^2}{E_\theta\left(\left(\frac{f_{\theta+\Delta}(X)}{f_\theta(X)}-1\right)^2\right)}$$

Hammersley – Chapmann – Robinson bounds.

Under suitable regularity conditions, we can show that

$$\frac{\left(\frac{g(\theta+\Delta)-g(\theta)}{\Delta}\right)^2}{E_\theta\left(\left(\frac{\{f_{\theta+\Delta}(X)-f_\theta(X)\}/\Delta}{f_\theta(X)}\right)^2\right)} \xrightarrow{\Delta\to 0} \left[\frac{(g'(\theta))^2}{E_\theta\left(\left(\frac{\partial f_\theta(X)/\partial\theta}{f_\theta(X)}\right)^2\right)}\right].$$

Fisher information

$$I(\theta) = E_\theta\left(\left(\frac{\partial \log f_\theta(X)}{\partial\theta}\right)^2\right)$$

(To be continued...)

Next lecture:
① Finish the remaining discussion on FI.
② Bayes estimator and avg risk optimality (TPE Ch.4).

31/Oct FINALISED. Lecture 1 – Fisher Information (Lecture 6 – FRR)

Final date

Last teaching day: 28/Nov. ↓ 5/Dec OR 12/Dec — Pending.

Lecture 7

17/Oct/2022.

Agenda: (1) Fisher Information
(2) Bayes Risk ...
~~– Equivariance (Ch. 3 of FPE) ...~~

[Cont'd] Fisher Information / C&R Low Bound

Define $I(\theta) = E_\theta\left(\left(\frac{\partial \log f_\theta(X)}{\partial \theta}\right)^2\right)$, ✓

where $\log f_\theta(x)$ is the log-likelihood function, as the Fisher information.

$$0 = \frac{\partial}{\partial\theta}(1) = \frac{\partial}{\partial\theta}\int f_\theta(x)\, d\mu(x) = \int \frac{\partial}{\partial\theta} f_\theta(x)\, d\mu(x) = \int \frac{\partial \log f_\theta(x)}{\partial\theta} f_\theta(x)\, d\mu(x)$$

($d\mu(x)$ = dx)

$$= E\left(\frac{\partial \log f_\theta(X)}{\partial\theta}\right)$$

and so, $$I(\theta) = E_\theta\left(\left(\frac{\partial \log f_\theta(X)}{\partial\theta}\right)^2\right) - \left\{E_\theta\left(\frac{\partial \log f_\theta(X)}{\partial\theta}\right)\right\}^2$$

$$= \mathrm{Var}_\theta\left(\frac{\partial \log f_\theta(X)}{\partial\theta}\right).$$

Furthermore, since

$$\int \frac{\partial^2 f_\theta(x)}{\partial\theta^2}\, d\mu(x) = E\left(\frac{\partial^2 f_\theta(X)/\partial\theta^2}{f_\theta(X)}\right) = 0$$

We can see that

$$\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2} = \frac{\partial^2 f_\theta(x)/\partial\theta^2}{f_\theta(x)} - \left(\frac{\partial \log f_\theta(x)}{\partial\theta}\right)^2$$

$$\Rightarrow \quad I(\theta) = -E\left(\frac{\partial^2 \log f_\theta(X)}{\partial\theta^2}\right). \text{ //}$$

[THEOREM]. Let $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ be a dominated family with Ω an open set in $\mathbb{R}$ and densities f_θ differentiable with respect to θ. If $E_\theta(\psi) = 0$ and $E_\theta(\delta^2) < \infty$, then

$$\mathrm{Var}_\theta(\delta) \geq \frac{\{g'(\theta)\}^2}{I(\theta)}, \quad \theta \in \Omega.$$

This is called the Cramér-Rao lower bound, or information bound.

(Example) Let $\mathcal{P}$ be a one-parameter exponential family in canonical form with densities f_η given by

$$f_\eta(x) = \exp\{\eta T(x) - A(\eta)\} h(x) .$$

Then

$$\frac{\partial \log f_\eta(x)}{\partial \eta} = T(x) - A'(\eta) .$$

By the previous results we have

$$I(\eta) = \mathrm{Var}_\eta\left(T(X) - A'(\eta)\right) = \mathrm{Var}_\eta(T(X)) = A''(\eta)$$

because

$$\frac{\partial^2 \log f_\eta(x)}{\partial \eta^2} = \frac{\partial^2}{\partial \eta^2}\left(\eta T(x) - A(\eta) + \log h(x)\right) = -A''(\eta) .$$

If the family is parameterised instead by $\mu = A'(\eta) = E_\eta(T(X))$, then $A''(\eta) = I(\mu)\{A''(\eta)\}^2$.

And, so, because $A''(\eta) = \mathrm{Var}(T)$, we have $I(\mu) = \frac{1}{\mathrm{Var}_\eta(T)}$

Observe also that because T is UMVUE of μ, the lower bound $\mathrm{Var}(\delta) \geq \frac{1}{I(\mu)}$ for an unbiased estimator of δ if μ is sharp. ||

(Example) Suppose $\mathcal{E}$ is an absolutely continuous random variable with density f. The family of distributions $\mathcal{P} = \{P_\theta : \theta \in \mathbb{R}\}$ with f_θ the distribution of $\theta + \mathcal{E}$ is called a <u>location family</u>.

$$\int g(x)\, dP_\theta(x) = E_\theta\left(g(X)\right) \overset{X=\theta+\mathcal{E}}{=} E_\theta\left(g(\theta+\mathcal{E})\right)$$
$$= \int g(\theta+e) f(e)\, de = \int g(x) f(x-\theta)\, dx .$$

So, P_θ has density $f_\theta(x) = f(x-\theta)$. The corresponding Fisher Information for this family is

$$I(\theta) = E_\theta\left(\left(\frac{\partial \log f(X-\theta)}{\partial \theta}\right)^2\right) = E_\theta\left(\left(-\frac{f'(X-\theta)}{f(X-\theta)}\right)^2\right)$$

$$= E\left(\left(\frac{f'(\xi)}{f(\xi)}\right)^2\right) = \int \frac{(f'(x))^2}{f(x)}\,dx \perp\!\!\!\perp \theta$$

So, for the location families, $I(\theta)$ is constant with respect to θ. ∎

If two (or more) independent vectors are observed, then the total Fisher information is the sum of the Fisher information provided by the individual observations:

Suppose X and Y are independent, X has density f_θ and Y has density g_θ.

The Fisher information from X is $I_X(\theta) = \mathrm{Var}_\theta\left(\frac{\partial \log f_\theta(X)}{\partial \theta}\right)$

$$I_Y(\theta) = \mathrm{Var}_\theta\left(\frac{\partial \log g_\theta(Y)}{\partial \theta}\right)$$

$$I_{(X,Y)}(\theta) = \mathrm{Var}_\theta\left(\frac{\partial \log (f_\theta(X) g_\theta(Y))}{\partial \theta}\right)$$

$$= \mathrm{Var}_\theta\left(\frac{\partial \log f_\theta(X)}{\partial \theta} + \frac{\partial \log g_\theta(Y)}{\partial \theta}\right)$$

$$= \mathrm{Var}_\theta\left(\frac{\partial \log f_\theta(X)}{\partial \theta}\right) + \mathrm{Var}_\theta\left(\frac{\partial \log g_\theta(Y)}{\partial \theta}\right)$$

$$= I_X(\theta) + I_Y(\theta).$$

More generically, for $X_i \overset{iid}{\sim} f_\theta$ $(i = 1, \dots, n)$, then

$$I_{\underset{\sim}{X}}(\theta) = \sum_{i=1}^n I_{X_i}(\theta) = n I_1(\theta).$$

$$\mathrm{Var}_\theta(\delta) \geq \frac{\{g'(\theta)\}^2}{n I(\theta)}.$$

How about $\underset{\sim}{\theta} \in \Omega \subset \mathbb{R}^k$?

Suppose $\underset{\sim}{\theta}$ takes values in $\mathbb{R}^k$, then the Fisher information will be a matrix, defined in regular case by

$$\{I(\theta)\}_{ij} = E_\theta\left(\frac{\partial \log f_\theta(X)}{\partial \theta_i} \cdot \frac{\partial \log f_\theta(X)}{\partial \theta_j}\right)$$

$$= \mathrm{Cov}_\theta\left(\frac{\partial \log f_\theta(X)}{\partial \theta_i}, \frac{\partial \log f_\theta(X)}{\partial \theta_j}\right)$$

$$= -E_\theta\left(\frac{\partial^2 \log f_\theta(X)}{\partial \theta_i \partial \theta_j}\right).$$

Note that $E_\theta\left(\nabla_\theta \log f_\theta(X)\right) = 0$,

$$I(\theta) = E_\theta\left(\{\nabla_\theta \log f_\theta(X)\}\{\nabla_\theta \log f_\theta(X)\}^T\right)$$
$$= \mathrm{Cov}_\theta\left(\nabla_\theta \log f_\theta(X)\right) = -E_\theta\left(\nabla_\theta^2 \log f_\theta(X)\right)$$

where ∇_θ is the gradient with respect to θ and ∇_θ^2 is the Hessian matrix of second order derivatives, i.e.

$\nabla_\theta = \left(\frac{\partial}{\partial\theta_1}, \dots, \frac{\partial}{\partial\theta_k}\right)^T$, $\nabla_\theta^2 = \begin{pmatrix} \frac{\partial^2}{\partial\theta_1^2} & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_k} \\ \vdots & & \\ \frac{\partial^2}{\partial\theta_k\partial\theta_1} & \cdots & \frac{\partial^2}{\partial\theta_k^2} \end{pmatrix}$.

$$\nabla_\theta = \left(\frac{\partial f_\theta(x)}{\partial\theta_1}, \dots, \frac{\partial f_\theta(x)}{\partial\theta_k}\right)^T \quad \text{and} \quad \underset{[k\times k \text{ matrix}]}{\nabla_\theta^2} = \begin{pmatrix} \frac{\partial^2 f_\theta(x)}{\partial\theta_1^2} & \frac{\partial^2 f_\theta(x)}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 f_\theta(x)}{\partial\theta_1\partial\theta_k} \\ \vdots & \ddots & & \\ & & & \frac{\partial^2 f_\theta(x)}{\partial\theta_k^2} \end{pmatrix}$$

The lower bound for the variance of an unbiased estimator δ of $g(\theta)$, where $g: \Omega \to \mathbb{R}^k$ is

$$\mathrm{Var}_\theta(\delta) \geq \left(\nabla_\theta g(\theta)\right)^T I^{-1}(\theta)\left(\nabla_\theta g(\theta)\right).$$

BAYES ESTIMATORS AND AVERAGE RISK OPTIMALITY

Explore an alternative approach to achieve optimality.

Notation: Data: X, $\mathcal{P} = \{P_\theta \in \Omega\}$, $L(\theta, d)$ loss function, $R(\theta, \delta) = E_\theta(L(\theta, d))$ Risk function.

We need to introduce a measure Λ over Ω. This measure Λ can be viewed as an assignment of weights to each parameter value $\theta \in \Omega$ a priori (prior dist.).

Before any data is observed, θ is no longer a constant. It is instead a random variable (Θ).

Given a measure Λ, our objective is to find an estimator δ_Λ which minimises the average risk:

$$r(\Lambda, \delta) = \int R(\theta, \delta)\,\Lambda(\theta) = E_\Theta\left(R(\theta, \delta)\right)$$

If Λ is a probability distribution on Θ, we call Λ the prior distribution. Correspondingly, the estimator δ_Λ, if exists, is called the Bayes estimator with respect to Λ, and the minimised average risk is called the Bayes risk.

We shall pay attention to $E\left(L(\Theta, \delta(X)) \,\middle|\, X=x\right)$, the conditional risk at (almost) every value of X. Note that the expectation here is taken w.r.t. the conditional distribution of Θ given X, i.e. $\Theta \,|\, X=x$.

[THEOREM] Suppose $\Theta \sim \Lambda$ and $X \,|\, \Theta=\theta \sim P_\theta$. If

1. there exists δ_0, and estimator of $g(\theta)$ with finite risk for all θ,
2. there exists a value $\delta_\Lambda(x)$ that minimises
   - $E\left(L(\Theta, \delta_\Lambda(X)) \,\middle|\, X=x\right)$ for almost every x, (*)

then δ_Λ is a Bayes estimator with respect to Λ.

· Proof. If (1) and (2) hold, for any other estimator δ', say, and for almost every x,

$$E\left(L(\Theta, \delta_\Lambda(X)) \,\middle|\, X=x\right) \leq E\left(L(\Theta, \delta'(X)) \,\middle|\, X=x\right)$$

Taking expectation over X, we obtain

$$E\left(L(\Theta, \delta_\Lambda(X))\right) \leq E\left(L(\Theta, \delta'(X))\right) \quad \forall \delta'$$

□

Remark: The almost sure statement in (*) is defined w.r.t. the marginal distribution of X: $P(X \in A) = \int P_\theta(X \in A)\, d\Lambda(\theta)$.

