STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: XU Weiwei

Lecture #1
07 September 2020

Different disciplines:
- Networks, images, recordings, ...
- Economics, Finance, ...
- Biology (Genetic), medical, earth science, ...

$\Big\} \rightarrow$ Data $\overset{\text{Statistics}}{\longrightarrow}$ Inference/Conclusions

Question Concerned:

- Modelling How to build a mathematical/statistical model that captures the *uncertainty* in our data?
- Methodology tools that allows us to deduce these statistical conclusions.
- Analysis: Optimal inference

  e.g. Mode: sample mean, sample median.

  * Finite sample optionality
  1. Given observations $(x_1, x_2, ..., x_n)$.
  2. Asymptotic properties $(n \to \infty)$.

1 Decision Theory (Wald, 1939)

Random element X takes values in a sample space χ. X can also be a vector (or matrix).

$$(X_1, X_2, ..., X_n) \overset{i.i.d}{\sim} f \text{ (distribution)}$$

I
- A statistical model is a family of distribution $\mathbb{P}$ indexed by a parameter θ. we denote

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

  where θ is the parameter, $\Omega \in \mathbb{R}^k$ is the parameter space and P_θ is a distribution.
- We assume that the data X come from some $P_\theta \in \mathbb{P}$ but the true value of θ is unknown.

Example 1 *Observe a sequence of coin flips $x_1, x_2, ..., x_n \in \{0, 1\}$. The objective is to estimate the probability of heads given the observations. (with 1 denotes a head). One can write*

$$\mathbb{P} = \Big\{ \textit{Bernoulli}(\theta) : \theta \in [0, 1] \triangleq \Omega \Big\}$$

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

Estimating Procedure:

$$\left.\begin{array}{r}\text{Estimator} \overset{\text{Observations}}{\longrightarrow} \text{Estimates} \\ +\text{Testing}\end{array}\right\} \rightarrow \text{Inference}$$

II A Decision Procedure

δ (estimator) is a map from χ to the decision space $\mathbb{D}$.

Example: Take $\mathbb{P} = \{\text{Bernoulli}(\theta)\}$ as shown above, we may be interested in estimating θ or testing θ based on:

(a) Estimating θ

The decision space is $\mathbb{D} = [0, 1]$ and the decision procedure might be

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

(b) Hypothesis Testing

Accepting/rejecting the hypothesis that $\theta = 0.5$; The corresponding decision space is $\mathbb{D} = \{\text{accept}, \text{reject}\}$, one possible decision procedure is

$$\delta(X) = \text{“Reject if } \bar{X}_n > 0.5'' \text{ and accept otherwise.}$$

(c) A loss function is a mapping $L : \Omega \times \mathbb{D} \rightarrow \mathbb{R}^+$, $L(\theta, d)$ represents the penalty for making the decision d when θ is in fact the true parameter for the distribution generating the data.

Example 2 *: [Squared-error Loss] For estimating θ with decision $d \in \mathbb{R} = \mathbb{D}$, a common loss function is the squared-error loss: $L(\theta, d) = (\theta - d)^2$.*

Risk Function:

$$\text{Average loss incurred} \rightarrow \; R(\theta, \delta) = E_\theta(L(\theta, \delta(X)))$$

Admissiability:

δ is inadmissible if there exists δ' such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all θ and $R(\theta', \delta') < R(\theta', \delta)$ for some θ'.

STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Ali Choo and Hon Ku To

Lecture 2
14 Sep 2020

2 10 Ways of Viewing a Random Variable

Define probability space (χ, ϕ, p) where

χ : Sample space with element ω
ϕ : $\sigma-$algebra with element
P : Probability measure which assigns probabilities to elements of $\mathcal{A}$ which satisfy

(i) $0 \leq p(A) \leq 1$.

(ii) $p(\chi) = 1$

(iii) If the element are disjoint i.e. $A_i \bigcap A_j = \{\emptyset\}$ for $i \neq j$, then

$$P(\bigcup_{i=1}^{n} A_i) = P(A_1 \bigcup A_2 ... \bigcup A_n) = \sum_{i=1}^{n} P(A_i).$$

2.1 Way #1. Random Variable

A function $X : \chi \to R$ such that image $X^{-1}(B)$ of any Borel set or elements of $\mathcal{A}$ is called a random variable. A $p-$tipple of r.v's is called random vector.

2.2 Way #2. Distribution Function

Associated with a random vector X on $(\chi, \mathcal{A}, P)$ is a distribution function $d.f.$: $F(\chi) = F_{x_1,...x_p}(x_1, x_2, ..., x_p) = P(\omega : X_1(\omega) \leq x_1, ..., X_p(\omega) \leq \chi_p)$ Note that, F is right-continuous with left limits (RCLL) [or càdlàg as in "*continue à droite, limite à gauche*"].

2.3 Way #3. τ^{th} Quantile $(0 < \tau < 1)$

For any scalar r.v X with $d.f.$ F, the quantity $\theta(\tau) = F^{-1}(\tau) = \inf\{x : F(x) \geq \tau\}$, $\tau \in (0, 1)$ is called the τ^{th} quartile of X or F. Specifically for $\tau = 1/2, \theta(1/2)$: Median, $\theta(1/4)$: Lower quartile, $\theta(3/4)$:upper quartile.

2.4 Way #4. Density Function

If the $d.f.F$ is absolutely continues with respect to the measure μ then F has a density function w.r.t μ. We interested in case where μ' is the league measure in which case can write $F(\chi) = \int_{-\infty}^{x} f(t)dt$, $f(t) = F'(t) = \partial F(t)/\partial t$.

Theorem 1 (Radon-Nikodym)**.** *If a finite measure P is absolute continuous w.r.t. a σ finite measure μ, then there exists a non-negative measurable function f such that*

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

This specific function f is called the Radon-Nikodym derivative of P w.r.t. μ (the density of p w.r.t. μ) denoted as $f = dp/d\mu$.

2.5 Way #5. Expectation

$$\begin{aligned} E(X) &= \int X(\omega) dp(\omega) = \int_{-\infty}^{\infty} x dF(x) = \int_{-\infty}^{\infty} x f(x)\, dx \\ E(aX + bY) &= aE(X) + bE(Y). \end{aligned}$$

2.6 Way #6: Moments

The kth central moment of a random variable X is

$$\mu_k = E(\{X - E(X)\}^k), \quad k = 1, 2 \ldots$$

In particular, μ_k for the cases $k = 2, 3, 4$ are closely related to the variance, skewness and kurtosis of X respectively as follows:

$$\begin{aligned} var(X) &= \mu_2. \\ \text{Skewness}(X) &= \mu_3/\sigma^3 \quad \text{, which measures the symmetry of } X. \\ \text{Kurtosis}(X) &= \mu_4/\sigma^4 \quad \text{, which measures the peakedness and tail length of } X. \end{aligned}$$

2.7 Way #7: Moment Generating Function (MGF)

The moment generating function of X is

$$m_X(t) = E(e^{tX}) = \int e^{tX}\, dF(x), \quad t \in \mathbb{R}.$$

When $m_X(t)$ and its derivatives exist in some neighbourhood of 0, we have

$$E(X^k) = \underbrace{m_X^{(k)}(0)}_{\text{the } k\text{th derivative of } m_X \text{ with respect to } t}, \quad k = 0, 1, 2, \ldots$$

Properties:

1. $m_{\mu+\sigma X}(t) = e^{\mu t} m_X(\sigma t)$.
2. $m_{X+Y}(t) = m_X(t) m_Y(t)$ if X and Y are independent.

Illustration

Suppose we have a discrete random variable on $\{0, 1, 2, \ldots\}$ with $\text{pr}(X = j) = a_j$, where $\text{pr}(X = j)$ is the probability mass function of X.

Define the "generating function" of X as

$$g(z) = \sum_{j=0}^{\infty} a_j z^j.$$

Since $\sum_{j=0}^{\infty} a_j = 1$, $|g(z)| \leq \sum_{j=0}^{\infty} |a_j||z|^j \leq \sum_{j=0}^{\infty} a_j = 1$ for any $|z| \leq 1$.

Consider the following derivatives:

$$\begin{aligned}
g'(z) = a_1 + 2a_2 z + 3a_3 z^2 + \cdots &= \sum_{j=1}^{\infty} j a_j z^{j-1}, \\
g''(z) = 2a_2 + 6a_3 z + \cdots &= \sum_{j=2}^{\infty} j(j-1) a_j z^{j-2}, \\
&\vdots \\
g^{(k)}(z) &= \sum_{j=k}^{\infty} \binom{j}{k} k! a_j z^{j-k}.
\end{aligned}$$

Thus

$$g^{(k)}(0) = k! a_k \quad \text{or} \quad a_k = (k!)^{-1} g^{(k)}(0).$$

So, all the information about a_k's are "contained" within the function g and is made accessible by simply differentiating it (repeatedly) and evaluating it at 0.

This means that the distribution of a non-negative integer valued random variable is uniquely defined by its generating function.

Restricting the absolute value of X between 0 and 1 can be quite restrictive.

Write $E(z^X) = E(e^{-\lambda X}), 0 \leq \lambda < \infty$.

So in the previous case,

$$E(e^{-\lambda X}) = \sum_{j=0}^{\infty} a_j e^{-\lambda x_j} = \begin{cases} \text{(discrete case)} \sum_j p_j e^{-\lambda x_j}, \\ \text{(continuous case)} \int e^{-\lambda u} f(u)\, du, \end{cases}$$

where x_j's are all possible values of X.

This formulation is the Laplace transform of X.

Example (c.f. Casella and Berger (2002) E.g. 2.3.10: Non-unique Moments)

Consider two probability density functions given by

$$\begin{aligned}
f_1(x) &= \frac{1}{\sqrt{2\pi} x} e^{-(\log x)^2/2}, \ 0 \leq x < \infty, \\
f_2(x) &= f_1(x)\{1 + \sin(2\pi \log x)\}, \ 0 \leq x < \infty.
\end{aligned}$$

(f_1 is the probability density function of a lognormal distribution.)

It can be shown that if $X_1 \sim f_1(x)$,

$$E(X_1^r) = e^{r^2/2}, \quad r = 0, 1, \ldots$$

Suppose $X_2 \sim f_2(x)$. We have

$$E(X_2^r) = \int_0^\infty x^r f_1(x)\{1 + \sin(2\pi \log x)\}\, dx = E(X_1^r) + \int_0^\infty x^r f_1(x) \sin(2\pi \log x)\, dx.$$

Consider the transformation: $y = \log x - r$. You can show that the transformed integral is an odd function over $(-\infty, \infty)$.

Hence $\int_0^\infty x^r f_1(x) \sin(2\pi \log x)\, dx = 0$ and $E(X_1^r) = E(X_2^r)$ for $r = 0, 1, \ldots$

Even though X_1 and X_2 have the same moments for all r, their probability density functions are different.

2.8 Way #8: Characteristic functions

The characteristic function of X is

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

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

For multivariate case,

$$\phi_X(t) = E(e^{it^T X}),$$

where $t = (t_1, \ldots, t_p)^T$, $X = (X_1, \ldots, X_p)^T$.

Existence: $|E(e^{itX})| \le E|e^{itX}| = E|\cos(tX) + i \sin(tX)| = E(\{\cos^2(tX) + \sin^2(tX)\}^{1/2}) = 1$.

(Because $|a + ib|^2 = (a + ib)(a - ib) = a^2 + b^2$).

Inversion Formula (See, for example, Billingsley, 1995)

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t)\, dt, \\ F_X(x) - F_X(y) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itx} - e^{-ity}}{-it} \phi_X(t)\, dt \quad \text{for points of continuity of } F \ x \text{ and } y. \end{aligned}$$

The inversion formula provides a correspondence between F (or f) and ϕ.

Any characteristic function is bounded by 1 (shown above) and is a uniformly continuous function on $\mathbb{R}^{(p)}$. [Exercise]

Theorem 2 (Uniqueness)**.** *Let X and Y be random k-vectors.*

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

(ii) If $m_X(t) = m_Y(t) < \infty$ for all t in a neighbourhood of 0, then $F_X = F_Y$. (c.f. Casella and Berger, 2002 Theorem 2.3.11)

Proof. (i) For any $a = (a_1, \ldots, a_k)^T \in \mathbb{R}^k$, $b = (b_1, \ldots, b_k)^T \in \mathbb{R}^k$, and $(a, b] = (a_1, b_1] \times \cdots \times (a_k, b_k]$ satisfying $\text{pr}_X(\text{the boundary of } (a, b]) = 0$,

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

(ii) (See next lecture's note)

□

References

Billingsley, P. (1995), *Probability and measure*, A Wiley-Interscience publication, Wiley, 3rd ed.

Casella, G. and Berger, R. L. (2002), *Statistical inference*, Pacific Grove, Calif.]: Duxbury/Thomson Learning, 2nd ed.

STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Yudan Zou and Tom Lee Cheuk Lam

Lecture 3

Recap last lecture

Proof of (ii): First, we consider the case $k = 1$. Since $e^{s|x|} \leq e^{sx} + e^{-sx}$, we conclude that $|X|$ has an mgf that is finite in the neighborhood of 0, say $(-c, c)$ for once $c > 0$.
By using the inequality:

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

We can write

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

which imples that for any $t \in \mathbb{R}$,

$$\phi_X(t+a) = \sum_{j=0}^{\infty} \frac{\phi_X^{(j)}(t)}{j!} a^j, \quad \text{for } |a| < c. \tag{*}$$

Similarly,(*) also holds for Y. That is, $\phi_Y(t+a) = \sum_{j=0}^{\infty} \{\phi_Y^{(j)}(t) a^j / j!\}$. Under the assumption that $m_X = m_Y < \infty$ in a neighbourhood of 0, X and Y have the same moment of all orders. Since $\phi_X^{(j)}(0) = \phi_Y^{(j)}(0)$ for all $j = 1, 2, ...$, which and * with $t = 0$ imply that ϕ_X and ϕ_Y are the same on the interval $(-c, c)$ and have the identical derivatives there.

Consider $t = c - \epsilon$ and $-c + \epsilon$ for an arbitrary small $\epsilon > 0$ in * and the result will follow in that ϕ_X and ϕ_Y will also agree on $(-2c + \epsilon, 2c - \epsilon)$ and hence on $(-2c, 2c)$. By 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 general case of $k > 2$, if $F_X \neq F_Y$, then part(i) concludes that there exists $t \in \mathbb{R}$ such that $\phi_X \neq \phi_Y$. Then $\phi_{t^T X}(1) \neq \phi_{t^T Y}(1)$, which implies that $F_{t^T X} \neq F_{t^T Y}$.But $m_X = m_Y < \infty$ in a neighborhood of $0 \in \mathbb{R}$ and by the result for $k = 1, F_{t^T X} = F_{t^T Y}$, this shows that $F_X = F_Y$.

2 10 ways of viewing a random variable (Cont'd)

2.9 Way # 9: Conditional probability

In our undergraduate study,

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

where by convention $P(B \mid A) = 0$ when $P(A) = 0$. But the definition breaks down for uncountable χ. If $\nu \ll \mu$, then there exists a non-negative function φ such that

$$\nu(A) = \int_A \varphi d\,\mu, \quad \text{for any } A \in \mathcal{A}.$$

For example, we have (X, Y) with joint density $f(x, y)$ and X with marginal density $g(x)$, then the conditional density

$$\varphi(y|x) = \frac{f(x, y)}{g(x)}$$

Alternatively, we write $\varphi(x) = E(Y|X)$, which can be interpret as a random variable which takes the value $E(Y|X = x)$ with $P(X = x)$ (see STAT 5050).

2.10 Way # 10: Tail behavior

For a scalar random variable $X \sim F$, we say X has an exponential tail if

$$\lim_{a\to\infty} \frac{-\log(1 - F(a))}{Ca^r} = 1, \quad \text{for some } C > 0, r > 0$$

and an algebraic tail if

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

Example 1. *Here are some examples:*

1. *Exponential:* $F(a) = 1 - e^{-\lambda a} \to c = \lambda, r = 1$
2. *Gaussian:* $F(a) = ... \to c = 2, r = 2$
3. *Student-t:* $m = \nu$ *(heavy-tail distributions/ extreme value theory)*

3 Sufficiency Principle

3.1 Introduction

Suppose $X_1, ...X_n \sim P_\theta$ for any unknown parameter $\theta \in \Omega, \Omega \subseteq \mathbb{R}^k$. Using n numbers $X_1, ...X_n$ to store the information and make inference about k features θ may waste storage space. Even worse, if n is large, the raw data $X_1, ..., X_n$ will become difficult to interpret. Therefore, we would like to produce a lower dimensional summary without losing information about θ (Data reduction).

3.2 Statistic and Sufficiency Principle

- **Statistic**: A statistic $T : \mathcal{X}^n \longrightarrow \mathcal{T}^m$ is a function of the data $X_1, ..., X_n$ and *free of any unknown parameter*.
- **Sufficiency Principle**: A statistic $T = T(X_1, ..., X_n)$ is sufficient for a model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ if for any $t = T(x_1, ..., x_n)$, the conditional distribution $X_{1:n} \mid T(x_{1:n}) = t$ is free of θ.
  * The n-dimensional statistic $X_{1:n} = (X_1, ..., X_n)^T$ is a *trivial sufficient statistic* for $\mathcal{P}$.

