1 Some Basic Stuff

1.1 Moment generating function and characteristic function

Given X a random k vector, its characteristic function is given by

$$\phi_X(t) = E(e^{it\cdot X}).$$

and its moment generating function is given by

$$\psi_X(t) = E(e^{t\cdot X}).$$

These two are used to find distributions.

Theorem 1.1 *(Uniqueness). Let X and Y be random k-vectors.*

1. *If $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}^k$, then $P_X = P_Y$.*
2. *If $\psi_X(t) = \psi_Y(t) < \infty$ for all t in a neighborhood of 0, then $P_X = P_Y$.*

1.2 Obtaining distribution

To obtain the joint distribution/probability density of some dependent random variable, if there is diffeomorphism that maps the depedent random variable to independent random variable, then we may make use of the Change of variable formula in the integration to obtain the joint distribution/probability density needed.

Theorem 1.2 *If $Y = (Y_1, ..., Y_n)$ dependent, but there exists F such that $F(X) = Y$, $X = (X_1, ..., X_n)$ independent, then the joint distribution of Y, will be*

$$P((Y_1, ..., Y_n) \in (B_1, ..., B_n)) = \int_{(B_1,...B_2)} f_X(F^{-1}(Y))|\frac{\partial F^{-1}(Y)}{\partial y}|dy.$$

Proof

$$\begin{aligned} P((Y_1, ..., Y_n) \in (B_1, ..., B_n)) &= P(X \in F^{-1}(B)) \\ &= \int_{F^{-1}(B)} f_X(x)dx \\ &= \int_{(B_1,...,B_n)} f_X(F^{-1}(y))|\frac{\partial F^{-1}(Y)}{\partial y}|dy. \end{aligned}$$

∎

Example If X follows standard exponential distribution, please show that

$$(X_{1,n}, ..., X_{n,n}) \stackrel{d}{=} (\frac{Z_1}{n}, \frac{Z_1}{n} + \frac{Z_2}{n-1}, ..., \sum_{i=1}^{n} \frac{Z_i}{n-i+1}).$$

where $X_{i,n}$ is the i-th order statics, and Z_i are i.i.d. followed with the standard exponential distribution.

Proof Pdf of $(Z_1, ..., Z_n) = e^{-x_1}1_{x_1\geq 0}...e^{-x_n}1_{x_n\geq 0}$.
Putting $U_j = \sum_{i=1}^{j} \frac{Z_i}{n-i+1}$, then $Z_1 = nU_1$, $Z_2 = (n-1)(U_2 - U_1)$,...,$Z_n = U_n - U_{n-1}$.
Hence $|\frac{\partial Z}{\partial U}| = n!$. Hence the pdf of $(U_1, ..., U_n)$ is

$$\begin{aligned} f_Z(nu_1, (n-1)(u_2-u_1), ..., u_n - u_{n-1})n! &= e^{-(u_1+...+u_n)}1_{0\leq u_1\leq ...\leq u_n}n! \\ &= \text{pdf of } (X_{1,n}, ..., X_{n,n}). \end{aligned}$$

∎

1.3 Gamma Distribution

We define the gamma distribution, $\Gamma(k, \beta)$ having the density

$$f_{k,\beta}(x) = \begin{cases} \frac{x^{k-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

For $k > 0$, $\beta > 0$, where k is called the shape parameter and β is called the scale parameter. We have the following properties for Gamma distribution.

Theorem 1.3 *The moment generating function for* $\Gamma(k, \beta)$ *is*

$$\frac{1}{(1 - t\beta)^k} \text{ for } u < \frac{1}{\beta}.$$

Theorem 1.4 *The following is true for Gamma function:*

1. $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$, $\Gamma(n + 1) = n!$.
2. $\Gamma(k, \beta) + \Gamma(k', \beta) = \Gamma(k + k', \beta)$.
3. $n\Gamma(k, \beta) = \Gamma(k, n\beta)$.

Proof For 1., use integration by part. For 2. and 3., investigate the moment generating function. ∎

Theorem 1.5 *(Relations with other distribution). The exponential distribution* $\exp(\lambda)$*, having density* $\frac{1}{\lambda}e^{-x/\lambda}1_{x\geq 0}$*,* $\exp(\lambda) = \Gamma(1, \lambda)$*, and the chi squared distribution of degree* k*,* $\chi^2_k = \sum_i Z_i^2$*,* $Z_i \sim N(0, 1)$*, is* $\Gamma(k/2, 2)$.

1.4 Order Statistics

Let $X_1, ..., X_n \sim F$, having density f. Then we define the order statistics

$$\begin{aligned} X_{(1)} &= \min X_i \\ X_{(2)} &= \text{ second smallest of } X_i \\ &... \\ X_{(n)} &= \max X_i. \end{aligned}$$

And we define the median

Definition M is called the median, where

$$M = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ odd.} \\ (X_{(n/2)} + X_{(n/2+1)} & \text{if } n \text{ even.} \end{cases}$$

Then

Theorem 1.6 *(Joint distribution of* $(X_{(1)}, ..., X_{(n)})$*).* $(X_{(1)}, ..., X_{(n)})$ *having density*

$$n!f(x_1)...f(x_n)1_{\{x_1<x_2<...<x_n\}}.$$

Proof For $x_1 < x_2 < ... < x_n$, we have $n!$ of ways to arranges this n items, and each of them is of $f(x_1)...f(x_n)$ since they are i.i.d., hence we have $n!f(x_1)...f(x_n)$ when $x_1 < x_2 < ... < x_n$. When $x_1 < x_2 < ... < x_n$ is not satisfied, we take $n = 2$ as an example, then if $x_1 > x_2$

$$\begin{aligned} &P(\min X_i < x_1, \max X_i < x_2) \\ =&P(\min X_i < \max X_i, \max X_i < x_2) \\ =&P(\max x_i < x_2) \end{aligned}$$

Hence differentiating with respect to x_1 gives zero. ∎

Theorem 1.7 *Let $X_{(1)}, ...X_{(n)}$ denote the order statistics of a random sample. $X_1, ..., X_n$ from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the pdf of $X_{(j)}$ is*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x)[F_X(x)]^{j-1}[1 - F_X(x)]^{n-j}$$

Proof To be filled in later. ∎

Theorem 1.8 *Let $X_{(1)}, ..., X_{(n)}$ denote the order statistics of a random sample, $X_1, ..., X_n$ from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the joint pdf of $X_{(i)}$ and $X_{(j)}$, $1 \le i < j \le n$, is*

$$f_{X_{(i)},X_{(j)}}(u, v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v)[F_X(u)]^{i-1}[F_X(v) - F_X(u)]^{j-1-i}[1 - F_X(v)]^{n-j}$$

1.5 Some useful inequalities

Theorem 1.9 *Let Z be a real random variable and g a nonnegative, nondecreasing function on the support of Z, i.e. a set B such that $P(Z \in B) = 1$, then*

$$P(Z \ge a) \le \frac{Eg(z)}{g(a)}.$$

Proof Observe that $g(a)I(Z \ge a) \le g(z)I(Z \ge a) \le g(z)$. Take expectation on both sides. ∎

Theorem 1.10 *(Cauchy Schwartz) Let $X = (X_1, ..., X_p)^T$ be a p-vector of real random variables and $U = E(XX^T)$. The matrix U is symmetric, nonnegative definite. It is with singularity ($\det(U) = 0$) if and only if there exists a p-vector $\alpha \neq 0$ such that $E(\alpha^T X)^2 = 0$.*

Proof Symmetric is trivial. To show that it is nonnegative definite, let β be a p-colomn vector, then $\beta^T U \beta = \beta^T E(XX^T \beta) = E(\beta X)^2 \ge 0$.
Note that a symmetric nonnegative definite U is singular if and only if there exists α such that $\alpha^T U \alpha = 0$. Thus conclude the proof. ∎

Corollary 1.10.1 *For random variables X_1 and X_2 (note that it is an entry not a vector),*

$$(E(X_1 X_2))^2 \le E(X_1^2)E(X_2^2)$$

and centering X_1 and X_2 at mean yields

$$(Cov(X_1, X_2))^2 \le Var(X_1)Var(X_2).$$

Proof Take $X = (X_1, X_2)$. By the U defined above, U is symmetric nonnegative definite. Hence its eigenvalue is nonnegative, since determinant is the product of its eigenvalues, so its determinant is nonnegative, and $\det(U) = E(X_1^2)E(X_2^2) - (E(X_1 X_2))^2$. ∎

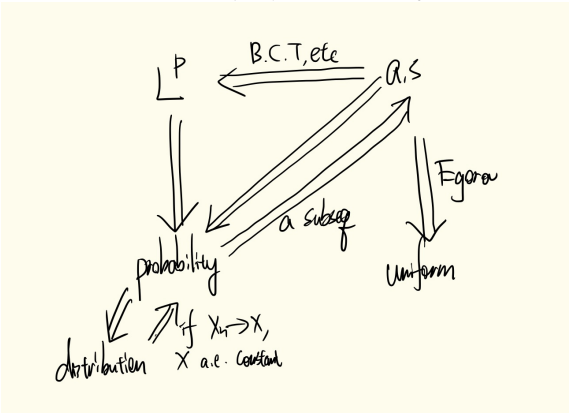
Theorem 1.11 *(Jenson) If X and $g(x)$ are integrable random variable's and $g(\cdot)$ is convex, then $g(E(X)) \le E(g(X))$.*

2 Convergence Results, O_p notation and delta method

Definition (Convergence in the q-th mean). $X_n \to X$ in q-th mean if $\lim_n E|X_n - X|^q = 0$.

Definition (Convergence in distribution). $X_n \to X$ in distribution if $\lim_n F_n(x) = F(x)$ at each point of continuity of F. We write $X_n \xrightarrow{d} X$.

Theorem 2.1 *(Interplay between different modes of convergence).*



Theorem 2.2 *(Skorohod's Theorem). If $X_n \xrightarrow{d} X$, then there are random vectors $Y, Y_1, Y_2, ...$ defined on a common probability space such that $P_Y = P_X$, $P_{Y_n} = P_{X_n}$, $n = 1, 2, ...$ and $Y_n \to Y$ a.s.*

Proof See Billingsley P.399-P.402. ∎

Theorem 2.3 *(Equivalence of convergence in distribution). Let $X, X_1, X_2...$ be random k-vectors. $X_n \xrightarrow{d} X$ if and only if*

1. *$E(h(X_n)) \to E(h(X))$ for every bounded continuous function h.*
2. *$\limsup_n P_{X_n}(C) \leq P_X(C)$ for any closed set $C \subseteq \mathbb{R}^k$.*
3. *$\liminf_n P_{X_n}(O) \geq P_X(O)$ for any open set $O \subset \mathbb{R}^k$*

Proof See Rick Durrect or Jun Shao. ∎

Theorem 2.4 *(Slusky). If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} y$, y a real constant. Then*

1. *$X_n + Y_n \xrightarrow{d} X + y$.*
2. *$X_n Y_n \xrightarrow{d} Xy$.*

Proof 1. First of all we have

$$P(X_n + Y_n \leq x) = P(X_n \leq x - Y_n) = P(X_n \leq x - Y_n, |Y_n - y| < \epsilon) + P(X_n \leq x - Y_n, |Y_n - y$$
$$\leq P(X_n \leq x - y + \epsilon) + P(|Y_n - y| \geq \epsilon)$$

Hence $P(X_n + Y_n \leq x) - P(X_n + y \leq x + \epsilon) \leq P(|Y_n - y| \geq \epsilon)$. Taking limit on both sides gives $\lim P(X_n + Y_n \leq x) - P(X + y \leq x + \epsilon) \leq 0$.
On the other hand,

$$P(X_n + Y_n \leq x) = P(X_n \leq x - Y_n) = P(X_n \leq x - Y_n, |Y_n - y| < \epsilon) + P(X_n \leq x - Y_n, |Y_n - y$$
$$\geq \epsilon) \geq P(X_n \leq x - y - \epsilon)$$

Hence $P(X_n + Y_n \leq x) - P(X_n \leq x - y - \epsilon) \geq 0$, taking limits on both sides gives $\lim P(X_n + Y_n \leq x) - P(X \leq x - y - \epsilon) \geq 0$. Combining the above and let $\epsilon \to 0$, since we are concerning only continuous points of F_{X+y}, we are done.

2. By continuous mapping theorem and the above suffices to consider $y > 0$, choose ϵ small such that $y - \epsilon > 0$ We have

$$\begin{aligned} P(X_n(y+\epsilon) \leq x) &\leq P(X_nY_n \leq x, |Y_n - y| \leq \epsilon) + P(X_nY_n \leq x, |X_n - y| > \epsilon) = P(X_nY_n \leq x) \\ &\leq P(X_n(y-\epsilon) \leq x) + P(|Y_n - y| \geq \epsilon) \end{aligned}$$

By argument as above we are done.

∎

Example In general we do not have $X_n \to X, Y_n \to Y$, then $X_n + Y_n \to X + Y$, simply consider $X_n \sim N(0,1)$, $Y_n := -X_n$, then $X_n + Y_n = 0$ will not converge to $N(0,2)$ in distribution.