(Example) If we consider $L(\theta, d) = (d-\theta)^2$, to find the Bayes estimator, we need to minimise $E\left(\{g(\Theta) - \delta(X)\}^2 \,|\, X=x\right)$

$$= E\left(\left\{g(\Theta) - E(g(\Theta)|X) + E(g(\Theta)|X) - \delta(X)\right\}^2 \,\middle|\, X=x\right)$$

$$= E\left(\{g(\Theta) - E(g(\Theta)|X)\}^2 \,\middle|\, X=x\right) + E\left(\{E(g(\Theta)|X) - \delta(X)\}^2 \,\middle|\, X=x\right),$$

or equivalently, minimising $E\left(\{E(g(\Theta)|X) - \delta(X)\}^2 \,|\, X=x\right)$.

The Bayes estimator is hence $E(g(\Theta)|X)$.

Posterior mean.

‖

Remark: By the Bayes' theorem,

$$\text{posterior} = \frac{\text{joint}}{\text{marginal}} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal}}$$

$$p(\theta \,|\, X) = \frac{p(\theta, x)}{\int p(\theta', x)\, d\theta'} = \frac{p(x|\theta) \times \pi(\theta)}{\int p(\theta', x)\, d\theta'}$$

↑ "normalizer"

(Example) Binomial mean estimation.

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (θ) given $\Theta = \theta$ and that Θ has a prior distribution Beta(α, β), where α and β are two fixed prior hyperparameters. The prior density is given by

$E(\Theta|X)$
$\pi(\theta|x)$

Conjugate prior $\longrightarrow$ $$\pi(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} I(0<\theta<1)$$ ✓

The model pmf of $X = X_1 + \ldots + X_n$ is

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

So the posterior distribution of Θ given X is

KEY

$$\pi(\theta|x) \propto \left[\binom{n}{x} \theta^x (1-\theta)^{n-x}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} I(0<\theta<1)\right]$$

$$\propto \underbrace{\theta^{(x-\alpha)-1}}_{\cdot} \underbrace{(1-\theta)^{(n-x+\beta)-1}}_{\text{This guy}} I(0<\theta<1) \Big/ \int_{\text{this guy } d\theta} \quad \int \to 1$$

From textbooks/wiki
$E(\text{Beta}(\alpha,\beta)) = \frac{\alpha}{\alpha+\beta}$ ✓

$$\sim \text{Beta}(x-\alpha,\ n-x+\beta).$$

We know that the posterior mean of $\Theta|X$ is

$$E_{\Theta}(\Theta \mid X=x) = \frac{x+\alpha}{(x+\alpha)+(n-x+\beta)} = \boxed{\frac{x+\alpha}{n+\alpha+\beta}}$$

$$= \underbrace{\frac{n}{n+\alpha+\beta}}_{\omega} \cdot \left(\frac{x}{n}\right) + \underbrace{\frac{\alpha+\beta}{n+\alpha+\beta}}_{(1-\omega)} \left(\frac{\alpha}{\alpha+\beta}\right)$$

"Let the data speak for themselves" ↗ — Sample mean — "mean of the prior."

(Example) Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\Theta, \sigma^2)$ with σ^2 known. Furthermore, let $\Theta \sim N(\mu, b^2)$, where μ and b^2 are two fixed prior hyperparameters. Then the posterior distribution of $\Theta|X$ is given by

$$\pi(\theta|x) \propto \prod_{i=1}^{n} \left\{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right)\right\} \times \frac{1}{\sqrt{2\pi b^2}} \exp\left(-\frac{(\theta-\mu)^2}{2b^2}\right)$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 - \frac{1}{2b^2}(\theta-\mu)^2\right\}$$

$$= \exp\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\theta - \frac{n\theta^2}{2\sigma^2} - \frac{1}{2b^2}\theta^2 + \frac{\mu\theta}{b^2}\right\}$$

$$= \cdots$$

$$= \exp\left\{ -\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{b^2}\right)\theta^2 + \left(\frac{n\bar{x}_n}{\sigma^2} + \frac{\mu}{b^2}\right)\theta \right\}$$

$$\overset{\Delta}{=} \exp\left\{ -\frac{1}{2\tilde{\sigma}^2}(\theta - \tilde{\mu})^2 \right\} \cdot \sim N(\tilde{\mu}, \tilde{\sigma}^2)$$

where $\tilde{\mu} = \dfrac{n\bar{x}_n/\sigma^2 + \mu/b^2}{n/\sigma^2 + 1/b^2}$ and $\tilde{\sigma}^2 = \dfrac{1}{n/\sigma^2 + 1/b^2}$.

Hence, the posterior mean of $\Theta \mid X = x$ is

$$E_\Theta(\Theta \mid X = x) = \frac{n\bar{x}_n/\sigma^2 + \mu/b^2}{n/\sigma^2 + 1/b^2} = \frac{n/\sigma^2}{n/\sigma^2 + 1/b^2}\bar{x}_n + \frac{1/b^2}{n/\sigma^2 + 1/b}\mu$$

$\bar{x}_n$: sample mean; μ: prior mean.

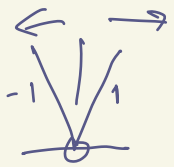
Next week:

1) How about other loss functions?
2) Relationship b/w $\bar{X}_n$ and Bayes estimator...
3) Admissibility.
4) Empirical Bayes...

Lecture 8

24 Oct 2022.

Recall: Average risk optimality: θ (fixed constant) $\Theta \sim \Lambda$ prior dist.



Bayes estimator (under the squared error loss function): $E(g(\Theta)|X)$ (posterior dist)

abs loss function → $\text{med}(g(\Theta)|X)$

Conjugate priors

$\frac{\partial}{\partial d}|\_| = [\text{sgn}(\_)]$

(Example) Assume that we consider instead $L(\theta, d) = \omega(\theta)(g(\theta)-d)^2$, where $\omega(\theta) \geq 0$, which can be interpreted as a weight function.

Our goal is to find the corresponding Bayes estimator, which minimises $\underbrace{E\left(\omega(\Theta)(g(\Theta)-d)^2 \middle| X=x\right)}_{\Xi}$ with respect to d.

the quantity Ξ can be rewritten as

$$d^2 E\left(\omega(\Theta)\middle|X=x\right) - 2d\,E\left(\omega(\Theta)g(\Theta)\middle|X=x\right) + E\left(\omega(\Theta)g(\Theta)^2\middle|X=x\right) \quad (*)$$

Taking derivative of $(*)$ w.r.t. d gives

$$2d^* E\left(\omega(\Theta)\middle|X=x\right) - 2E\left(\omega(\Theta)g(\Theta)\middle|X=x\right) = 0$$

$$d^*\left(\equiv \delta_\Lambda(X)\right) = \frac{E\left(\omega(\Theta)g(\Theta)\middle|X=x\right)}{E\left(\omega(\Theta)\middle|X=x\right)}.$$

In particular, if $\omega(\cdot) \equiv 1$, $\delta_\Lambda(X) = E\left(g(\Theta)\middle|X=x\right)$

[consistent with our previous result].

Q: Can $\bar{X}_n \left(= n^{-1}\sum_{i=1}^n X_i\right)$ be a Bayes estimator?

No. Except that it degenerates to 0, which is a rare case.

[THEOREM] (TPE 4.2.3) If δ is unbiased for $g(\theta)$ with $r(\Lambda, \delta) < \infty$ and $E(g(\Theta)^2) < \infty$, then δ is not Bayes under the squared error loss function unless its average risk is zero, i.e.

$$\rightarrow \quad E_{(X,\Theta)}\left(\{\delta(X) - g(\Theta)\}^2\right) = 0 .$$

· Proof. Let δ be an unbiased estimator under the squared error loss function. Then, we know that δ is the posterior mean, i.e.

$$\delta(X) = E\left(g(\Theta) \mid X\right) \quad \text{a.s.} \quad \forall$$

Thus we have

$$E\left(\delta(X)\, g(\Theta)\right) = E\left(E\left(\delta(X) g(\Theta) \mid X\right)\right)$$
$$= E\left(\delta(X)\, E\left(g(\Theta) \mid X\right)\right) = E\left(\{\delta(X)\}^2\right). \quad \text{— (1)}$$

Also,

$$E\left(\delta(X) g(\Theta)\right) = E\left(E\left(\delta(X) g(\Theta) \mid \Theta\right)\right)$$
$$= E\left(g(\Theta)\, E\left(\delta(X) \mid \Theta\right)\right) = E\left(\{g(\Theta)\}^2\right). \quad \text{— (2)}$$

Observe that

$$E\left(\{\delta(X) - g(\Theta)\}^2\right) = E\left(\delta^2(X)\right) - 2E\left(\delta(X) g(\Theta)\right) + E\left(g^2(\Theta)\right)$$
$$= E\left(\delta^2(X)\right) - E\left(\delta(X) g(\Theta)\right) + E\left(g^2(\Theta)\right) - E\left(\delta(X) g(\Theta)\right)$$
$$= E\left(\delta^2(X)\right) - E\left(\delta^2(X)\right) + E\left(g^2(\Theta)\right) - E\left(g^2(\Theta)\right)$$
(By (1)) (By (2))
$$= 0$$

Thus we have $E_{X,\Theta}\left(\{\delta(X) - g(\Theta)\}^2\right) = 0$.

$N(\theta, 0)$ & no variation

(Example) Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\underset{\varphi}{\theta}, \sigma^2)$ with σ^2 known.

$\theta \sim N(\mu, b^2)$

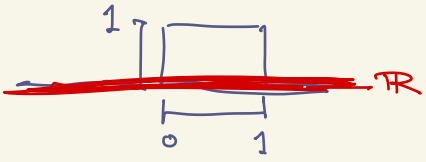
Is $\bar{X}_n$ Bayes under the squared error loss function for some choice of the prior?

Observe that $E(\bar{X}_n \mid \theta) = \theta$. Hence $\bar{X}_n$ is unbiased for θ.

The corresponding average risk under the squared error loss function is given by

$$E_{(X,\theta)}\left(\{\underset{\varphi}{\bar{X}_n} - \underset{\varphi}{\theta}\}^2\right) = \frac{\sigma^2}{n} \neq 0$$

$\Rightarrow$ $\bar{X}_n$ is not a Bayes estimator under "any prior distribution".



Remark: $\bar{X}_n$ does minimize a form of average risk: it minimizes the average risk with respect to the Lebesgue measure, that is, with respect to the density $\pi(\theta) = 1$ for all θ. We call this choice of π an improper prior since the integral $\int \pi(\theta) d\theta = \infty$ [in which $\pi(\theta)$ does not define a proper prob. distribution]

non-informative prior.

DBA

We may define a formal posterior for this improper prior:

Formal posterior $\propto$ Likelihood $\times$ Improper Prior.

(e.g. Normal:

$$\text{Formal posterior} \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(X_i - \theta)^2\right) \times \underbrace{1}_{\text{improper prior}}$$

$$\propto \exp\left(\frac{n\bar{X}_n}{\sigma^2}\theta - \frac{n}{2\sigma^2}\theta^2\right)$$

$$\sim N(\bar{X}_n, \sigma^2/n) \checkmark$$

We call $\bar{X}_n$ as a generalized Bayes estimator)

[THEOREM] A unique Bayes estimator (a.s. for all P_θ) is admissible.
(TPE 5.2.4)

An estimator is admissible if it is not uniformly dominated by some other estimator. δ is said to be inadmissible if and only if there exists δ' such that

$$R(\theta,\delta') \le R(\theta,\delta) \quad \forall \theta \in \Omega \text{ and}$$
$$R(\theta,\delta') < R(\theta,\delta) \quad \text{for some } \theta \in \Omega.$$

• Proof • Suppose δ_Λ is Bayes for Λ, and for some δ', $R(\theta,\delta') \le R(\theta,\delta_\Lambda)$ for all $\theta \in \Omega$. If we take the expectation with respect to Θ, the inequality above is preserved and can be written as

$$\int_{\theta\in\Omega} R(\theta,\delta')\, d\Lambda(\theta) \le \int_{\theta\in\Omega} R(\theta,\delta_\Lambda)\, d\Lambda(\theta).$$

This implies that δ' is also Bayes because δ' has less (or equal) risk than δ_Λ which minimizes the average risk. Hence $\delta' = \delta_\Lambda$ with prob. 1 for all P_θ.

Question: When will a Bayes estimator be unique?