Example 2. $T(X_1, ..., X_n) = \bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ *(sample mean), and* $T(X_1, ..., X_n) = S_n^2 = (n - 1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ *(sample variance) are a statistic.*
** If μ is unknown, then the population variance $\sigma^2 = n^{-1} \sum_{i=1}^n (X_i - \mu)^2$ is <u>not</u> a statistic.*

Example 3. *Let* $X_1, ..., X_n \overset{iid}{\sim} Bern(\theta)$ *for any* $\theta \in (0, 1)$. *Let* $T = T(X_{1:n}) = \sum_{i=1}^n X_i$. *Consider*

- *Case 1:* $\sum_{i=1}^n x_i \neq t, P_\theta(x_{1:n} \mid t) = 0$.

- *Case 2:* $\sum_{i=1}^n x_i = t$. *Consider* $\{X_{1:n} = x_{1:n}, T = t\} = \{X_{1:n} = x_{1:n}\}$ *as knowing all data* $x_{1:n}$ *gives more information than knowing* $t = T(x_{1:n})$. *Note that* $T \sim Bin(n, \theta)$, *we have*

$$\begin{aligned} P_\theta(x_{1:n} \mid t) &= \frac{P_\theta(x_{1:n}, t)}{P_\theta(t)} \\ &= \frac{P_\theta(x_{1:n})}{P_\theta(t)} \qquad \frac{\textit{A likelihood function}}{\textit{Binomial distribution}} \\ &= \frac{\prod_{i=1}^n \{\theta^{x_i}(1-\theta)^{1-x_i}\}}{\binom{n}{t}\theta^t(1-\theta)^{1-t}} = \binom{n}{t}^{-1} \end{aligned}$$

Hence, for any cases, $P_\theta(x_{1:n} \mid t)$ *is free of* θ, *so* $T(x_{1:n}) = \sum_{i=1}^n x_i$ *is a sufficient statistic for* $\mathcal{P} = Bern(\theta)$.

Example 4. *(Order Statistics) Let* $X_{1:n} \overset{iid}{\sim} P_\theta \in \mathcal{P}$ *for any model* $\mathcal{P}$, *then the order statistics* $T = (x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)})^T$ *are sufficient. To see why* T *is sufficient, note that given* T, *the possible values of* X *are in* $n!$ *permutations of* T. *By symmetry, we can see that each of their permutation has an equal probability of* $\frac{1}{n!}$

$$\begin{aligned} p_\theta(X_1 = X_{(1)}, X_2 = X_{(2)}, ..., X_n = X_{(n)}) &= \frac{1}{n!} \\ p_\theta(X_1 = X_{(2)}, X_2 = X_{(1)}, ..., X_n = X_{(n)}) &= \frac{1}{n!} \\ &... \\ p_\theta(X_1 = X_{(n)}, X_2 = X_{(n-1)}, ..., X_n = X_{(1)}) &= \frac{1}{n!} \end{aligned}$$

Hence $X_{1:n} = x_{1:n} \mid T = t = \frac{1}{n!} \perp\!\!\!\perp \theta$ *thus* $T = (x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)})^\top$ *is a sufficient statistic.*

Theorem 1. *If* $X \sim P_\theta \in \mathcal{P}$ *and* $T = T(X)$ *is a sufficient statistic for* $\mathcal{P}$, *then for any decision procedure* δ, *there exists a (possibly randomized) decision procedure of equal risk that depends on* X *only through* $T = T(X)$ *only.*

To illustrate the concept of randomization, suppose, given an independent source of randomness, say $U \sim Unif(0, 1)$, we can always generate a new data set $X' = f(T(X), U)$ from the conditional distribution $p(X \mid T(X))$ and define a randomized procedure

$$\delta^*(X, U) \equiv \delta\{f(T(X), U)\} - \delta(X') \overset{d}{=} \delta(X)$$

Example 5. *Suppose X and Y are independent with common density*

$$f_\theta(x) = \begin{cases} \theta \exp(-\theta x) & \textit{for } x \geq 0 \\ 0 & \textit{otherwise}, \end{cases}$$

and let $U \sim unif(0, 1)$ *and independent of* X, Y. *Define* $T = X + Y$ *and define*

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

Let us find the joint density of $\tilde{X}$ and $\tilde{Y}$. The density of T is needed, and this can be found by smoothing. Because X and Y are independent,

$$\begin{aligned} P(T \le t \mid Y = y) &= P(X + Y \le t \mid Y = y) \\ &= \mathbb{E}\Big\{ I(X + Y \le t) \mid Y = y \Big\} \\ &= \int I(X + Y \le t) dF_X(x) \\ &= F_X(t - y). \end{aligned}$$

So $P(T \le t \mid Y) = F_X(t - Y)$ and

$$F_T(t) = P(T \le t) = \mathbb{E}\Big\{ F_X(t - Y) \Big\}.$$

This formula holds generally. Specializing to our specific problem, $F_X(t-Y) = 1 - \exp\big\{ -\theta(t-Y)\big\}$ on $Y < t$ and is zero on $Y \ge t$. Writing the expected value of this variable as an integral against the density of Y, for $t \ge 0$,

$$F_T(t) = \int_0^t \Big[1 - \exp\big\{ -\theta(t-y)\big\} \Big] \theta \exp(-\theta y) dy = 1 - \exp(-\theta t) - t\theta \exp(-\theta t)$$

Taking derivative, T has density

$$p_T(t) = F_T'(t) = \begin{cases} t\theta^2 \exp(-\theta t) & \text{for } t \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

Because T and U are independent, they have the joint density

$$p_\theta(t,u) = \begin{cases} t\theta^2 exp(-\theta t) & \text{for } t \ge 0, u \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

From this,

$$p\bigg(\begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} \in B \bigg) = \int \int I\{tu, t(1-u)\} p_\theta(t,u) du dt$$

Changing variables to $x = ut, du = dx/t$ in the inner integral, and reversing the order of integration using Fubini's theorem,

$$p\bigg(\begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} \in B \bigg) = \int \int I\{x, t-x\} t^{-1} p_\theta(t, \tfrac{x}{t}) dt dx$$

Now a change of variables to $y = t - x$ in the inner integral gives

$$p\bigg(\begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} \in B \bigg) = \int \int I\{x,t\} (x-y)^{-1} p_\theta(x+y, \tfrac{x}{x+y}) dy dx$$

Thus $\tilde{X}$ and $\tilde{Y}$ have joint density

$$\frac{p_\theta(x+y, \frac{x}{x+y})}{x+y} = \begin{cases} \theta^2 exp\big\{ -\theta(x+y)\big\} & \text{for } x, y \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

This density is the same as the joint density of X and Y, and so this calculation shows that the joint distribution of $\tilde{X}$ and $\tilde{Y}$ is the same as the joint distribution of X and Y. Considered as data that provide information about θ, the pair $(\tilde{X}, \tilde{Y})$ should be just as informative as (X, Y).

3.3 Neyman-Fisher Factorization Theorem

Suppose each $P_\theta \in \mathcal{P}$ has density $p(x_{1:n};\theta)$ with respect to a common σ -finite measure μ. That is, $dP_\theta/d\mu = p(x_{1:n};\theta)$, then $T = T(X_{1:n})$ is sufficient if and only if for any $\theta \in \Theta, x_{1:n} \in \mathcal{X}^n$,

$$p(x_{1:n};\theta) = g_\theta(T(x_{1:n}))h(x_{1:n})$$

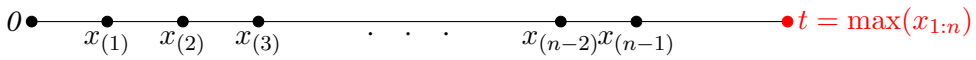
for some functions g_θ, h.
* A necessary and sufficient condition for $T(x_{1:n})$ to be sufficient is that the density $p(x_{1:n};\theta)$ can be factorized into two components, one of which depends on both $\theta, T(x_{1:n})$, and another one is free of θ.

Example 6. *Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$ with unknown $\theta = (\mu, \sigma^2)^T$, then we have*

$$\begin{aligned} p(x_{1:n};\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x_i-\mu)^2}{2\sigma^2} \right\} \\ &= \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2}(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2) \right\} \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \\ &= g_\theta(T(x_{1:n}))h(x_{1:n}) \end{aligned}$$

By the Neyman-Fisher factorization theorem, $T(X_{1:n}) = (\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i)^T$ is a sufficient statistic for $\mathcal{P} = N(\mu, \sigma^2)$.

Example 7. *Let $X_1, ..., X_n \overset{iid}{\sim} Unif(0, \theta)$ for any $\theta > 0$. $T = T(X_{1:n}) = \max(X_{1:n})$ is a sufficient statistic for $\mathcal{P} = Unif(0, \theta)$.*
The intuition: think of $x_1, ..., x_n$ as n numbers on the real line $\mathbb{R}$, then the remaining $n-1$ numbers, given the maximum is fixed at $t = \max(x_{1:n})$, behave like $n-1$ iid random samples drawn from $Unif(0, t)$.



for some order statistics $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ of $x_1, ..., x_n$. To show that $t = \max(x_{1:n})$ is a sufficient statistic,

$$\begin{aligned} p(x_{1:n};\theta) &= \prod_{i=1}^{n} \left\{ \frac{1}{\theta} I(0 < x_i < \theta) \right\} \\ &= \frac{1}{\theta^n} I(x_{(n)} < \theta) I(0 < x_{(1)}) \\ &= g_\theta(T(x_{1:n}))h(x_{1:n}) \end{aligned}$$

By the Neyman-Fisher factorization theorem, $T = T(X_{1:n}) = \max(X_{1:n})$ is a sufficient statistic for $\mathcal{P} = Unif(0, \theta)$.

Proof of the Neyman-Fisher factorization theorem

Proof. *To begin, suppose $p_\theta \in p$ and $\theta \in \Omega$*

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

With respect to μ. Modifying h, we can assume without loss of generality that μ us a probability measure equivalent to the family $P = \{p_\theta : \theta \in \Omega\}$ [Equivalence referes to the situation where $\mu(N) = 0$ iff $p_\theta(N) = 0 \quad \forall \theta \in \Omega$].

Let E^* and P^* be the expectation and probability where $X \sim \mu$. Let G^* and G_θ denote marginal distribution for $T(x)$ where $X \sim \mu$ and $X \sim P_\theta$ respectively. Let Q be the conditional distribution for X given T where $X \sim \mu$.

To find the densities for T,

$$\begin{aligned} E_\theta f(T) &= \int f(T(x)) g_\theta(T(x)) h(x) d\mu(x) \\ &= E^*\{f(T) g_\theta(T) h(X)\} \\ &= \int\int f(t) g_\theta(T) h(x) dQ_t(x) dG^*(t) \\ &\triangleq \int f(t) g_\theta(t) \omega(t) dG^*(t), \end{aligned}$$

where $\omega(t) = \int h(x) dQ_t(x)$. If f is an indicator function this shows that G_θ has the density $g_\theta \omega(t)$ with respect to G^*. Next we define $\widetilde{Q}$ to have density $h/\omega(t)$ with respect to $Q(t)$, so that

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

the conditional distribution of X given T under P_θ is independent of Q.

$$\begin{aligned} E_\theta \int (X, T) &= E^*\{f(X, T) g_\theta(T) h(x)\} \\ &= \iint f(x, t) g_\theta(t) h(x) dQ_t(x) dG^*(t) \\ &= \iint f(x, t) d\widetilde{Q}_t(x) dG_\theta(t) \end{aligned}$$

By the definition of conditional distribution, it shows that $\widetilde{Q}$ is a conditional distribution of X given under P_θ. Because $\widetilde{Q}$ does not depend on Q, it is sufficient statistic. (TBC)

STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Bowen Jia and Zheng Zhang

Lecture # 4

---

1 Sufficiencies

Recap: Neyman-Fisher Factorization criterion. T(X) is sufficient is sufficient iff $p(x;\theta) = g_\theta(T(x))h(x)$ prove for the discrete cases,$p(x|T)$ is independent of θ We will look at the proof for the continuous case (Ref. Keener 6.4).

To begin, suppose $p_\theta \in p$ and $\theta \in \Omega$

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

With respect to μ. Modifying h, we can assume without loss of generality that μ us a probability measure equivalent to the family $P = \{p_\theta : \theta \in \Omega\}$ [Equivalence referes to the situation where $\mu(N) = 0$ iff $p_\theta(N) = 0 \quad \forall\theta \in \Omega$].

Let E^* and P^* be the expectation and probability where $X \sim \mu$. Let G^* and G_θ denote marginal distribution for $T(x)$ where $X \sim \mu$ and $X \sim P_\theta$ respectively. Let Q be the conditional distribution for X given T where $X \sim \mu$.

To find the densities for T,

$$\begin{aligned} E_\theta f(T) &= \int f(T(x))g_\theta(T(x))h(x)d\mu(x) \\ &= E^*\{f(T)g_\theta(T)h(X)\} \\ &= \int\int f(t)g_\theta(T)h(x)dQ_t(x)dG^*(t) \\ &\triangleq \int f(t)g_\theta(t)\omega(t)dG^*(t), \end{aligned}$$

where $\omega(t) = \int h(x)dQ_t(x)$. If f is an indicator function this shows that G_θ has the density $g_\theta\omega(t)$ with respect to G^*. Next we define $\widetilde{Q}$ to have density $h/\omega(t)$ with respect to $Q(t)$, so that

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

the conditional distribution of X given T under P_θ is independent of Q.

$$\begin{aligned} E_\theta \int (X,T) &= E^*\{f(X,T)g_\theta(T)h(x)\} \\ &= \iint f(x,t)g_\theta(t)h(x)dQ_t(x)dG^*(t) \\ &= \iint f(x,t)d\widetilde{Q}_t(x)dG_\theta(t) \end{aligned}$$

By the definition of conditional distribution, it shows that $\widetilde{Q}$ is a conditional distribution of X given under P_θ. Because $\widetilde{Q}$ does not depend on Q, it is sufficient statistic.
2nd part: T is sufficient statistic $\rightarrow$ factorization holds(tutorial)

2 Sufficiency

Data reduction $\rightarrow$ all information about θ is stored in $\Theta \rightarrow$ improves data interpretability. (c.f. example

$$\begin{cases} \widetilde{X} = TU, \\ \widetilde{Y} = T(1-U), \end{cases} \tag{1}$$

where U is a uniform (0,1) independent of T.
Question: how much data compression/reduction can be achieved while the inference for θ is not impaired (in any sense)? what is the optimal data reduction strategy?

3 Exponential families

3.1 Basics

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

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

where $\eta_i(\theta) \in \mathbb{R}$ are called the natural parameters, $T_i(X) \in \mathbb{R}$ are its sufficient statistics, $B(\theta)$ is the log-partition function, which means that it is the logarithm of a normalising factor:

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

and $h(x) \in \mathbb{R}$ is the base measure (e.g. $I(x \in \mathbb{R})$ or $I(x) \geq 0$).

Remark: Many common distributions are exponential families. Examples include Normal, Binomial, Poisson distribution to name but a few. Exponential families are also closely related to the motions of sufficiency and optimal data reduction.

Example 1. *Exponential distribution $P = \{\exp(\theta) : \theta > 0\}$ the densities take the form:*

$$p(x; \theta) = \theta e^{-\theta x} = \exp(-\theta x + \log\theta) I_{(x \geq 0)},$$

which means that the family is a one-dimensional exp family with $\eta_i(\theta) = -\theta$, $T_i(x) = x$, $B(\theta) = -log(\theta)$ and $h(x) = I_{(x \geq 0)}$. It is noteworthy that the parameterization is not unique.

Example 2. *Beta distribution $P = \{Beta(\alpha, \beta) : \alpha, \beta > 0\}$, $\theta = (\alpha, \beta)$ the densities take the form*

$$\begin{aligned} p(x; \theta) &= x^{\alpha-1}(1-x)^{\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} I_{(0<x<1)} \\ &= \exp\left\{(\alpha-1)\log x + (\beta-1)\log(x-1) + \log\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right\} I(0<x<1) \end{aligned}$$

which means that the beta distribution belongs to a 2-dimensional exponential family with $\eta_i(\theta) = \alpha - 1$, $\eta_2(\theta) = \beta - 1$, $T = (T_1, T_2) = (\log x, \log(1-x))$, $B(\theta) = -\log(\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta)))$ *and* $h(x) = I(0 < x < 1)$. *One may also rewrite* $p(x;\theta)$ *as:*

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

which change the natural parameter from $\eta_1(\theta)$ *to* $\eta_1^*(\theta) = \alpha$ *and* $\eta_2(\theta)$ *to* $\theta_2^* = \beta$ *with* $h^*(x)$ *becomes* $I(0<x<1)/\{x(1-x)\}$.