Theorem 2.5 *(Results about convergence in distribution implies converges in probability). For a sequence of random variables $\{X_n\}_{n\geq 1}$, $X_n \xrightarrow{d} c$ then $X_n \xrightarrow{p} c$, where c is a constant.*

Proof Consider

$$\begin{aligned} P(|X_n - c| \geq \epsilon) &= P(X_n \geq \epsilon + c) + P(X_n \leq c - \epsilon) \\ &= 1 - P(X_n < \epsilon + c) + P(X_n \leq c - \epsilon) \end{aligned}$$

Take $n \to \infty$, since $X_n \xrightarrow{d} c$, $P(|X_n - c| \geq \epsilon) \to 0$. ∎

Theorem 2.6 *(Continuous mapping theorem). If $X_n \xrightarrow{d} X$ and g a continuous mapping, $g(X_n) \xrightarrow{d} g(X)$.*

Remark Note that from continuous mapping theorem we get that if $(X_n, Y_n) \xrightarrow{d} (X, Y)$, then $X_n + Y_n \xrightarrow{d} X + Y$, $X_nY_n \xrightarrow{d} XY$, this implies Slusky if X_n, Y_n are independent, however the power of Slusky is that we don't need indedependency.

Definition (O_p, o_p) $X_n = O_p(Y_n)$ if for any $\epsilon > 0$, there is a constant $C_\epsilon > 0$ such that $\sup_n P(\|X_n\| \geq C_\epsilon |Y_n|) < \epsilon$. $X_n = o_p(Y_n)$ if $\frac{X_n}{Y_n} \xrightarrow{p} 0$.

Theorem 2.7 *(δ method). Suppose $a_n(X_n - b) \xrightarrow{d} X$, where a_n is a sequence of constants tending to ∞ and b is a fixed number. Let $g \in C^1(\mathbb{R})$, then*

$$a_n(g(X_n) - g(b)) \xrightarrow{d} g'(b)X.$$

Proof

$$a_n(g(X_n) - g(b)) = a_n \frac{g(X_n) - g(b)}{X_n - b}(X_n - b) \to g'(b)X$$

By Slusky's theorem. ∎

Theorem 2.8 *(Second order δ method). Suppose $a_n(X_n - b) \xrightarrow{d} X$, where a_n is a sequence of constants tending to ∞ and b is a fixed number.. For a given function $g \in C^2$, suppose that $g'(b) = 0$ and $g''(b) \neq 0$. Then*

$$a_n^2(g(X_n) - g(b)) \xrightarrow{d} g''(b)X^2$$

Proof

$$a_n^2(g(X_n) - g(b)) = a_n^2(X_n - b)^2 \frac{g(X_n) - g(b)}{(X_n - b)^2} \to g''(b)X^2$$

By continuous mapping theorem and Slusky's theorem. ∎

Example Let $X_i, i = 1, 2, ...$, be independent Bernoulli(p) random vaiables and let $Y_n = \frac{1}{n}\sum_{i=1}^n X_i$.

1. Show that $\sqrt{n}(Y_n - p) \xrightarrow{d} N(0, p(1-p))$.
2. Show that for $p \neq 1/2$, the estimate of variance $Y_n(1 - Y_n)$ satisfies $\sqrt{n}[Y_n(1 - Y_n) - p(1-p)] \xrightarrow{d} N(0, (1-2p)^2 p(1-p))$.
3. Show that for $p = 1/2$, $n[Y_n(1 - Y_n) - \frac{1}{4}] \xrightarrow{d} -\frac{1}{4}\chi_1^2$.

Proof
1. By CLT, $\frac{Y_n - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} N(0,1)$ hence $\sqrt{n}(Y_n - p) \xrightarrow{d} N(0, p(1-p))$ by Slusky and constant times a normal is still a normal.
2. We use the delta method, setting $g(y) := y(1-y)$,

$$\sqrt{n}(g(Y_n) - g(p)) \xrightarrow{d} g'(p)N(0, p(1-p)) = N(0, (1-2p)^2 p(1-p)).$$

3. Since $g(1/2) = 0$ we use the second order delta method to get our desired conclusion.

∎

3 Asymptotic behavior of sample means

3.1 Law of Large Number

First of all we have weak law of large number:

Theorem 3.1 *(Weak Law of Large Number). Suppose $X_1, X_2, ...$ are i.i.d., with mean μ and variance σ^2, and let $\bar{X}_n = (X_1 + ...X_n)/n$. Then*

$$\bar{X}_n \xrightarrow{p} \mu \text{ as } n \to \infty$$

Proof $E(\bar{X}_n - \mu)^2 = \text{Var}(\bar{X}_n) = \sigma^2/n \to 0$, and L^2 convergence implies convergence in probability. ∎

Theorem 3.2 *(Strong Law of Large Number). If $X_1, X_2, ...$ are i.i.d. with finte mean $\mu = EX_i$, and if $\bar{X}_n = (X_1 + ... + x_n)/n$, then $\bar{X}_n \to \mu$ almost surely as $n \to \infty$.*

Proof See Billingsley. ∎

Remark Both the weak and strong law of large numbers hold with only $E|X_i| < \infty$.

A special case of SLLN is taking $Z_i = I_{[X_i,\infty)}(x)$ for r.v.s $X_i \sim F$ and fixed x. Observe that

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

We can infer that

$$F_n(x) := n^{-1} \sum_{i=1}^{n} I_{[X_i,\infty)}(x) \to F(x) \text{ as } n \to \infty \text{ a.s.}$$

Actually we have a much stronger version of this. Which is due to the nondecreasing property of the probability distribution so we may choose finite points to capture the whole F.

Theorem 3.3 *(Gilvenko Cantelli). With settings as above,*

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

Proof Let $\epsilon > 0$ and find an integer $k \geq \frac{1}{\epsilon}$ and numbers $-\infty = x_0 < x_1 \leq ... \leq x_{k-1} \leq x_k = \infty$ such that

$$F(x_j^-) \leq \frac{j}{k} \leq F(x_j), j = 1, ..., k-1.$$

where $F(x^-) = P(X < x)$. Note that if $x_{j-1} < x_j$ then $F(x_j^-) - F(x_{j-1}) \leq \epsilon$. By SLLN, $F_n(x_j) \to F(x_j)$ a.s. and $F_n(x_j^-) \to F(x_j^-)$ for $j = 1, ..., k-1$. Now let x be arbitrary and find j such that $x_{j-1} < x \leq x_j$, then

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

Then $\sup_x |F_n(x) - F(x)| \to \epsilon$ as $n \to \infty$ since there is only finite number of j. ∎

3.2 Central Limit Theorem

Theorem 3.4 *Suppose $X_1, ...,$ are i.i.d. with $E(X_1) = \mu$, Var$(X_1) = \sigma^2$, then*

$$\frac{\bar{X}_n \to \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0,1).$$

Proof Later. ∎

Why is it normal? One can do the following formal deviation. Suppose $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} X$, what would X look like? Consider

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

Clearly $Z_{2n} \xrightarrow{d} X$, but $Z_{2n} = \frac{X_1+...+X_n}{\sqrt{2n}} + \frac{X_{n+1}+...+X_{2n}}{\sqrt{2n}} := Z_{a_n} + Z_{b_n}$, where $Z_{a_n}, Z_{b_n} \xrightarrow{d} \frac{X}{\sqrt{2}}$, which means Z_{a_n}, Z_{b_n} satisfy the stable law, therefore X will be something very similar to the Gaussian.

Theorem 3.5 *(Lyapunov) Suppose $X_1, X_2, ...$ is a sequence of independent random variables, each with finite expected value μ_i and variance σ_i^2. Define*

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2$$

If for some $\delta > 0$, Lyapunov's condition

$$\lim_n \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} E[|X_i - \mu_i|^{2+\delta}] = 0$$

is satisfied, then

$$\frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} N(0, 1).$$

For the dependent cases, we have the α-mixing.

Definition (α mixing). Given a sequence of $X_1, X_2, ...$ and sets $A \in \sigma(X_1, ..., X_k)$, $B \in \sigma(X_{k+n+1}, X_{k+n+1})$ for $k \geq 1$ and $n \geq 1$, then if there exists a sequence of real numbers $\alpha_n \to 0$ such that

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

then $\{X_n\}$ is α-mixing.

Then we have the CLT for α mixing sequences.

Theorem 3.6 *(CLT for α mixing sequences). Suppose $X_1, X_2, ...$ is stationary and α-mixing with $\alpha_n = O(n^{-5})$, $E(X_n) = 0$ and $E(X_n^2) < \infty$. Set $S_n = X_1 + ... + X_n$. If $n^{-1}Var(S_n) \to \sigma^2 = E(X_1^2) + \sum_{k=1}^{\infty} E(X_1 X_{k+1})$ convege absolutely with $\sigma^2 > 0$, then $\frac{S_n}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1)$.*

Theorem 3.7 *With the assumptions in CLT, If f is differentiable at μ, then*

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

Proof Just a simple application of the Delta method. ∎

4 Asymptotic of Medians and Percentiles

Let $X_1, ..., X_n$ be random variables. We arrange these random variables in increasing order, $X_{(1)} \leq ... \leq X_{(n)}$ are called order statistics. The first order is the smallest value while the largest order is the largest value. The median is the middle order statistic when n is odd, or the average of the two middle order statistics when n is even. We notate the median by $\tilde{X}$. One advantage of median compared to the mean is that it will not affect by the extreme values.

Theorem 4.1 *Let $X_1, X_2, ...$ i.i.d. with common cumulative distribution function F, let $\gamma \in (0, 1)$, and let $\gamma \in (0, 1)$ and let $\tilde{\theta}_n$ be the $\lfloor \gamma n \rfloor$th order statistic for $X_1, ..., X_n$ (or a weighted average of the $\lfloor \gamma n \rfloor$ and $\lceil \gamma n \rceil$th order statistics). If $F(\theta) = \gamma$, and if $F'(\theta)$ exists and is finite and positive, then*

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N(0, \frac{\gamma(1-\gamma)}{[F'(\theta)]^2}),$$

as $n \to \infty$.

Proof We consider the case $\gamma = 1/2$. Assume now that $X_1, X_2, ...$ are i.i.d. with common cumulative distribution function F, and let $\hat{X}_n$ be the median of the first n observations. Assume that F has a unique median θ, so $F(\theta) = 1/2$, and that $F'(\theta)$ exists and is finite and positive. We approximate

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

Define

$$S_n = \#\{i \leq n : X_i \leq \theta + \frac{a}{\sqrt{n}}\}.$$

Note taht $\hat{X}_n \leq \theta + a/\sqrt{n}$ if and only if $S_n \geq \lceil \frac{n}{2} \rceil$, and

$$S_n \sim \text{Binomial}(n, F(\theta + a/\sqrt{n})).$$

Hence from the CLT we have

$$\begin{aligned}
P(S_n \geq \lceil n/2 \rceil) &= P(\frac{\frac{S_n}{n} - F(\theta + \frac{a}{\sqrt{n}})}{\sqrt{\frac{F(\theta+\frac{a}{\sqrt{n}})(1-F(\theta+\frac{a}{\sqrt{n}}))}{n}}} \geq \frac{\frac{\lceil \frac{n}{2} \rceil}{n} - F(\theta + \frac{a}{\sqrt{n}})}{\sqrt{\frac{F(\theta+\frac{a}{\sqrt{n}})(1-F(\theta+\frac{a}{\sqrt{n}}))}{n}}}) \\
&\to 1 - \Phi(\frac{\frac{\lceil \frac{n}{2} \rceil}{n} - F(\theta + \frac{a}{\sqrt{n}})}{\sqrt{\frac{F(\theta+\frac{a}{\sqrt{n}})(1-F(\theta+\frac{a}{\sqrt{n}}))}{n}}}) = \Phi(\frac{\frac{F(\theta+\frac{a}{\sqrt{n}}) - \lceil \frac{n}{2} \rceil}{n}}{\sqrt{\frac{F(\theta+\frac{a}{\sqrt{n}})(1-F(\theta+\frac{a}{\sqrt{n}}))}{n}}}) \\
&= \Phi(\frac{a\frac{F(\theta+\frac{a}{\sqrt{n}}) - \lceil \frac{n}{2} \rceil}{\frac{a}{\sqrt{n}}}}{\sqrt{F(\theta + \frac{a}{\sqrt{n}})(1 - F(\theta + \frac{a}{\sqrt{n}}))}}) \\
&\to \Phi(2aF'(\theta))
\end{aligned}$$

Hence

$$\sqrt{n}(\hat{X}_n - \theta) \xrightarrow{d} N(0, \frac{1}{4[F'(\theta)]^2}).$$

∎

5 Sufficiency

Definition (Statistic). A statistic T is a function of the data, $T(X) = T(X_1, ..., X_n)$.

The information within the statistic $T(X)$ concerning the unknown distribution of X is contained in $\sigma(T(X))$, and $\sigma(T(X)) \subseteq \sigma(X)$, therefore $T(X)$ can be seen as a reduction of the data.

Definition (Sufficient Statistic). A statistic is sufficient for a mode $P = \{P_\theta : \theta \in \Theta\}$ if for all t, the conditional distribution $X|T(x) = t$ doesn't depend on θ, in other words, $T(x) = t$ contains all information about θ.

Example (Weighted Coin Flips). Let $X_1, ..., X_n$ be i.i.d. according to Bernoulli random variables. Is the number of heads, i.e., $\sum_{i=1}^n X_i$, is it sufficient? Denote $X = (X_1, ..., X_n)$,

$x = (x_1, ..., x_n)$.
We have

$$P_\theta(X = x) = \Pi_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

So the conditional distribution is

$$\begin{aligned} P(X = x|T(x) = t) &= \frac{P_\theta(X = x, T(x) = t)}{P_\theta(T(x) = t)} \\ &= \frac{I(\sum_{i=1}^n X_i = t)\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{I(\sum_{i=1}^n X_i = t)}{\binom{n}{t}}. \end{aligned}$$

which is independent of θ, so the sum of heads is a sufficient statistic. The example is saying that given that we know the total number of heads, then a particular sequence appear will be independent of the probability of obtaining one head.