[THEOREM] (TPE 4.1.4) Let Q be the marginal distribution of X, that is

$$Q(E) = \int P(X \in E \mid \Theta)\, d\Lambda(\Theta).$$

Then under a strictly convex loss function, δ_Λ is unique (a.s. for all P_θ) if

(a) $r(\Lambda,\delta_\Lambda)$ is finite and ← finiteness for comparison

(b) $P_\theta \ll Q$ (Ω is defined on the support of Λ.)
↑ abs. cont.

• Why do we need to consider Bayes estimators?

1). All admissible estimators are limits of Bayes estimators.

[explain/elaborate further later when we discuss sequences of priors Λ_m] $\to N(\mu, \frac{b^2}{m})$ minimax

2) Prior information

↓ How can we choose a prior?

a. subjective – prior knowledge / experience

b. objective – maximally non-informative prior or reference prior (Jeffreys priors etc).

c. conjugate prior – for computational convenience

d. Hierarchical Bayes → TPE §4.5.

1)

2) Empirical Bayes (*).

we begin with the following Bayes model:

hyperparameter

$$X_i \mid \theta \sim f(x \mid \theta) \qquad i = 1, \dots, n$$

$$\theta \mid \gamma \sim \pi(\theta \mid \gamma) \qquad \text{(prior distribution)}$$

One can calculate the marginal distribution of $\underset{\sim}{X}$ with density

$$m(\underset{\sim}{x} \mid \gamma) = \int \underbrace{\prod_{i=1}^{n} f(x_i \mid \theta)}_{\text{likelihood}} \underbrace{\pi(\theta \mid \gamma)}_{\text{prior}} \, d\theta$$

Based on $m(\underset{\sim}{x} \mid \gamma)$, we can obtain an estimate, say $\hat{\gamma}(x)$ of γ.

We can then substitute $\hat{\gamma}(x)$ for γ in $\pi(\theta \mid \gamma)$ and determine the estimator that minimises the empirical posterior loss

$$\int L(\theta, \delta(x)) \, \pi(\theta \mid x, \hat{\gamma}(x)) \, d\theta$$

$\delta(x)$: empirical Bayes estimator (minimizer)

(Example). Suppose there are K different groups of patients, where each group has n patients. Each group is given a different treatment for the same illness, and in the k^{th} group ($k=1,\dots,K$), we count X_k as the number of successful treatments out of these n patients.

$$X_k \sim \text{binomial}(n, p_k)$$

$$p_k \sim \text{Beta}(a,b) \qquad k=1,2,\dots,K$$

instead of (a_k, b_k).

Share a common prior distribution

The single-prior Bayes estimator p_k under the squared error loss is

$$\delta^{\pi}(x_k) = E(p_k \mid x_k, a, b) = \frac{a+x_k}{a+b+n} \quad (6.7) \quad \text{[done before]}.$$

In the empirical Bayes model, we consider these hyperparameters unknown and we want to estimate them.

To construct an EBE, we calculate

$$m(\underset{\sim}{x} \mid a,b) = \int_0^1 \cdots \int_0^1 \prod_{k=1}^{K} \binom{n}{x_k} p_k^{x_k} (1-p_k)^{n-x_k} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p_k^{a-1} (1-p_k)^{b-1} \, dp_k$$

$$= \left[\prod_{k=1}^{K} \binom{n}{x_k} \frac{\Gamma(a+b)\Gamma(a+x_k)\Gamma(n-x_k+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)} \right]$$

product of beta-binomial distribution

Check Table 6.1 (on Page 265 of TPE §4.6) $K=10$, $n=20$

Prior parameters		Bayes Risk		
a	b	δ^{π} (6.7)	$\delta^{\hat{\pi}}$ (6.9)	X/n
2	2	0.0833	0.0850	0.1000
⋮				

$\hat{a}, \hat{b} \overset{\delta}{=} \underset{a,b}{\text{argmax}}$

The corresponding empirical Bayes estimator is given by:

$$\delta^{\hat{\pi}}(x_k) = E(p_k \mid x_k, \hat{a}, \hat{b}) = \frac{\hat{a}+x_k}{\hat{a}+\hat{b}+n} \quad (6.9)$$

(check)

[THEOREM] (TPE 4.6.2) For the situation of Corollary 3.3 with the prior distribution $\pi(\eta \mid \lambda)$. Suppose $\hat{\lambda}(x)$ is the MLE of $m(\underset{\sim}{x} \mid \lambda)$, then the empirical Bayes estimator is

$p_\eta(x) = e^{\sum_i \eta_i x_i - A(\eta)} h(x)$ $\quad E(\eta_i \mid x, \hat{\lambda}) = \frac{\partial}{\partial x_i} \log m\left(x \mid \hat{\lambda}(x)\right) - \frac{\partial}{\partial x_i} \log h(x)$

(You may verify this result)

(Example) Consider the estimation of

$$X_i \mid \theta_i \sim N(\theta_i, \sigma^2) \quad i = 1, \dots, p, \text{ independent}$$
$$\Theta_i \sim N(\mu, \tau^2) \quad i = 1, \dots, p$$

Here μ is unknown. We can use the previous theorem to calculate the empirical Bayes estimator,

$$E(\Theta_i \mid \underset{\sim}{X}, \hat{\mu}) = \sigma^2 \left(\frac{\partial}{\partial X_i} \log m(\underset{\sim}{X} \mid \hat{\mu}) - \frac{\partial}{\partial X_i} \log h(X) \right),$$

where $\hat{\mu}$ is the MLE of μ from

$$m(\underset{\sim}{X} \mid \mu) = \frac{1}{\{2\pi(\sigma^2+\tau^2)\}^{p/2}} e^{-\frac{\sum_{i=1}^p (X_i - \mu)^2}{2(\sigma^2+\tau^2)}}$$

Here $\hat{\mu} = \bar{x}$ and

$$\frac{\partial}{\partial X_i} \log m(\underset{\sim}{X} \mid \hat{\mu}) = \frac{\partial}{\partial X_i} \left(-\frac{1}{2(\sigma^2+\tau^2)} \sum_{i=1}^n (X_i - \bar{X})^2 \right),$$

which gives the empirical Bayes estimator

$$E(\Theta_i \mid \underset{\sim}{X}, \hat{\mu}) = \frac{\tau^2}{\sigma^2+\tau^2} X_i + \frac{\sigma^2}{\sigma^2+\tau^2} \bar{X}.$$

↳ conv. to the prior $\pi(\theta \mid \bar{x})$.

... (You may also take a look of

| Risk calculation for the emp. Bayes estimator |)

$$R(\eta, E(\eta \mid X, \hat{\lambda}(X))) = R(\eta, -\nabla \log h(X)) + \sum_{i=1}^p E_\eta \left\{ 2 \frac{\partial^2}{\partial X_i^2} \log m(X \mid \hat{\lambda}(X)) + \left(\frac{\partial}{\partial X_i} \log m(X \mid \hat{\lambda}(X)) \right)^2 \right\}$$

Lecture 9

7/Nov/2022

Worst-case Optimality

Given $X \sim P_\theta$, where $\theta \in \Omega$ and a loss function $L(\theta, d)$, we want to minimise the maximum risk: $\sup_{\theta \in \Omega} R(\theta, \delta)$. The minimiser is known as a minimax estimator.

Recall the definition of Bayes risk under an arbitary prior distribution Λ:

$$r_\Lambda = \inf_\delta r(\Lambda, \delta) = \inf_\delta \int_{\theta \in \Omega} \underbrace{R(\theta, \delta)}_{E(L(\Theta, \delta(X)) | \Theta)} d\Lambda(\theta)$$

Definition A prior distribution Λ is said to be a least favourable prior if $r_\Lambda \geq r_{\Lambda'}$ for any other prior distribution Λ'.

[THEOREM] TPE 5.1.4 Suppose δ_Λ is Bayes for Λ with $r_{\Lambda \in \Omega} = \sup_\theta R(\theta, \delta_\Lambda)$

Bayes risk is the max risk of δ_Λ

1) δ_Λ is minimax;
2) Λ is a least favourable prior and
3) If δ_Λ is the unique Bayes estimator for Λ, it is also unique minimax.

· Proof:

1) $$\sup_{\theta \in \Omega} R(\theta, \delta) \geq \int R(\theta, \delta) d\Lambda(\theta)$$
$$\geq \int R(\theta, \delta_\Lambda) d\Lambda(\theta)$$
$$= \sup_\theta R(\theta, \delta_\Lambda)$$

$\Rightarrow \delta_\Lambda$ is minimax. 3) ...

2) Let Λ' be any other prior distribution. Then we have

$$r_{\Lambda'} = \inf_\delta R(\theta, \delta) d\Lambda'(\theta)$$
$$\leq \int R(\theta, \delta_\Lambda) d\Lambda'(\theta) \leq \underbrace{\sup_\theta R(\theta, \delta_\Lambda)}_{\text{cond.}} = r_\Lambda$$

$\Rightarrow \Lambda$ is a least favourable prior.

□

[Corollary] TPE 5.1.5 If a Bayes estimator δ_Λ has constant risk: $R(\theta, \delta_\Lambda) = R(\theta', \delta_\Lambda)$ for $\theta, \theta' \in \Omega$, then δ_Λ is minimax.

* This is a sufficient but not necessary cond.

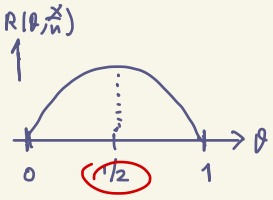
Strategy: Find a support set ω such that $\Lambda(\omega) = 1$ and for which $R(\theta, \delta_\Lambda)$ is maximum for all $\theta \in \omega$.

[Corollary] TPE 5.1.6. Define $\omega_\Lambda = \left\{\theta : R(\theta, \delta_\Lambda) = \sup_{\theta'} R(\theta', \delta_\Lambda)\right\}$

then a Bayes estimator δ_Λ is minimax iff $\Lambda(\omega) = 1$. (xiff)

(Example) Suppose $X \sim$ Binomial (n, θ) for some $\theta \in (0,1)$ and we adopt the squared error loss function. Is the sample portion X/n minimax?

$$R(\theta, \tfrac{X}{n}) = E\left(\left(\tfrac{X}{n} - \theta\right)^2\right) = \frac{\theta(1-\theta)}{n}$$



$$\sup_\theta R(\theta, \tfrac{X}{n}) = R(\tfrac{1}{2}, \tfrac{X}{n}) = \frac{1}{4n}.$$

TPE 5.1.5 / 5.1.6 is not helpful here as $\Lambda(\{\tfrac{1}{2}\}) = 1 \Rightarrow \delta_\Lambda(X) = \tfrac{1}{2} \left(\neq \tfrac{X}{n}\right)$

With Beta(a,b) as the prior, $\delta_{a,b}(X) = \dfrac{X+a}{n+a+b}$ (see Lecture _).

For any a and b,

$$R(\theta, \delta_{a,b}(X)) = E_\theta\left(\frac{X+a}{n+a+b} - \theta\right)^2$$

$$= \frac{1}{(n+a+b)^2} E_\theta\left(\{X + a - \theta(n+a+b)\}^2\right)$$

$$= \frac{1}{(n+a+b)^2}\left[n\theta(1-\theta) + \{a(\theta-1) + \theta b\}\right]^2, \checkmark$$

which is a quadratic function of θ. To eliminate θ, we need to set

$$\begin{cases} -n + (a+b)^2 = n - 2a(a+b) = 0 \cdots & \text{[coeff. of } \theta^2\text{]} \\ \cdots\cdots\cdots\cdots \\ n - 2a(a+b) = 0 \cdots & \text{[coeff. of } \theta\text{]} \end{cases}$$

$\Rightarrow$ Note that $a, b > 0$

$$a + b = \sqrt{n}, \quad n - 2a\sqrt{n} = 0 \cdots$$

$$\Rightarrow a = \sqrt{n} - b \Rightarrow n - 2(\sqrt{n} - b)\sqrt{n} = n - 2n + 2b\sqrt{n}$$

$$\Rightarrow \cdots \quad n = 2b\sqrt{n} \Rightarrow \begin{pmatrix} b = \sqrt{n}/2 \\ a = \sqrt{n}/2 \end{pmatrix}.$$

$\Rightarrow$ Beta$\left(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}\right)$ is a least favourable prior with constant risk. Then our Bayes estimator $\delta_{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}(X) = \dfrac{X + \sqrt{n}/2}{n + \sqrt{n}}$ ✓ is minimax with constant risk of $\dfrac{1}{4(\sqrt{n}+1)^2}$. (CHECK).