Definition 1. *An exponential family is in canonical form when the density has the form*

$$p(x;\eta) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x). \tag{2}$$

This parameterises the densities in terms of the natural parameters η *instead of* θ.

Definition 2. *The set of all valid natural parameters* Θ *is called the natural parameter space: for each* $\eta \in \Theta$, *there exists a normalising constant* $A(\eta)$ *such that* $\int p(x;\eta)\, dx$ = *1, Equivalently,*

$$\Theta = \left\{\eta : 0 < \int \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x) d\mu x < \infty\right\} \tag{3}$$

For any canonical exponential family $P = p_\eta : \eta \in H$, *we have* $H \in \Theta$. *One can show that* Θ *is convex. The differences between canonical and non-canonical one is that for the non-canonical one, there is other parametrisations.*

3.2 Dimension reduction

There are two cases when the superficial dimension of an s-dimensional exponential family $P = p_\eta : \eta \in H$ can be reduced.

3.2.1 Case 1

The $T_i(x)$'s satisfy an affine equality constraint for all $x \in X$. In other words, $\{T_i\}$ are linearly dependent and ~~we call η unidentifiable.~~

Definition 3. *If* $\mathcal{P} = \{p_\theta; \theta \in \Omega\}$, *then* θ *is unidentifiable if for two parameters* $\theta_1 \neq \theta_2$, $p_{\theta_1} = p_{\theta_2}$.

Example 3. *Let* $X \sim \exp(\eta_1, \eta_2)$ *with*

$$p(x;\eta_1,\eta_2) = \exp\{-\eta_1 x - \eta_2 x + \log(\eta_1+\eta_2)\} I(x \geq 0) \tag{4}$$

Here $T_1(x) = T_2(x) = x$ *(they are linearly dependent).We can actually combine* (η_1, η_2) *into* $\eta_1 + \eta_2$ *and write*

$$p(x;\eta_1,\eta_2) = \exp\{-(\eta_1+\eta_2)x + \log(\eta_1+\eta_2)\} I(x \geq 0) \tag{5}$$

Besides, η *is unidentifiable since* $p(x;\eta_1+c,\eta_2-c) = p(x;\eta_1,\eta_2)$ *for all* $c < \eta_2$.

3.2.2 Case 2

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

Example 4. *Let* $p(x;\eta) = c(\eta_1,\eta_2)\exp(\eta_1 x + \eta_2 x^2)$ *for all* (η_1, η_2) *satisfying* $\eta_1 + \eta_2 = 1$. *Then we can rewrite*

$$p(x;\eta) = c(\eta_1,\eta_2)\exp(\eta_1(x-x^2) + x^2) \tag{6}$$

3.2.3 Minimal

When neither of the above two cases hold, we call the exponential family minimal.

Definition 4. *A canonical exponential family $P = p_\eta : \eta \in H$ is minimal if*
(1) $\sum_{i=1}^{s} \lambda_i T_i(x) = \lambda_0, \forall x \in X \Longrightarrow \lambda_i = 0 \; \forall i \in \{0, ..., s\}$
(2) $\sum_{i=1}^{s} \lambda_i \eta_i = \lambda_0, \forall \eta \in H \Longrightarrow \lambda_i = 0 \; \forall i \in \{0, ..., s\}$

Definition 5. *Suppose is $P = p_\eta : \eta \in H$ a s-dimensional exponential family. If H contains an open s-dimensional rectangle, then P is called full-rank, otherwise P is called curved, which means that the η_i's are related non-linearly.*

Example 5. *Consider $N(\mu, \sigma^2)$ where in this case $\eta_1 = 1/(2\sigma^2)$, $\eta_2 = \mu/\sigma^2$,$T_1(x) = -x^2$,$T_2(x) = x$.*

1. *Take $\mu = \sigma^2$, then $\eta_1 = 1/(2\sigma^2)$, $\eta_2 = 1$, then $1/(2\sigma^2)\eta_2 - \eta_1 = 0$. Therefore, the family is non-minimal in this case.*
2. *Take $\mu = \sqrt{\sigma^2}$, then $\eta_1 = 1/(2\sigma^2)$, $\eta_2 = 1/\sqrt{\sigma^2}$, then $\eta_2 = \sqrt{2\eta_1}$. Therefore, the family is minimal and curved in this case.*
3. *When there's no constraint on (μ, σ^2), H contains an open rectangle: $\mathbb{R} \times (0, \infty)$. Therefore, the family is minimal and full-rank in this case.*

3.3 Properties of exponential families

1. If $X_1, X_2, ..., X_n \overset{i.i.d}{\sim} p(x; \theta) = \exp\{\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - B(\theta)\} h(x)$.
   Then by NFFC, $(\sum_{j=1}^{n} T_1(x), ..., \sum_{j=1}^{n} T_s(x))$ is a sufficient statistic. Hence the exponential family is exceptionally compressible.

2. If f is integrable and $\eta \in \Theta$, then

$$G(f, \eta) = \int f(x) \exp\left\{\sum_{i=1}^{s} \eta_i T_i(x)\right\} h(x) d\mu(x) \tag{7}$$

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

3. The moments of T_i's can be directly calculated by taking $f(x) = 1$:

$$G(f, \eta) = \int \exp\left\{\sum_{i=1}^{s} \eta_i T_i(x)\right\} h(x) d\mu(x) = \exp(A(\eta)) \tag{8}$$

$$\frac{\partial G(f, \eta)}{\partial \eta_i} = \int T_i(x) \exp\left\{\sum_{i=1}^{s} \eta_i T_i(x)\right\} h(x) d\mu(x) = \frac{\partial A(\eta)}{\partial \eta_i} \exp(A(\eta)). \tag{9}$$

   Therefore,

$$\frac{\partial A(\eta)}{\partial \eta_i} = \int T_i(x) \exp\{\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\} h(x) d\mu(x) = E_\eta\{T_i(x)\} \tag{10}$$

   Besides, it can be shown that

$$\frac{\partial^2 A(\eta)}{\partial \eta_i \partial \eta_j} = \text{Cov}_\eta(T_i(x), T_j(x)) \tag{11}$$

3.4 Minimal Sufficiency

Definition 6. *A sufficient statistic T is minimal if for every sufficient statistics T' and for every $x, y \in X$, $T(x) = T(y)$ 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 $x \in X$.*

The following theorem allows us to verify whether a sufficient statistic is minimal or not.

Theorem 7. *Let $p(x;\theta) : \theta \in \Omega$ be a family of densities with respect to some measure μ(usually lebesgue measure for continuous distribution and counting measure for discrete distribution).Suppose that there exists a statistic T such that for every $x, y \in X$*

$$p(x;\theta) = c(x,y)p(y;\theta) \longleftrightarrow T(x) = T(y) \tag{12}$$

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

Proof. *First prove that T is sufficient and then T is minimal.*

1. *(T is sufficient) For all $t \in T(X)$(the image of T), consider the preimage $A_t = T^-1(t)$. For each A_t, we denote x_t as a representative. Then for any $y \in X$, we have $y \in A_{T(y)}$ and $x_{T(y)} \in A_{T(y)}$. From the assumption of T, we have*

$$p(y;\theta) = c(y, x_{T(y)})p(x_{T(y)};\theta) = h(y)g_\theta(T(y)) \tag{13}$$

   Therefore, by NFFC, T is sufficient.

2. *(T is minimal) Consider another sufficient statistic T'. By NFFC,*

$$p(x;\theta) = \tilde{g}_\theta(T'(x))\tilde{h}(x) \tag{14}$$

   Take any x and y such that $T'(x) = T'(y)$, then

$$p(x;\theta) = \tilde{g}_\theta(T'(x))\tilde{h}(x) = \tilde{g}_\theta(T'(y))\tilde{h}(y)\frac{\tilde{h}(x)}{\tilde{h}(y)} = p(y;\theta)C(x,y) \tag{15}$$

   By the assumption of T, $T(x) = T(y)$. Therefore, we've proved that for any sufficient statistics T' and any x and y, $T'(x) = T'(y)$ implies $T(x) = T(y)$. T is minimal.

STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Dominic Leung and Bencong Zhu

Lecture 5

5 Anciallarity and Completeness

5.1 Recap: Minimal Sufficiency

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

$$p(x;\theta) = C_{x,y}p(y;\theta), \quad \Leftrightarrow \quad T(x) = T(y).$$

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

Reference from books: Theorem 6.2.3 **?**, Theorem 3.11 **?**.

Example 1 (Normal minimal sufficient statistic) *Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, with μ and σ^2 unknown. Let x and y be two sample points, and let $(\overline{x}, S_x^2)$ and $(\overline{y}, S_y^2)$ be the sample means and variances corresponding to the x and y samples respectively.*

$$\begin{aligned}\frac{f(x;\mu,\sigma^2)}{f(y;\mu,\sigma^2)} &= \frac{(2\pi\sigma^2)^{n/2}\exp\left[-\left\{n(\overline{x}-\mu)^2 + (n-1)S_x^2\right\}/(2\sigma^2)\right]}{(2\pi\sigma^2)^{n/2}\exp\left[-\left\{n(\overline{y}-\mu)^2 + (n-1)S_y^2\right\}/(2\sigma^2)\right]} \\ &= \exp\left[\left\{-n(\overline{x}^2-\overline{y}^2) + 2n\mu(\overline{x}-\overline{y}) - (n-1)(S_x^2 - S_y^2)\right\}/(2\sigma^2)\right].\end{aligned}$$

This ratio will be constant as a function of $\theta = (\mu, \sigma^2)$ if and only if $\overline{x} = \overline{y}$ and $S_x^2 = S_y^2$. Thus, by the above theorem, $(\overline{X}, S^2)$ is a minimum sufficient statistic for θ, where $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \overline{X})^2$.

Example 2 (Curved exponential family) *Let $X_1, \ldots, X_N \overset{iid}{\sim} N(\sigma, \sigma^2), \sigma > 0$. Denote $\theta = \sigma$, then*

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

Hence, $T(X) = (T_1(X), T_2(X)) = \left(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i\right)$ is minimal sufficient.

Remark 1 *You should be reminded that if $p(x;\theta) = C_{x,y}p(y;\theta)$, x and y must be supported by the same θ (support of X : $\{x \in \mathcal{X} : p(x;\theta) > 0\}$). Otherwise, the 'constant' $C_{x,y}$ will be θ-dependent.*

Example 3 *Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Uniform$(0, \theta)$ and $T(X) = \max_{1 \le i \le n} X_i = X_{(n)}$. In that case for $x = (x_1, \ldots, x_n)$ such that $x_i > 0$, $i = 1, \ldots, n$,*

$$p(x;\theta) = \prod_{i=1}^n \frac{1}{\theta} I(x_i < \theta) = \frac{1}{\theta^n} I(T(X) < \theta).$$

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

Conversely, if $x, y > 0$ (i.e. $x_i, y_i > 0$, $i = 1, \ldots, n$) are supported by the same θ's, then

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

Therefore, it implies $T(x) = T(y)$ and is a minimal sufficient statistic.

Theorem 2 *For any minimal, s-dimensional exponential family, the statistic $(\sum_{i=1}^n T_1(X_I), \ldots, \sum_{i=1}^n T_s(X_i))$ is a minimal sufficient statistic. [Example 3.12 **?**]*

Proof 1 *Let $p(x;\theta) = \exp\{\eta(\theta)T(x) - B(\theta)\}\, h(x)$ be the density of an s-dimensional exponential family, where $\theta \in \Omega$. By NFFC, T is sufficient.*

Suppose $p(x;\theta) \propto_\theta p(y;\theta)$, then

$$e^{\eta(\theta)T(x)} \propto_\theta e^{\eta(\theta)T(y)},$$

which imples that

$$\eta(\theta) \cdot T(x) = \eta(\theta) \cdot T(y) + C,$$

where the constant C may depend on both x and y (but is independent of θ).

If θ_0 and θ_1 are any two points in Ω,

$$\{\eta(\theta_0) - \eta(\theta_1)\} \cdot T(x) = \{\eta(\theta_0) - \eta(\theta_1)\} \cdot T(y)$$

if and only if

$$\{\eta(\theta_0) - \eta(\theta_1)\} \cdot \{T(x) - T(y)\} = 0 \tag{1}$$

This shows that $T(x) - T(y)$ is orthogonal to eery vector in

$$\eta(\Omega) \ominus \eta(\Omega) \equiv \{\eta(\theta_0) - \eta(\theta_1) : \theta_0 \in \Omega, \theta_1 \in \Omega\},$$

so it must lie in the orthogonal complement of the linear span of $\eta(\Omega) \ominus \eta(\Omega)$. In particular, if the linear span of $\eta(\Omega) \ominus \eta(\Omega)$ is all of $\mathbb{R}^s$, then $T(x)$ must equal $T(y)$ in which case T is minimal sufficient.

5.2 Anciallarity and Completeness

Illustration:

Exponential families $\rightarrow$ significant data compression (without losing any information about θ).

Example 4 *Consider $X_1, \ldots, X_n \overset{iid}{\sim}$ Cauchy(θ), with densities*

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

*Based on **?** §1.5 result, we know that $(X_{(1)}, \ldots, X_{(n)})$ is minimal sufficient. The similar conclusion/observation can also be found for the double exponential location model: $p(x;\theta) \propto \exp(|x - \theta|)$. [The density is $f(x;\theta) = e^{|x-\theta|}/2$].*

IDEA: Determine the amount of 'ancillarity' information stored in its minimal sufficient statistics.

Definition 3 *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 4 (Continued) *Again $X_1, \ldots, X_n \overset{iid}{\sim}$ Cauchy(θ), then*

$$A(X) = X_{(n)} - X_{(1)} \text{ is ancillary},$$

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

To see this, observe that $X_i = Z_i + \theta$ for $Z_i \overset{iid}{\sim}$ Cauchy(0), so $X_{(i)} = Z_{(i)} + \theta$ and $A(X) = A(Z)$, which does not depend on θ.

Definition 4 (First-order ancillary statistic) *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 5 (Complete statistic) *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 probability 1 for all θ.*

Remark 2

1. *If T is complete sufficient, then T is minimal sufficient. [Bahadur's theeorem].*
2. *Complete sufficient statistic yield optimal unbiased estimators.*

Example 5 (Discrete Case) *Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(θ), $\theta \in (0, 1)$. Then $T(X) = \sum_{i=1}^n X_i$ is sufficient.*

Suppose that $E_\theta[f(T(X))] = 0$ for all $\theta \in (0, 1)$, then

$$\sum_{j=0}^{n} f(j) \binom{n}{j} \theta^j (1-\theta)^{n-j} = 0, \qquad \forall \theta \in [0, 1]. \tag{2}$$

Dividing both sides by θ^n and substituting $\beta = \theta/(1-\theta)$, we can rewrite (2) *as*

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

If f are non-zero, then the quantity on 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 equals 0 for every $\beta > 0$ unless $f = 0$. So, T is complete.

Example 6 (Continuous Case) *Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ with unknown $\theta \in \mathbb{R}$ and $\sigma^2 > 0$. We can verify if $T(X) = \overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ is complete for the model. [Note that T is minimal sufficient.]*

Let's consider the case wtih $n = 1$ and assume WLOG $\sigma^2 = 1$, $T(X) \sim N(\theta, 1)$. Suppose

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

We then decompose f into its positive and negative parts as

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

where $f_+(x) = \max(f(x), 0)$ and $f_-(x) = \max(-f(x), 0)$. Then $f_+(x) \geq 0$ and $f_-(x) \geq 0$ for all $x \in \mathbb{R}$.

Observation: *$f_+(x) = f_-(x)$ if and only if $f_+(x) = f_-(x) = 0$.*

1. *If $f(x) \geq 0$ almost everywhere (a.e.) or $f(x) \leq 0$ a.e., then (†) implies that $f(x) = 0$ a.e. because setting $\theta = 0$ because setting $\theta = 0$ gives us an integral of a nonnegative (resp. non-positive) function of zero. This gives/shows completeness.*

2. *Suppose f_+ and f_- have non-zero components, we may write*

$$\frac{\int_{-\infty}^{\infty} f_+(x)e^{-\frac{x^2}{2}}e^{\theta x}dx}{\int_{-\infty}^{\infty} f_+(x)e^{-\frac{x^2}{2}}dx} = \frac{\int_{-\infty}^{\infty} f_-(x)e^{-\frac{x^2}{2}}e^{\theta x}dx}{\int_{-\infty}^{\infty} f_-(x)e^{-\frac{x^2}{2}}dx}, \tag{††}$$

since (†) shows that the denominator of (††) are both equal. The quantity

$$\frac{f_+(x)e^{-\frac{x^2}{x}}}{\int_{-\infty}^{\infty} f_+(x)e^{-x^2}xdx}$$

defines a probability density and the LHS of (††) is the moment generating function of this density. Similarly, the RHS is the moment generating function of the density

$$\frac{f_-(x)e^{-\frac{x^2}{x}}}{\int_{-\infty}^{\infty} f_-(x)e^{-x^2}xdx}$$

It implies that $f_+(x) = f_-(x)$ a.e.. Then $f_+(x) = f_-(x) = 0$ a.e., or in other words, $f(x) = 0$ a.e..