Example (Maximum of uniform). Let $X_1, ..., X_n$ be i.i.d. uniform $(0, \theta)$. Then $T(X) = \max(X_1, ..., X_n)$ is sufficient. The intuition is that once $T(X) = t$ is known, then the remaining part behaves like $n-1$ i.i.d. of uniform $(0, t)$. To make it rigorous,later.

Example (Order Statistic). Let $X_1, ..., X_n$ be i.i.d. with any model. Then the order statistics $T = \{X(1) \leq ... \leq X(n)\}$ are sufficient.

Theorem 5.1 *(TPE 1.6, Theorem 6.1). If $X \sim P_\theta \in \mathcal{P}$ and T is sufficient for $\mathcal{P}$, then for any decision procesure δ, there is a decision procedure of equal risk that depend on X only through $T(X)$.*

Theorem 5.2 *(Neyman Fisher Factorization Criterion). Suppose each $P_\theta \in \mathcal{P}$ has density $p(x, \theta)$ with respect to a common σ finite measure μ, i.e., $\frac{dP_\theta}{d\mu} = p(x, \theta)$, or $P_\theta(X \in A) = \int 1_A(X(\omega))dP_\theta = \int 1_B(x)dP_\theta \circ X^{-1} = \int 1_B(x)p(x, \theta)d\mu$. Then T is sufficient if and only if $p(x, \theta) = g_\theta(T(x))h(x)$ for some g_θ and h.*

Proof We prove the case for discrete case.
($\Rightarrow$): Say, T is sufficient.

$$\begin{aligned} p(x, \theta) &= P_\theta(X = x) = P_\theta(X = x|T(X) = T(x))P_\theta(T(X) = T(x)) \\ &= h(x)g_\theta(T(x)), \end{aligned}$$

where $h(x) = P_\theta(X = x|T(X) = T(x))$ and $g_\theta(y) = P_\theta(T(X) = y)$.

($\Leftarrow$): Suffice to consider $T(X) = T(x)$ only.

$$\begin{aligned}
P_\theta(X = x|T(X) = T(x)) = &\frac{P_\theta(X = x, T(X) = T(x))}{P_\theta(T(X) = T(x))} \\
=&\frac{P_\theta(X = x)}{P_\theta(T(X) = T(x))} \\
=&\frac{g_\theta(T(x))h(x)}{P_\theta(T(X) = T(x))} \\
=&\frac{g_\theta(T(x))h(x)}{\sum_{x'} P_\theta(X = x')1_{T(x)}(T(x'))} \\
=&\frac{g_\theta(T(x))h(x)}{\sum_{x'} g_\theta(T(x'))h(x')1_{T(x)}(T(x'))} \\
=&\frac{g_\theta(T(x))h(x)}{\sum_{x'} h(x')1_{T(x)}(T(x'))}
\end{aligned}$$

is therefore independent of θ. ∎

6 Exponential Families

6.1 Basics

Definition A dominated family (dominated by a measure, usually Lebsgue) $\{P_\theta : \theta \in \Theta\}$ is said to form a k dimensional exponential family if the corresponding density functions $\{p_\theta(x)\}_{\theta\in\Theta}$ are of the form

$$p_\theta(x) = \exp\{\sum_{i=1}^{k} \eta_i(\theta)T_i(x) - B(\theta)\}h(x).$$

where $h, T_1, ..., T_k : \chi \to \mathbb{R}$, $B, \eta_1, ..., \eta_k : \Theta \to \mathbb{R}$.

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

Example Let $X_1, ..., X_n \sim N(\theta, \sigma^2)$ be i.i.d., $\Theta = \mathbb{R} \times (0, \infty)$.

$$\begin{aligned}
p_\theta(x) &= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp(-\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2\sigma^2}) \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp(-\sum_{i=1}^{n} \frac{X_i^2}{\sigma^2} + \frac{\mu\sum_{i=1}^{n} X_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2})
\end{aligned}$$

where

$$\begin{aligned}
T_1(x) = \sum_{i=1}^{n} X_i^2, \eta_1(\theta, \sigma^2) = -\frac{1}{2\sigma^2} \\
T_2(x) = \sum_{i=1}^{n} X_i, \eta_2(\theta, \sigma^2) = \frac{\mu}{\sigma^2} \\
B(\theta, \sigma^2) = \frac{n\mu^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2) \\
h(x) = 1
\end{aligned}$$

Example $X_1, ..., X_n \sim$ Cauchy, i.e., $p_\theta = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$ is the density of p_θ with respect to Lebsgue measure. In this case, $T(X) = (X(1), ..., X(n))$ is sufficient, where $X(i)$ is the i-th order statistic. Indeed, $T(X)$ is minimal sufficient.

Theorem 6.1 *(Pittman-Koopman-Darmois). Suppose* $(X_1, ..., X_n)$ *are i.i.d. with density* $\{p_\theta : \theta \in \Theta\}$ *with respect to the Lebsgue measure, which are continuous in* x *for* θ *fixed and support on an interval* $I \subseteq \mathbb{R}$. *Suppose there exists a sufficient statistic* $c(X) = (T_1(X), ..., T_k(X))$ *which are continuous.*

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

Remark Note that for $k = n$, it is trivial/useless. If $1 \leq k < n$, this is saying that as long as there is a sufficient statistic which reduces some dimension, and the sufficient statistic and p_θ are nice enough, then it is just an exponential family.

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

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

Which parametrize the density in terms of the natural parameters η instead of θ.

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

$$\Theta = \{\eta; 0 < \int \exp(\sum_{i=1}^k \eta_i T_i(x)h(x)d\mu(x) < \infty)\}.$$

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

The differences between canonical and non-canonical one is that for the non-canonical one, there is other parametrizations. Let η be a function from some space Ω into Θ and define

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

for $\theta \in \Omega, x \in \mathbb{R}^n$, where $B(\theta) = A(\eta(\theta))$, note the $\eta_i(\theta)$.

6.2 Reducing the dimension

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

6.2.1 Case 1

The $T_i(x)'s$ satisfy an affine equality constraint ($AT = c$ for some linear functional T and constant c) $\forall x \in X$.

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

Example Let $X \sim \exp(\eta_1, \eta_2)$, i.e., $p(x; \eta_1, \eta_2) = \exp(-(\eta_1 + \eta_2)x + \log(\eta_1 + \eta_2))1_{x>0}$, there $T_1(X) = T_2(X) = x$, i.e., they are linearly dependent.
Also, P is unidentiable, since $p(x; \eta_1 + a; \eta_2 - a) = p(x; \eta_1, \eta_2)$ for any $a < \eta_2$.

6.2.2 Case 2

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

Example 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 $p(x; \eta) = \exp(\eta_1(x - x^2) + x^2)$.

Definition (Minimal canonical exponential family). A canonical exponential family $P = \{p_\eta; \eta \in H\}$ is minimal if

1. $\sum_{i=1}^k \lambda_i T_i(x) = \lambda_0 \forall x \in X \implies \lambda_i = 0$, for $i = 0, ..., k$.
2. $\sum_{i=1}^k \lambda_i \eta_i = \lambda_0 \forall \eta \in H \implies \lambda_i = 0$, for $i = 0, ..., k$.

Definition Suppose $P = \{p_\eta; \eta \in H\}$ is a k dimensional exponential family. If H contains an open k dimensional rectangle, then P is called full rank, otherwise P is called curved.

Consider 3 types of exponential family with normal distribution $N(\mu, \sigma^2)$, where in this case, $\eta_1 = \frac{1}{2\sigma^2}, \eta_2 = \frac{\mu}{\sigma^2}, T_1(x) = -x^2, T_2(x) = x$.

(Nonminimal). Take $\mu = \sigma^2$, then $\eta_1 = \frac{1}{2\sigma^2}, \eta_2 = 1$, therefore $\frac{1}{2\sigma^2}\eta_2 - \eta_1 = 0$. Then the graph corresponding to η_1, η_2 will be a straight line.

(Minimal and Curved). Take $\mu = \sqrt{\sigma^2}$, so $\eta_1 = \frac{1}{2\sigma^2}, \eta_2 = \frac{1}{\sqrt{\sigma^2}}$, hence $\eta_2^= \eta_1$. Then the graph of η_1, η_2 will contain a curved line.

(Minimal and Full Rank). We have no extra constraint on (μ, σ^2) where the natural parameter is $(0, \infty) \times \mathbb{R}$ which contains an open rectangle.

Remark If there exists a function $F(\eta) = 0$, then by implicit function theorem, we can locally solve this, therefore on every compact subsets, the graph of $\eta_i's$ will be of measure zero.

6.3 Properties of Exponential Family

1. If $X_1, ..., X_n \sim p(x, \theta) = \exp(\sum_{i=1}^k \eta_i(\theta) T_i(x) - B(\theta))h(x)$, then

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

By NFFC, $(\sum_{j=1}^n T_1(x_j), ..., \sum_{j=1}^n T_k(x_j))$ is therefore a sufficient statistic. Hence exponential family data is highly compressible.

2. If f is integrable and $\eta \in \Theta$, then $G(f,\eta) = \int f(x)\exp(\sum_{i=1}^{k}\eta_i T_i(x))h(x)d\mu(x)$ is infinitely differentiable with respect to η and the derivatives can be obtained by differentiating under the integral sign.

3. We have the moments of $T_i's$ by direct calculations, i.e., $E_\eta T_j(X) = \frac{\partial A(\eta)}{\partial \eta_j}$.

7 Minimal Sufficiency

Definition A sufficient statistic T is minimal if for every statistic T', T is a function of T', i.e., $T(X) = f(T'(X))$ for some function f. Under f there is some loss of information, hence T contains less information and hence minimal. Equivalently, T is minimal if for every sufficient statistic T', $T'(X) = T'(Y) \implies T(X) = T(Y)$.

Theorem 7.1 *Let $\{p_\theta(x)\}_{\theta\in\Theta}$ be a family of densities with respect to some measure (usually Lebsgue), and the support $\{x \in \chi; p_\theta(x) > 0\}$ is independent of θ. Suppose that there exists a statistic T such that $\forall x, y \in \chi$, for every $\theta \in \Theta$ and some $c(x,y) \in \mathbb{R}$*

$$p_\theta(x) = c(x,y)p_\theta(y) \text{ if and only if } T(x) = T(y),$$

then T is a minimal sufficient statistic.

Proof (T is sufficient). For all $t \in T(\chi)$(the image of T), consider the preimage $A_t = T^{-1}(t)$. For each A_t, we denote x_t a representative from A_t. Then for all $y \in \chi$, from the assumption of T, we have

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

which implies the sufficiency by NFFC.
(T is minimal). Consider another sufficient statistic T', by NFFC,

$$p_\theta(x) = g_\theta(T'(X))h(x).$$

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

$$\begin{aligned} p_\theta(x) &= g_\theta(T'(x))h(x) \\ &= g_\theta(T'(y))h(y)\frac{h(x)}{h(y)} \\ &= p_\theta(y)c(x,y) \end{aligned}$$

Hence $T(x) = T(y)$ by the assumption of the theorem. So T is minimal statistic. ∎

Remark For any k dimensional exponential family, the statistic $(\sum_{j=1}^{n} T_1(x_j), ..., \sum_{j=1}^{n} T_k(x_j))$ is a minimal sufficient statistic. (See Keener Ex3.12).

Remark For example, We cannot apply the above result to $U(0,\theta)$ since the support of $U(0,\theta)$ depends on θ.

Example (Casella Problem 6.9). Find a minimal sufficient statistic for θ.

1. $f(x|\theta) = \frac{1}{\sqrt{2\pi}}e^{-(x-\theta)^2/2}$, $x, \theta \in \mathbb{R}$.

2. $f(x|\theta) = e^{-(x-\theta)}, x > \theta, \theta \in \mathbb{R}$.
3. $f(x|\theta) = \frac{e^{-(x-\theta)}}{(1+e^{-(x-\theta)})^2}, x \in \mathbb{R}$,$\theta \in \mathbb{R}$.
4. $f(x|\theta) = \frac{1}{\pi[1+(x-\theta)^2]}, x, \theta \in \mathbb{R}$.
5. $f(x|\theta) = \frac{1}{2}e^{-|x-\theta|}, x, \theta \in \mathbb{R}$.

A minimal sufficient statistics for them are

1. $T(X) = \sum_{i=1}^{n} X_i$.
2. $T(X) = \min_i X_i$.
3. $T(X) = (X_{(1)}, ..., X_{(n)})$.
4. $T(X) = (X_{(1)}, ..., X_{(n)})$.
5. $T(X) = (X_{(1)}, ..., X_{(n)})$.

8 Ancillary and Completeness

Consider $X_1, X_2, ... \sim$ CauchyLoc(θ) whose distribution is given by

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

then $(X_{(1)}, ..., X_{(n)})$ is minimal sufficient (See TPE 1.5).

Definition A statistic A is ancillary for $X \sim p_\theta \in P$ if the distribution of $A(X)$ does not depend on θ.

Example Consider the above example, then $A(X) = X_{(n)} - X_{(1)}$ is ancillary even though $(X_{(1)}, ..., X_{(n)})$ is minimal sufficient. To see this, $X_i = Z_i + \theta$ for $Z_i \sim$Cauchy(0) so $X_{(i)} = Z_{(i)} + \theta$ and $A(X) = A(Z)$ which doesn't depend on θ.

Theorem 8.1 *Let $f(\cdot)$ be any pdf. Let μ be any real number, and let σ be any positive real number. Then X is a random variable with pdf $\frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$ if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.*

Proof (Casella 3.5.6). ∎

Corollary 8.1.1 *For this kind of X, $X_{(i)} - X_j$ will be ancillary for μ.*

Proof Trivial. ∎

Definition A statistic A is first order ancillary for $X \sim p_\theta \in P$ if $E_\theta(A(X))$ doesn't depend on θ.