We conclude that $\frac{X}{n}$ is NOT minimax. $\Leftarrow$ $\dfrac{1}{4n + (8\sqrt{n} + 4)} < \dfrac{1}{4}$. //

(Example) Minimax for iid Normal r.v.s. with unknown mean θ : $X_1, \dots, X_n \sim N(\theta, \sigma^2)$ (known)

Observe that $\bar{X}_n$ has constant risk of $\frac{\sigma^2}{n}$. HOWEVER, $\bar{X}_n$ is not Bayes for any prior. [Recall TPE 4.2.3 : unbiased estimator are Bayes only in the degenerate situations of zero risk]

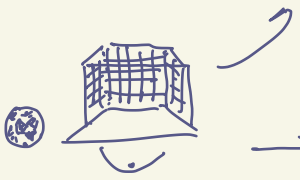
We cannot conclude via TPE 5.1.5 that $\bar{X}_n$ is minimax.

Consider instead : $\delta_{a,\mu_0}(X) = a\bar{X}_n + (1-a)\mu_0$, for $a \in (0,1)$, $\mu_0 \in \mathbb{R}$

The corresponding worst-case risk:

$$\sup_\theta E_\theta\left((\theta - \delta_{a,\mu_0}(X))^2\right) = \sup_\theta \left(a^2 \mathrm{Var}_\theta(\bar{X}_n) + (1-a)^2(\theta-\mu_0)^2\right)$$

$$= \frac{a^2\sigma^2}{n} + (1-a)^2 \sup_\theta (\theta-\mu_0)^2 = +\infty.$$



→ Extend our minimax results to the limits of Bayes estimators rather than restricting our attention to "regular" Bayes estimator only.

Definition (Least favourable sequences of priors) $\Lambda \sim N(\mu_0, \tau^2)$

Let $\{\Lambda_m\}$ be a sequence of priors with minimal average risk $r_{\Lambda_m} = \inf_\delta \int R(\theta,\delta)\, d\Lambda_m(\theta)$. Then, $\{\Lambda_m\}$ is a least favourable sequence of priors if there is a real number r such that $r_{\Lambda_m} \to r < \infty$ and $r_0 \geq r_{\Lambda'}$ for any other prior Λ'.

[THEOREM] 5.1.12 (Extend 5.1.4) Suppose there is a real number r such that $\{\Lambda_m\}$ is a sequence of priors with $r_m \to r < \infty$. Let δ be any estimator such that $\sup_\theta R(\theta,\delta) = r$. Then i) δ is minimax and ii) $\{\Lambda_m\}$ is a least favourable sequence of priors.

· Proof. i) Let δ' be any other estimator. Then for any m,

$$\sup_\theta R(\theta,\delta') \geq \int R(\theta,\delta')\, d\Lambda_m(\theta) \geq r_{\Lambda_m}.$$

Sending $m \to \infty$ yields

$$\sup_\theta R(\theta,\delta') \geq r = \sup_\theta (R,\delta_0)$$

$\Rightarrow \delta$ is minimax.

ii) Let Λ' be any prior, then

$$r_{\Lambda'} = \int R(\theta, \delta_{\Lambda'}) \, d\Lambda'(\theta) \leq \int R(\theta, \delta) \, d\Lambda'(\theta) \leq \sup_\theta R(\theta, \delta) = r.$$

$\Rightarrow \{\Lambda_m\}$ is least favourable.

❗ No uniqueness statement is guaranteed here because of the limit.

✔ Generalise the scope so that Bayes estimators whose Bayes risks converge to the maximum risk can serve as a possible candidate.

(Example) (Minimax for iid Normal r.v.s. cont'd)

By the result above, it suffices to find a sequence $\{\Lambda_m\}$ such that $r_{\Lambda_m} \to \frac{\sigma^2}{n} = r$.

Let $\{\Lambda_m\}$ be conjugate priors $\{N(0, m^2)\}$ in which case Λ_m tends to the uniform prior on $\mathbb{R}$ (also improper with $\pi(\theta) = 1, \ \forall \theta \in \mathbb{R}$).

Recall that

$$\theta \mid (X_1, \dots, X_n) \sim N\left(\frac{n\bar{X}_n}{\frac{n}{\sigma^2} + \frac{1}{m^2}}, \underbrace{\frac{1}{\frac{n}{\sigma^2} + \frac{1}{m^2}}}_{\perp\!\!\!\perp (X_1, \dots, X_n)} \right)$$

$$r_{\Lambda_m} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{m^2}} \xrightarrow{m \to \infty} \frac{\sigma^2}{n} = \sup_\theta R(\theta, \bar{X}_n).$$

$\Rightarrow \bar{X}_n$ is minimax and $\{\Lambda_m\}$ is least favourable. ∎

- Minimaxity via submodel restriction.

Another technique of deriving a minimax estimator for a general family of models is to restrict the problem/discussion to a subset of that family.

(Example) Let $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$, where neither θ nor σ^2 is known.

Note that $\sup_{(\theta, \sigma^2)} R\left((\theta, \sigma^2), \bar{X}_n\right) = \sup_{\sigma^2} \frac{\sigma^2}{n} = \infty$.

Restrict our attention to $\Omega = \{(\theta, \sigma^2) : \theta \in \mathbb{R}^2, \ \sigma^2 \leq B\}$, where B is a known constant.

- $\sup_{\theta \in \mathbb{R}, \sigma^2 \leq B} R\left((\theta, \sigma^2), \bar{X}_n\right) = \frac{B}{n}$

$$= \sup_{\theta \in \mathbb{R}, \sigma^2 = B} R\left((\theta, \sigma^2), \bar{X}_n\right) \leq \sup_{\theta \in \mathbb{R}, \sigma^2 = B} R\left((\theta, \sigma^2), \delta\right)$$

$$\leq \sup_{\theta \in \mathbb{R}, \sigma^2 \leq B} R\left((\theta, \sigma^2), \delta\right)$$

∎

(Example). (Weighted squared error loss)

Let $X \sim \text{Binomial}(n, \theta)$, with the loss function $L(\theta, d) = \frac{(d-\theta)^2}{\theta(1-\theta)}$ ← (hence $\omega(\theta) = \frac{1}{\theta(1-\theta)}$). Note that here, for any θ, $R(\theta, \frac{X}{n}) = \frac{1}{n}$; that is to say, the risk is constant in θ.

Recall that with the loss function $L(d, \theta) = \omega(\theta)(d-\theta)^2$, the associated Bayes estimator is given by $E_{\Theta|X}(\Theta\,\omega(\Theta)|X) \,/\, E_{\Theta|X}(\omega(\Theta)|X)$.

Applying this result, we find that the Bayes estimator has the form

$$\delta_\Lambda(X) = \frac{E_{\Theta|X}\left(\frac{1}{1-\Theta}\middle|X\right)}{E_{\Theta|X}\left(\frac{1}{\Theta(1-\Theta)}\middle|X\right)}.$$

We use a prior conjugate to the binomial likelihood: $\Theta \sim \Lambda_{a,b} \overset{d}{=} \text{Beta}(a,b)$ for some $a, b > 0$

Suppose we observe $X = x$. If $a+x > 1$, and $b+n+x > 1$ (a, b: posterior), then we can claim that the resulting estimator is given by

$$\delta_{a,b}(x) = \frac{a+x-1}{a+b+n-2} \qquad \text{(to be proved in a min)},$$

and it minimises the posterior risk.

In particular, the estimator $\delta_{1,1}(x) = \frac{x}{n}$ minimises the posterior risk with respect to the uniform prior after observing $0 < x < n$. ← verify

If we can verify also that this form remains unchanged when $x \in \{0, n\}$

→ then we can claim that $\delta_{1,1}(X) = \frac{X}{n}$ is Bayes with constant risk and hence minimax.

→ Recall that the Beta function can be evaluated as

$$\int_0^1 x^{k_1-1}(1-x)^{k_2-1}\,dx = \frac{\Gamma(k_1)\Gamma(k_2)}{\Gamma(k_1+k_2)}, \quad \text{where } k_1, k_2 > 0.$$

Therefore, if $Y \sim \text{Beta}(a,b)$, where $a, b > 0$, we can evaluate the expectation

$$E\left(\frac{1}{1-Y}\right) = \int_0^1 \frac{1}{1-y}\left\{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}\right\} dy$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \left\{y^{a-1}(1-y)^{b-2}\right\} dy$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a)\Gamma(b-1)}{\Gamma(a+b-1)}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a+b-1)} \cdot \frac{\Gamma(b-1)}{\Gamma(b)} \cdot \frac{\Gamma(a)}{\Gamma(a)} = \frac{a+b-1}{b-1}.$$

Similarly $E\left(\frac{1}{Y(1-Y)}\right) = \ldots = \frac{(a+b-2)(a+b-1)}{(a-1)(b-1)}$, where $a>1$.

Combining these results/identities, we have, for $a, b > 1$,

$$\frac{E\left(\frac{1}{1-Y}\right)}{E\left(\frac{1}{Y(1-Y)}\right)} = \frac{a-1}{a+b-2}.$$

To see this is the case, note that the posterior risk under the prior $\Lambda_{1,1}$ after observing $X=x$ and deciding $\delta(x)=d$ is

$$\int_0^1 \underbrace{\frac{(d-\theta)^2}{\theta(1-\theta)}}_{\text{loss}} \cdot \frac{\Gamma(x+1+n-x+1)}{\Gamma(x+1)\Gamma(n-x+1)} \cdot \theta^x (1-\theta)^{n-x} \, d\theta,$$

which, in the case of $X=0$, simplifies to $\int_0^1 \frac{(d-\theta)^2}{\theta}(1-\theta)^{n-1} d\theta$.

This integral converges for $d=0$ and diverges otherwise, so the posterior risk is minimised by choosing $\delta(0) = 0 \; (= 0/n)$. Similarly in the case of $X=n$, the posterior risk is minimised by choosing $\delta(n) = 1 = n/n$. This confirms that $\delta_{1,1}(X) = X/n$ minimises the posterior for any outcome X and is indeed Bayes. This estimator has constant risk, we conclude that X/n is minimax.

||

Randomised Minimax Estimators.

Randomised estimators are functions of the form $\delta(X, U)$, where $U \sim \text{uniform}(0,1)$

Non-random estimators usually achieve better or equal average risk and they are oftentimes available.

Observation: With convex losses, we can "ignore" randomised estimators because we can find a better of equally good non-randomised/deterministic estimator.

Recall that the data X is sufficient, so by Rao-Blackwell theorem, the non-random estimator $\tilde{\delta}(X) = E(\delta(X,U) \mid X)$ is no worse than $\delta(X,U)$.

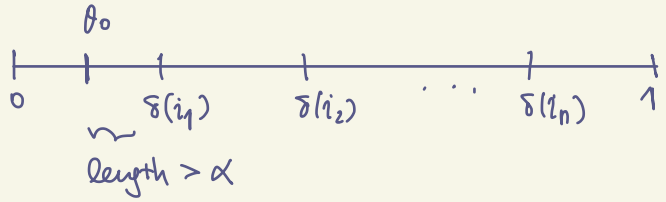
(Example) (Randomised minimax estimator).

Let $X \sim \text{Binomial}(n, \theta)$, where $\theta \in (0,1)$, and consider estimation of θ under the 0-1 loss:

$$L(\theta, d) = \begin{cases} 0 & \text{if } |d-\theta| < \alpha \\ 1 & \text{otherwise} \end{cases}.$$

First consider an arbitrary non-random estimator δ. Since X can only take the $n+1$ values $\{0, 1, \ldots, n\}$, the estimator $\delta(X)$ can also only take $n+1$ values $\{\delta(0), \delta(1), \ldots, \delta(n)\}$.

If $\alpha < \frac{1}{2(n+1)}$, then we can always find θ_0 such that $|\delta(x) - \theta_0| \geq \alpha$ for every $x \in \{0, 1, \ldots, n\}$.



$R(\theta_0, \delta(X)) = 1$, which corresponds to the maximum risk of any non-random δ.

Consider instead the estimator $\delta'(X, U) = U$, which is completely random and independent of the data X. Then for any $\theta \in (0,1)$,

$$\begin{aligned} R(\theta, \delta') &= E\left(L(\theta, \underbrace{\delta'(X,U)}_{U})\right) \\ &= P(|U-\theta| \geq \alpha) \\ &= 1 - P(\theta - \alpha < U < \alpha + \theta) \\ &\leq 1 - \alpha < 1 < \text{maximum risk of any non-random estimator } \delta. \end{aligned}$$

||

Next lecture:

? $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ (σ^2 known), $L(\theta, d) = (\theta - d)^2$, when will $\bar{X}_n$ be admissible?

Lecture 10

14 Nov 2022

* Admissibility of minimax est.
* Simultaneous estimation: James-Stein estimator.
* Testing

Minimax Estimators and submodels.