Hence T is copmlete (and sufficient).

Example 7 *(Example 3.16 of Keenen (2010))*

Exercise 1 *If $X_1, \ldots, X_n \overset{iid}{\sim} p(x, \theta) \propto h(x)e^{\theta x}$, then the Statistics $T(x) = X$ is complete*
$\rightarrow$ Suppose $\int f(x)h(x)e^{\theta x}dx = 0$ for all $\theta \in \Omega$
$\rightarrow$ decompose $f(x) = f_+(x) - f_-(x)$ with $f_+ \geq 0, f_- \geq 0$
$\rightarrow$ f_+ and f_- can be viewed as unnormalised densities $p_+(x)$ and $p_-(x)$, respectively.
$\rightarrow$ argue that the $m.g.f.$'s of p_+ and p_- are equal

Theorem 6 (Theorem 4.3.1 ?) *$(T_1, \ldots, T_n)$ is complete for any s-dimensional full rank exponential family. [see P. 117 of TSH]*

Theorem 7 (Basu's Theorem) *If T is complete and sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ and V is ancillary, then $T(X) \perp\!\!\!\perp V$.*

Proof 2 *Define $q_A(t) = P_\theta(V \in A \mid T = t)$ or $q_A(T) = P_\theta(V \in A \mid T)$ and $p_A = P_\theta(V \in A)$. By sufficiency and ancillarity, neither p_A nor $q_A(t)$ depends θ. By smoothing,*

$$(P_A = P_\theta(V \in A) = E_\theta\left(P_\theta(V \in A \mid T)\right) = E_\theta(q_A(T))$$

and so by completeness, $q_A(T) = p_A$ a.e. for $\mathcal{P}$. Again, by smoothing/tower expecation,

$$\begin{aligned}
P_\theta(T \in B, V \in A) &= E_\theta\left(1_B(T)1_A(V)\right) \\
&= E_\theta\left(E_\theta\left(1_B(T)1_A(v) \mid T\right)\right) \\
&= E_\theta\left(1_B(T)E_\theta\left(1_A(v) \mid T\right)\right) \\
&= E_\theta\left(1_B(T)q_A(T)\right) \\
&= E_\theta\left(1_B(T) \cdot p_A\right) \\
&= P_\theta(T \in B) \cdot P_\theta(V \in A)
\end{aligned}$$

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

Example 8 *Let* $X_1, \ldots, X_n \overset{iid}{\sim} N\left(\mu, \sigma^2\right)$, *where both of* μ *and* σ^2 *are unknown. Then* $\bar{X}_n \perp\!\!\!\perp n^{-1}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$ *with* $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$.

Fix any $\sigma > 0$ and consider the submodel $\mathcal{P}_\sigma = \left\{N\left(\mu, \sigma^2\right) : \mu \in \mathbb{R}\right\}$ In each submodel, $\bar{X}_n$ is complete and sufficient, and $n^{-1}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2$ is ancillary

$$X_i = Z_i + \mu \qquad X_i - \bar{X}_n \to Z_i - \bar{Z}$$

By Basu's theorem, $\bar{X}_n \perp\!\!\!\perp \sum_{i=1}^n \left(x_i - \bar{x}_n\right)^2$ under $N\left(\mu, \sigma^2\right)$ for any μ. Since σ is arbitrary, we can conclude that $\bar{X}_n \perp\!\!\!\perp \sum_{i=1}^n \left(x_i - \bar{x}_n\right)^2$ hold for the full model $\mathcal{P} = \left\{N\left(\mu, \sigma^2\right) : \mu \in \mathbb{R}, \sigma^2 > 0\right\}$

From data reduction to optimal Inference

Lossless data reduction (sufficiency)	$\rightarrow$	Optimal data reduction (minimal sufficiency + completeness)	$\rightarrow$	Optimal inference estimator

Definition 8 *A function* $f : C \to \mathbb{R}$ *with* C *convex is a convex function if* $x \neq y \in C$ *and* $\gamma \in (0, 1)$*:*

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

The function f *is said to be strictly convex if the above inequality holds strictly (ie. " < ")*

Example 9 *For any* $\theta \in \Omega$*, the function* $f(d) = (d-\theta)^2$ *is strictly convex on* $\mathbb{R}$.

Example 10 *For any* $\theta \in \Omega$*, the function* $f(d) = |d - \theta|$ *is convex, but not strictly convex.*

Theorem 9 (Jensen's Inequality) *If* $f : C \to \mathbb{R}$ *is convex on any open set* C, $P(x \in C) = 1$ *and* $E(X)$ *exists, then*

$$f(E(x)) \leqslant E(f(x))$$

If f *is strictly convex, then the above inequality holds strictly unless* $X = E(X)$ $w.p.1$

Theorem 10 *Suppose that* 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 (as a function of* $d \in \mathcal{D}$*), then*

$$R(\theta, \eta) \leqslant R(\theta, \delta) \quad for \quad \eta(T(x)) = E(\delta(x) \mid T(x))$$

If $L(\theta, \cdot)$ *is strictly convex, then* $R(\theta, \eta) < R(\theta, \delta)$ *for any* θ *unless* $\eta(T'(x)) = \delta(x)$

Example 11 *Let* $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(\theta), \theta \in (0, 1)$*, and consider* $L(\theta, d) = (\theta - d)^2$*. Suppose we start with an unresaonable estimator* $\delta(X) = X_1$*. We know that* $T(X) = \bar{x}_n$ *is sufficient, so we can apply Rao-Blackwell theorem to improve our eatimator* δ

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

Recall that in lecture I, we showed already that $R(\theta, \eta) = \frac{\theta(1-\theta)}{n} < \theta(1-\theta) = R(\theta, \delta)$.

$$(\delta_{naive}(X) = 1/2) \qquad R(1/2, \delta_{naive}) < R(1/2, \delta')$$

References

STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Kaiser Fan and Chuchu Wang (Reviser)

Lecture 6

1 Unbiased Estimation

Definition 1 *An estimator is said to be unbiased if* $\mathbb{E}_\theta\{\delta(X)\} = g(\theta)$ *for all* θ.

Although it is very challenging to obtain uniformly best estimator, we can find an unbiased estimator with uniformly minimum risk, *i.e.* an unbiased estimator δ satisfying $R(\theta, \delta) \leq R(\theta, \delta')$ for all $\theta \in \Omega$ and any other unbiased estimator δ'.

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

When $L(\theta, d) = (\theta - d)^2$, an UMRUE becomes a uniformly minimum variance unbiased estimator (UMVUE) because

$$\underbrace{\mathbb{E}_\theta\{(g(\theta) - \delta(X))^2\}}_{\text{Mean squared error (MSE)}} = \underbrace{\{\mathbb{E}_\theta\{\delta(X)\} - g(\theta)\}^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}_\theta\left\{(\delta(X) - \mathbb{E}_\theta\{\delta(X)\})^2\right\}}_{\text{Variance}}$$

The left-hand side of the equation is called the Mean Squared Error (MSE) of the estimator $\delta(X)$ for $g(\theta)$. Moreover, if $\delta(X)$ is unbiased, *i.e.* $\mathbb{E}_\theta\{\delta(X)\} - g(\theta) = 0$, then the MSE is reduced to

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

Definition 2 *If an unbiased estimator exists, then* $g(\cdot)$ *is called* U*-estimable.*

Example 1 *Suppose* $X \sim U(0, \theta)$. *Then* δ *is unbiased if*

$$\int_0^\theta \delta(x)\theta^{-1}\, dx = g(\theta), \qquad \forall\, \theta > 0$$

or if

$$\int_0^\theta \delta(x)\, dx = \theta g(\theta), \qquad \forall\, \theta > 0 \tag{1}$$

So, g *cannot be* U*-estimable unless* $\theta g(\theta) \to 0$ *as* $\theta \downarrow 0$. *If* g' *exists, then differentiating Equation 1, by fundamental theorem of calculus, we have*

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

For, say $g(\theta) = \theta$, *then* $\delta(X) = X + X \cdot 1 = 2X$.

Example 2 *Suppose $X \sim \mathrm{Bin}(n, \theta)$. If $g(\theta) = \sin\theta$, then δ will be unbiased if*

$$\sum_{k=0}^{n} \delta(k) \binom{n}{k} \theta^k (1-\theta)^{n-k} = \sin\theta, \qquad \forall\, \theta \in (0,1) \tag{2}$$

The LHS of Equation 2 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 3 *An unbiased estimator δ is uniformly minimum variance unbiased (UMVU) if*

$$\mathrm{Var}_\theta(\delta) \leq \mathrm{Var}_\theta(\delta'), \qquad \forall\, \theta \in \Omega$$

for any other competing unbiased estimator δ'.

Theorem 4 ***(Lehmann-Scheffé Theorem)*** *If T is a complete and sufficient statistic, and $\mathbb{E}_\theta\{h(T(X))\} = g(\theta)$, 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 $\mathcal{P}$-null set, under any strictly convex loss function.

Proof

(a) Suppose $\mathbb{E}_\theta\{\tilde{h}(T(X))\} = g(\theta)$, then

$$\mathbb{E}_\theta\{\tilde{h}(T(X)) - h(T(X))\} = 0, \qquad \forall\, \theta \in \Omega$$

Thus, $\tilde{h}(T(X)) = h(T(X))$ almost surely for all $\theta \in \omega$ by completeness ($\mathbb{E}_\theta(f(T)) = 0 \Rightarrow f(T) = 0$).

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

$$\mathbb{E}_\theta\{\tilde{h}(T(X))\} = \mathbb{E}_\theta\{\mathbb{E}_\theta\{\delta(X) \mid T(X)\}\} = \mathbb{E}_\theta\{\delta(X)\} = g(\theta)$$

by tower property of conditional expectation ("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 $R(\theta, \tilde{h}(T(X))) \leq R(\theta, \delta)$ for all $\theta \in \Omega$ if the loss function is convex. It follows that

$$R(\theta, h(T(X))) \leq R(\theta, \delta),\ \forall \theta \in \Omega$$

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

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

Remark 1 *For $\mathcal{P}$-null set, it is with measure 0 s.t. we need not to consider that set, because it would not affect the expectation nor the risk. This theorem provides us with some useful strategies ("educated guesses") for finding UMRUEs under convex loss functions:*

1. *Rao-Blackwellisation.*
2. *Solve for the (unique) δ satisfying $\mathbb{E}_\theta\{\delta(T(X))\} = g(\theta)$ for all $\theta \in \Omega$.*
3. *Guess (the right form of unbiased function of $T(X)$)*

Example 3 ***(Rao-Blackwellisation)*** *Suppose $X_1, \cdots, X_n \overset{i.i.d.}{\sim} \text{Bernoulli}(\theta)$. 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.*

$$\mathbb{E}_\theta\left(\frac{T(X)}{n}\right) = \frac{1}{n}\mathbb{E}_\theta\left(\sum_{i=1}^n X_i\right) = \frac{n\theta}{n} = \theta.$$

Therefore, $n^{-1}T(X)$ is an UMRUE for θ under any convex loss function. Suppose that, instead, we want to estimate θ^2. Let's examine: $\delta(X) = \mathbb{1}\{X_1 = X_2 = 1\} = X_1 \cdot X_2$ (GUESS). Consider $\mathbb{E}_\theta\{\delta(X)\}$, we have

$$\mathbb{E}_\theta\{\delta(X)\} = \mathbb{E}_\theta\{\mathbb{1}\{X_1 = X_2 = 1\}\} = \mathbb{E}_\theta(X_1 \cdot X_2) \overset{i.i.d.}{=} \{\mathbb{E}_\theta(X_1)\}^2 = \theta^2.$$

By conditioning on $T(X)$, we can find the UMRUE via

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

Note that in this case $\mathbb{1}\{t \geq 2\}$ is redundant as for $t = 0$ or 1, the term $t(t-1) = 0$. Hence, we conclude that the UMRUE for θ^2 is

$$\frac{T(X)\{T(X)-1\}}{n(n-1)}.$$

Example 4 ***(Rao-Blackwellisation)*** *Suppose $X_1, \cdots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$. 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.*

$$\mathbb{E}_\theta(2X) = 2 \times \frac{\theta}{2} = \theta.$$

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

$$\mathbb{P}_\theta(X_1 = x_1 \mid T(X)) = \frac{1}{n} \times \mathbb{1}\{T(X) = x_1\} + \left(1 - \frac{1}{n}\right) \times \frac{1}{T(X)}\mathbb{1}\{0 < x_1 < T(X)\}$$

To find the UMVUE, we calculate

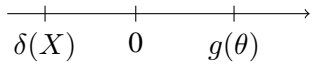
$$\begin{aligned}\mathbb{E}_\theta\{\delta(X) \mid T(X)\} = 2\mathbb{E}_\theta\{X_n \mid T(X)\} &= 2\left\{\frac{1}{n}T(X) + \left(1-\frac{1}{n}\right)\int_0^{T(X)} \frac{x_1}{T(X)}\,dx_1\right\} \\ &= 2\left\{\frac{T(X)}{n} + \left(1-\frac{1}{n}\right)\frac{T(X)}{2}\right\} \\ &= \left(\frac{n+1}{n}\right)T(X),\end{aligned}$$

which gives the desired result.

Example 5 ***(Solve for unique δ)*** *Let $X \sim \text{Poisson}(\theta)$. X is a complete and sufficient statistic, X is also unbiased and therefore UMVU for θ. Suppose we are interested in estimating $g(\theta) = e^{-a\theta}$ for $a \in \mathbb{R}$, known instead. We need to find an estimator δ such that $\mathbb{E}_\theta\{\delta(X)\} = g(\theta)$ for all $\theta > 0$. Under this model, we write*

$$\begin{aligned} & \mathbb{E}_\theta\{\delta(X)\} = \sum_{x=0}^{\infty}\delta(x)\frac{e^{-\theta}\theta^x}{x!} = e^{-a\theta}, && \forall\,\theta > 0 \\ \Leftrightarrow \quad & \sum_{x=0}^{\infty}\frac{\delta(x)\theta^x}{x!} = e^{(1-a)\theta} = \sum_{x=0}^{\infty}\frac{(1-a)^x\theta^x}{x!}, && \forall\,\theta > 0 \\ \Rightarrow \quad & \delta(X) = (1-a)^X \text{ is the UMVUE for } g(\theta) \end{aligned}$$

Note that the estimator is not ideal in the sense that if $a = 2$, the estimator $\delta(X) = (-1)^X$ will change sign according to X's "evenness" even though the estimand $e^{-a\theta}$ is non-negative. The estimator is hence inadmissible when $a > 1$ and dominated by $\max\{\delta(X), 0\}$.



If $\delta(X)$ is negative, then it will be far from the unbiased estimator $g(\theta)$.

Example 6 ***(Guess)*** *Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. Consider the case where $\theta = (\mu, \sigma^2)$ is unknown.*

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

(b) How about the UMVUE for σ?

(c) What is the UMVUE for μ^2?

(b) Observe that

$$X_i - \bar{X}_n \sim N\left(0,\, \frac{n-1}{n}\sigma^2\right) \qquad \Rightarrow \qquad \mathbb{E}\left\{|\, X_i - \bar{X}_n|\right\} = \sigma\sqrt{\frac{2}{\pi}} \times \sqrt{\frac{n-1}{n}}$$

This implies

$$\delta' = \frac{\sqrt{\pi n}}{\sqrt{2(n-1)}}\mid X_i - \bar{X}_n|$$

is unbiased for σ. At this point, we can Rao-Blackwellise this term. But the calculation is very tedious. Instead, we can observe another fact that

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

We know that $S_*^2 \sim \sigma^2 \chi_{n-1}^2$. Hence $\mathbb{E}(S_*) = \sigma \mathbb{E}(\chi_{n-1})$, which in turns implies that

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

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

(c) Taking the expecatation of the UMVUE for μ and squaring it, we obtain

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

So

$$\delta_n(X) = \bar{X}_n^2 - \frac{S_*^2}{n(n-1)}$$

is the UMVUE. However, $\delta_n(X)$ can be negative even though the estimand is a non-negative quantity. The estimator is in fact inadmissible and dominated by the biased estimator $\max(0, \delta_n(X))$ since

$$\mathbb{E}\{\max(0, \delta_n(X))\} \neq \max\{\mathbb{E}(\delta(X)), 0\}.$$

Remark 2 *From the above example, with MLE,*

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \left(\frac{n-1}{n}\right) S_*^2$$

STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Xiaoxuan Xia

Lecture # 7

---

1 Fisher Information

Recap:
Example (UMVUE for normal population variance)
Let $X_1, ..., X_n \sim^{i.i.d} N(\mu, \sigma^2)$ with both μ and σ^2 are unknown. Define $s^2 = \Sigma_{i=1}^n (X_i - \bar{X})^2$.

In this setting, $s^2/(n-1)$ is the UMNUE for σ^2. The MLE for σ^2 is s^2/n which has a lower mean squared error. In fact, the shrunk estimator $s^2/(n+1)$ has an even lower mean squared error. Therefore, neither UMVUE nor the MLE is admissible.