Definition A statistic T is complete for $X \sim p_\theta \in P$ if every nonconstant measurable function g of T is first order ancillary. In other words, if $E_\theta(f(T(X))) = 0 \forall \theta \in \Theta$, then $f(T(X)) = 0$ with probability 1 for all θ.

Remark If T is complete sufficient then T is minimal sufficient. (Bahader's Theorem, TPE)

Remark Complete sufficient statistics gives optimal unbiased estimate.

Example Let $X_1, X_2, ..., X_n \sim$ Bernoulli(θ), $\theta \in (0,1)$. Then $T(X) = \sum_{i=1}^n X_i$ is sufficient. Suppose $E_\theta(f(T(X))) = 0, \forall \theta \in (0,1)$, then

$$\sum_{j=0}^{n} f(j)\binom{n}{j}\theta^j(1-\theta)^{n-j} = 0 \ \ \forall (0,1).$$

By linearly independency of β^j we draw that $f(j) = 0 \, \forall j = 0, ..., n$. Hence T is complete.

Example Let $X_1, ..., X_n \sim N(\theta, \sigma^2)$ with unknown $\theta \in \mathbb{R}$ and a known $\sigma^2 > 0$. We examine $\bar{X_n} = \frac{\sum_{i=1}^n X_i}{n}$ is complete. To simplify we consider $n = 1$ and $\sigma = 1$, so that $T(X) = X \sim N(\theta, 1)$. Suppose

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

Using fourier transform gives you $f = 0$ a.e. Or you express it in terms of moment generating functions.

Theorem 8.2 *(Basu's Theorem). If T is complete and sufficient for $P = \{p_\theta : \theta \in \Theta\}$ and A is ancillary, then $T(X)$ is independent to $A(X)$ with respect to all p_θ.*

Proof Fix p_θ, suffices to show $P_\theta(A \in \mathcal{A}|T) = P_\theta(A \in \mathcal{A})$ a.s. p_θ. Note that $P_\theta(A \in \mathcal{A}|T)$ is free of θ by sufficient of T and $P_\theta(A \in \mathcal{A})$ is free of θ since A is ancillary hence

$$E_{\theta'}(P_\theta(A \in \mathcal{A}|T)) = E_{\theta'}(P_\theta(A \in \mathcal{A})) \ \ \forall \theta' \in \Theta$$

and hence $P_\theta(A \in \mathcal{A}|T) = P_\theta(A \in \mathcal{A})$ by completeness of T. ∎

Example $X_1, ..., X_n \sim N(\mu, \sigma^2)$, (μ, σ^2) both unknown. We claim that $\bar{X_n}$ is independent to $n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

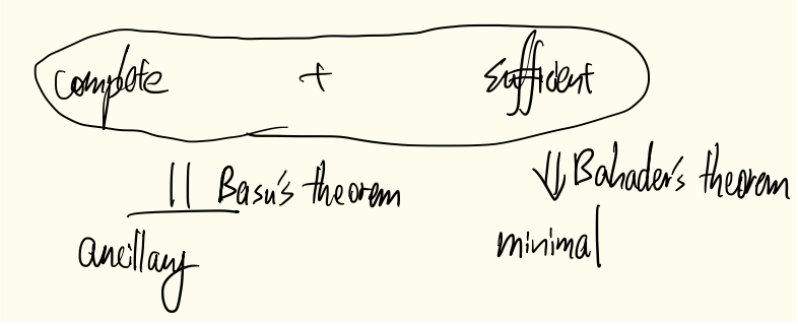
Proof Fix $\sigma > 0$, consider $P_\sigma = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}\}$, then $\bar{X_n}$ is complete and sufficient by above example. Hence $n^{-1}\sum_{i=1}^n (X_i - \bar{X_n})^2$ is ancillary. By Basu' theorem and σ is arbitrary, we are done. ∎

We often show completeness by definition, however, if the distribution family is exponential, then the following theorem helps:

Definition An exponential family with densities $p_\theta = \exp\{\eta(\theta) \cdot T(x) - B(\theta)\}h(x), \theta \in \Omega$, is said to be of full rank if the interior of $\eta(\Omega)$ is not empty and if $T_1, ..., T_s$ do not satisfy a linear constraint of the form $v \cdot T = c$ (a.e. μ).

Theorem 8.3 *In an exponential family of full rank, T is complete.*

The relation can be concluded in a graph:



9 From Data Compression to Risk Reduction/optimal estimation

Definition We call $L(\theta, \delta(X))$ the loss function and $R(\theta, \delta) = E_\theta(L(\theta, \delta(X)))$ the risk function.

Theorem 9.1 *(Rao Blackwell Theorem). Let g be a known function. Suppose that T is sufficient for $P = \{p_\theta : \theta \in \Theta\}$, that $\delta(X)$ is an estimator for $g(\theta)$ for which $E(\delta(X))$, $R(g(\theta), \delta) := E_\theta(L(g(\theta), \delta(X))) < \infty$. If $L(\theta, \cdot)$ is convex then*

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

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

Proof By Jenson's inequality,

$$E_\theta(L(g(\theta), \delta(X))|T) \geq L(g(\theta, E_\theta(\delta(X)|T))) = L(g(\theta), \eta(T)).$$

Taking expectation,

$$E_\theta(L(g(\theta), \delta(X)) \geq E_\theta(L(g(\theta), \eta(T))) \implies R(g(\theta), \delta) \geq R(g(\theta), \eta).$$

∎

Remark This is saying that when a complete sufficient statistic is known, the risk becomes smaller which makes much sense.

Example Let $X_1, ..., X_n \sim$Bernouli(θ), $\theta \in (0, 1)$. Consider the loss function $L(\theta, d) = (\theta - d)^2$. First consider an unreasonable estimator $\delta(X) = X_1$, we have shown that $T(X) = \bar{X}_n$ is sufficient, so we may apply Rao Blackwell Theorem to improve δ. In particular,

$$\eta(T(X)) = E(\delta(X)|T(X)) = \frac{1}{n}\sum_{i=1}^{n} E(X_1|\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i|\bar{X}_n) = E(\bar{X}_n|\bar{X}_n) = \bar{X}_n.$$

Observe that $R(\theta, \eta) = \frac{\theta(1-\theta)}{n} < \theta(1-\theta) = R(\theta, \delta)$. It gives a strict improvement.

Remark Rao Blackwell Theorem do not necesarily lead to a uniformly optimal estimator. For example, consider the above, $\delta_{naive} = \frac{1}{2}$, then $E(T(X)) = E(\delta_{naive}(X)|\bar{X}_n) = \frac{1}{2}$ as well. Since $R(\theta, \eta) = (\frac{1}{2} - \theta)^2$, neither Rao Blackwellized outcome is uniformly better across θ.

10 Unbiased Estimation

Definition (UMRUE, UMVUE). Let g be a known function which can infer θ. An estimator is unbiased if $E_\theta(\delta(X)) = g(\theta)$. We attempt to find an unbiased estimator with uniformly minimum risk, i.e., an unbiased δ satisfying $R(\theta, \delta) \leq R(\theta, \delta')\, \forall \theta \in \Theta$ and unbiased estimator δ'. Such an estimator is called Uniformly Minimum Risk Unbiased Estimator (UMRUE).
In particular, $L(\theta, \delta) = (\theta - \delta)^2$ is the chosen loss function, then UMRUE becomes UMVUE. Note that we have

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

Theorem 10.1 *(Lehmann Scheefe's Theorem). If T is a complete and sufficient statistic, and $E_\theta(h(T(X))) = g(\theta)$, i.e., $h(T(X))$ is unbiased for $g(\theta)$ then $h(T(X))$ is*

1. *The only function of $T(X)$ that is unbiased for $g(\theta)$.*
2. *an UMRUE under any convex loss function.*
3. *the unique UMRUE (a.e.) under any strictly convex loss function.*
4. *the unique UMVUE (a.e.).*

Proof 1. Suppose $E(\tilde{h}(T(X))) = g(\theta)$, then $E_\theta(\tilde{h}(T(X)) - h(T(X))) = 0\ \forall\theta \in \Theta$. Then $\tilde{h}(T(X)) = h(T(X))$ by completeness of T.

2. Consider an unbiased estimator $\delta(X)$, and let $\tilde{h}(T(X)) = E(\delta(X)|T(X))$ and by the Rao Blackwell Theorem, $R(g(\theta), h(T(X))) = R(g(\theta), \tilde{h}(T(X))) \leq R(g(\theta), \delta), \forall\theta$ if the loss function is convex. Therefore, $h(T(X))$ is an UMRUE under any convex loss function.

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

4. By (3).

∎

11 Strategies for obtaining UMVUEs

Let T be a complete sufficient statistics. Mainly we have two strategies for obtaining UMVUEs, namely

(A) Rao Blackwellization (Conditioning). For this strategy, we first find an unbiased estimator δ, and the taking conditional expectation of δ with respect to T, then $E(\delta|T(X)) = \varphi(T(X))$ for some measurable φ and $E(E(\delta|T(X))) = E(\delta) = g(\theta)$, then by Lehmann Scheffe we get the desired UMVUE.

(B) Solve for h satisfying $E_\theta(h(T(X))) = g(\theta), \forall\theta \in \Theta$ by Lehmann Scheffe.

Example Suppose $X_1, ..., X_n \sim$Bernoulli(θ). $T(X) = \sum_{i=1}^n X_i$ is a complete and sufficient statistic and

$$E(\frac{T(X)}{n}) = E(\frac{1}{n}\sum_{i=1}^n X_i) = \theta$$

Therefore $\bar{X}_n$ is an UMRUE for θ under any convex loss function. Suppose now we are interestd in estimating $g(\theta) = \theta^2$. If we choose $\delta(X) = I(X_1 = X_2 = 1) = X_1X_2$, then $E_\theta(\delta(X)) = \theta^2$ is unbiased. Apply (A) to obtain

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

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

Example Suppose $X_1, ..., X_n \sim$ Uniform$(0, \theta)$. In this case, $T(X) = X_{(n)} = \max_{1\le i\le n} X_i$ is a complete and sufficient statistic, and $\delta(X) = 2X_1$ is an unbiased estimator of θ. Given the knowledge of $X_{(n)}$, X_1 is equal to $X_{(n)}$ with probability $\frac{1}{n}$ and distributed according to Uniform$(0, X_{(n)})$ with proability $1 - \frac{1}{n}$. So

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

Hence our UMVUE,

$$\begin{aligned} E(\delta(X)|T(X)) = 2E(X_1|T(X)) = 2\frac{1}{n}T(X) + (1 - \frac{1}{n})\int_0^{T(X)} \frac{x_1}{T(X)}dx_1 \\ = (\frac{n+1}{n})T(X). \end{aligned}$$

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

$$\begin{aligned} \sum_{x=0}^{\infty} \frac{\delta(x)e^{-\theta}\theta^x}{x!} = e^{-a\theta} \ \ \forall\theta \\ \sum_{x=0}^{\infty} \frac{\delta(x)\theta^x}{x!} = e^{(1-a)\theta} = \sum_{x=0}^{\infty} \frac{(1-a)^x\theta^x}{x!} \\ \delta(x) = (1-a)^x \text{ is the UMVUE of } g(\theta). \end{aligned}$$

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

Example When using (B). We often come to the situation of solving an ODE. Let $X_1, ..., X_n$ be i.i.d. absolutely continuous variables with common density $f_\theta, \theta \in \mathbb{R}$, given by

$$f_\theta(x) = \begin{cases} \frac{\phi(x)}{\Phi(x)}, & x < \theta, \\ 0 & x \ge \theta. \end{cases}$$

(This is the density for the standard normal distribution truncated above at θ). We now derive a formula for the UMVUE for $g(\theta)$.

$$p(x_1, ..., x_n|\theta) = \frac{1}{\Phi(\theta)^n}\phi(x_1)...\phi(x_n)1_{\max X_i<\theta}.$$

Hence $T = \max X_i$ is sufficient complete. We will come to the point that

$$\int_{-\infty}^{\theta} h(y)\frac{\phi^n(y)}{\Phi^n(\theta)}dy = g(\theta).$$

Putting $\Phi^n(\theta)$ to the right, and differentiating we get

$$h(\theta) = \frac{n\Phi^{n-1}(\theta)\phi(\theta)g(\theta) + \Phi^n(\theta)g(\theta)}{\phi^n(\theta)}.$$

Suppose we have δ_i UMVUE for $g_i(\theta)$ for $i = 1, 2$. Is $\delta_1 + \delta_2$ UMVUE for $g_1(\theta) + g_2(\theta)$?

Theorem 11.1 *(Characterization of UMVUE (TPE 2.1.7)). Let* $\Delta = \{\delta : E_\theta(\delta^2) < \infty\}$. *Then* $\delta_0 \in \Delta$ *is UMVUE for* $g(\theta) = E_\theta(\delta_0)$ *if and only if* $E_\theta(\delta_0 U) = 0$ *for all* $U \in \mathcal{U}$, $\mathcal{U} = \{$*unbiased estimator of* $0\} = \{U : X \to \mathbb{R}$ *such that* $E_\theta(U(X)) = 0\}$.