Recall that an estimator δ^M is minimax if its maximum risk is minimal:

$$\inf_{\delta} \sup_{\theta \in \Omega} R(\theta, \delta) = \sup_{\theta} R(\theta, \delta^M).$$

[Lemma] TPE 5.1.15 Suppose that δ is minimax for a submodel $\theta \in \Omega_0 \subset \Omega$ and

$$\sup_{\theta \in \Omega_0} R(\theta, \delta) = \sup_{\theta \in \Omega} R(\theta, \delta)$$

Then δ is minimax for the full model $\theta \in \Omega$.

→ Allows us to find a minimax est. for a particular tractable submodel, then we show that the worst-case risk doesn't rise as we expand the model.

(Example) Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Our parameter vector $\theta = (\mu, \sigma^2)$ and our parameter space $\Omega = \mathbb{R} \times \mathbb{R}^+$. Our goal is to estimate μ. Our loss function is the rel. squared error loss, which is given by

$$L((\mu, \sigma^2), d) = \frac{(d-\mu)^2}{\sigma^2} \quad \leftarrow \quad (\text{c.f. } \underbrace{\tilde{L}((\mu, \sigma^2), d) = (\mu - d)^2}_{\text{unbounded risk,}})$$

We consider the submodel where $\sigma^2 = 1$, i.e. $\Omega_0 = \mathbb{R} \times \{1\}$ and our loss function reduces to $L((\mu, 1), d) = (d-\mu)^2$.

Recall that $\bar{X}_n$ is minimax for Ω_0 because $R((\mu, \sigma^2), \bar{X}) = \frac{1}{n}$ $\forall (\mu, \sigma^2) \in \Omega$

Thus the risk does not depend on σ^2.

$R((\mu, 1), \bar{X}_n) = R((\mu, \sigma^2), \bar{X}_n)$, we have that the maximum risks are equal. [i.e. $\sup_{\theta \in \Omega_0} R(\theta, \delta) = \sup_{\theta \in \Omega} R(\theta, \delta)$]. It follows from Lemma 5.1.15 that $\bar{X}_n$ is minimax on Ω.

Admissibility of minimax estimators.

1) Admissibility can give rise to minimaxity: If δ is admissible with constant risk, then δ is also minimax.

2) Minimaxity does not guarantee admissibility: It only ensures that the worst case risk is optimal.

(See also TPE 5.2.4)

(Example) Let $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ with σ^2 known, and θ is the estimand. Then the minimax estimator is $\bar{X}_n$ under the squared error loss and we would like to determine whether or not $\bar{X}_n$ is admissible.

Q: Let's try to answer a more general question: When is $a\bar{X}_n + b$, for $a, b \in \mathbb{R}$ (i.e. affine function of $\bar{X}_n$) admissible?

Case 1 $0 < a < 1$: In this case $a\bar{X}_n + b$ is a convex combination of $\bar{X}_n$ and b. By our previous results, it is a Bayes estimator with respect to some Gaussian prior on θ. Since we adopt the squared error loss function, which is strictly convex, the Bayes estimator is unique. By Theorem 5.2.4 of TPE (a unique Bayes est. is always admissible), $a\bar{X}_n + b$ is admissible.

Case 2 $a = 0$: In this case b is also a unique Bayes estimator w.r.t. a degenerate prior distribution with the unit mass at $\theta = b$. So, by Theorem 5.2.4, b is also admissible.

Case 3 $a = 1, b \neq 0$: In this case $\bar{X}_n + b$ is not admissible because it is dominated by $\bar{X}_n$. Note that $\bar{X}_n + b$ and $\bar{X}_n$ have the same variance, but $\bar{X}_n$ has a strictly smaller bias.

In general, the risk of $a\bar{X} + b$ is

$$E\left((a\bar{X}_n + b - \theta)^2\right) = E\left(\{a(\bar{X}_n - \theta) + b + \theta(a-1)\}^2\right) \quad (+a\theta - a\theta)$$

$$= \dots = \frac{a^2\sigma^2}{n} + \{b + \theta(a-1)\}^2 \qquad (*)$$

Case 4 $a > 1$: We have by $(*)$

$$E\left((a\bar{X}_n + b - \theta)^2\right) \geq \frac{a^2\sigma^2}{n} > \frac{\sigma^2}{n} = R(\theta, \bar{X}_n)$$

Hence, we conclude that $\bar{X}_n$ dominates $a\bar{X}_n + b$ when $a > 1$ in which case $a\bar{X}_n + b$ is inadmissible.

Case 5 $a < 0$: $E\left((a\bar{X}_n + b - \theta)^2\right) > \{b + \theta(a-1)\}^2$

$$= (a-1)^2\left(\theta + \frac{b}{a-1}\right)^2$$

$$> \left(\theta + \frac{b}{a-1}\right)^2, \quad \text{cf.} \quad \left(\theta - \left(\widetilde{-\frac{b}{a-1}}\right)\right)^2$$

est.

which is the risk of predicting with the constant $\frac{-b}{a-1}$. So, $-\frac{b}{a-1}$ dominates $a\bar{X}_n + b$ and therefore $a\bar{X}_n + b$ is again inadmissible.

Case 6 $a = 1, b = 0$: We shall proceed with a limiting Bayes argument.

Suppose $\bar{X}_n$ is inadmissible, then, WLOG, we let $\sigma^2 = 1$, we have $R(\theta, \bar{X}_n) = \frac{1}{n}$.

By our hypothesis, there must exist an estimator δ' such that $R(\theta, \delta') \leq 1/n$ for all θ and $R(\theta_0', \delta') < 1/n$ for at least one $\theta_0' \in \Omega$.

Because $R(\theta, \delta)$ is continuous in θ, there must exist $\varepsilon > 0$ and an interval (θ_0, θ_1) containing θ_0' such that

$$\rightarrow \quad R(\theta, \delta') < \frac{1}{n} - \varepsilon \quad \forall \theta \in (\theta_0, \theta_1). \qquad (\#)$$

Let r_τ' be the average risk of δ' w.r.t. the prior distribution $N(0, \tau^2)$ on θ.

(we used the same prior to prove that $\bar{X}_n$ was the limit of a Bayes estimator, and hence minimax. We did this by letting $\tau \to \infty$ (improper prior: $\pi(\theta) = 1 \ \forall \theta$).

Let r_τ be the average risk of a Bayes estimator δ_τ under the same prior.

Note that $\delta_\tau \neq \delta'$ because $R(\theta, \delta_\tau) \to \infty$ as $\theta \to \infty$, which is not consistent with $R(\theta, \delta') \leq 1/n \ \forall \theta \in \mathbb{R}$. So $r_\tau < r_\tau'$ because the Bayes estimator is unique almost surely with respect to the marginal distribution of θ.

We shall look into the following ratio, which we will show to become arbitrarily large and form a contradiction that

$r_\tau < r_{\tau'}$.

Using the fact of the Bayes risk r_τ we can write

$$\frac{\frac{1}{n} - r_{\tau'}'}{\frac{1}{n} - r_\tau} = \frac{\frac{1}{\sqrt{2\pi}\tau}\int_{-\infty}^{\infty}\left\{\frac{1}{n} - R(\theta,\delta')\right\}\exp\left(\frac{\theta^2}{2\tau^2}\right)d\theta}{\frac{1}{n} - \frac{1}{n+1/\tau^2}}$$

prior $N(0,\tau^2)$

Applying (#), we find:

$$\frac{\frac{1}{n} - r_{\tau'}'}{\frac{1}{n} - r_\tau} \geq \frac{\frac{1}{\sqrt{2\pi}\tau}\int_{\theta_0}^{\theta_1}\varepsilon e^{-\frac{\theta^2}{2\tau^2}}d\theta}{\frac{1}{n(n\tau^2+1)}}$$

$$= \frac{n(1+n\tau^2)}{\tau\sqrt{2\pi}}\varepsilon\int_{\theta_0}^{\theta_1} e^{-\frac{\theta^2}{2\tau^2}}d\theta \quad \rightsquigarrow (\theta_1 - \theta_0)$$

$\to \infty$

As $\tau \to \infty$, the first expression $\frac{n(1+n\tau^2)\varepsilon}{\tau\sqrt{2\pi}} \to \infty$ and since the integrand converges monotonically to 1, Lebesgue's MCT ensures that the integral approaches the positive quantity $\theta_1 - \theta_0$. So, for sufficiently large τ, we must have

$$\frac{\frac{1}{n} - r_{\tau'}'}{\frac{1}{n} - r_\tau} > 1. \iff r_{\tau'}' < r_\tau$$

However this is a contradiction because r_τ is the optimal avg. risk (since it is the Bayes risk). So our assumption that there was a dominating estimator δ' is wrong and in this case $a\bar{X}_n + b = \bar{X}_n$ is admissible. □

Simultaneous estimation

So far, we consider only one parameter.

(Example). Let $X_1, X_2, \ldots, X_p$ be independent with $X_i \sim N(\theta_i, \sigma^2)$ for $1 \leq i \leq p$. For the sake of simplicity / laziness, let $\sigma^2 = 1$. Our goal is to estimate $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_p)^T$ under the loss function $L(\underset{\sim}{\theta}, \underset{\sim}{d}) = \sum_{i=1}^p (d_i - \theta_i)^2$.

A natural estimator for θ is $X = (X_1, X_2, \ldots, X_p)^T$.

It can be shown that X is the UMRUE, the MLE, a generalized Bayes and a minimax estimator for θ.

So it would be natural for us to think that X is admissible.
HOWEVER, it turns out that this is not the case when $p \geq 3$.

When $p \geq 3$, X is dominated by the James-Stein estimator: (J-S)

$$\delta(\underset{\sim}{X}) = (\delta_1(\underset{\sim}{X}), \dots, \delta_p(\underset{\sim}{X})),$$

where $\delta_i(X) = \left(1 - \frac{p-2}{\|X\|_2^2}\right) X_i$.

Motivation for the J-S estimator:

View it from the empirical Bayes framework.

a) Suppose $\theta_i \overset{iid}{\sim} N(0, A)$, then the Bayes estimator for θ_i is

$$\delta_{A,i}(X) = \frac{X_i}{1 + \frac{1}{A}} = \left(1 - \frac{1}{A+1}\right) X_i \checkmark$$

b) We need to choose A.
Marginalising over θ, we can show that X has the distribution

$$X_i \overset{iid}{\sim} N(0, A+1) \leftarrow \int p(x;\theta)\,\pi(\theta;A)\,d\theta.$$

We may use X and the knowledge of this marginal distribution to find an estimate of $1/(A+1)$. Instead of using the MLE, we shall used an unbiased estimate:

it can be shown that $E\left(\frac{1}{\|X\|_2^2}\right) = \frac{1}{(p-2)(A+1)}$ (Exercise)

($\frac{1}{A+1}\|X\|_2^2$ follows a χ_p^2 distribution)

So, $1 - \frac{p-2}{\|X\|_2^2}$ is unbiased for $1 - \frac{1}{A+1}$.

↳ J-S: $\delta(X_i) = \left(1 - \frac{p-2}{\|X\|_2^2}\right) X_i$.

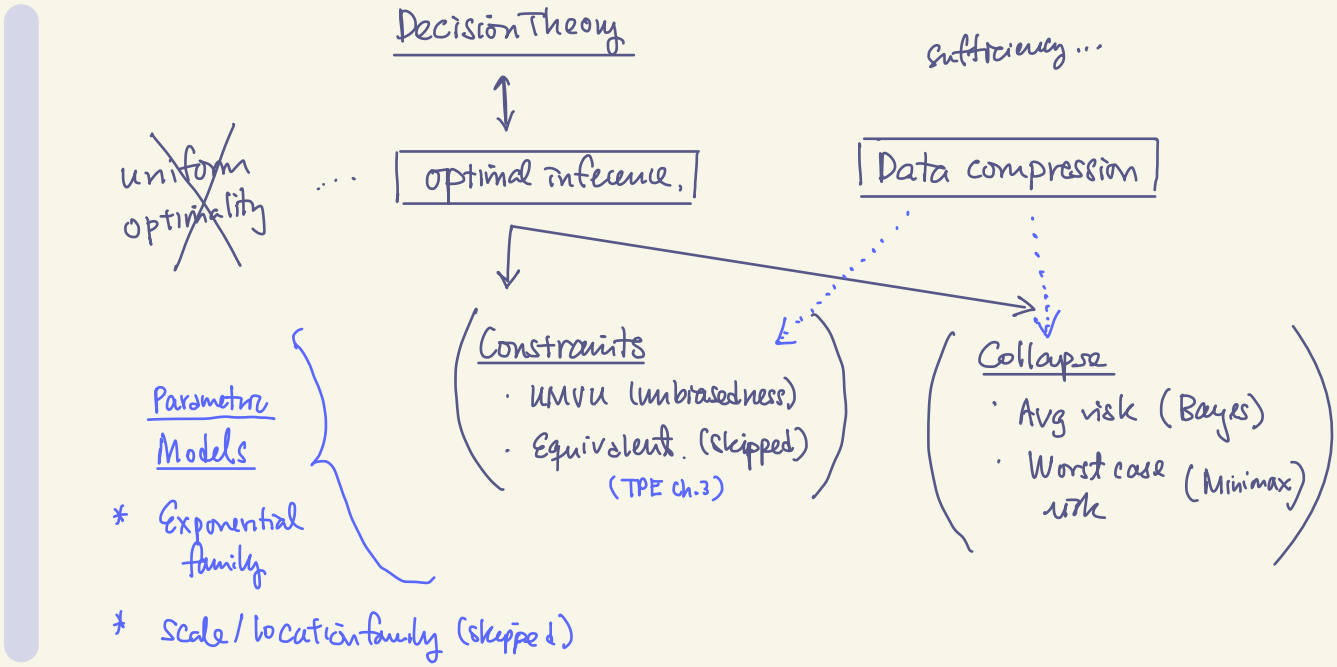
Note that:

$$E\left(\|X\|_2^2\right) = E\left(\sum_{i=1}^p X_i^2\right) = p + \sum_{i=1}^n \theta_i^2$$
$$= p + \|\theta\|_2^2 > p. \quad \text{PO}^2$$

⇒ We may view the J-S estimator as a method for correcting the bias in the size of X. (See TPE 5.5.1)

Keener 11.2

Point Estimation



↓ Testing (Decision problems)

· Simple vs simple; simple vs comp.; comp vs comp } Cox & Hinkley (74)

Model Setup

We assume that the data are sampled from $X \sim P_\theta$, where $P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Omega\}$.