Question 1:
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 $g_1(\theta) + g_2(\theta)$?

If our underlying family of distributions has a complete sufficient statistic, then the answer if yes. (Because Lehman-scheffe Theorem).
Otherwise,...

Theorem 1: (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 an d only if $E(\delta_0(\theta), u) = 0$ for every $u \in \mathcal{U} = \{E(u) = 0\}$.

Proof 1: If δ_0 is an UMVUE, let's consider $\delta_\lambda = \delta_0 + \lambda u$ for $\lambda \in \mathbb{R}$ and $u \in \mathcal{U}$.
Since δ_0 has minimal variance,

$$\begin{aligned} Var(\delta_\lambda) &= Var(\delta_0) + \lambda^2 Var(U) + 2\lambda cov(\delta_0, u) \\ &\geqslant Var(\delta_0) \end{aligned}$$

Consider the quadratic form $q(\lambda) = \lambda^2 Var(U) + 2\lambda cov(\delta_0, u)$.
The form q has the roots $\lambda = 0$ and $-2cov(\delta_0, u)/var(u)$.
If the roots are distinct, then the form must be negative at some point, which would violate the inequality above.
Hence, $-2cov(\delta_0, u)/var(u) = 0$ in which case, $E(u\delta_0) = cov(\delta_0, u) = 0$.

To prove the converse result, we assume that $E(u\delta_0) = 0$ for all $u \in \mathcal{U}$ and consider any δ unbiased for $g(\theta)$. It follows that $\delta - \delta_0 \in \mathcal{U}$. So $E(\delta_0(\delta - \delta_0)) = 0$.
This implies that $E(\delta_0\delta) = E(\delta_0^2)$ and subtracting $E(\delta_0)E(\delta_0)$ on both sides, we obtain

$$Var(\delta_0) = cov(\delta_0, \delta) \leqslant \sqrt{Var(\delta_0)Var(\delta)}$$

by Cauchy-Schwarz inequality. Hence, $Var(\delta_0) \leqslant Var(\delta)$ for any arbitrary unbiased estimator δ and δ_0. Hence, δ_0 is an UMVUE for $g(\theta)$.

Answer for Question 1: $\forall u \in \mathcal{U}$, $E((\delta_1 + \delta_2)u) = E(\delta_1 u) + E(\delta_2 u) = 0$. Therefore, $\delta_1 + \delta_2$ is an UMVUE for $g_1(\theta) + g_1(\theta))$.

2 Variance Bound and Information

Recall: $Cov(X, Y) \leq \sqrt{Var(X)Var(Y)}$

Using the covariance inequality, if δ is an unbiased estimator for $g(\theta)$ and ψ is an arbitrary random variable, then

$$Var_\theta(\delta) \geq \frac{Cov_\theta^2(\delta, \psi)}{Var_\theta(\psi)} \tag{1}$$

The trick here is to choose a suitable ψ so that the bound is meaningful in the sense that $Cov_\theta(\delta, \psi)$ is the same for all δ that are unbiased for $g(\theta)$.

Question 2: How to find proper ψ?

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

To begin, $E_{\theta+\Delta}(\delta) - E_\theta(\delta)$ gives the same value $g(\theta + \Delta) - g(\theta)$, for any unbiased δ.
Here, Δ must be chosen so that $\theta + \Delta \in \Omega$.
Next, we write $E_{\theta+\Delta}(\delta) - E_\theta(\delta)$ as a covariance under p_θ.
This step involves the use of the "likelihood ratio". We assume here that $p_{\theta+\Delta}(x) = 0$ whenever $p_\theta(x) = 0$.
Define $L(x) = \frac{p_{\theta+\Delta}(x)}{p_\theta(x)}$ when $p_\theta(x) > 0$, and $L(x) = 1$ otherwise. We have

$$L(x)p_\theta(x) = \frac{p_{\theta+\Delta}(x)}{p_\theta(x)} p_\theta(x) = p_{\theta+\Delta}(x), a.e.x$$

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

$$\begin{aligned} E_{\theta+\Delta}h(x) &= \int h p_{\theta+\Delta} d\mu = \int h L p_\theta d\mu \\ &= E_\theta(L(x)h(x)). \end{aligned}$$

Take $h = 1$, $E_\theta L = 1$ (because $\int \frac{p_{\theta+\Delta}(x)}{p_\theta(x)} p_\theta(x) dx = \int_{\theta+\Delta} dx = 1$).
Take $h = \delta$, $E_{\theta+\Delta}\delta = E_\theta(L\delta)$, so if we define $\boldsymbol{\psi(x) = L(x) - 1}$ (**answer for Question 2**), then we can see that

$$E_\theta(\psi(x)) = E_\theta(L - 1) = 1 - 1 = 0$$

and

$$E_{\theta+\Delta}(\delta) - E_\theta(\delta) = E_\theta(L\delta) - E_\theta(\delta) = E_\theta(\psi\delta) = Cov_\theta(\delta, \psi)$$

($E_\theta(\psi\delta) = Cov_\theta(\delta, \psi)$ because $\psi = L - 1$).

As a result,

$$Cov_\theta(\delta, \psi) = g(\theta + \Delta) - g(\theta)$$

for any unbiased estimator δ. With this particular choice of ψ, the inequality of Equation 1 can be written as:

$$Var_\theta(\delta) \geq \frac{\{g(\theta + \Delta) - g(\theta)\}^2}{Var_\theta(\psi)} = \frac{\{g(\theta + \Delta)\}^2}{E_\theta \left(\frac{p_{\theta+\Delta}(x)}{p_\theta(x)} - 1\right)^2}, \tag{2}$$

which is known as the *Hammersley−Chapman−Robbins inequality*.

Under suitable conditions, we can show that

$$\lim_{\Delta \to 0} \frac{\{\frac{g(\theta+\Delta)-g(\theta)}{\Delta}\}^2}{E_\theta \left(\frac{\{p_{\theta+\Delta}(x)-p_\theta(x)\}/\Delta}{p_\theta(x)}\right)^2} = \frac{(g^{'}(\theta)^2)}{E_\theta \left(\frac{\partial p_\theta(x)/\partial\theta}{p_\theta(x)}\right)^2}. \tag{3}$$

The denominator here is known as **Fisher Information**, denoted as $I(\theta)$ and is given by

$$I(\theta) = E_\theta \left(\frac{\partial \log p_\theta(x)}{\partial \theta}\right)^2 \tag{4}$$

With enough regularity to interchange integration and differentiation,

$$\begin{aligned} 0 = \frac{\partial}{\partial\theta}(1) = \frac{\partial}{\partial\theta} \int p_\theta(x) d\mu(x) = \int \frac{\partial}{\partial\theta} p_\theta(x) d\mu(x) \\ = \int \frac{\partial \log p_\theta(x)}{\partial\theta} p_\theta(x) d\mu(x) = E_\theta \left(\frac{\partial \log p_\theta(x)}{\partial\theta}\right) \end{aligned}$$

and so

$$I(\theta) = E_\theta \left(\frac{\partial \log p_\theta(x)}{\partial\theta}\right)^2 - \left\{E_\theta \left(\frac{\partial \log p_\theta(x)}{\partial\theta}\right)\right\}^2 = Var \left(\frac{\partial \log p_\theta(x)}{\partial\theta}\right). \tag{5}$$

Furthermore, since

$$\int \frac{\partial^2 \log p_\theta(x)}{\partial\theta^2} d\mu(x) = E_\theta \left(\frac{\partial^2 p_\theta(x)/\partial\theta^2}{p_\theta(x)}\right) = 0$$

We can see that

$$\begin{aligned} \frac{\partial^2 \log p_\theta(x)}{\partial\theta^2} = \frac{\partial^2 p_\theta(x)/\partial\theta^2}{p_\theta(x)} - \left(\frac{\partial \log p_\theta(x)}{\partial\theta}\right)^2 \\ \Rightarrow I(\theta) = -E_\theta \left(\frac{\partial^2 \log p_\theta(x)}{\partial\theta^2}\right) \end{aligned} \tag{6}$$

Therefore,

$$Var_\theta(\delta) \geq \frac{\{g^{'}(\theta)\}^2}{I(\theta)}, \theta \in \Omega$$

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

$$Var_\theta(\delta) \geq \frac{\{g^{'}(\theta)\}^2}{I(\theta)}, \theta \in \Omega \tag{7}$$

This result is called the **Cramer−Rao**, or **information bound**.

Example (Exponential Families)
Let $\mathcal{P}$ be a one parameter exponential family in canonical form and density p_η given by

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

Then,

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

By the previous results, we have

$$I(\eta) = Var_\eta(T(x) - A^{'}(\eta)) = Var_\eta(T(x)) = A^{''}(\eta) \tag{8}$$

because $\frac{\partial^2 \log p_\eta(x)}{\partial \eta^2} = -A^{''}(\eta)$.

If the family is parameterized instead by $\mu = A^{'}(\eta) = E_\eta(T(x))$.
Then,

$$A^{''}(\eta) = I(\mu)\{A^{''}(\eta)\}^2$$

and so, because $A^{''}(\eta) = Var(T)$, we have

$$I(\mu) = \frac{1}{Var_\eta(T)} \tag{9}$$

observe also that because T is UMVUE for μ. The lower bound variance $Var_\mu(\delta) \geq 1/I(\mu)$ for an unbiased estimator δ of μ is sharp.

Example (Location Family)
Suppose q is an absolutely continuous random variable with density f. The family of distributions $\mathcal{P} = \{p_\theta : \theta \in \mathbb{R}\}$. With p_θ the distribution of $\theta + \varepsilon$ is called a **location family**.

$$\begin{aligned}\int g(x)dP_\theta(x) &= E_\theta(g(x)) = E_\theta(g(\theta + \varepsilon)) \\ &= \int g(\theta + \varepsilon)f(\varepsilon)d\varepsilon = \int g(x)f(x - \theta)dx\end{aligned}$$

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

$$
\begin{aligned}
I(\theta) &= E_\theta \left(\frac{\partial \log f(x-\theta)^2}{\partial \theta} \right)^2 = E \left(-\frac{f^{'}(x-\theta)}{f(x-\theta)} \right)^2 \\
&= E \left(\frac{f^{'}(\varepsilon)}{f(\varepsilon)} \right)^2 = \int \frac{\{f^{'}(x)\}^2}{f(x)} dx
\end{aligned}
\tag{10}
$$

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

Question 3: If two (or more) independent vectors are observed, what is the total Fisher Information?

Answer to Question 3:
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, and that X has density p_θ and Y has density q_θ. The Fisher Information from X is

$$
I_X(\theta) = Var_\theta \left(\frac{\partial \log p_\theta(x)}{\partial \theta} \right).
$$

Correspondingly, the Fisher Information from Y is

$$
I_Y(\theta) = Var_\theta \left(\frac{\partial \log q_\theta(y)}{\partial \theta} \right).
$$

Then

$$
\begin{aligned}
I_{X,Y}(\theta) &= Var_\theta \left(\frac{\partial \log\{p_\theta(x) q_\theta(y)\}}{\partial \theta} \right) \\
&= Var_\theta \left(\frac{\partial \log\{p_\theta(x)\}}{\partial \theta} + \frac{\partial \log\{q_\theta(y)\}}{\partial \theta} \right) \\
&= Var_\theta \left(\frac{\partial \log\{p_\theta(x)\}}{\partial \theta} \right) + Var_\theta \left(\frac{\partial \log\{q_\theta(y)\}}{\partial \theta} \right) \\
&= I_X(\theta) + I_Y(\theta).
\end{aligned}
$$

Suppose we have $X_1, ..., X_n \overset{i.i.d}{\sim} p_\theta$, $I_{\mathbf{x}} = I_{X_1}(\theta) + \cdots + I_{X_n}(\theta) = nI_{X_1}(\theta)$.
Then

$$
Var_\theta(\delta) \geq \frac{g^{'}(\theta)}{nI(\theta)}.
\tag{11}
$$

Multi-dimensional Fisher Information:
Suppose $\boldsymbol{\theta}$ takes values in $\mathbb{R}^k$, then the Fisher Information will become a matrix defined in regular case by

$$
\begin{aligned}
\{\boldsymbol{I}(\theta)\}_{i,j} &= E_\theta\left(\frac{\partial \log p_\theta(x)}{\partial \theta_i}\frac{\partial \log p_\theta(x)}{\partial \theta_j}\right) \\
&\quad (E_\theta(\nabla_\theta \log p_\theta(x)) = 0) \\
&= Cov_\theta\left(\frac{\partial \log p_\theta(x)}{\partial \theta_i}, \frac{\partial \log p_\theta(x)}{\partial \theta_j}\right) \\
&= -E_\theta\left(\frac{\partial^2 \log p_\theta(x)}{\partial \theta_i \partial \theta_j}\right).
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{I}(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}(\{\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x)\}\{\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x)\}^\top) \\
&= Cov(\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x)) = -E_{\boldsymbol{\theta}} \nabla^2_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x)
\end{aligned} \tag{12}
$$

Where $\nabla_{\boldsymbol{\theta}}$ is the gradient with respect to $\boldsymbol{\theta}$ and $\nabla^2_{\boldsymbol{\theta}}$ is the Hessian matrix for the second order derivatives.

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

$$
Var_{\boldsymbol{\theta}}(\delta) \geq \{\nabla g(\boldsymbol{\theta})\}^T I^{-1}(\boldsymbol{\theta})\{\nabla g(\boldsymbol{\theta})\}. \tag{13}
$$

3 Next Lecture

Average Risk Optimality:
Originally, we have

$$
R(\theta, \delta) = E_\theta(L(\theta, \delta(x)))
$$

The average risk:

$$
r(\Lambda, \delta) = \int R(\theta, \delta) d\Lambda(\theta)
$$

where, $\Lambda(\theta)$ is the prior distribution on θ.
We aim to find

$$
r(\Lambda, \delta^*) \leq r(\Lambda, \delta).
$$

STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Huan Cheng; Ip Man Fai

Lecture 8

---

1 Bayes Estimators and Average Risk Optimality

We need to introduce a measure Λ over the parameter space Ω. This measure Λ can be viewed as an assignment if weights to each of the parameters values $\theta \in \Theta$ a priori. [i.e. before any data is observed]

Remark 1 *The parameter of interest θ is not fixed and unknown constant.*

Given a measure Λ, our objective is to find an estimator δ_Λ which minimizes the average risk, which is given by

$$r(\Lambda, \delta) = \int R(\theta, \delta) d\Lambda(\theta) = E_\theta(R(\theta, \delta)). \tag{1}$$

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

$$r(\Lambda, \delta) = E_{(X,\Theta)}(L(\Theta, \delta(X))) = E_\Theta(E_X(L(\Theta, \delta(X) \mid \Theta))) = E_\Theta(R(\Theta, \delta)). \tag{2}$$

We shall pay attention to $E(L(\Theta, \delta(X)) \mid X = x)$, the conditional risk at (almost) every value of X. Notice that the expectation have is taken with respect to the conditional distribution of Θ given X, i.e. $(\Theta \mid X = x)$.

Theorem 1 *Suppose $\Theta \sim \Lambda$ and $X \mid \Theta = \theta \sim P_\theta$. If*

(a) There exists δ_0, an estimator of $g(\theta)$ with finite risk for all θ, and

(b) There exists a value $\delta_\Lambda(X)$ that minimizes $E(L(\Theta, \delta_\Lambda(X)) \mid X = x)$ for almost every X,

then δ_Λ is a Bayes estimator with respect to Λ.

Note that the almost sure statement is defined with respect to the marginal distribution of X, which is given by

$$P(X \in A) = \int P_\theta(X \in A) d\Lambda(\theta) \tag{3}$$

Proof 1 *Under the assumptions of theorem (a) and (b), for any other estimator δ', say, and for almost surely X, $E(L(\Theta, \delta_\Lambda(X)) \mid X = x) \leq E(L(\Theta, \delta'_\Lambda(X)) \mid X = x)$. After taking expectation over X, we obtain $E(L(\Theta, \delta_\Lambda(X))) \leq E(L(\Theta, \delta'_\Lambda(X)))$ for all δ'.*

Example 1 (Bayes estimator of L^2 loss) *If we consider the squared loss function $L(\theta, d) = (\theta - d)^2$, to find the Bayes estimator. We need to minimize $E((g(\Theta) - \delta(X))^2 \mid X = x)$ and in this case, the Bayes estimator is $\delta_\Lambda(X) = E(g(\Theta) \mid X)$, the posterior mean of $g(\Theta)$ given $X = x$*

Consider the Risk function, $E(L(\Theta, \delta(X)) \mid X = x)$, we can observe that

$$
\begin{aligned}
&E(\{g(\Theta) - E(g(\Theta) \mid X) + E(g(\Theta) \mid X) - \delta(X)\}^2 \mid X = x) \\
&= E(\{g(\Theta) - E(g(\Theta) \mid X)\}^2 \mid X = x) + E(\{E(g(\Theta) \mid X) - \delta(X)\}^2 \mid X = x)
\end{aligned}
$$

which shows the risk function could be minimized by posterior mean if it is the Bayes estimator.