Proof If δ_0 is UMVUE, let us consider

$$\delta_\lambda = \delta_0 + \lambda U, \text{ for } \lambda \in \mathbb{R}, U \in \mathcal{U}$$

Since δ_0 has the minimal variance,

$$\text{Var}(\delta_\lambda) = \text{Var}(\delta_0) + \lambda^2 \text{Var}(U) + 2\lambda \text{Cov}(\delta_0, U) \geq \text{Var}(\delta_0)(UMVUE)$$

Consider the quadratic form $q(\lambda) = \lambda^2 \text{Var}(U) + 2\lambda \text{Cov}(\delta_0, U)$. Then q has roots 0 and $-2\text{Cov}(\delta_0, U)/\text{Var}(U)$. If the roots are distinct then the form must be negative at some point which violates the inequality above, hence $-2\text{Cov}(\delta_0, U)/\text{Var}(U) = 0$ and thus $E(U\delta_0) = \text{Cov}(\delta, U) + g\theta E_\theta(U) = 0 + 0 = 0$.

For the converse part, we assume $E(\delta_0(U)) = 0, \forall U \in \mathcal{U}$ and consider any unbiased estimator δ for $g(\theta)$. Then $\delta - \delta_0 \in \mathcal{U}$ so $E(\delta_0(\delta - \delta_0)) = 0$. This implies that $E(\delta_0\delta) = E(\delta^2)$ and subtracting $E(\delta_0)E(\delta)$ on both sides we have

$$\text{Var}(\delta_0) = \text{Cov}(\delta_0, \delta) \leq \sqrt{\text{Var}\delta_0 \text{Var}\delta}$$

Hence $\text{Var}\delta_0 \leq \text{Var}\delta$ for any arbitrary estimator δ and δ_0 is the UMVUE. ∎

Hence $\forall U \in \mathcal{U}$, $E((\delta_1 + \delta_2)U) = E(\delta_1 U) + E(\delta_2 U) = 0 \implies \delta_1 + \delta_2$ is UMVUE for $g_1(\theta) + g_2(\theta)$.

12 Cramer Rao Lower bound

12.1 One variable version

Assume the following:

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

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

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

Define $I(\theta) = E_\theta(\frac{\partial}{\partial \theta} \log p_\theta(x))^2 = \int_A (\frac{(p'_\theta(x))^2}{p_\theta(x)}) d\mu$. Note that log gives information in the sense that, for example, for an i.i.d., $\log p_\theta$ breaks dowm to sum, and we do Taylor expansion to get some information about θ. $\frac{\partial}{\partial \theta} \log p_\theta(x)$ gives how sensitive it is for the density to the θ.

Lemma 12.1 *1. Assume (a)-(c) hold, and (d)* $\frac{\partial}{\partial \theta} \int_A p_\theta(x) d\mu = \int_A \frac{\partial}{\partial \theta} p_\theta(x) d\mu$, *then*

$$I(\theta) = Var_\theta(\frac{\partial}{\partial \theta} \log p_\theta(x)).$$

2. *In addition, $p''_\theta(x)$ exists for all $\theta \in \Theta$, $x \in A$ and (e):*

$$\frac{\partial^2}{\partial\theta^2}\int_A p_\theta(x)d\mu = \int_A \frac{\partial^2}{\partial\theta^2}p_\theta(x)d\mu$$

then

$$I(\theta) = -E_\theta(\frac{\partial^2}{\partial\theta^2}\log p_\theta(x)).$$

Proof 1. Suffices to show $E_\theta(\frac{\partial}{\partial\theta}\log p_\theta(x)) = 0$. We have

$$E_\theta(\frac{\partial}{\partial\theta}\log p_\theta(x)) = \int_A \frac{p'_\theta(x)}{p_\theta(x)}p_\theta(x)d\mu = \int_A \frac{\partial}{\partial\theta}p_\theta(x)d\mu = \frac{\partial}{\partial\theta}\int_A p_\theta(x)d\mu = 0.$$

2. $\frac{\partial^2}{\partial\theta^2}\log p_\theta(x) = \frac{p''_\theta(x)}{p_\theta(x)} - (\frac{p'_\theta(x)}{p_\theta(x)})^2$. Then

$$\begin{aligned} E(\frac{\partial^2}{\partial\theta^2}\log p_\theta(x)) &= \int_A p''_\theta(x)d\mu - E_\theta(\frac{\partial}{\partial\theta}\log p_\theta(x))^2 \\ &= \int_A \frac{\partial^2}{\partial\theta^2}p_\theta(x)d\mu - I(\theta) \\ &= \frac{\partial^2}{\partial\theta^2}\int_A p_\theta(x)d\mu - I(\theta) = -I(\theta). \end{aligned}$$

∎

Remark Information depends on parametrization. If $\eta = \tau(\theta), \tau \in C^1, \tau'(\theta) \neq 0, I(\tau) = \frac{I(\theta)}{(\tau'(\theta))^2}$ because

$$E_\tau(\frac{\partial}{\partial\tau}\log p_{\tau(\theta)}(x))^2 = \frac{I(\theta)}{(\tau'(\theta))^2}.$$

by direct calculation.

Example (Information of normal). Suppose $X \sim N(\theta, 1), \theta > 0$.

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

Therefore

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

However, if we let $\eta = \theta^2$, $X \sim N(\sqrt{\eta}, 1)$. Then we have

$$\frac{\partial}{\partial\eta}\log p_\theta(\eta)(x) = \frac{1}{4\eta} = \frac{1}{4\theta^2} = \frac{1}{(\tau'(\theta))^2}.$$

Theorem 12.2 *(Cramer Rao Lower Bound). Suppose (a)-(d) hold, let $\delta(X)$ be an estimator such that $E_\theta(\delta(X))^2 < \infty$, $I(\theta) \in (0, \infty)$. Assume further that*

$$\int_A \delta(x) \frac{\partial}{\partial \theta} p_\theta(x) d\mu = \frac{\partial}{\partial \theta} E_\theta(\delta(X)).$$

Then

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

Remark For unbiased estimator δ, then $\text{Var}_\theta(\delta(X)) \geq \frac{(g'(\theta))^2}{I(\theta)}$.

Remark This theorem gives another way of showing the UMVUE. If there is an unbiased estimator such that it attains the lower bound, then it must be the UMVUE. However, the UMVUE may not always attain the lower bound, which we will see in the forthcoming examples.

Proof Let $V = \frac{\partial}{\partial \theta} \log p_\theta(x)$, so $E_\theta(V^2) = I(\theta) = \text{Var}_\theta(V)$. Also $E_\theta(V) = 0$ by the above lemma.
Then by Cauchy Schwartz,

$$\text{Var}_\theta(V)\text{Var}_\theta(\delta(X)) \geq (\text{Cov}(V, \delta(X)))^2 \implies \text{Var}_\theta(\delta(X)) \geq \frac{(\text{Cov}(V, \delta(X)))^2}{I(\theta)}.$$

Now it suffices to show $\text{Cov}(V, \delta(X)) = \frac{\partial}{\partial \theta} E_\theta(\delta(X))$. Observe that

$$\begin{aligned}
\text{Cov}(V, \delta(X)) &= E_\theta(V[\delta(X) - E_\theta(\delta(X))]) = E_\theta(V\delta(X)) = \int_A \frac{\partial}{\partial \theta} \log p_\theta(x) \delta(x) p_\theta(x) d\mu \\
&= \int_A \frac{\partial p_\theta}{\partial \theta} \delta(X) d\mu = \frac{\partial}{\partial \theta} \int_A p_\theta(x) \delta(x) d\mu \\
&= \frac{\partial}{\partial \theta} E_\theta(\delta(X)), \text{ where } A := \{x \in X : p_\theta(x) > 0\}.
\end{aligned}$$

∎

Example $X_1, ..., X_n \sim N(\theta, 1)$. We are interested in estimating $g(\theta) = \theta$. Take $\delta(X) = \overline{X_n}$, then $\text{Var}\overline{X_n} = \frac{1}{n}$. We have $\text{Var}_\theta(\delta(X)) \geq \frac{1}{I_n(\theta)} = \frac{1}{n}$, where $I_n(\theta) = E_\theta(\frac{\partial}{\partial \theta} \log p_\theta(x_1, ..., x_n))^2 = n$. Hence $\overline{X_n}$ is the UMVUE. Note that

$$\begin{aligned}
I_n(\theta) &= E(\frac{\partial}{\partial \theta} \log p_\theta(x_1) + ... + \frac{\partial}{\partial \theta} \log p_\theta(x_n))^2 \\
&= \sum_{i=1}^n E(\frac{\partial}{\partial \theta} \log p_\theta(x)) = n.
\end{aligned}$$

as calculated in the example information of normal (12.1).

Example $X_1, ..., X_n \sim \text{Poisson}(\lambda)$, $\lambda > 0$. We want to verify that $\overline{X_n}$ is the UMVUE for λ. Observe that

$$\text{Var}(\overline{X_n}) = \frac{\text{Var}(X_1)}{n} = \frac{\lambda}{n},$$

also $\text{Var}_\lambda(\delta(X)) \geq \frac{1}{nI_1(\lambda)}$. It suffies to show that $I_1(\lambda) = \frac{1}{\lambda}$. Recall that $p_\lambda(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $\log p_\lambda(x) = -\lambda + x\log\lambda - \log x!$, which gives you

$$\frac{\partial}{\partial\lambda}\log p_\lambda(x) = -1 + \frac{x}{\lambda} \implies I_1(\lambda) = E(\frac{x}{\lambda} - 1)^2 = \frac{\text{Var}(X)}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}.$$

Remark Assume (a)-(d) hold, then $\text{Var}_\theta(\delta(X)) \geq \frac{(\frac{\partial}{\partial\theta}E_\theta(\delta(X)))^2}{I(\theta)}$. If the equality hold, then

$$\frac{\partial}{\partial\theta}\log p_\theta(x) = a(\theta)\delta(x) + b(\theta) \text{ a.s. } p_\theta$$

Since it is just Cauchy Schwartz.
Also, if we assume that $\frac{\partial}{\partial\theta}\log p_\theta(x)$ is continuous in θ, then $\theta \mapsto a(\theta)$ and $\theta \mapsto b(\theta)$ are also continuous, provided δ is not a degenerated random variable. This can be seen as the following: for fixed $\theta_0 \in \Theta$,

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

p_θ is a 1 parameter exponential family and $\delta(X)$ is the natural sufficient statistic. Let $A := \{x : \frac{\partial}{\partial\theta}\log p_\theta(x) = a(\theta)\delta(x) = b(\theta)\}$. For there exists $x_1 \neq x_2$ such that $\delta(x_1) \neq \delta(x_2)$. For these x_1, x_2, equality holds in $\frac{\partial}{\partial\theta}\log p_\theta(x) = a(\theta)\delta(x) + b(\theta) =: h(x,\theta)$. Hence

$$\begin{aligned} h(x_1,\theta) &= a(\theta)\delta(x_1) + b(\theta) \\ h(x_2,\theta) &= a(\theta)\delta(x_2) + b(\theta) \\ \implies \frac{h(x_1,\theta) - h(x_2,\theta)}{\delta(x_1) - \delta(x_2)} &= a(\theta). \end{aligned}$$

Hence a is continous in θ and hence so is b.

Theorem 12.3 *Let $p_\theta(x) = e^{\eta(\theta)T(x) - B(\theta)}\tilde{h}(x)$, $\theta \in \Theta$ an open interval. Assume $\tau(\theta) = E_\theta(T)$. Assume T is not a constant random variable, then*

1. *$\tau'(\theta) \neq 0$ and $I(\tau(\theta)) = \frac{1}{Var_\theta(T)}$.*
2. *$I(h(\theta)) = (\frac{\eta'(\theta)}{h'(\theta)})^2 Var_\theta(T)$.*

12.2 Multiparameter Cramer Rao Lower Bound

Suppose the following conditions hold:

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

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

(c) $\frac{\partial p_\theta(x)}{\partial\theta_i}$ exists for all $i = 1, ..., k$ for all $x \in I$ and is finite.

(d) $\frac{\partial}{\partial\theta_i}\int p_\theta(x)d\mu = \int \frac{\partial}{\partial\theta_i}p_\theta(x)d\mu$ for $i = 1, ..., k$.

(e) $\frac{\partial}{\partial\theta_i}\int \delta(x)p_\theta(x)d\mu = \int \frac{\partial}{\partial\theta_i}\delta(x)p_\theta(x)d\mu$ for all $i = 1, ..., k$.

Define the $k \times k$ information matrix, $I(\theta)$ by $I_{ij}(\theta) = E_\theta[(\frac{\partial}{\partial\theta_i}\log p_\theta(x))(\frac{\partial}{\partial\theta_j}\log p_\theta(x))]$. In particular, if $k = 1$, $I(\theta) = E_\theta(\frac{\partial}{\partial\theta}\log p_\theta(x))^2$. Assume $I(\theta)$ is finite and positive definite, then

$$\text{Var}(\delta(X)) \geq \alpha^T I^{-1}(\theta)\alpha,$$

where $\alpha = (\alpha_1, ..., \alpha_k)^T = (\frac{\partial}{\partial\theta_1}E_\theta(\delta(X)), ..., \frac{\partial}{\partial\theta_k}E_\theta(\delta(X)))^T$. In particular, if $\delta(\lambda)$ is unbiased for $g(\theta)$, then $\text{Var}_\theta(\delta(X)) \geq \alpha^T I^{-1}(\theta)\alpha$, $\alpha_i = \frac{\partial}{\partial\theta_i}g(\theta)$, $i = 1, ..., k$.