We divide the models in $\mathcal{P}$ into two disjoint subclasses known as hypotheses:

H_0: $\theta \in \Omega_0 \subset \Omega$ (null hypothesis)

H_1: $\theta \in \Omega_1 \subset \Omega \setminus \Omega_0$ (alternative hypothesis)
(Ha)

Our goal is to infer (from the data) which hypothesis is "correct".

Our decision space is $\mathcal{D} = \{$accept H_0, reject $H_0\}$.
(do not reject)

Decision \ Truth	existing drug is better. $\theta \in \Omega_0$ (H_0 is correct)	new drug is better. $\theta \in \Omega_1$ (H_1 is correct)
Reject H_0	(α) Type I error	0 ✓
Accept H_0	0 ✓	(β) 1 - power = Type II error.

Clinical trials (sample size)

Loss function $L(\theta, d)$

Test function $\phi(X) \in [0,1]$. (critical function)

This test function ϕ indicates that $\delta_\phi(X, U)$ rejects H_0 with probability $\phi(X)$. i.e. (U: randomized component $\perp\!\!\!\perp X$.)

$$\phi(X) = P\left(\delta_\phi(X, U) = \text{Reject } H_0 \,\middle|\, X\right)$$

<u>Definition</u> The power function of a test ϕ is $\beta(\theta) = E_\theta(\phi(X))$

$$= P_\theta(\text{Reject } H_0)$$

If $\theta_0 \in \Omega_0$ then $\beta(\theta_0) = R(\theta_0, \delta_\phi)$ = Type I error

If $\theta_1 \in \Omega_1$, then $\beta(\theta_1) = 1 - R(\theta_1, \delta_\phi) = 1 -$ Type II error,

Our ideal goal is to minimize $\beta(\theta_0)$ uniformly for all $\theta \in \Omega_0$ and maximize $\beta(\theta_1)$ uniformly for all $\theta_1 \in \Omega_1$.

Lecture 11

21/Nov/2022.

Neyman-Pearson Paradigm

Fix $\alpha \in (0,1)$ to control the Type I error. (Level of significance)

We require that our procedure can satisfy the following risk bound:

$$\sup_{\theta_0 \in \Omega_0} E_{\theta_0}(\phi(X)) = \sup_{\theta_0 \in \Omega_0} \beta(\theta_0) \leq \alpha$$

↑ size of the test.

Optimality goal: Find a level α test that maximizes the power $\beta(\theta_1) = E_{\theta_1}(\phi(X))$ for each $\theta_1 \in \Omega_1$. Such a test is called uniformly most powerful test (UMP).

MP test for the "simple vs simple" case

Definition A hypothesis H_0 is called simple if $|\Omega_0| = 1$, otherwise it is called composite. The same is true for H_1.

For the simple versus simple case, we will adopt the notation:

$H_0: X \sim p_0$ (density corr. to P_{θ_0}) $\rightarrow \theta = \theta_0$

$H_1: X \sim p_1$ (density corr. to P_{θ_1}) $\rightarrow \theta = \theta_1$ $\quad \theta_0 \neq \theta_1$

Our goal can be compactly described as:

$$\max_{\phi} E_{p_1}(\phi(X))$$

such that $E_{p_0}(\phi(X)) \leq \alpha$. ($\alpha$: predefined value chosen by user)

[Lemma] (Neyman-Pearson)

(i) Existence. For testing $H_0: p_0$ vs $H_1: p_1$, there is a test $\phi(x)$ and a constant k such that

a) $E_{p_0}\phi(x) = \alpha$ (size = level)

b) $\phi(x) = \begin{cases} 1 & \text{if } \frac{p_1(x)}{p_0(x)} > k_\alpha \\ 0 & \text{if } \frac{p_1(x)}{p_0(x)} < k_\alpha \end{cases}$ — rejection region / critical

reject if $LR > k_\alpha$

[This test is called a likelihood ratio test]

(ii) Sufficiency. If a test satisfies (a) and (b) for some constant k_α, it is most powerful for testing $H_0: p_0$ vs $H_1: p_1$ at level α [ϕ is MP].

(iii) Necessity. If a test ϕ is MP at level α, then it satisfies (b) for some k_α and it also satisfies (a) unless there exists a test of size $< \alpha$ with power 1.

(Example). Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, with σ^2 known. Consider the following two hypotheses:

$H_0: \mu = 0$ vs $H_1: \mu = \mu_1$ ($\mu_1 > 0$) (known).

$$L(x) = \frac{p_1(x)}{p_0(x)} = \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu_1)^2}{2\sigma^2}\right\}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x_i^2}{2\sigma^2}\right\}} \quad \begin{matrix} -\mu=\mu_1 \\ -\mu=0. \end{matrix}$$

$$= \exp\left(\frac{1}{\sigma^2}\mu_1 \sum_{i=1}^n x_i - \frac{n\mu_1^2}{2\sigma^2}\right)$$

$\sum_{i=1}^n x_i$: suff. stat. ; $\frac{n\mu_1^2}{2\sigma^2}$: const.

We have the following observation:

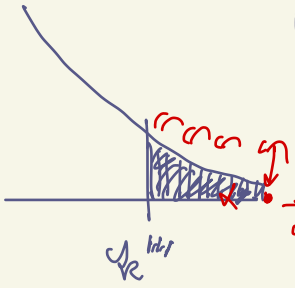
$$L(x) > k \iff \mu_1 \frac{\sum_{i=1}^n x_i}{\sigma^2} - \frac{n\mu_1^2}{2\sigma^2} > \log k$$

$$\Leftrightarrow \quad \mu_1 \sum_{i=1}^n X_i > k' \qquad \checkmark$$

$$\Leftrightarrow \quad \begin{cases} \sum_{i=1}^n X_i > k'' & \text{if } \mu_1 > 0 \\ \sum_{i=1}^n X_i < k'' & \text{if } \mu_1 < 0, \end{cases}$$

Let's focus on the first case where $\mu_1 > 0$. We can rewrite our test in a different form as follows:

$$\frac{\sqrt{n}\,\bar{X}_n}{\sigma} > k''' \qquad (\Sigma X_i)$$



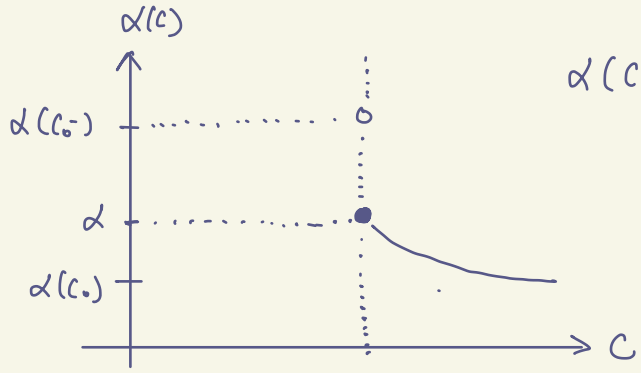
By NP Lemma, we reject H_0 iff $\sqrt{n}\bar{X}_n/\sigma > k_\alpha$ is MP where $k''' = k(\alpha) = z_{1-\alpha}$ is the $(1-\alpha)$ quantile of $N(0,1)$.

$$E_{P_0}\phi(X) = \alpha = P_{\mu=0}\left(\frac{\sqrt{n}\bar{X}_n}{\sigma} > k'''\right) \qquad \leq z_{1-\alpha}.$$

1.Ex

• Proof. Let $r(X) = \frac{p_1(X)}{p_0(X)}$ be the likelihood ratio and denote the cumulative distribution function of $r(X)$ under H_0 by F_0.

(i) Existence. Let $\alpha(c) = P_0(r(X) > c) = 1 - F_0(c)$. Then $\alpha(c)$ is a non-increasing, right continuous function of c (i.e. $\lim_{\varepsilon \downarrow 0} \alpha(c+\varepsilon) = \alpha(c)$). Note that $\alpha(c)$ is not necessarily left-continuous at every value of c, but the left-hand limits exist; denote them as $\alpha(c^-) = \lim_{\varepsilon \downarrow 0} \alpha(c-\varepsilon)$



$$\alpha(c_0) \leq \alpha \leq \alpha(c_0^-).$$

We define our test function to be

$$\phi(x) = \begin{cases} 1 & \text{if } r(x) > c_0 \\ \gamma & \text{if } r(x) = c_0 \\ 0 & \text{if } r(x) < c_0, \end{cases}$$

for some constant $\gamma \in (0,1)$. We note that ϕ only depends on x through $r(x)$. The test ϕ always rejects if the likelihood ratio strictly exceeds the threshold c_0 and never rejects it if the ratio falls strictly below c_0.

By construction, if we take $k = c_0$, then ϕ satisfies condition (b) of part (i). We choose γ such that ϕ also satisfies part (a). The size of ϕ is given by

Type I error. →

$$\begin{aligned} E_0(\phi(X)) &= P_0(r(X) > c_0) + \gamma P_0(r(X) = c_0) \\ &= \alpha(c_0) + \gamma \{\alpha(c_0^-) - \alpha(c_0)\} \end{aligned}$$

If $\alpha(c_0^-) = \alpha(c_0)$, then $\alpha(c_0) = \alpha$ and we automatically have $E_0(\phi(X)) = \alpha$ for any choice of γ.

otherwise, set $\gamma = \dfrac{|\alpha - \alpha(c_0)|}{\alpha(c_0^-) - \alpha(c_0)}$

that gives $E_0\phi(X) = \alpha$.

(ii) Let ϕ satisfy (i)(a) & (b), and let ϕ' be any other level α test so that $E_0(\phi'(X)) = \int \phi'(x) p_0(x) d\mu(x) \le \alpha$.

In order to show ϕ is most powerful, we bound the power difference $E_1(\phi(X)) - E_1(\phi'(X))$ ✓ from below by the size difference $E_0(\phi(X)) - E_0(\phi'(X))$ up to a constant multiple.

To do this, we claim that the following holds:

$$\int \{\phi(x) - \phi'(x)\}\{p_1(x) - k p_0(x)\} d\mu(x) \geq 0 . \qquad (*)$$

($\phi(x)$: +ve, $\phi'(x)$: -ve, $p_1(x) - k p_0(x)$: +ve, $-k p_0(x)$: -ve)

To see this, we consider the following three cases:

1) If $p_1(x) > k p_0(x)$, then $\phi(x) = 1$ because of our construction. Since $\phi'(x) \leq 1$, the integrand is non-negative.

2) If $p_1(x) < k p_0(x)$, then $\phi(x) = 0$ by construction. Since $\phi'(x) \geq 0$, the integrand is non-negative too.

3) If $p_1(x) = k p_0(x)$, then the integrand is zero.

By exhaustion, the inequality $(*)$ holds. Hence, we have

$$\int \{\phi(x) - \phi'(x)\} p_1(x) \, d\mu(x) \geq k \int \{\phi(x) - \phi'(x)\} p_0(x) \, d\mu(x)$$

$$= k\{E_0(\phi(x)) - E_0(\phi'(x))\} \geq 0$$

($E_0(\phi(x)) = \alpha$, $E_0(\phi'(x)) \leq \alpha$)

We conclude that $E_1(\phi(x)) \geq E_1(\phi'(x))$; ϕ is the most powerful at level α.