Remark 2 *To calculate the posterior mean $E(g(\Theta) \mid X)$, we should find out the posterior distribution first. Since posterior = joint / marginal = priori $\times$ likelihood / marginal , which is equivalent to $p(\theta \mid X) = p(\theta, X)/\int p(\theta', X)d\theta' = p(X \mid \theta) \times \pi(\theta)/\int p(\theta', X)d\theta'$ by Bayes's Theorem , posterior distribution could be derived as posterior $\propto$ prior $\times$ likelihood.*

Example 2 (Binomial-Beta) *Suppose $X \sim Binomial(n, \theta)$ given $\Theta = \theta$ and that Θ has a prior distribution $Beta(\alpha, \beta)$, with hyperparameters α and β. The prior density is given by*

$$
\pi(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1}(1 - \theta)^{\beta - 1} \mathbf{1}_{\{0 < \theta < 1\}}. \tag{4}
$$

Obviously, the model density is $f(X; \theta) = \binom{n}{x}\theta^x(1 - \theta)^{n-x}$, in which case the posterior distribution of Θ given X is

$$
\begin{aligned}
\pi(\theta \mid X) &\propto \underbrace{\binom{n}{x}\theta^x(1 - \theta)^{n-x}}_{\textit{Likelihood}} \underbrace{\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha - 1}(1 - \theta)^{\beta - 1}}_{\textit{Prior}} \\
&\propto \underbrace{\theta^{(x+\alpha)-1}(1 - \theta)^{(n-x+\beta)-1}}_{\textit{Kernel part}} \\
&\sim Beta(x + \alpha, n - x + \beta),
\end{aligned}
$$

where $\int p(\theta', X)d\theta'$ (the denominator part of posterior) is normalising constant, meaning that the posterior of $\Theta \mid X = (x + \alpha)/(n + \alpha + \beta)$.

Remark 3 *The posterior mean can be rewritten as:*

$$
\overbrace{\frac{X + \alpha}{n + \alpha + \beta}}^{\textit{Shrink the estimate from prior mean}} = \underbrace{\frac{n}{n + \alpha + \beta}}_{\omega}\left(\frac{X}{n}\right) + \overbrace{\underbrace{\frac{\alpha + \beta}{n + \alpha + \beta}}_{1-\omega}\left(\frac{\alpha}{1 + \beta}\right)}^{\textit{Contribution from prior}}
$$

ω and $1 - \omega$ can be treated as the weight average of the sample mean $\bar{X}_n$ and the prior mean $\alpha/(\alpha + \beta), correspondingly.$ As $n \to \infty$ (by empirical evidence and observations), $E(\Theta \mid X) \to \bar{X}_n$. (Let the data "speak for themselves.")

Example 3 (Normal Mean Estimation) *Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\Theta, \sigma^2)$, with σ^2 known. Let $\Theta \sim N(\mu, b^2)$*

where μ and b^2 are two fixed prior hyperparameters. Then the posterior distribution of $\Theta \mid X$ is

$$\begin{aligned}
\pi(\theta \mid X) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(X_i-\theta)^2\right\} \times \frac{1}{\sqrt{2\pi b^2}} \exp\left\{-\frac{1}{2b^2}(\theta-\mu)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2 - \frac{1}{2b^2}(\theta-\mu)^2\right\} \\
&\propto \cdots \\
&\propto \exp\left\{-\frac{1}{2}\Big(\frac{n}{\sigma^2}+\frac{1}{b^2}\Big)\theta^2 + \Big(\frac{n\bar{X}}{\sigma^2}+\frac{\mu}{b^2}\Big)\theta\right\} \\
&\propto \exp\left\{-\frac{1}{2\tilde{\sigma}^2}(\theta-\tilde{\mu})^2\right\}.
\end{aligned}$$

The posterior distribution of Θ given X is $N(\tilde{\mu}, \tilde{\sigma}^2)$ where

$$\tilde{\mu} = \frac{n\bar{X}/\sigma^2 + \mu/b^2}{n/\sigma^2 + 1/b^2} \qquad \textit{and} \qquad \tilde{\sigma}^2 = \frac{1}{n/\sigma^2+1/b^2}$$

Hence, the posterior mean of $\Theta \mid X$ is $\dfrac{n\bar{X}/\sigma^2 + \mu/b^2}{n/\sigma^2 + 1/b^2}$ and similarly we can rewrite as

$$\underbrace{\frac{n/\sigma^2}{n/\sigma^2+1/b^2}}_{1 \text{ as } n\to\infty}\bar{X} + \underbrace{\frac{1/b^2}{n/\sigma^2+1/b^2}}_{0 \text{ as } n\to\infty}\mu$$

Thus, Bayes estimator δ_Λ is $\tilde{\mu}$ if we adopt the squared loss function.

Example 4 (Bayes estimator of weighted L^2 loss) *Assume that we consider $L(\theta, d) = \omega(\theta)\{d - g(\theta)\}^2$, where $\omega(\theta) \geqslant 0$, which can be interpreted as a weight function. Our goal is to find the corresponding Bayes estimator, which minimizes $E(\omega(\Theta)\{g(\Theta) - d\}^2 \mid X = x)$ $(*)$with respect to d.*
$(*)$ can be rewritten as

$$d^2 E(\omega(\Theta) \mid X = x) - 2dE(\omega(\Theta)g(\Theta) \mid X = x) + E(\omega(\Theta)g(\Theta)^2 \mid X = x). \qquad (\dagger)$$

Tanking derivative of $(\dagger)$ with respect to d, we obtain

$$2d^* E(\omega(\Theta) \mid X = x) - 2E(\omega(\Theta)g(\Theta) \mid X = x) = 0.$$

Thus

$$\delta_\Lambda(x) = d^* = \frac{E(\omega(\Theta)g(\Theta) \mid X = x)}{E(\omega(\Theta) \mid X = x)}. \qquad (5)$$

In particular, if $\omega(\cdot) \equiv 1$, $\delta_\Lambda(x)$ (with $\omega(\cdot) \equiv 1$)=$E(g(\Theta) \mid X = x)$.

Theorem 2 *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 loss function unless its average risk is zero, which is*

$$E_{(X,\Theta)}(\{\delta(X) - g(\Theta)\}^2) = 0. \qquad (6)$$

Proof 2 *Let δ be an unbiased estimator under the squared loss function. Then we know that δ is the posterior mean, which is*

$$\delta(X) = E(g(\Theta) \mid X),$$

almost surely. Thus, we have

$$\begin{aligned} E(\delta(X)g(\Theta)) &= E(E(\delta(X)g(\Theta) \mid X)) \\ &= E(\delta(X)E(g(\Theta) \mid X)) \\ &= E(\delta^2(X)). \end{aligned} \tag{7}$$

Also,

$$\begin{aligned} E(\delta(X)g(\Theta)) &= E(E(\delta(X)g(\Theta) \mid \Theta)) \\ &= E(g(\Theta)E(\delta(X) \mid \Theta)) \\ &= E(g^2(\Theta)). \end{aligned} \tag{8}$$

Observe that

$$\begin{aligned} E(\{\delta(X) - g(\Theta)\}^2) &= E(\delta^2(X)) - 2E(\delta(X)g(\Theta)) + E(g^2(\Theta)) \\ &= E(\delta^2(X)) - E(\delta(X)g(\Theta)) + E(g^2(\Theta)) - E(\delta(X)g(\Theta)) \\ &= E(\delta^2(X)) - E(\delta^2(X)) + E(g^2(\Theta)) - E(g^2(\Theta)) \ (due\ to\ (7)\ and\ (8)) \\ &= 0. \end{aligned}$$

Thus we have that $E(\{\delta(X) - g(\Theta)\}^2) = 0$, which means the average risk is zero. The claim is thus proved.

Example 5 (Application of Theorem 2) *Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\Theta, \sigma^2)$, with σ^2 known. Is $\bar{X}$ Bayes under the squared loss function for some choice of the prior distribution?*
Observe that $E(\bar{X} \mid \theta) = \theta$, hence $\bar{X}$ is unbiased for θ. The corresponding average risk under the squared loss function is given by

$$E_{(X,\Theta)}(\{\bar{X} - \Theta\}^2) = \frac{\sigma^2}{n} \neq 0.$$

So $\bar{X}$ is not Bayes estimator under any prior distribution.

Theorem 3 (Admissibility) *A unique Bayes estimator (almost surely for all P_θ) is admissible.*
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

$$\begin{cases} R(\theta, \delta') \leq R(\theta, \delta), \ for\ any\ \theta \in \Omega \\ R(\theta, \delta') < R(\theta, \delta), \ for\ some\ \theta \in \Omega \end{cases}$$

Proof 3 *Suppose δ_Λ is Bayes for Λ, and for some δ', $R(\theta, \delta') \leq R(\theta, \delta_\Lambda)$ for all $\theta \in \Omega$. If we take expectation with respect to Θ, the inequality above is preserved and we can write*

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

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 probability one for all P_θ.

Question: When is a Bayes estimator unique?

Theorem 4 (Uniqueness) *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 (almost surely for all P_θ) if
(a) $r(\Lambda, \delta_\Lambda)$ is finite and
(b) $P_\theta \ll Q$ (absolute continuity)

$$\text{Benefits of Bayes} \begin{cases} (i)\ \text{Admissible} \\ (ii)\ \text{Incorporate} \underbrace{\text{prior information}}_{\text{domain knowledge}} \longrightarrow \text{frequentist} \\ (iii) \ldots \end{cases}$$

2 Next Lecture

1. Minimax Estimator

   Considering

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

2. Worst-case Scenario/Optimality

3. Testing of Statistical Hypothesis (UMP, UMPU...)

STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Qinghua (Devon) Ding and Dongrong Li

Lecture 9

1 Minimax Estimators and Worst-Case Optimality

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

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

$$r_\Lambda = \inf_\delta r(\Lambda, \delta) = \inf_\delta \int_{\theta\in\Omega} R(\theta, \delta) d\Lambda(\theta)$$

Definition 1. *A prior distribution is said to be a least favorable prior if $r_\Lambda \geq r_{\Lambda'}$, for any other prior distribtion Λ'.*

Following the definition is the theorem:

Theorem 2 (TPE 5.1.4)**.** *Suppose δ_Λ is Bayes for Λ with*

$$r_\Lambda = \sup_\theta R(\theta, \delta_\Lambda)$$

i.e. the Bayes risk of δ_Λ is the maximum risk of δ_Λ, then:

(i) *δ_Λ is minimax,*

(ii) *Λ is a least favorable prior,*

(iii) *If δ_Λ is the unique Bayes estimator for Λ almost surely, for all P_θ, then it is a unique minimax estimator.*

Proof. (i) Let δ be any other estimator, then we have that:

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

This implies that δ_Λ is minimax.

(ii) If $\delta\Lambda$ is the unique Bayes estimator, then the ineqality above $(*)$ is strict for $\delta \neq \delta_\Lambda$, which implies that δ_Λ is the unique minimax.

(iii) Let Λ' be any other prior distribution, then

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

Since the worst case risk of δ_Λ is its Bayes risk over Λ, we know that Λ is a least favorable prior distribution.

□

An implication is that we can find a minimax estimator by finding a Bayes estimator with Bayes risk equals its maximum risk, which gives the following corollary:

Corollary 3 (TPE 5.1.5)**.** *If a Bayes estimator of δ_Λ has constant risk, i.e. $R(\theta, \delta_\Lambda) = R(\theta', \delta_\Lambda)$ for any $\theta, \theta' \in \Omega$, then δ_Λ is minimax.*

An implication of this corollary is that, if a Bayes estimator has constant risk, it is minimax too. We may find a prior support set ω such that $\Lambda(\omega) = 1$ and for which $R(\theta, \delta_\Lambda)$ is maximum for any $\theta \in \Omega$.

Corollary 4 (TPE 5.1.6)**.** *Define $\omega_\Lambda = \{\theta : R(\theta, \delta_\Lambda) = \sup_{\theta'} R(\theta', \delta_\Lambda)\}$. A Bayesian estimator δ_Λ is minimax if $\Lambda(\omega_\Lambda) = 1$.*

Example 1. *Suppose $X \sim Binomial(n, \theta)$ for some $\theta \in (0, 1)$ and we adopt the squared loss function, is $\frac{x}{n}$ minimax?*

Notice that the corresponding risk is $R(\theta, \frac{x}{n}) = \frac{\theta(1-\theta)}{n}$. Observe that the risk has a unique maximum at $\theta = \frac{1}{2}$. The worst risk is:

$$\sup_{\theta \in \Omega} R(\theta, \frac{x}{n}) = R(\frac{1}{2}, \frac{1}{2}) = \frac{1}{4n}$$

In this case, [TPE 5.1.6] is not helpful because if $\Lambda(\{\frac{1}{2}\}) = 1$, then $\delta_\Lambda(X) = \frac{1}{2} \neq \frac{x}{n}$.

However, [TPE 5.1.5] can be helpful instead. To find a minimax estimator, we will need to search for a prior such that the Bayes estimator has constant risk.

Recall that if the prior is $Beta(\alpha, \beta)$, the Bayes estimator under the squared loss is:

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

for any α, β.

$$\begin{aligned} R(\theta, \delta_{\alpha,\beta}) &= \mathbb{E}_\theta \left(\left\{ \frac{x + \alpha}{n + \alpha + \beta} - \theta \right\}^2 \right) \\ &= \frac{1}{(n + \alpha + \beta)^2} \mathbb{E}_\theta \left(\{x - n\theta - \alpha(\theta - 1) - \theta\beta\}^2 \right) \\ &= \frac{1}{(n + \alpha + \beta)^2} \left[n\theta(1 - \theta + \{\alpha(\theta - 1) + \theta\beta\}^2) \right] \end{aligned}$$

To eliminate the θ dependence in $R(\theta, \delta_{\alpha,\beta})$, we need to set the coefficients of θ^2 and θ be zero, that is:

$$\begin{aligned} -n + (\alpha + \beta)^2 &= 0 \\ n - 2\alpha(\alpha + \beta) &= 0, \end{aligned}$$

which solves $\alpha = \beta = \frac{\sqrt{n}}{2}$. The Bayes estimator $\delta_{\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}}(X) = \frac{X + \sqrt{n}/2}{n + \sqrt{n}}$ is minimax (TPE 5.1.4) with constant risk of $\frac{1}{4(\sqrt{n}+1)^2}$, we can conclude that $\frac{X}{n}$ is not minimax.

2 Generalization of Minimax-Bayes Theorems

We remark that minimax estimators may not be Bayes estimators. This is illustrated in the following example.

Example 2 (minimax for normal with unknown mean θ)**.** *Let $X_1, X_2, ..., X_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$ with σ^2 unknown. Our goal is to estimate θ under the squared loss function. Our candidate is $\overline{X}$, which has constant risk $R(\theta, \overline{X}) = \mathbb{E}_\theta[(\overline{X} - \theta)^2] = \frac{\sigma^2}{n}$. This suggests that $\overline{X}$ can be a minimax estimator (TPE 5.1.4 and 5.1.5). However, $\overline{X}$ is not Bayes for any prior (Example 5, Lecture 8 and TPE 4.2.3).*

We also recall *TPE 4.2.3* here.

Theorem 5 (TPE 4.2.3)**.** *Unbiased estimators are Bayes only in the degenerate case of zero risk, i.e.,*

$$\mathbb{E}_{\Theta,X}\left[\{\delta(X) - g(\Theta)\}^2\right] = 0.$$

Thus we cannot yet conclude that $\overline{X}$ is minimax. We now consider the family of estimators with the form $\delta_{\omega,\mu_0}(X) = \omega\overline{X} + (1-\omega)\mu_0$, where $\omega \in (0,1)$ and $\mu_0 \in \mathbb{R}$. However, the worst case risk for this family of estimators is infinite.

$$\begin{aligned}
\sup_\theta \ \mathbb{E}_\theta\left[(\theta - \delta_{\omega,\mu_0}(X))^2\right] &= \sup_\theta \ \mathbb{E}_\theta\left[(\theta - \omega\overline{X} - (1-\omega)\mu_0)^2\right] \\
&= \sup_\theta \ \mathbb{E}_\theta\left[(\omega(\overline{X} - \theta) + (1-\omega)(\mu_0 - \theta))^2\right] \\
&= \sup_\theta \ \omega^2 \text{Var}\left(\overline{X}\right) + (1-\omega)^2(\mu_0 - \theta)^2 \\
&= \sup_\theta \ \frac{\omega^2\sigma^2}{n} + (1-\omega)^2(\mu_0 - \theta)^2 \\
&= +\infty
\end{aligned}$$

These estimators have much poorer worst-case risk than $\overline{X}$, hence they are certainly not minimax. To prove that $\overline{X}$ is indeed a minimax estimator, we need to generalize the previous definitions and theorems in the following way.