Proof Let $\psi_i(x) = \frac{\partial}{\partial\theta_i}\log p_\theta(x)$, then

$$\begin{aligned} E_\theta\psi_i(x) &= \int \frac{\partial}{\partial\theta_i}\log p_\theta(x)d\mu = \int \frac{\frac{\partial}{\partial\theta_i}p_\theta(x)}{p_\theta(x)}p_\theta(x)d\mu \\ &= \int \frac{\partial}{\partial\theta_i}p_\theta(x)d\mu = \frac{\partial}{\partial\theta_i}\int p_\theta(x)d\mu = 0 \end{aligned}$$

For a nonzero vector $(a_1, ..., a_k)$, then $E_\theta(\sum_{i=1}^k a_i\psi_i(x)) = 0$. We now claim that $\text{Var}(\sum_{i=1}^k a_i\psi_i(x)) = a^T I(\theta)a$. Observe that

$$\begin{aligned} \text{Var}(\sum_{i=1}^k a_i\psi_i(x)) &= \sum_{i,j} a_i a_j \text{Cov}(\psi_i(x), \psi_j(x)) \\ &= \sum_{i,j} a_i a_j E(\psi_i(x)\psi_j(x)) \\ &= \sum_{i,j} a_i a_j I_{ij}(\theta) = a^T I(\theta)a. \end{aligned}$$

Finally,

$$\begin{aligned} \text{Cov}(\delta(X), \sum_{i=1}^k a_i\psi_i(x)) &= \sum_{i=1}^k a_i \text{Cov}(\delta(X), \psi_i(x)) = \sum_{i=1}^k a_i E(\delta(X)\psi_i(X)) \\ &= \sum_{i=1}^k a_i \int \delta(x)\frac{\partial}{\partial\theta_i}\log p_\theta(x)p_\theta(x)d\mu \\ = \sum_{i=1}^k a_i \int \delta(x)\frac{\partial}{\partial\theta_i}p_\theta(x)d\mu &= \sum_{i=1}^k a_i \frac{\partial}{\partial\theta_i}\int \delta(x)p_\theta(x)d\mu = \sum_{i=1}^k a_i\alpha_i(\theta). \end{aligned}$$

By CS inequality,

$$\begin{aligned} &\text{Var}(\delta(X))\text{Var}(\sum_{i=1}^n a_i\psi_i(x)) \geq \text{Cov}(\sum_{i=1}^k a_i\psi_i(x), \delta(X)) \\ \implies &\text{Var}(\delta(X)) \geq \sup_{a\neq 0}\frac{(\sum_{i=1}^k a_i\alpha_i(\theta))^2}{a^T I(\theta)a} = \alpha^T I^{-1}(\theta)\alpha. \end{aligned}$$

∎

We now give an example of the above theorem, and at the same time an example such that its UMVUE do not attain the lower bound.

Example For $X_1, ..., X_n \sim N(\mu, \sigma^2)$, $\mu > 0$, $\sigma^2 > 0$. We wish to estimate $g_1(\mu, \sigma^2) = \mu$ and $g_2(\mu, \sigma^2) = \sigma^2$. We first claim that $S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})$ is UMVUE for σ^2. This is because

1. $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is complete sufficient.
2. $\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{1}{n-1}(\sum_{i=1}^n X_i^2 - n\overline{X}^2)$.
3. $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X})^2 \sim \chi^2_{n-1} \implies E(\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2) = \frac{\sigma^2}{n-1}(n-1) = \sigma^2$.

We second claim that $\overline{X}$ is UMVUE for μ. $E(\overline{X} - \mu)^2 = \frac{\sigma^2}{n}$. Note that $E(\sum_{i=1}^n \frac{(X_i - \overline{X})^2}{n-1} - \sigma^2)^2 = \text{Var}(\sum_{i=1}^n \frac{(X_i-\overline{X})^2}{n-1}) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}$.
$\log p_{\mu,\sigma^2}(x) = -\frac{n}{2}\log\sigma^2 - \sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^2} + C_n$.
$\frac{\partial}{\partial\mu}\log p_{\mu,\sigma^2}(x) = \sum_{i=1}^n \frac{(X_i-\mu)}{\sigma^2}$.
$\frac{\partial}{\partial\mu^2}(\log p_{\mu,\sigma^2}(x)) = -\frac{n}{\sigma^2}$.
$\frac{\partial}{\partial\sigma^2}(\log p_{\mu,\sigma^2}(x)) = -\frac{n}{2\sigma^2} + \sum_{i=1}^n \frac{(X_i-\mu)^2}{2\sigma^4}$.
$\frac{\partial^2}{(\partial\sigma^2)^2}\log p_{\mu,\sigma^2}(x) = \frac{n}{2\sigma^4} - \sum_{i=1}^n \frac{(X_i-\mu)^2}{\sigma^4}$.
$\frac{\partial^2}{\partial\mu\partial\sigma^2}\log p_{\mu,\sigma^2}(x) = -\sum_{i=1}^n \frac{(X_i-\mu)}{\sigma^4}$.
$I_{22} = -\frac{n}{2\sigma^4} + E\sum_{i=1}^n \frac{(X_i-\mu)^2}{\sigma^6} = -\frac{n}{2\sigma^4} + \frac{n\sigma^2}{\sigma^6} = \frac{n}{2\sigma^4}$, and we have $I = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^2} \end{pmatrix}$.
We therefore have the CRLB for $\mu = \begin{pmatrix} 1 & 0 \end{pmatrix} I^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{I_{11}} = \frac{\sigma^2}{n}$.
CRLB for $\sigma^2 = \begin{pmatrix} 0 & 1 \end{pmatrix} I^{-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{I_{22}} = \frac{2\sigma^4}{n}$. The CRLB is attained by $\overline{X}$ but not S^2.

13 Average Risk Optimality (Bayes Estimator)

Suppose $\{p_\theta : \theta \in \Theta\}$ is a collection of probability measure on X dominated by a σ finite measure μ. Assume that now θ is a random variable on Θ with distribution π, which is regarded as the prior distribution.
Suppose we want to estimate $g(\theta)$, where $g : \Theta \to \mathbb{R}$. For an estimator $\delta(X)$, let the loss incurred be $L(g(\theta), \delta(X))$. Then the risk function, as defined before, is $R(g(\theta), \delta) = E_{X\sim p_\theta}(L(g(\theta)), \delta(X)) = E(L(g(\theta)), \delta(X)|\theta)$. Define the Bayes risk of δ by $r(\pi, \delta) = E_{\theta\sim\pi}(R(g(\theta), \delta(X)))$. An estimator δ_0 is said to be a Bayes estimator if it minimize the Bayes risk, i.e., for any other estimator, we have

$$r(\pi, \delta_0) \le r(\pi, \delta).$$

The conditional distribution of $\pi(\theta|x)$ is called the posterior distribution, and $\pi(\theta)$ is called prior distribution. Define the marginal distribution of X as M (which has the density m w.r.t. μ),

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

Example Let $X_1, ..., X_n \sim N(\theta, \sigma^2)$, where σ is known, $\Theta = \mathbb{R}$. Assume that $\theta \sim N(\mu, \tau^2)$, then

$$p_\theta(x) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{-\sum_{i=1}^n \frac{(X_i-\theta)^2}{2\sigma^2}}$$
$$\pi_\theta(x) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}$$

Then the joint density will be $p_\theta(x)\pi_\theta(x) \propto e^{-\sum_{i=1}^n \frac{(X_i-\theta)^2}{2\sigma^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}$.
The posterior density will be $\propto e^{-\sum_{i=1}^n \frac{(X_i-\theta)^2}{2\sigma^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}} = ... = e^{-\frac{\theta^2}{2}(\frac{n}{\sigma^2}+\frac{1}{\tau^2})+\theta(\frac{\sum_i X_i}{\sigma^2}+\frac{\mu}{\tau^2})}$.
Now if $\pi(\theta|x) \sim N(a, b)$ (Since $P(A|B) = \frac{P(AB)}{P(B)}$ hence we use proportional to in this case) this has density proportional to

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

Which solves like $b^2 = \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}, a = \frac{\mu\sigma^2+n\overline{X}\tau^2}{\sigma^2+n\tau^2}$, hence posterior distribution is $N(\frac{\mu\sigma^2+n\overline{X}\tau^2}{\sigma^2+n\tau^2}, \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2})$.
Observe that as $n \to \infty$, $a \to \overline{X}$ which makes sense about the Bayesian thing.

Remark If the prior has density π, the posterior has density $\pi(\theta|x)$ w.r.t. the dominating measure.

$$\pi(\theta|x)m(x) = p_\theta(x)\pi(\theta) \implies \pi(\theta|x) \propto p_\theta(x)\pi(\theta) \propto g_\theta(T(X))\pi(\theta).$$

Which says that the conditional distribution depends on X through sufficiency.

We first define conditional variance, by

$$\text{Var}(X|Y) = E((X - E(X|Y))^2|Y).$$

Remark In the following, whenever we write $E(f(\theta, X))$, we mean $\int_\Omega f(\theta, X)dP = \int_{\Theta\times X} f(\theta, X)g_{\theta,x}d\theta dx$, where $g_{\theta,x}$ is the joint density of (θ, X). Under this circumstances,

$$\begin{aligned}
\int_{X\le b} E(g(\theta)|X)dP &= \int_{X\le b} g(\theta)dP \\
&= \int_{-\infty}^b \int_\Theta g(\theta)p(x,\theta)\pi(\theta)d\theta dx \\
&= \int_{-\infty}^b \Big[\Big(\int_\Theta g(\theta)\frac{p(x,\theta)\pi(\theta)}{\int_\Theta p(x,\psi)\pi(\psi)d\psi}d\theta\Big)\int_\Theta p(x,\psi)\pi(\psi)d\psi\Big]dx \\
&= \int_{X\le b}\Big(\int_\Theta g(\theta)\frac{p(X,\theta)\pi(\theta)}{\int_\Theta p(X,\psi)\pi(\psi)d\psi}d\theta\Big)dP.
\end{aligned}$$

Hence $E(g(\theta)|X) = \int_\Theta g(\theta)\frac{p(X,\theta)\pi(\theta)}{\int_\Theta p(X,\psi)\pi(\psi)d\psi}d\theta$, and $E(g(\theta)|x) = \int_\Theta g(\theta)\frac{p(x,\theta)\pi(\theta)}{\int_\Theta p(x,\psi)\pi(\psi)d\psi}d\theta$.

And we have the following theorem for obtaining the Bayes estimator, and its Bayes risk.

Theorem 13.1 *If $L(g(\theta), \delta(X)) = \{g(\theta) - \delta(X)\}^2$ then $\delta_0(X) = E(g(\theta)|X)$ is a Bayes estimator with Bayes risk $E(Var(g(\theta)|X))$. If $\delta(X)$ is another Bayes estimator, then $\delta_0 = \delta$ with probability* 1.

Proof Let δ be another estimator, then we have

$$\begin{aligned}
&E(\delta(X)-g(\theta))^2\\
=&E(\delta(X)-\delta_0(X)+\delta_0(X)-g(\theta))^2\\
=&E(\delta(X)-\delta_0(X))^2+E(\delta_0(X)-g(\theta))^2+2E((\delta(X)-\delta_0(X))(\delta_0(X)-g(\theta)))\\
\geq&E(\delta(X)-\delta_0(X))^2+E(\delta_0(X)-g(\theta))^2
\end{aligned}$$

Since

$$\begin{aligned}
E[(\delta(X)-\delta_0(X))(\delta_0(X)-g(\theta))] &= E[E\Big((\delta(X)-\delta_0(X))(\delta_0(X)-g(\theta))|X\Big)]\\
&= E[(\delta_0(X)-\delta(X))(\delta_0(X)-E(g(\theta)|X))]=0
\end{aligned}$$

Hence δ_0 is a Bayes estimator, and any Bayes estimator is Bayes estimator if and only if $\delta(x)=\delta_0(x)$ with probability 1.
Finally, the Bayes risk of

$$\begin{aligned}
\delta_0 &= E\Big(g(\theta)-E(g(\theta)|X)\Big)^2\\
&= E(E(\Big(g(\theta)-E(g(\theta)|X)\Big)^2|X))\\
&= E(\text{Var}(g(\theta)|X))
\end{aligned}$$

∎

Remark (Procedure for finding the Bayes estimator). Note that the joint density will be $p(x|\theta)\pi(\theta)$, where $p(x|\theta)$, π is always given. Hence $\pi(\theta|x)=\frac{p(x|\theta)\pi}{\text{marginal of x}}$.

Remark 1. $\delta_0(x)=\delta(x)$ with probability 1 refers to the joint probability when X and θ are both random. This also means that $\delta_0(x)=\delta(x)$ with probability 1 under the marginal distribution of X.

2. This does not imply $P(\delta(X)=\delta_0(X)|\theta)=1$.

3. If however the marginal distribution of X dominates P_θ, $\theta\in\Theta$, then we have $\delta_0(X)$ is the unique Bayes estimate in the sense that

$$P_\theta(\delta(X)=\delta_0(X))=1 \text{ for all } \theta.$$

Example Suppose $X\sim$Binomial(n,θ), $\theta\in(0,1]$.

$$\pi_1(\theta)=U(0,1),\ \ \pi_2(0)=\pi_2(1)=\frac{1}{2}.$$

Case 1:

$$P(X=x)=\int_0^1\binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta=\binom{n}{x}\int_0^1\theta^x(1-\theta)^{n-x}d\theta=\frac{\binom{n}{x}}{Beta(x+1,n-x+1)}=\frac{1}{n+1}.$$

In this case marginal dominates conditional, i.e., $P(X\in A)=0 \implies P(X\in A|\theta)=0$ hence the Bayes estimate is unique. In this case

$$\begin{aligned}
\pi(\theta|x)\propto\theta^x(1-\theta)^{n-x} &= Beta(x+1,n-x+1)\\
\implies \text{Bayes Estimator} &= \frac{x+1}{n+2}.
\end{aligned}$$

Case 2:

$$\begin{aligned} P(X=x) =& \frac{1}{2}P(X=x|\theta=0)+\frac{1}{2}P(X=x|\theta=1) \\ =& \frac{1}{2}(I(x=0)+I(x=n)). \end{aligned}$$

Hence $P(X=0)=P(X=n)=\frac{1}{2}$ and $P(X=x)=0$ for $x\in\{1,2,...,n-1\}$. Therefore the marginal does not dominate the conditional probability, for example, ? Hence Bayes estimate is not unique, say for $E(\theta|x)$,

$$E(\theta|1)=0.$$

$$E(\theta|n)=P(\theta|n)=1$$

Then the class of all Bayes estimate is given by

$$\delta_0(0)=0,\delta_0(n)=1,\delta_0(x)= \text{ anything for } x\in\{1,...,n-1\}.$$

Lemma 13.2 *A Bayes estimate with respect to squared error cannot be unbiased, unless* $\delta_0(x)=g(\theta)$ *with probability* 1.