(iii) Necessity. Suppose ϕ^* is most powerful at level α, and let ϕ be a likelihood ratio test satisfying (i)(a) & (b). We have to show that $\phi^*(x) = \phi(x)$ except possibly when $p_1(x)/p_0(x) = k$ for μ-a.e. x. Define the sets

$$S^+ = \{x : \phi(x) > \phi^*(x)\}$$

$$S^- = \{x : \phi(x) < \phi^*(x)\}$$

$$S_0 = \{x : \phi(x) = \phi^*(x)\}$$

and $S = (S^+ \cup S^-) \cap \{x : p_1(x) \neq k p_0(x)\}$.

We want to show that $\mu(S) = 0$.

Suppose it is not the case, i.e. $\mu(S) > 0$. As in part (ii), we have $(\phi - \phi^*)(p_1 - k p_0) > 0$ on S. Therefore,

$$\int_X (\phi - \phi^*)(p_1 - k p_0)\, d\mu(x)$$
$$= \int_{S^+ \cup S^-} (\phi - \phi^*)(p_1 - k p_0)\, d\mu(x)$$
$$= \int_S (\phi - \phi^*)(p_1 - k p_0)\, d\mu(x) .$$

By the hypothesis, $E_0(\phi(X)) = \alpha$ and $E_0(\phi^*(X)) \le \alpha$, so the previous inequality implies

$$E_1(\phi(X)) - E_1(\phi^*(X)) > k\{E_0(\phi(X)) - E_0(\phi^*(X))\} \ge 0$$

that is $E_1(\phi(X)) > E_1(\phi^*(X))$. Hence contradiction.

Hence $\mu(S) = 0$.

It remains to show that the size of ϕ^* is α unless there exists a test of size strictly less that α and power 1. For this, note that if size $< \alpha$ and power < 1, we can add points to the rejection region until either the size is α or power is 1. □

Sample size

Definition For simple $H_0: p_0$ vs $H_1: p_1$, we call $\beta_\phi(P_1) = E_{P_1}(\phi(X))$ the power of ϕ, i.e. the prob. of rejecting H_0 under the alternative hypothesis.

[corollary] TSH 3.2.1 Suppose β is the power of a most powerful level α test of $H_0: P_0$ vs $H_1: P_1$ with $\alpha \in (0,1)$, then $\alpha < \beta$ (unless $P_0 = P_1$)

(This MP test rejects more often under H_1 than H_0)

· Proof. Consider the test $\phi_0(x) \equiv \alpha$, which always rejects with probability α. Since ϕ_0 is level α and β is the max power, we have

$$\beta \geq E_{P_1}(\phi_0(X)) = \alpha.$$

Suppose $\beta = \alpha$. Then $\phi_0(x) = \alpha$ is a most powerful level α test. As a result,

$$\phi_0(x) = \begin{cases} 1 & \text{if } p_1(x)/p_0(x) > k \\ 0 & \text{if } p_1(x)/p_0(x) < k. \end{cases}$$ by NP(iii) for same k

Since $\phi_0(x)$ never equals 0 or 1, it must be the case that $p_1(x) = k\, p_0(x)$ with prob. 1. Note that $\int p_1(x)\, d\mu(x) = k \int p_0(x)\, d\mu(x) = 1$, this implies that $k = 1$ and hence $P_0 = P_1$. □

Exponential Families and UMP one-sided tests

↑ uniformly most powerful

Goal: Extend the MP tests for simple alternatives up to UMP tests for composite alternatives.

(Example): (one-parameter exponential family). Consider $X_1, \ldots, X_n \overset{iid}{\sim} P_\theta$, where $p_\theta(x) \propto h(x) \exp(\theta T(x))$, and we are interested in testing

$$H_0: \theta = \theta_0 \quad \text{vs} \quad H_1: \theta = \theta_1$$

We want to construct an MP test at level α. The corr. likelihood ratio is

$$\frac{\prod_{i=1}^n p_{\theta_1}(x_i)}{\prod_{i=1}^n p_{\theta_0}(x_i)} \propto \exp\left\{ (\theta_1 - \theta_0) \sum_{i=1}^n T(x_i) \right\}$$

Assuming that $\theta_1 > \theta_0$, we shall reject for large value of $\sum_{i=1}^n T(x_i)$.

That is, an MP test has the following form:

H_1: $\theta_1 < \theta_0$

$$\phi(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n T(x_i) < k \\ \gamma & \text{if } \sum_{i=1}^n T(x_i) = k \\ 0 & \text{if } \sum_{i=1}^n T(x_i) > k \end{cases}$$

Where k and γ are chosen to satisfy the size constraint

$$\alpha = E_{\theta_0}(\phi(X)) = P_{\theta_0}\left(\sum_{i=1}^n T(X_i) < k\right) + \gamma P_{\theta_0}\left(\sum_{i=1}^n T(X_i) = k\right)$$

Note that $\sum_{i=1}^n T(X_i)$ has no explicit θ dependence and that k and γ do not depend on θ_1 (with $\theta_1 > \theta_0$). This means that ϕ is in fact a UMP test for testing

$$H_0: \theta = \theta_0 \quad \text{vs} \quad H_1: \theta > \theta_0. \qquad (\theta \neq \theta_0)$$

Hence H_1 is an example of a one-sided alternative, which arises when the parameter values of interest lie on only one side of the real-valued parameter θ_0.

Exponential family

Definition (Families with monotone likelihood ratio, MLR)

We say that the family of densities $\{p_\theta : \theta \in \mathbb{R}\}$ has monotone likelihood ratio in $T(X)$ if

(i) $\theta \neq \theta'$ implies $p_\theta \neq p_{\theta'}$ [identifiability].

(ii) $\theta < \theta'$ implies $p_{\theta'}(x) / p_\theta(x)$ is a non-decreasing function of $T(X)$ [monotonicity]

(Example)

1) $T(X) = \sum_{i=1}^n T(X_i)$ in the one-parameter exp. family.

2) (Double exponential). Let $X \sim$ Double exponential (θ) with density $p_\theta(x) = \frac{1}{2}\exp(-|x-\theta|)$. It is easy to see that the model is identifiable, so we need to check only the 2nd condition.

Fix any $\theta' > \theta$, and consider the likelihood ratio

$$\frac{p_{\theta'}(x)}{p_\theta(x)} = \exp\left(|x-\theta| - |x-\theta'|\right).$$

Note that

$$|x-\theta| - |x-\theta'| = \begin{cases} \theta - \theta' & \text{if } x < \theta \\ 2x - \theta - \theta' & \text{if } \theta \le x \le \theta' \\ \theta' - \theta & \text{if } x > \theta' \end{cases},$$

which is non-decreasing in x. Therefore, the family has MLR in $T(X) = X$.

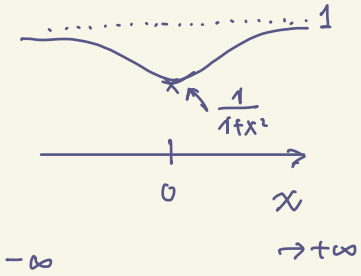
(Example) (Cauchy location model). Let X has density

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

We find two points for which the MLR cond. fails. For any fixed $\theta > 0$,

$$\frac{p_\theta(x)}{p_0(x)} = \frac{1+x^2}{1+(x-\theta)^2} \to 1 \quad \text{as } x \to \pm\infty.$$

But $p_\theta(0)/p_0(0) = \frac{1}{1+\theta^2} < 1$. Thus the ratio must increase at some values of x and decrease at others. In particular, it is not monotone in x. Hence, we conclude that the LR in $T(X) = X$ is not MLR.

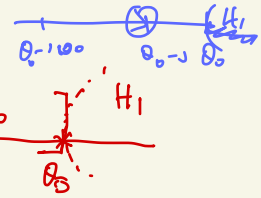


[THEOREM] TSH 3.4.1 Suppose $X \sim P_\theta$ has MLR in $T(X)$ and we test $H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$, then

$\theta = \theta_0 - 5$

$\theta = \theta_0 - 100$

(i) There exists a UMP test at level α of the form



$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > k \\ \gamma & \text{if } T(x) = k \\ 0 & \text{if } T(x) < k \end{cases}$$

where k, γ are determined by the condition $E_{\theta_0}(\phi(X)) = \alpha$

(ii) The power function $\beta(\theta) = E_\theta(\phi(X))$ is strictly increasing when $0 < \beta(\theta) < 1$.

Proof --> TSH.

Possible directions for constructing UMPs:

1) Reduce the composite alternative to a simple alternative

* 2) Collapse the composite null to a simple null.

3) Apply NP Lemma.

least favorable → optimal tests for composite nulls.

→ duality b/w CI & testing

→ p-values.

↘ ...

Lecture 12

28/Nov/2022.

Optimal Tests for Composite Nulls

H_0: $X \sim f_\theta$. $\theta \in \Omega_0$

H_1: $X \sim g$, where g is known.

We may impose a prior distribution Λ on Ω_0. So, we consider the new hypothesis

H_Λ: $X \sim h_\Lambda(x) = \int_{\Omega_0} f_\theta(x)\, d\Lambda(\theta)$, "simple"

where $h_\Lambda(x)$ is the marginal distribution of X induced by Λ.

Let β_Λ be the power of the MP level-α test Φ_Λ for testing H_Λ vs g.

(§3.8 of TSH)

Definition (Least favorable distribution) Λ is a least favorable distribution if $\beta_\Lambda \leq \beta_{\Lambda'}$ for any prior Λ'.

[THEOREM] TSH 3.8.1 Suppose Φ_Λ is an MP level α-test for testing H_Λ against g. If Φ_Λ is level-α for the original hypothesis H_0 (i.e. $E_{\theta_0}(\Phi_\Lambda) \leq \alpha \ \forall\, \theta_0 \in \Omega_0$), then

1) The test Φ_Λ is MP for the original test on $H_0: \theta \in \Omega_0$ vs g
2) The prior distribution Λ is least favorable.

· Proof. Let ϕ^* be any other level α test of $H_0: \theta \in \Omega_0$ vs g. Then ϕ^* is also a level-α test for H_Λ vs g because

$$E_\theta(\phi^*(X)) = \int \phi^*(x) f_\theta(x)\, d\mu(x) \leq \alpha \qquad \forall \theta \in \Omega_0$$

which implies that

$$\int \phi^*(x)\, h_\Lambda(x)\, d\mu(x) = \iint \phi^*(x) f_\theta(x)\, d\mu(x)\, d\Lambda(\theta) \leq \int \alpha\, d\Lambda(\theta) = \alpha$$

Since ϕ_Λ is MP for H_Λ vs g. we have

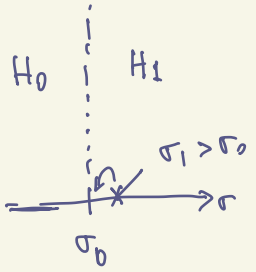
$$\int \phi^* g(x)\, d\mu(x) \le \int \phi_\Lambda(x) g(x)\, d\mu(x)$$

Hence ϕ_Λ is a MP test for H_0 vs g because ϕ_Λ is also level α.

2. Let Λ' be any distribution on Ω_0. Since $E_\theta(\phi_\Lambda(X)) \le \alpha \ \forall \theta \in \Omega_0$, we know that ϕ_Λ must be level-α for $H_{\Lambda'}$ vs g. Thus $\beta_\Lambda \le \beta_{\Lambda'}$. So Λ is the least favorable distribution. □

(Example) (Testing in the presence of missance parameters) Let $X_1, \dots, X_n$ be iid $N(\theta, \sigma^2)$ where both θ and σ^2 are unknown. We consider testing $H_0: \sigma \le \sigma_0$ against $H_1: \sigma > \sigma_0$.

To find a UMP test, we follow the strategy discussed in the previous lecture:



1) First, we fix a single alternative (θ_1, σ_1) for some arbitrary θ_1 and $\sigma_1 > \sigma_0$

2) We choose a prior distribution Λ to collapse our null hypothesis over. Intuitively, the least favorable prior should make the alternative hypothesis hard to distinguish.

One sensible choice is concentrating Λ on the boundary between H_1 and H_0 i.e. the line $\{\sigma = \sigma_0\}$. Thus Λ will be a probability distribution over $\theta \in \mathbb{R}$ for the fixed $\sigma = \sigma_0$.

Given any test function $\phi(x)$ and a sufficient statistic T, there exists a test function η that has the same power as ϕ but depends on x only through T:

$$\eta(T(x)) = E\left(\phi(X) \mid T(X)\right).$$

Hence, we can restrict our attention to the sufficient statistic (Y, U), where $Y = \bar{X}$ and $U = \sum_{i=1}^n (X_i - \bar{X})^2$. We know that $Y \sim N(\theta, \sigma^2/n)$, $U \sim \sigma^2 \chi^2_{n-1}$ and Y is independent of U by Basu's theorem.