Definition 6 (Least Favourable Sequence of Priors)**.** *Let $\{\Lambda_m\}$ be a sequence of priors with minimal average risk*

$$r_{\Lambda_m} = \inf_\delta \int_\Omega R(\theta, \delta)\, \mathrm{d}\Lambda_m(\theta).$$

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 $r \geq r_{\Lambda'}$ for any prior Λ'.

Theorem 7 (TPE 5.1.12)**.** *Suppose there is a real number r such that $\{\Lambda_m\}$ is a sequence of priors with $r_{\Lambda_m} \to r < \infty$. Let δ be any estimator such that $\sup_\theta R(\theta, \delta) = r$. Then we have*

(i) δ is minimax;

(ii) $\{\Lambda_m\}$ is least-favourable.

Proof. (i) Let δ' be any other estimator. Then for any m, we have

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

Then sending $m \to \infty$ yields

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

which implies that δ is minimax.

(ii) Let Λ' be any prior, then

$$r_{\Lambda'} = \int_\Omega R(\theta, \delta_{\Lambda'})\, \mathrm{d}\Lambda'(\theta) \leq \int_\Omega R(\theta, \delta)\, \mathrm{d}\Lambda'(\theta) \leq \sup_\theta R(\theta, \delta) = r,$$

which means that $\{\Lambda_m\}$ is least favourable.

□

Remark.

1. *Unlike Theorem 5.1.4 (TPE), this theorem does not guarantee the uniqueness of the minimax estimator even if the Bayes estimators δ_{Λ_m}'s are unique. The problem arises from the step that we send m to the limit.*

2. *This theorem allows us to consider a much wider class of estimators, instead of limiting our attentions to Bayes estimators only. Specifically, we may also consider the estimators that comes from a sequence of priors.*

Example 3 (cont'd)**.** *If we manage to find a sequence of priors $\{\Lambda_m\}$ such that $r_{\Lambda_m} \to \frac{\sigma^2}{n} = r$, then we can obtain a minimax estimator for θ. Let consider the sequence of priors $\Lambda_m \sim \mathcal{N}(0, m^2)$ (Λ_m will tend to the uniform prior over $\mathbb{R}$ which is improper with $\pi(\theta) = 1$ for any $\theta \in \mathbb{R}$). This will yield the following posterior distribution.*

$$\begin{aligned} f(\theta|x_1, ..., x_n) &\propto \pi(\theta) \cdot f(x_1, ..., x_n|\theta) \\ &\propto \exp\left(-\frac{\theta^2}{2m^2} - \frac{\sum_{i=1}^n (x_i - \theta)^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{m^2} + \frac{n}{\sigma^2}\right)\theta^2 + \frac{n\overline{x}}{\sigma^2} \cdot \theta\right) \\ &\sim \mathcal{N}\left(\frac{\frac{n\overline{x}}{\sigma^2}}{\frac{1}{m^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{m^2} + \frac{n}{\sigma^2}}\right) \end{aligned}$$

Note that the posterior variance does not depend on $(X_1, ..., X_m)$, hence

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

It now follows from Theorem 5.1.12 (TPE) that $\overline{X}$ is minimal and $\{\Lambda_m\}$ is least favourable.

We remind that the choice of loss function will also influence the corresponding minimax estimators. Specially, we consider the following example.

Example 4 (weighted squared loss)**.** *Let $X \sim$ Binomial(n, θ) with the loss function $L(\theta, d) = \frac{(d-\theta)^2}{\theta(1-\theta)}$. We may view this loss function as the weighted squared loss function with weights $w(\theta) = \frac{1}{\theta(1-\theta)}$.*

Note that for any θ, $R(\theta, X/n) = \frac{1}{n}$, which is constant in θ. This suggests that X/n can be minimax. But be reminded that we cannot directly apply TPE 4.2.3 because L is not the vanilla squared loss function.

Consider the prior $\Theta \sim \Lambda_{\alpha,\beta} =$ Beta(α, β), for some $\alpha, \beta > 0$. By results in Lecture 8, we have $\Theta|X \sim$ Beta$(X + \alpha, n - X + \beta)$ and we can find the Bayes estimator as

$$\boxed{\delta_\Lambda(X) = \frac{\mathbb{E}_{\Theta|X}\left(\frac{1}{1-\Theta}\middle|X\right)}{\mathbb{E}_{\Theta|X}\left(\frac{1}{\Theta(1-\Theta)}\middle|X\right)}}$$

Suppose we have observed $X = x$ with $\alpha + x > 1$ and $n + \beta + x > 1$, then the resulting Bayes estimator is

$$\delta_{\alpha,\beta}(x) = \frac{\alpha + x - 1}{\alpha + \beta + n - 2}.$$

In particular, when $\alpha = \beta = 1$, we have $\delta_{1,1}(x) = x/n$ minimizes posterior risk under prior $\Lambda_{1,1}$ after observing $0 < x < n$.

When $x \in \{0, n\}$, then the posterior risk under the prior $\Lambda_{1,1}$ after observing $X = x$ and deciding $\delta(x) = d$ is

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

which for $x = 0$ reduces to $\int_0^1 \frac{(n+1)(1-\theta)^{n-1}(d-\theta)^2}{\theta} \, \mathrm{d}\theta$. Note this converges only when $\delta(0) = 0$. Similarly, one can deduce that $\delta(n) = 1$.

Now we may conlcude that X/n minimizes the posterior risk under prior distribution $\Lambda_{1,1}$ for any outcome X. Hence X/n is indeed minimax under such weighted squared loss function.

3 Next Lecture

1. Admissibility of minimax estimators;
2. Hypothesis testing (NP lemma/UMP).

STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Chuchu Wang and Gan Wu

Lecture #10

---

1 Admissibility of Minimax Estimator

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

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

Admissibility implies maximaxity: if δ is admissible with constant risk, then δ is also minimax.

Proof 1 (Argument) *Let the constant risk of δ be r, then r is also the worst case risk of δ, as the risk is constant. Now if we assume that δ is not minimax, there exists a different estimator, say δ', such that it is minimax (with the corresponding risk as $r' < r$). But since this is the worst case risk of δ', it implies that the risk of δ' is lower than r throughout, and thus δ' dominates δ. This is a contradiction, as δ is admissible, which implies that δ is minimax.*

Note, however, that minimaxity does not guarantee admissibility (need to check case-by-case).

Example 1 *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\theta, \sigma^2)$, where σ^2 is known, and the parameter θ is the estimand. The minimax estimator is $\bar{X}$ under the squared error loss function.*

Question: is $\bar{X}$ admissible?
[A more general question: when is $a\bar{X} + b, a, b \in \mathbb{R}$ (any affine function of $\bar{X}$) admissible?]

Case 1 *$0 < a < 1$:*
Observe that $a\bar{X} + b$ is a convex combination of $\bar{X}$ and b. It is a Bayes estimator with respect to some Gaussian prior on θ. Since we are considering the squared error loss function, which is strictly convex, the Bayes estimator is unique. By Theorem 5.2.4 (TPE), $a\bar{X} + b$ is therefore admissible.

Case 2 *$a = 0$:*
In this case, b is also a unique Bayes estimator with respect to a degenerate prior distribution with unit mass at $\theta = b$. ($\Lambda(\theta) = N(b, 0^2)$). So by Theorem 5.2.4, b is admissible.

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

Case 4 *$a > 1$:*

$$\textit{Risk of } a\bar{X} + b = \mathbb{E}[(a\bar{X} + b - \theta)^2] = \mathbb{E}[(a(\bar{X} - \theta) + b + \theta(a-1))^2] = \frac{a^2\sigma^2}{n} + (b + \theta(a-1))^2$$

So, when $a > 1$,

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

Hence, $\bar{X}$ dominates $a\bar{X} + b$ when $a > 1$, and so in this case $a\bar{X} + b$ is inadmissible.

Case 5 $a < 0$*:*

$$\mathbb{E}[(a\bar{X} + b - \theta)^2] > (b + \theta(a-1))^2 = (a-1)^2(\theta + \frac{b}{a-1})^2 > (\theta + \frac{b}{a-1})^2$$

and this is the risk of predicting the constant $-\frac{b}{a-1}$*. So* $-\frac{b}{a-1}$ *dominates* $a\bar{X} + b$*. Hence,* $a\bar{X} + b$ *is inadmissible.*

Case 6 $a = 1, b = 0$*:*
We use a limiting Bayes argument. Suppose $\bar{X}$ *is inadmissible. WLOG, we assume that* $\sigma^2 = 1$*, and have*

$$R(\theta, \bar{X}) = \frac{1}{n}$$

By our hypothesis, there must exist an estimator δ' *such that* $R(\theta, \delta') \leq \frac{1}{n}$ *for all* θ*, and* $R(\theta', \delta') < \frac{1}{n}$ *for some* $\theta' \in \Omega$ *[at least one]. Because* $R(\theta, \delta)$ *is continuous in* θ*, there must exist* $\epsilon > 0$ *and an interval* (θ_0, θ_1) *containing* θ' *such that*

$$R(\theta, \delta') < \frac{1}{n} - \epsilon, \forall \theta \in (\theta_0, \theta_1) \tag{*}$$

Let r'_τ *be the average risk of* δ' *with respect to the prior distribution* $N(0, \tau^2)$ *on* θ*. 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* $R(\theta, \delta') \leq \frac{1}{n}$ *for all* $\theta \in \Omega = \mathbb{R}$*. So,* $r_\tau < r'_\tau$ *because the Bayes estimator is unique almost surely w.r.t. the marginal distribution of* θ*.*

$$\frac{\frac{1}{n} - r'_\tau}{\frac{1}{n} - r_\tau} = \frac{\frac{1}{\sqrt{2\pi\tau^2}} \int_{-\infty}^{\infty} \{\frac{1}{n} - R(\theta, \delta')\} \exp(-\frac{\theta^2}{2\tau^2}) d\theta}{\frac{1}{n} - \frac{1}{n + \frac{1}{\tau^2}}} \tag{\#}$$

By (*), we can simplify (#) as follows

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

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

$$\frac{\frac{1}{n} - r'_\tau}{\frac{1}{n} - r_\tau} > 1$$

This means that $r'_\tau < r_\tau$*. But this is a contradiction because* r_τ *is the optimal average risk. So our assumption that there was a dominating estimator is incorrect, in which case* $a\bar{X} + b = \bar{X}$ *is admissible.*

James-Stein (JS) Estimator (empirical Bayes)

Reference: TPE 5.4 - 5.5, simultaneous estimation
Let $X_1, \ldots, X_n$ *with* $X_i \sim N(\theta_i, \sigma^2)$ *for* $1 \leq i \leq n$*. Our goal is to estimate* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ *under the loss function*

$$L(\boldsymbol{\theta}, \boldsymbol{d}) = \sum_{i=1}^{n} (d_i - \theta_i)^2$$

Then,

$$\delta_i(\boldsymbol{X}) = \max\left(1 - \frac{p-2}{\|\boldsymbol{X}\|_2^2}, 0\right) X_i$$

"OPTIMAL" INFERENCE (no uniform optimality) $\Longleftrightarrow$ Decision Theory

$$\begin{cases} \text{I. Constraint} \begin{cases} \text{UMVU (unbiasedness)} \\ \text{Equivariance (Ch.3, TPE)} \end{cases} & \textit{Related to data compression, sufficient statistics, testing, } \ldots \\ \text{II. Collapse} \begin{cases} \text{Average risk (Bayes)} \\ \text{Worst case (Minimax)} \end{cases} & \textit{Related properties: admissibility, uniqueness, } \ldots \end{cases}$$

Considered in finite sample case.

2 Testing of Statistical Hypotheses

2.1 Another decision problem: hypothesis testing

We assumme that the data is sampled according to $X \sim P_\theta$, where $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$.

Two disjoint subclasses of θ (hypotheses):

$$H_0 : \theta \in \Omega_0 \subset \Omega \ \text{(null hypothesis)}$$
$$H_1(\text{or } H_a) : \theta \in \Omega_1 = \Omega/\Omega_0 \ \text{(alternative hypothesis)}$$

Decision space $\mathcal{D} = \{\text{Reject } H_0, \text{ not to reject } H_0 \text{ (or accept } H_0)\}$.
e.g. H_0 : no change in infection rate (default).

H_1 : improvements in lowering infection rate.

Truth / Decision	$\theta \in \Omega_0$	$\theta \in \Omega_1$
Reject H_0	1 Type I error (More severe)	0 (Good)
Accept H_0	0 (Good)	1 Type II error

Terminologies

⋆ Test function/critical function : $\phi(x) \in [0, 1]$

$$\phi(x) = P\left(\delta_\phi(x, u) = \text{Reject } H_0 | x\right)$$

where u is a uniform random variable independent of X.

⋆ Power function of a test ϕ is $\beta(\theta) = \mathbb{E}_\theta\left(\phi(X)\right) = P_\theta(\text{Reject } H_0)$.

Note:

If $\theta_0 \in \Omega_0$, then $\beta(\theta_0) = R(\theta_0, \delta_\phi)$ = Type I error.

For $\theta_1 \in \Omega_1$, then $\beta(\theta_1) = 1 - R(\theta_1, \delta_\phi)$ = 1-Type II error.

Our "ideal" optimality goal is to minimize $\beta(\theta_0)$ uniformly for all $\theta_0 \in \Omega_0$ and maximize $\beta(\theta_1)$ uniformly for all $\theta_1 \in \Omega_1$.

Neyman-Pearson Framework

Control the level of significance

$$\sup_{\theta_0 \in \Omega_0} \mathbb{E}_{\theta_0} \phi(X) = \sup_{\theta_0 \in \Omega_0} \beta(\theta_0) \leq \alpha$$

where $\sup_{\theta_0 \in \Omega_0} \beta(\theta_0)$ is called the size of the test.

Optimality Goal: Find a level α test that maximizes the power $\beta(\theta_1) = \mathbb{E}_{\theta_1}(\phi(X))$ for each $\theta_1 \in \Omega_1$. Such a test is called a uniformly powerful (UMP) test.

MP for the "simple-vs-simple" case

Definition 7 *A hypothesis H_0 is called **simple** if $|\Omega_0| = 1$, otherwise it is called **composite**. This applies to H_1 as well.*

Hence a simple-vs-simple test:

$$H_0 : X \sim p_0 \ \ (p_0 = P_{\theta_0})$$
$$H_1 : X \sim p_1 \ \ (p_1 = P_{\theta_1})$$

Our goal is to find ϕ:

$$\max_{\phi} \mathbb{E}_{p_1}(\phi(X)) \quad \text{subject to} \ \ \mathbb{E}_{p_0}(\phi(X)) \leq \alpha$$

Lemma 1 (Neyman-Pearson Lemma) .

(i) Existence. For testing $H_0 : p_0$ vs $H_1 : p_1$, there exists a test $\phi(X)$ and a constant k such that

(a) $E_{p_0}(\phi(X)) = \alpha$ (size = level)

(b)

$$\phi(x) = \begin{cases} 1, & \text{if } \frac{p_1(x)}{p_0(x)} > k \ \ [\text{Rejection}] \\ 0, & \text{otherwise } \ [\text{Acceptance}] \end{cases}$$

such a test is called a likelihood ratio test.

(ii) Sufficiency: If a test satisfies (a) and (b) for some constant k, it is most powerful for testing $H_0 : p_0$ vs $H_1 : p_1$ at level α.

(iii) Necessity: If a test ϕ is MP at level α, then it satisfies (b) for some k, and it also satisfies (a) unless there exists a test if $size < \alpha$ with power 1.