Proof Let $\delta_0(X)=E(g(\theta)|X)$ be the Bayes estimates. Assume that $E(\delta_0(X)|\theta)=g(\theta)$(unbiased). We claim that $I:=E(\delta_0(X)-g(\theta))^2=0$.

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

where

$$\begin{aligned} E(\delta_0(X)g(\theta)) &= E(E(\delta_0(X)g(\theta)|X)) \\ &= E(\delta_0(X)E(g(\theta)|X))=E(\delta_0(X))^2 \end{aligned}$$

Or

$$\begin{aligned} E(\delta_0(X)g(\theta)) &= E(E(\delta_0(X)g(\theta)|\theta)) \\ &= E(g(\theta)E(\delta_0(X)|\theta))=E(g(\theta))^2. \end{aligned}$$

Hence

$$I=-I \implies I=0.$$

∎

Definition (Conjugate). A class of probability distribution F is said to be a conjugate family of prior for a model $\{p_\theta:\theta\in\Theta\}$ if the posterior distribution $\pi(\theta|x)$ also belongs to F.

Example $X_1,...,X_n\sim N(\theta,\sigma^2),\theta\sim N(\mu,\tau^2)$, then $\pi(\theta|x)\sim N(a,b)$ as discussed before.

Example $X_1, ..., X_n \sim Binomial(1, p), p \sim Beta(\alpha, \beta)$. This has density

$$\pi_{\alpha,\beta}(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{Beta(\alpha,\beta)}, \text{ where } Beta(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp$$

Note that

$$Beta(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \text{ where } \Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}dx,\ \Gamma(k+1) = k!$$

and we have

$$f_p(x) = p^{\sum x_i}(1-p)^{n-\sum x_i} \implies \pi(p|x) \propto p^{\sum x_i+\alpha-1}(1-p)^{n-\sum x_i+\beta-1} = Beta(\sum x_i + \alpha, n - \sum x_i +$$

and

$$E(p|x) = \frac{\sum x_i + \alpha}{n+\alpha+\beta}$$

Example $X_1, ..., X_n \sim p_\theta(\lambda), \lambda \sim \Gamma(\alpha, \gamma)$, then

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

Example $X_1, ..., X_n \sim U(0, \theta), \theta \sim Pereto(a, c)$,

$$\pi_{a,c}(\theta) = \frac{ac^a}{\theta^{a+1}}1_{\theta>c}.$$

14 Minimaxity

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

$$\sup_{\theta\in\Theta} R(g(\theta), \delta_0(X)) \leq \sup_{\theta\in\Theta} R(g(\theta), \delta(X)).$$

Definition (Bayes Risk of a Prior). Given a probability distribution π (prior) on Θ, define the Bayes risk of the prior π by

$$r(\pi) = r(\pi, \delta_\pi) = \int R(\theta, \delta_\pi)d\pi(\theta).$$

where δ_π is the Bayes Estimator with respect to π.

Definition A prior π is said to be least favourible if $r(\pi) \geq r(\pi')$ for all π', other prior distribution on Θ.

A minimax procedure, by minimizing the maximum risk, tries to do as well as possible in the worst case.

Theorem 14.1 *Suppose π is a distribution on Θ such that $r(\pi) = r(\pi, \delta_\pi) = \sup_\theta R(\theta, \delta_\pi)$. Then*

1. *δ_π is minimax.*
2. *If δ_π is unique Bayes with respect to π, then δ_π is unique minimax.*
3. *π is least favourable.*

Proof 1. Let δ be arbitrary, then

$$\begin{aligned}\sup_{\theta\in\Theta} R(\theta,\delta) &\geq \int_\Theta R(g(\theta),\delta)\pi(d\theta)\\ &= r(\pi,\delta) \geq r(\pi,\delta_\pi) = \sup_{\theta\in\Theta} R(\theta,\delta_\pi).\end{aligned}$$

Hence δ_π is minimax.

2. Let $\delta \neq \delta_\pi$, i.e., there exists θ such that $P_\theta(\delta(X) \neq \delta_\pi(X)) > 0$. Note that we have $r(\pi,\delta) > r(\pi,\delta_\pi)$ in this case hence δ_π is the unique minimax.
3. Let π' be any distribution, need to show that

$$r(\pi') \leq r(\pi)$$

. We observe that

$$\begin{aligned}r(\pi') = \int_\Theta R(g(\theta),\delta_{\pi'})\pi'(d\theta) &\leq \int_\Theta R(g(\theta),\delta_\pi)\pi'(d\theta)\\ &\leq \sup_{\theta\in\Theta} R(g(\theta),\delta_\pi) = r(\pi).\end{aligned}$$

Hence π is least favourable.

∎

Corollary 14.1.1 *A Bayes estimator with constant risk is minimax.*

Proof This means $R(\theta,\delta_\pi) = \alpha$ (free of θ). Hence

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

Then we apply the theorem. ∎

Example Let $X_1,...,X_n \sim B(1,p)$. We find a minimax estimator for p. Let the prior on p be $Beta(\alpha,\beta)$, i.e., $\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$. Then the Bayes estimator is

$$\delta_\pi(x) = \frac{\sum_{i=1}^n x_i + \alpha}{n+\alpha+\beta}.$$

We want to make use of the corollary.

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

To make this free of p, we have

$$\begin{cases} n = 2\alpha(\alpha+\beta) \\ n = (\alpha+\beta)^2 \end{cases}$$

which implies

$$\begin{cases} \alpha + \beta = \sqrt{n} \\ 2\alpha\sqrt{n} = n \end{cases}$$

Hence

$$\alpha = \frac{\sqrt{n}}{2}, \beta = \frac{\sqrt{n}}{2}.$$

Hence $\delta_\pi = \frac{\sum_{i=1}^n x_i + \frac{\sqrt{n}}{2}}{n+\sqrt{n}}$ is the unique minimax estimator.

Definition A sequence of priors $\{\pi_n\}_{n=1}^\infty$ is least favourable if

$$\lim_n r(\pi_n) = \sup_\pi r(\pi).$$

That is, it is a sequence tending to the largest Bayes Risk.

Theorem 14.2 *If there exists a δ_0 such that $\{\pi_n\}_{n\geq 1}$ is a sequence of priors such that*

$$\lim_n r(\pi_n) = \sup_{\theta\in\Theta} R(g(\theta), \delta_0),$$

then

1. *δ_0 is minimax.*
2. *$\{\pi_n\}$ is least favourable.*

Proof 1. Let δ be any other estimator, then

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

Take limit to get $\sup_{\theta\in\Theta} R(g(\theta), \delta_0) = \lim_n r(\pi_n)$ and hence δ_0 is minimax.

2. Need to show $\sup_\pi r(\pi) = \lim_n r(\pi_n)$. Observe that

$$\sup_\pi r(\pi) \geq r(\pi_n) \implies \sup_\pi r(\pi) \geq \lim_n r(\pi_n).$$

for any π. Also

$$\lim_n r(\pi_n) = \sup_{\theta\in\Theta} R(g(\theta), \delta_0) \geq r(\pi, \delta_0) \geq r(\pi)$$

Hence

$$\sup_\pi r(\pi) = \lim_n r(\pi_n).$$

∎

Example Let $X_1, ..., X_n \sim N(\theta, \sigma^2)$. Find a minimax estimate for θ with the squared loss function. We claim that $\overline{X_n}$ is minimax.
Let $\pi_{\mu,\tau^2}(\theta) = N(\mu, \tau^2)$. The Bayes estimator is

$$\delta_\pi = \frac{\frac{n\overline{X_n}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}.$$

Where the Bayes risk is $r(\pi) = r(\pi, \delta_\pi) = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$. There $\delta_0 = \overline{X_n}$, $R(\theta, \overline{X_n}) = E(\overline{X} - \theta)^2 = \frac{\sigma^2}{n} \implies \sup_{\theta\in\Theta} R(\theta, \overline{X_n}) = \frac{\sigma^2}{n}$. Also,

$$\lim_{\tau\to\infty} r(\pi_\tau) = \frac{\sigma^2}{n} = \sup_{\theta\in\Theta} R(\theta, \overline{X_n}).$$

Therefore $\overline{X_n}$ is minimax. Also, $\{\pi_\tau\}_{\tau\in\mathbb{N}}$ is a least favourable distribution.

Lemma 14.3 *Suppose $\delta(X)$ is minimax for $g(\theta)$ on the parameter set $\theta \in \Theta_0$, where $\Theta_0 \subseteq \Theta$. If $\sup_{\theta\in\Theta_0} R(g(\theta), \delta) = \sup_{\theta\in\Theta} R(g(\theta), \delta)$ then δ is the minimax for $\theta \in \Theta$.*

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

Example Assume $\mu \in \mathbb{R}$, $0 < \sigma \leq M$, $\Theta = \mathbb{R} \times [0, M]$. In this case $\overline{X}$ is again minimax. This is because let $\Theta_0 = \mathbb{R} \times \{M\}$. In this case we know that $\overline{X_n}$ is minimax and $\sup_{\theta\in\Theta_0} R(\mu, \overline{X_n}) = \frac{M^2}{n}$. Also, $R(\mu, \overline{X_n}) = \frac{\sigma^2}{n} \implies \sup_{\theta\in\Theta} R(\mu, \overline{X_n}) = \sup_{\sigma\in[0,M]} \frac{\sigma^2}{n} = \frac{M^2}{n} = \sup_{\theta\in\Theta_0 R(\mu,\overline{X_n})}$.
Hence $\overline{X_n}$ is minimax on Θ.

15 Admissibility

Since this section will not be included in the final exam, we skip it first.

16 Asymptotic Optimality

Let $\{X_1, ..., X_n\}$ be i.i.d. from $\{p_\theta : \theta \in \Theta\}$ with pdf $p_\theta(\cdot)$ with respect to some σ-finite measure. Suppose we want to estimate $g(\theta)$ and a candidate estimator $\delta_n(x_1, ..., x_n)$.

Definition (Consistent). We say $\delta_n(x)$ is consistent for $g(\theta)$ if $\delta_n(x) \xrightarrow{p,p_\theta} g(\theta) \forall\theta \in \Theta$.

Example $X_1, ..., X_n \sim Bin(1, \theta)$, UMVUE for $g(\theta) = \theta$ is $\overline{X_n}$. We have

$$\overline{X_n} \xrightarrow{p,p_\theta} \theta \text{ by WLLN.}$$

Hence $\overline{X_n}$ is consistent for θ.

Remark For i.i.d.s $X_1, ..., X_n \sim F$.

1. Assume $E_F|X_1| < \infty$, then $\frac{1}{n}\sum_{i=1}^n x_i \xrightarrow{p} E_F X_1$ by WLLN.
2. Assume $E_F X_1^2 < \infty$, then $W_n := \frac{\sum X_i - nE_F X_1}{\sqrt{n\text{Var}(X_1)}} \xrightarrow{d} N(0, 1)$ by CLT, i.e., $\lim_n P(W_n \leq t) \to \Phi(t)\ \forall t \in \mathbb{R}$.

Definition Let $L(\theta|x_1,...,x_n)=\Pi_{i=1}^n p_\theta(x_i)$ be the likelihood function, that is given $x_1,...,x_n$, the plausibility of seeing θ, hence it is a function of θ, and $l(\theta|x_1,...,x_n)=\log L(\theta|x_1,...,x_n)$ be the log likelihood function. If there exists a unique $\widehat{\theta_n}$ which is a global maximizer of $\theta\mapsto L(\theta|x)$ or $\theta\mapsto l(\theta|x)$, then define $\widehat{\theta_n}$ as the maximum likelihood estimation (MLE) for θ.

Example Suppose $X_1,...,X_n\sim Bin(1,\theta)$, $p_\theta(x)=\theta^x(1-\theta)^x$. Therefore we have

1. $L_n(\theta|x)=\Pi_{i=1}^n p_\theta(x_i)=\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$
2. $l_n(\theta|x)=\sum_{i=1}^n x_i\log\theta+(n-\sum_{i=1}^n x_i)\log(1-\theta)$.