Thus, for Λ supported on $\sigma = \sigma_0$, we have the joint density of (Y, U) under H_Λ as

$$C_0 u^{\frac{n-3}{2}} \exp\left(-\frac{u}{2\sigma_0^2}\right) \int \exp\left(-\frac{n}{2\sigma_0^2}(y-\theta)^2\right) d\Lambda(\theta) \quad (*)$$

and the joint density under the alternative hypothesis (θ_1, σ_1) as

$$C_0 u^{\frac{n-3}{2}} \exp\left(-\frac{u}{2\sigma_1^2}\right) \exp\left(-\frac{n}{2\sigma_1^2}(y-\theta_1)\right)^2. \quad (**)$$

We see from (*) and (**) that the choice of Λ only affects the distribution of Y. To achieve the maximum power against the alternative, we need to choose Λ such that the two distributions become as close as possible.

Under H_1: $Y \sim N\left(\theta_1, \frac{\sigma_1^2}{n}\right)$. Under H_Λ, the distribution of Y is in a convolution form: $Y = Z + \textcircled{H}$ for $Z \sim N\left(0, \frac{\sigma_0^2}{n}\right)$ and $\textcircled{H} \sim \Lambda$, where $\textcircled{H}$ and Z are independent.

Hence, if we choose $\textcircled{H} \sim N\left(\theta_1, \frac{\sigma_1^2 - \sigma_0^2}{n}\right)$, Y will have the same distribution under the null and the alternative, which is $N\left(\theta_1, \frac{\sigma_1^2}{n}\right)$. Under this choice of prior, the LRT rejects for large values $\exp\left(-\frac{u}{2\sigma_1^2} + \frac{u}{2\sigma_0^2}\right)$, i.e. it rejects for large values of u (since $\sigma_1 > \sigma_0$). So, the MP test rejects H_Λ if $\sum_{i=1}^n (X_i - \bar{X})^2$ lies above some threshold determined by the size constraint. In particular, it rejects if $\sum_{i=1}^n (X_i - \bar{X})^2 > \sigma_0^2 C_{n-1, 1-\alpha}$, where $C_{n-1,1-\alpha}$ is the $(1-\alpha)^{th}$ quantile of χ^2_{n-1}.

3) Next, we check if the MP test is level α for the composite null. For any (θ, σ) with $\sigma \le \sigma_0$, the prob. of rejection is

$$P_{\theta,\sigma}\left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} > \frac{\sigma_0^2 C_{n-1,1-\alpha}}{\sigma^2}\right) = P\left(\chi^2_{n-1} > \frac{\sigma_0^2}{\sigma^2} C_{n-1,1-\alpha}\right)$$

while the equality holds iff $\sigma = \sigma_0$ $\le \alpha$

Hence from TSH 3.8.1 that our test is MP for testing the original null H_0 vs $N(\theta_1, \sigma_1)$.

4) Finally, the MP level α test for testing the composite null H_0 versus an arbitrarily chosen (θ_1, σ_1) does not depend on the choice of (θ_1, σ_1). Hence, it is UMP for testing the original composite null versus the composite alternative. ||

Duality between Testing and Interval Estimation.

CI { Recall that a random set $S(X)$ is a $1-\alpha$ confidence region (interval) for a parameter $\xi = \xi(\theta)$ if

$$P_\theta(\xi \in S(X)) \geq 1-\alpha \quad \forall \theta \in \Omega$$

For every ξ_0, let $A(\xi_0)$ be the acceptance region for non-randomised level α test of $H_0: \xi(\theta) = \xi_0$ versus $H_1: \xi(\theta) \neq \xi_0$ so that

$$P_\theta\left(X \in A(\xi(\theta))\right) \geq 1-\alpha, \quad \forall \theta \in \Omega$$

Define $$S(X) = \{\xi : X \in A(\xi)\}$$

then $\xi(\theta) \in S(X)$ if and only if $X \in A(\xi(\theta))$, so

$$P_\theta(\xi(\theta) \in S(X)) = P_\theta\left(X \in A(\xi(\theta))\right) \geq 1-\alpha.$$

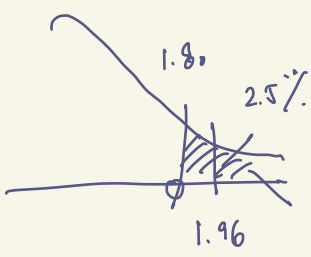
This shows that $S(X)$ is a $1-\alpha$ confidence interval for ξ.

§3.3 TSH.

p-values

For varying α, the resulting tests provide an example of the typical situation in which the rejection regions S_α are nested in the sense that

$$S_\alpha \subset S_{\alpha'} \text{ if } \alpha < \alpha'.$$



the quantity

$$\hat{p} = \hat{p}(X) = \inf\{\alpha : X \in S_\alpha\}$$

↑ Smallest significant level

is called a p-value.

It gives an idea of how strongly the data contradict the null hypothesis.

(Example) Let Φ denote the standard normal cdf. Then the rejection region can be written as

$$S_\alpha = \left\{ X : X > \sigma z_{1-\alpha} \right\} = \left\{ X : \Phi\left(\frac{X}{\sigma}\right) > 1-\alpha \right\} = \left\{ X : 1-\Phi\left(\frac{X}{\sigma}\right) < \alpha \right\}.$$

For a given observed value of X, the inf over all α where the last inequality holds is $\hat{p} = 1 - \Phi\left(\frac{X}{\sigma}\right)$.

Alternatively, the p-value is $P_0\left(X \geq x\right)$, where x is the observed value of X. Note that under $\xi = 0$, the distribution of $\hat{p}$ is given by

$$P_0\left(\hat{p} \leq u\right) = P_0\left(1 - \Phi\left(\frac{X}{\sigma}\right) \leq u\right) = P_0\left(\Phi\left(\frac{X}{\sigma}\right) \geq 1-u\right) = u.$$

$\Rightarrow \hat{p}$ is uniformly distributed on $(0,1)$. ||

(Lemma 3.3.1 of TSH + Example 3.3.2)

discrete

↑ EOF

Things to know..

UMP X

→ impose a reasonable restriction on the tests to be considered and the "optimal" tests within the class of tests under the restriction.

→ Two typical strategies:
① unbiasedness → UMP unbiased test.
② invariance.

Recall that a UMP test ϕ of size α has the property that

$$E_\theta(\phi(X)) \leq \alpha \quad \theta \in \Omega_0 \quad \text{and} \quad E_\theta(\phi(X)) \geq \alpha \quad \theta \in \Omega_1. \quad (\dagger)$$

This means that ϕ is at least as good as the naïve test $\phi \equiv \alpha$

.

Definition Let α be a given level of significance. A test ϕ for $H_0: \theta \in \Omega_0$ versus $H_1: \theta \in \Omega_1$ is said to be unbiased of level α if and only if (†) holds. A test of size α is called a uniformly most powerful unbiased (UMPU) test if and only if it is UMP within the class of unbiased tests of level α.

UMP is UMPU

uniform. (θ, θ) $\theta = \theta_0$, $\theta \neq \theta_0$ $\{\theta > \theta_0 \rightarrow$ $\theta < \theta_0 \leftarrow$

(See Keener § 12.7, Thm 12.26 & Example 12.29) — Exponential dist.

two-sided

$H_0: \theta = 1$ vs $H_1: \theta \neq 1$.

Final Exam

12/Dec. 12:30 – 15:00 (same classroom)

Coverage: Everything

combining "2" cheat sheets

Consistency / Asymptotic Normality of MLE

(Keener ch. 9) → EE, EM algo.

$\{\tilde{\theta}_n\}$ for $\theta \in \Omega$.

weakly consistent : $\tilde{\theta}_n \xrightarrow{P} \theta_0$

strongly consistent : $\tilde{\theta}_n \xrightarrow{a.s.} \theta_0$

Let $X_1, \ldots, X_n$ be iid with density $f(x|\theta)$ w.r.t. some σ-finite measure ν, where $\theta \in \Omega$.

The likelihood function: $L_n(\theta) = L_n(\theta \mid x_1, \ldots, x_n) \overset{iid}{=} \prod_{i=1}^{n} f(x_i \mid \theta)$

A MLE of θ (denoted by $\hat{\theta}$) is defined as $\hat{\theta} = \underset{\theta \in \Omega}{\operatorname{argmax}} L_n(\theta)$.

Kullback-Leibler information number is defined as

$$K(f_0, f_1) = E_0\left(\log \frac{f_0(X)}{f_1(X)}\right) = \int \log \frac{f_0(x)}{f_1(x)} f_0(x)\, d\nu(x)$$

[Lemma] (Shannon-Kolmogorov Information Inequality). Let $f_0(x)$ and $f_1(x)$ be densities w.r.t. ν. Then

$$K(f_0, f_1) = E_0\left(\log \frac{f_0(x)}{f_1(x)}\right) \geq 0,$$

with equality holds if and only if $f_1(x) = f_0(x)$ (a.e. $d\nu$).

(Proof: skipped) → use Jensen's inequality.

Let θ_0 denote the true value of θ. Then, the MLE is the value of θ that maximises

$$\ell_n(\theta) - \ell_n(\theta_0) = \log L_n(\theta) - \log L_n(\theta_0)$$
$$= \sum_{j=1}^{n} \left(\log f(X_j|\theta) - \log f(X_j|\theta_0)\right).$$

By the SLLN and the lemma (SK)

$$\frac{1}{n} \log \frac{L_n(\theta)}{L_n(\theta_0)} = \frac{1}{n} \sum_{j=1}^{n} \log \frac{f(X_j|\theta)}{f(X_j|\theta_0)} \xrightarrow{a.s.} E_{\theta_0} \log \frac{f(X|\theta)}{f(X|\theta_0)} = -K(\theta_0, \theta) \leq 0$$

unless $f(x|\theta) = f(x|\theta_0)$.

→ $\frac{L_n(\theta)}{L_n(\theta_0)}$ (LR) converges exp. fast to zero at the rate $\exp(-nK(\theta_0, \theta))$.

This implies already that if Ω is finite, the MLE is strongly consistent.

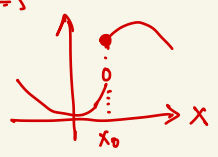
[THEOREM] Let $X_1, \ldots, X_n$ be iid with density $f(x|\theta)$, $\theta \in \Omega$, and let θ_0 denote the true value of θ. If

(1) Ω is compact

(2) $f(x|\theta)$ is upper semicontinuous in θ for all x

(3) there exists a function K such that $E_{\theta_0}|K(X)| < \infty$

A fn $f: X \to \overline{\mathbb{R}}$ is called USC at x_0 if for every real $y > f(x_0)$, there exists a neighbourhood U of x_0 such that $f(x) < y \ \forall x \in U$
($\Leftrightarrow$ $\limsup_{x \to x_0} f(x) \leq f(x_0)$)



and $u(x,\theta) = \log f(x|\theta) - \log f(x|\theta_0) \le k(x)$, for all x & θ

(4) for all $\theta \in \Omega$ and sufficiently small $\epsilon > 0$

$\sup_{|\theta'-\theta|<\epsilon} f(x|\theta')$ is measurable in x

(5) (identifiability) $f(x|\theta) = f(x|\theta_0)$ (a.e. $d\nu$) $\Rightarrow \theta = \theta_0$.

then, $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$. (See proof in Keener for example)...

AN.:

$\sqrt{n}$ CLT.

$\Psi(x,\theta) = \left(\frac{\partial}{\partial\theta} \log f(\theta|x)\right)^T$

iid

$\sqrt{n}$

$u(x,\hat{\theta}) = u(x,\theta_0) + \Psi(x,\theta_0)^T(\hat{\theta}_{MLE} - \theta_0)$

$+ (\theta-\theta_0)^T \int_0^1 \int_0^1 \lambda \dot{\Psi}(x, \theta_0 + \lambda\mu(\theta-\theta_0))\, d\lambda\, d\mu(\theta-\theta_0)$

remainder term...

$\Psi^{-1}(\cdot)$ $\left[\frac{1}{n}\sum \log f(x_i|\theta)\right] = \sqrt{n}(\hat{\theta} - \theta_0)$

CLT. N.

Normal.

near 0.

$f(x) = f(x_0) + f'(x_0)(x-x_0) + \dots$

$N(0, I(\theta_0)^{-1})$ (Cramér) $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$

Suff. ...

Estimation

↳ unbiasedness (constraint)

↳ avg risk optimality.

↳ worst-case optimality.

$\sup_x |\hat{F}_n(x) - F(x)| \to 0$

Testing

↳ MP (NP Lemma)

↳ UMP

↳ Least favorable prin.

↳ unbiased test

↳ CI, p-values.

finite-sample