Example 2 *Let $X_1, ..., X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$ with σ^2 known. We want to test:*

$$H_0 : \mu = 0 \quad vs \quad H_1 : \mu = \mu_1$$

where μ_1 is given. We calculate the likelihood ratio:

$$r(x) = \frac{p_1(x)}{p_0(x)} = \cdots = \exp\left(\frac{1}{\sigma^2}\mu_1 \sum_{i=1}^{n} x_i - \frac{n\mu_1^2}{2\sigma^2}\right)$$

Suppose

$$\begin{aligned} r(x) > k &\iff \mu_1 \frac{\sum_{i=1} x_i}{\sigma^2} - \frac{n\mu_1^2}{2\sigma^2} > \log k \\ &\iff \mu_1 \sum_{i=1}^{n} x_i > k' \\ &\iff \begin{cases} \sum_{i=1}^{n} x_i > k'', & \text{if } \mu_1 > 0 \\ \sum_{i=1}^{n} x_i < k''', & \text{if } \mu_1 < 0 \end{cases} \end{aligned}$$

Let's focus on the case where $\mu_1 > 0$*,*

$$\begin{aligned} r(x) > k &\iff \sum_{i=1}^{n} x_i > k'' \\ &\iff \frac{\sqrt{n}\bar{x}}{\sigma} > k'''' \end{aligned}$$

To calculate k''''*(critical value), we need to evaluate*

$$\mathbb{E}_{p_0}(\phi(X)) = \alpha = P_{\mu=0}\left(\frac{\bar{X}}{\sigma/\sqrt{n}} > k''''\right)$$

where $\frac{\bar{X}}{\sigma/\sqrt{n}}$ *is normally distributed with zero mean.*

STAT 5010: Advanced Statistical Inference

Lecturer: Tony Sit
Scribe: Yuhan HU; Yunjie Liang

Lecture 11

---

11 Testing of Statistical Hypothesis

Lemma 1 *Let P_0 and P_1 be probability distributions possessing densities p_0 and p_1 respectively with respect to a measure μ.*

1. *Existence. For testing $H_0 : p_0$ against the alternative $H_1 : p_1$ there exists a test $\phi(X)$ and a constant k such that*

   (a) $E_{p_0}(\phi(X)) = \alpha$

   (b) $\phi(x) = \begin{cases} 1 & p_1(x) > kp_0(x)\ [rejection] \\ 0 & p_1(x) < kp_0(x)\ [acceptance] \end{cases}$ *, such a test is called a likelihood ratio test.*

2. *Sufficiency: If a test satisfies (a) and (b) for some constant k, it is most powerful for testing $H_0 : p_0$ against $H_1 : p_1$ at level α.*

3. *Necessity: If a test ϕ is MP at level α, then it satisfies (b) for some k, and it also satisfies (a) unless there exists a test of size $< \alpha$ with power 1.*

Proof 1 *Let $\gamma(x) = \frac{p_1(x)}{p_0(x)}$ be the likelihood ratio. Denote the cumulative distribution function of $r(x)$ under H_0 as F_0.*

1. *Existence. Let $\alpha(c) = P_0(\gamma(x) > c) = 1 - F_0(c)$ for $c \in R$. Then $\alpha(c)$ is a non-increasing, right-continuous function of c $[\alpha(c) = lim_{\epsilon \to 0}\alpha(c + \epsilon)]$. Observe also that $\alpha(c)$ is not necessarily left-continuous at every value of c, but the left-hand limits exist.*
   There exists a value c_0 such that $\alpha(c_0) \le \alpha \le \alpha(c_0^-)$. Note that $F(c_0) \ge 1 - \alpha \ge F(c_0^-)$, ie. c_0 is the $1 - \alpha$ quantile of r(x).
   we define our test function to be:

$$\phi(x) = \begin{cases} 1 & if\ \gamma(x) > c_0, \\ \gamma & if\ \gamma(x) = c_0, \\ 0 & if\ \gamma(x) < c_0, \end{cases}$$

   for some constant γ. The test function always rejects the null if the $\gamma(x) > c_0$, and does not reject the null if $\gamma(x) < c_0$.
   The size of the test ϕ is given by

$$\begin{aligned} E_0(\phi(X)) &= P_0(\gamma(X) > c_0) + \gamma P_0(\gamma(X) = c_0) \\ &= \alpha(c_0) + \gamma\{\alpha(c_0^-) - \alpha(c_0)\} \end{aligned}$$

   If $\alpha(c_0) = \alpha(c_0^-)$ [continuous case], then $\alpha(c_0) = \alpha$ and we automatically have $E_0(\phi(X)) = \alpha$ for any choice of γ. Otherwise, we can set

$$\gamma = \frac{\alpha - \alpha(c_0)}{\alpha(c_0^-) - \alpha(c_0)}$$

   which gives also $E_0(\phi(X)) = \alpha$.

2. *Sufficiency: Let ϕ satisfies (a) and (b) in part(1), and let ϕ' be any other level α test, which satisfies*

$$E_0(\phi'(X)) = \int \phi'(x)p_0(x)d\mu(x) \leq \alpha$$

We need to 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.
we claim the following inequality holds:

$$\int \{\phi(x) - \phi'(x)\}\{p_1(x) - kp_0(x)\}d\mu(x) \geq 0$$

To see this, we consider the following three cases:

(a) if $p_1(x) > kp_0(x)$ [$\equiv \gamma(x) > k$], then $\phi(x) = 1$. Since $\phi'(x) \leq 1$, the the integrand is non-negative.

(b) $p_1(x) < kp_0(x)$, ...

(c) $p_1(x) = kp_0(x)$, ...

It implies that

$$\int \{\phi(x) - \phi'(x)\}p_1(x)d\mu(x) \geq k \int \{\phi(x) - \phi'(x)\}p_0(x)d\mu(x) \geq 0$$

meaning that $E_1(\phi(x)) > E_1(\phi'(x))$, ie. ϕ is most powerful at level α.

3. *Necessity: Suppose ϕ^* is most powerful at level α, and let ϕ be a likelihood ratio test satisfying (a) and (b), we want to show that $\phi^*(x) = \phi(x)$ except possibly cases where $\frac{p_1(x)}{p_0(x)} = k$ for μ-a.l. x. Define*

$$S^+ = \{x : \phi(x) > \phi^*(x)\}$$
$$S^- = \{x : \phi(x) < \phi^*(x)\}$$
$$S_0 = \{x : \phi(x) = \phi^*(x)\}$$

and $S = (S^+ \cup S^-) \cap \{x : p_1(x) \neq kp_0(x)\}$. We want to show that $\mu(x) = 0$. Suppose $\mu(x) > 0$, as in part 2, we have $(\phi(x) - \phi'(x))(p_1(x) - kp_0(x)) > 0$ on S. Therefore,

$$\begin{aligned}\int_x (\phi(x) - \phi^*(x))(p_1(x) - kp_0(x))d\mu x &= \int_{S^+\cup S^-} (\phi(x) - \phi^*(x))(p_1(x) - kp_0(x))d\mu x \\ &= \int_S (\phi(x) - \phi^*(x))(p_1(x) - kp_0(x))d\mu x > 0\end{aligned}$$

By hypothesis, $E_0(\phi(x)) = \alpha$ and $E_0(\phi^*(x)) \leq \alpha$, so the previous inequality implies that

$$E_1(\phi(x)) - E_1(\phi^*(x)) > k\{E_0(\phi(x)) - E_0(\phi^*(x))\}$$

$\Rightarrow E_1(\phi(x)) \geq E_1(\phi^*(x))$, which contradicts the assumption that ϕ^* is most powerful. Hence $\mu(S) = 0$. It remains to show that the size of ϕ^* is α unless there exists a test of size which is strictly less than α and power 1. For this, note that if size < α and power < 1, we can add points to rejection region until either the size = α or the power is 1.

Definition 1 *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. [prob(rejecting $H_0|H_1$]).*

Corollary 2 *(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 $P_0 = P_1$).*

Proof 2 *Consider the test $\phi_0(x) \equiv \alpha$, which rejects the null with probability α. Since ϕ_0 is level α, and β is the max power, we have*

$$\beta > E_1(\phi_0(x)) = \alpha$$

Suppose $\beta = \alpha$, then $\phi_0(x) = \alpha$ is a most powerful level α test. As a result,

$$\phi_0(x) = \begin{cases} 1 & if\ \ \gamma(x) > k, \\ 0 & if\ \ \gamma(x < k, \end{cases}$$

as by NP lemma 3 for same k.
Since $\phi_0(x)$ never equals 0 or 1, it must be the case that $p_1(x) = kp_0(x)$ w.p. 1. Note that

$$\int p_1(x)d\mu(x) = k \int p_0(x)d\mu(x) = 1$$

This implies that k=1. Hence $P_0 = P_1$.

Example 1 *(One parameter exponential family) Consider $X_1, ..., X_n \overset{iid}{\sim} p_\theta(x) \propto h(x)\exp(\theta T(x))$. We are interested in the testing*

$$H_0 : \theta = \theta_0 \ vs\ H_1 : \theta = \theta_1$$

The likelihood ratio is given by

$$\frac{\prod_{i=1}^n p_{\theta_1}(x_i)}{\prod_{i=1}^n p_{\theta_0}(x_i)} \propto exp((\theta_1 - \theta_0)\sum_{i=1}^n T(x_i))$$

Assuming that $\theta_1 > \theta_0$, we shall reject H_0 for large $\sum_{i=1}^n T(x_i)$. In other words, an MP test is of the form:

$$\phi(x) = \begin{cases} 1 & if\ \ \sum_{i=1}^n T(x_i) > k, \\ \gamma & if\ \ \sum_{i=1}^n T(x_i) = k, \\ 0 & if\ \ \sum_{i=1}^n T(x_i) < k, \end{cases}$$

Of course, the quantities k and γ are chosen to satisfy

$$\alpha = E_{\theta_0}\phi(x) = P_{\theta_0}(\sum_{i=1}^n T(x_i) > k) + \gamma P_{\theta_0}(\sum_{i=1}^n T(x_i) = k)$$

Note also that $\sum_{i=1}^n T(x_i)$ has no θ dependence and that k and γ do not depend on θ_1 (assuming that $\theta_1 > \theta_0$ only). This means that ϕ is in fact uniformly MP for testing:

$$H_0 : \theta = \theta_0 \ vs\ H_1 : \theta > \theta_0$$

Monotone Likelihood Ratios (MLR) and UMP one-sided Tests.

Definition 3 *We say that the family of densities $\{p_\theta : \theta \in R\}$ has monitone likelihood ratio in T(x) if*

1. *$\theta \neq \theta'$ implies $p_\theta \neq p_{\theta'}$ (identifiability)*
2. *$\theta < \theta'$ implies $p_{\theta'}(x) / p_\theta(x)$ is a non-decreasing function of T(x) (Monotonicity)*

Example 2 *(Double exponential) Let X ~ Double Exponential (θ) with density $p_\theta(x) = \frac{1}{2}e^{-|x-\theta|}$. To check the second cond, fix any $\theta' > \theta$ and consider*

$$\frac{p_{\theta'}(x)}{p_\theta(x)} = e^{|x-\theta|-|x-\theta'|}$$

observe that

$$|x-\theta| - |x-\theta'| = \begin{cases} \theta - \theta' & if \;\; x < \theta \\ 2x - \theta - \theta' & if \;\; \theta \le x \le \theta' \\ \theta' - \theta & if \;\; x > \theta', \end{cases}$$

Which is non-decreasing in x. Therefore the family has MLR in T(x)=x.

Example 3 *(Cauchy location model). Let X have density $p_\theta = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$. We consider the likelihood ratio if $\frac{p_\theta(x)}{p_{\theta'}(x)} = \frac{1+x^2}{1+(x-\theta)^2} \to 1$ as $x \to \infty$ or $x \to -\infty$ for $\theta > 0$. But $p_\theta(0)/p_{\theta'}(0) = \frac{1}{1+\theta^2} < 1$. This ratio must increase at some value of x and decrease at other locations. Hence this family does not satisfy the MLR property.*

Theorem 4 *(TSH 3.4.1) Suppose $X \sim p_\theta(x)$ has MLR is T(x) and we test $H_0 : \theta \le \theta_0$ vs $H_1 : \theta > \theta_0$. Then,*

1. *there exists a UMP test at level α of the form*

$$\phi(x) = \begin{cases} 1 & if \;\; T(x) > k \\ \gamma & if \;\; T(x) = k \\ 0 & if \;\; T(x) < k, \end{cases}$$

   where k and γ are determined by $E_{\theta_0}\phi(x) = \alpha_j$.

2. *the power function $\beta(\theta) = E_\theta\phi(X)$ is strictly increasing when $0 < \beta(\theta) < 1$.*

Outline 5 *To show ϕ is UMP at level α for testing $H_0 : \theta \le \theta_0$ vs $H_1 : \theta > \theta_0$, we have to show that the constraint for $\theta < \theta_0$.*

Optinal Test for composite Nulls. Consider the case with a simple alternative:

$$H_0 : X \sim f_0, \theta \in \Omega_0$$

$$H_1 : X \sim g(unknown), [simple]$$

We impose a prior distribution Λ on Ω_0. So we consider the new hypothesis:

$$H_\Lambda : X \sim h_\Lambda(x) = \int_{\Omega_0} f_0(x) d\Lambda(\theta)$$

, where $h_\Lambda(x)$ is the marginal distribution of X induced by Λ. We shall test H_Λ vs H_1.
Let β_Λ be the power of the MP level-α test Φ_Λ for testing H_Λ vs. $H_1(g)$.

Definition 6 *The prior Λ is a least favourable distribution if $\beta_\Lambda \le \beta_{\Lambda'}$ for any prior Λ'.*

Theorem 7 *(TSH 3.8.1) Suppose Φ_Λ is a MP level-α test for testing H_Λ against g. If ϕ_Λ is level-α for the original hypothesis H_0 (i.e $E_{\theta_0}\Phi_\Lambda(x) \le \alpha \forall \theta \in \Omega_0$), then*

1. *The test Φ_Λ is MP for the original: $H_0 : \theta \in \Omega_0$ vs $H_1 : g$*
2. *The distribution Λ is least favourable.*

Proof 3 *:*

1. *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*
$$E_\theta(\Phi^*(X)) = \int \Phi^*(x) f_\theta(x) d\mu(x) \leq \alpha \quad \forall \theta \in \Omega_0$$
which implies that
$$\int \Phi^*(x) h_\Lambda(x) d\mu(x) = \int \int \Phi^*(x) f_\theta(x) d\mu(x) d\Lambda(\theta) \leq \int \alpha d\Lambda(\theta) = \alpha$$
Since Φ_Λ is MP for H_Λ vs g, we have
$$\int \Phi * (x) g(x) d\mu(x) \leq \int \Phi_\Lambda(x) g(x) d\mu(x),$$
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 $E_\theta(\Phi_\Lambda(x) < \alpha \quad \forall \theta \in \Omega_0$, we know that Φ_Λ must be level α for $H_{\Lambda'}$ vs g. Thus $\beta_\Lambda \leq \beta_{\Lambda'}$, so Λ is the least favourable distrubution.*

Example 4 *Let $X_1, ..., X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ with both σ^2 and θ unknown. We consider teting $H_0 : \sigma \leq \sigma_0$ against $H_1 : \sigma > \sigma_0$. Our goal is to find an UMP test.*

1. *Fix a simple alternative(θ_1, σ_1) for some arbitrary θ_1 and $\sigma_1 > \sigma_0$*
2. *Choose a prior Λ to collapse one null hypothesis. The least favourable prior should make the alt. Hypothesis hand to distinguish for the null. Hence a rule of thumb: to concentrate Λ on the boundary between H_1 and H_0 (i.e. the $\{\sigma = \sigma_0\}$). In this case, we assign Λ to be a prob distribution on $\theta \in R$ and a fixed $\sigma = \sigma_0$.*
Given away test function $\Phi(x)$ and a sufficient statistic T, there exists a test function η that has the same power as Φ but depends only on X through T:
$$\eta(T(x)) = E(\Phi(x)|T(x)).$$
We restrict our attention to (Y, u) where $Y = \bar{X}$ and $u = \sum_{i=1}^n (x_i - \bar{x})^2$. We know that $Y \sim N(0, \frac{\sigma^2}{n})$, $u \sim \sigma^2 \chi^2_{n-1}$ and Y $\perp\!\!\!\perp$ u by Basu Theorem.
Thus for Λ supported on $\sigma = \sigma_0$, we obtain the joint density of (Y,u) under H_Λ as
$$C_0 u^{\frac{n-3}{2}} e^{-\frac{u}{2\sigma_0^2}} \int e^{-\frac{n}{2\sigma_0^2}(y-\theta)^2} d\Lambda(\theta)$$
and the joint density under alternative (θ_1, σ_1) as
$$C_1 u^{\frac{n-3}{2}} e^{-\frac{u}{2\sigma_1^2}} \int e^{-\frac{n}{2\sigma_1^2}(y-\theta)^2} d\Lambda(\theta_1)$$
We can see that the choice of Λ only affects the distribution of Y. To achieve the minimal max, power against the alternative, we need to choose Λ that the two distribution because as close as possible.

Under the alternative hypothesis, $Y \sim N(\theta_1, \frac{\sigma_1^2}{n})$ under H_0, the distribution of Y is a convolution from, i.e. $Y = Z + \Theta$, where $Z \sim N(0, \frac{\sigma_1^2}{n})$, $\Theta \sim \Lambda$, with $Z \perp\!\!\!\perp \Theta$, Hence if we choose $\Theta \sim N(\theta_1, \frac{\sigma_1^2 - \sigma_0^2}{n})$, Y will have the same dist, under the null and the alternative which is $N(\theta_1, \frac{\sigma_1^2}{n})$.
Under this choice of prior, the LRT rejects for large value of $\exp\{-\frac{u}{2\sigma_1^2} + \frac{u}{2\sigma_0^2}\}$ (hence large value of u). So the MP test rejects H_Λ if $\sum_{i=1}^n (x_i - \bar{x})^2 > \sigma_0^2 C_{n-1,1-\alpha}$

3. *We need to check if the MP test is for the composite null. For any (θ, σ) with $\sigma < \sigma_0$, the prob of rejection is*

$$P_{\theta,\sigma}\left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2}\right) = P\left(\chi^2_{n-1} > \frac{\sigma_0^2}{\sigma^2} C_{n-1,1-\alpha}\right) \leq \alpha$$

   with equality holds $\sigma = \sigma_0$. Hence our test is MP for the testing original null H_0 vs $N(\theta_1, \sigma_1)$.

4. *The test does not depend on (θ_1, σ_1). Hence the test above is NMP for testing the original null vs the composite alternative.*