Hence

$$l_n'(\theta|x)=\frac{\sum_{i=1}^n x_i}{\theta}-\frac{n-\sum_{i=1}^n x_i}{1-\theta}.$$
$$l_n''(\theta|x)=-\frac{\sum_{i=1}^n x_i}{\theta^2}-\frac{n-\sum_{i=1}^n x_i}{(1-\theta)^2}<0.$$

Hence $l_n(\cdot|x)$ is strictly concave.
Observe that

$$l_n'(\theta|x)\Big|_{\theta=\widehat{\theta_n}}=0\implies\frac{\sum_{i=1}^n x_i}{\widehat{\theta_n}}=\frac{n-\sum_{i=1}^n x_i}{1-\widehat{\theta_n}}$$
$$\widehat{\theta_n}=\frac{\sum_{i=1}^n x_i}{n}=\overline{X_n}.$$

Hence MLE exists and equals $\overline{X_n}$. Also, $\overline{X_n}\xrightarrow{p,p_\theta}\theta\,\forall\theta\in(0,1)$ hence we have the consistency and $\sqrt{n}(\overline{X_n}-\theta)\xrightarrow{d,p_\theta}N(\frac{1}{\theta(1-\theta)})$ (asymptotic normality via CLT).

Theorem 16.1 *Suppose $X_1,...,X_n$ are i.i.d. from p_θ for some $\theta\in\Theta$, with pdf $p_\theta(\cdot)$.*

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

(A1) $\{p_\theta,\theta\in\Theta\}$ have common support.

Then, $p_{\theta_0}(l_n(\theta_0|x)>l_n(\theta|x))\to 1\,\forall\theta\neq\theta_0$.

Proof Let $T_n=\frac{1}{n}\sum_{i=1}^n\log\frac{p_\theta(x_i)}{p_{\theta_0}(x_i)}$, then $T_n\xrightarrow{p,p_\theta}E_{\theta_0}(\log\frac{p_\theta(x_i)}{p_{\theta_0}(x_i)})$. Now

$$E_{\theta_0}(\log\frac{p_\theta(x_i)}{p_{\theta_0}(x_i)})=\int\log\frac{p_\theta(x)}{p_{\theta_0}(x)}p_{\theta_0}(x)d\mu=-D(\theta_0||\theta)<0\text{ for }\theta\neq\theta_0.$$

Hence

$$P_{\theta_0}(T_n<0)\xrightarrow{n}1$$

but

$$\begin{aligned}T_n<\infty\iff&\frac{1}{n}\sum_{i=1}^n\log\frac{p_\theta(x_i)}{p_{\theta_0}(x_i)}<0\\\iff&\log\Pi_{i=1}^n p_\theta(x_i)<\log\Pi_{i=1}^n p_{\theta_0}(x_i)\\\iff&l_n(\theta|x)<l_n(\theta_0|x).\end{aligned}$$

∎

Corollary 16.1.1 *Suppose (A0) and (A1) hold. If Θ is finite, then the MLE $\widehat{\theta_n}$ exists with high probability (prob$\to$ 1) and $p_{\theta_0}(\widehat{\theta_n} = \theta_0) \to 1$.*

Proof Let $\Theta = \{\theta_0, ..., \theta_k\}$, $p_{\theta_0}(l_n(\theta_0|x) > l_n(\theta_j|x)) \xrightarrow{n} 1$ for $1 \leq j \leq k$, and hence

$$p_{\theta_0}(l_n(\theta_0|x) > \max_{i \leq j \leq k} l_n(\theta_j|x)) \xrightarrow{n} 1$$

Let $A_n = \{x : l_n(\theta_0|x) > \max_{1 \leq j \leq k} l_n(\theta_j|x)\}$. If $x \in A_n$, then $\widehat{\theta_n}(x) = \theta_0$ and $p_{\theta_0}(A_n) \to 1$. ∎

Now we talk about things concerning the MLE expansion, that is, what is $\sqrt{n}(\widehat{\theta_n} - \theta_0) \to ?$.

$$0 = l'n(\widehat{\theta_n}) = l'_n(\theta_0) + (\widehat{\theta_n} - \theta_0)l''_n(\theta_0) + \frac{1}{2}(\widehat{\theta_n} - \theta_0)^2 l'''_n(\xi_n), \;\; \xi_n \in (\theta_0, \widehat{\theta_n}).$$

Hence

$$(\widehat{\theta_n} - \theta_0)\Big(l''_n(\theta_0) + \frac{1}{2}(\widehat{\theta_n} - \theta_0)l'''_n(\xi_n)\Big) = -l'_n(\theta_0)$$

and we have

$$\sqrt{n}(\widehat{\theta_n} - \theta_0) = \frac{-\frac{l'_n(\theta_0)}{\sqrt{n}}}{-\frac{l''_n(\theta_0)}{n} - \frac{\frac{1}{2}(\widehat{\theta_n} - \theta_0)l'''_n(\xi_n)}{n}}$$

Now it suffices to show

1. $\frac{1}{\sqrt{n}} l'_n(\theta_0) \xrightarrow{d} N(0, I(\theta_0))$.
2. $-\frac{1}{n} l''_n(\theta_0) \xrightarrow{p} I(\theta_0)$.
3. $\frac{1}{n}(\widehat{\theta_n} - \theta_0)l'''_n(\xi_n) \xrightarrow{p} 0$.

where $I(\theta_0)$ denotes the fisher information, then

$$\sqrt{n}(\widehat{\theta_n} - \theta_0) \xrightarrow{d} \frac{N(0, I(\theta_0))}{I(\theta_0)} = N(0, I(\theta)^{-1}).$$

Observe that

$$\frac{1}{\sqrt{n}} l'_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log p_\theta(x_i)\Big|_{\theta=\theta_0} \xrightarrow{d,\theta_0} N(0, I(\theta_0)) \text{ By CLT.}$$

$$\frac{1}{n} l''_n(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial^2 \theta} \log p_\theta(x_i)\Big|_{\theta=\theta_0} \xrightarrow{p,\theta_0} -I(\theta_0) \text{ by WLLN.}$$

and

$$|\frac{1}{n}(\widehat{\theta_n} - \theta_0)l'''_n(\xi_n)| \leq |\widehat{\theta_n} - \theta_0| \frac{1}{n} \sum_{i=1}^{n} M_i(x_i) \xrightarrow{p} 0 \cdot E_\theta(M(X_1)) = 0.$$

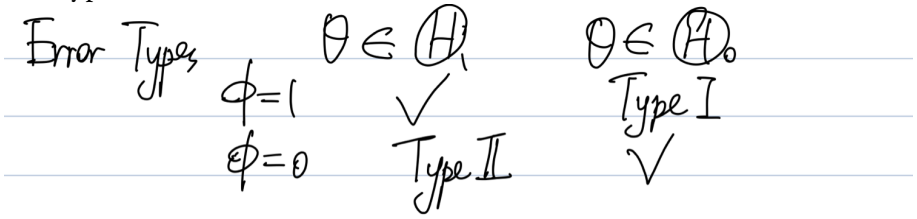
By consistency and WLLN.

17 Hypothesis Testing

Let $\{p_\theta : \theta \in \Theta\}$ be a collection of probability measure on $\mathcal{X}$, dominated by a σ-finite measure μ. Let Θ_0 and Θ_1 be two disjoint subsets such that $\Theta = \Theta_0 \cup \Theta_1$. Given $X \sim p_\theta$ for some $\theta \in \Theta$ we want to decide whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$. We call Θ_0 null and Θ_1 alternative, we always want to test null.

Definition A function $\phi : \mathcal{X} \to \{0, 1\}$ is called a nonrandomized test function.

The types of error is concluded as follows:



Where type I means null is true but we reject it, and type II means it is actually alternative but we accept it as null.

Definition (Power function of ϕ). The performance of this test is described by its power funciton $\beta(\cdot)$, which gives the chances of rejecting Θ_0 as a function of $\theta \in \Theta$.

$$\beta(\theta) = 1 - \text{Probability of Type II error under } P_\theta = P_\theta(\phi = 1).$$

In words it means the ability to reject Θ_0.
(Size of a test ϕ): It is defined as

$$\sup_{\theta \in \Theta_0} p_\theta(\phi = 1)$$

i.e., the maximum error it rejects null while it is actually null.

Let $\alpha \in (0, 1)$, a test ϕ is called level α if $\sup_{\theta \in \Theta_0} p_\theta(\phi = 1) = \alpha$.

Definition A test ϕ is called uniformly most powerful level α test if given any other level α test ψ, we have $P_\theta(\phi = 1) \geq P_\theta(\psi = 1)$ for all $\theta \in \Theta_1$.

Definition A function $\phi : \mathcal{X} \to [0, 1]$ is called a randomized test function or just a test function, if $\phi(x) = p$, and we toss a coin with probability of heads p. If head we choose Θ_1, if tail we choose Θ. We replace all the previous definition $P_\theta(\phi = 1)$ by $E_\theta \phi$, i.e.,
The power function:

$$\beta(\theta) = E_\theta(\phi), \theta \in \Theta$$

The size function:

$$\sup_{\theta \in \Theta_0} E_\theta(\phi)$$

Theorem 17.1 *(Neyman-Pearson). Suppose that we want to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ at level α. Then*

1. *There exists test ϕ satisfying*

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

(b) There exists $K \in [0, \infty]$ *such that*

$$\begin{cases} 1 \text{ if} \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k \\ 0 \text{ if} \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < k \end{cases}$$

2. *If a test satisfies (a), (b) above then* ϕ *is a most powerful test for testing* $\theta = \theta_0$ *versus* $\theta = \theta_1$ *at level* α.

3. *If* ϕ *is most powerful at level* α*, it must satisfy (b). It also satisfies (a) unless* $E_{\theta_1}(\phi) = 1$.

This theorem means for the case simple versus simple, there must exists most powerful test, and is of NP form, and NP form means most powerful test.

Proof 1. The idea of the proof is we cutoff at level α and choose appropriate number such that we have the desired size. If $\alpha = 0$, take $K = \infty$, $\phi = 0$. If $\alpha = 1$, take $k = 0$, $\phi = 1$. For $\alpha \in (0, 1)$, let

$$\begin{aligned} \alpha(c) &:= p_{\theta_0}(p_{\theta_1}(x) > cp_{\theta_0}(x)), c > 0 \\ &= p_{\theta_0}(\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > c) = 1 - p_{\theta_0}(\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \leq c). \end{aligned}$$

Hence α is decreasing and right continuous. Also, $\alpha(c-) - \alpha(c) = p_{\theta_0}(\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = c), \alpha(\infty) = 0, \alpha(0-) = 1$.
There exists c_0 such that $\alpha(c_0) \leq \alpha \leq \alpha(c_0-)$. Let

$$\phi = \begin{cases} 1 & p_{\theta_1}(x) > c_0 p_{\theta_0}(x) \\ \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} & p_{\theta_1(x)} = c_0 p_{\theta_0}(x) \\ 0 & p_{\theta_1}(x) < c_0 p_{\theta_0}(x) \end{cases}$$

with if $\alpha(c_0-) = \alpha(c_0)$, set $\phi = 1$ on $p_{\theta_1}(x) = c_0 p_{\theta_0}(x)$. Hence

$$\begin{aligned} E_{\theta_0}(\phi) &= P_{\theta_0}(p_{\theta_1}(x) > c_0 p_{\theta_0}(x)) + \frac{\alpha - \alpha(c_0)}{\alpha(c_0-) - \alpha(c_0)} P_{\theta_0}(p_{\theta_0} = c_0 p_{\theta_1}(x)) \\ &= \alpha. \end{aligned}$$

2. Let ϕ be of the NP form, i.e., there exists K such that $E_{\theta_0}(\phi) = \alpha$,

$$\phi(x) = \begin{cases} 1 \text{ if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k \\ 0 \text{ if } \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < k \end{cases}$$

Let $\phi^*(x)$ be such that $E_{\theta_0}(\phi^*(x)) \leq \alpha$. We need to show that $E_{\theta_1}(\phi^*(X)) - E_{\theta_1}(\phi^*(X)) \geq 0$.
Consider the integral:

$$\begin{aligned} &\int (\phi(x) - \phi^*(x))(p_{\theta_1}(x) - kp_{\theta_0}(x))d\mu \\ &= \int_{\phi > \phi^*} (\phi(x) - \phi^*(x))(p_{\theta_1}(x) - kp_{\theta_0}(x))d\mu + \int_{\phi < \phi^*} (\phi(x) - \phi^*(x))(p_{\theta_1}(x) - kp_{\theta_0}(x))d\mu \end{aligned}$$

Observe that if $\phi > \phi^* \geq 0 \implies p_{\theta_1}(x) \geq kp_{\theta_0}(x)$, if $\phi < \phi^* \leq 1 \implies p_{\theta_1}(x) \leq kp_{\theta_0}(x)$.
Therefore

$$\begin{aligned} 0 &\leq \int (\phi(x) - \phi^*(x))(p_{\theta_1}(x) - kp_{\theta_0}(x))d\mu \\ &= E_{\theta_1}(\phi(x)) - E_{\theta_1}(\phi^*(x)) - k(E_{\theta_0}(\phi(x)) - E_{\theta_0}(\phi^*)(x)) \end{aligned}$$

$$\implies E_{\theta_1}\phi(x) - E_{\theta_1}\phi^*(x) \geq k(E_{\theta_0}(\phi(x)) - E_{\theta_0}(\phi^*(x))) \geq K(\alpha - \alpha) = 0$$

3. Let ϕ^* be a MP test. Let ϕ be the test from 1. Let $S = (\{\phi < \phi^*\} \cup \{\phi > \phi^*\}) \cap \{x : p_1(x) \neq kp_0(x)\}$, suppose that $\mu(S) > 0$, then by the argument above we have

$$\int_{(\{\phi<\phi^*\}\cup\{\phi>\phi^*\})} (\phi - \phi^*)(p_1 - kp_0)d\mu = \int_S (\phi - \phi^*)(p_1 - kp_0)d\mu > 0.$$

and hence ϕ is more powerful against p_1 than ϕ^*, contradiction arises, which means whenever $(\{\phi < \phi^*\} \cup \{\phi > \phi^*\})$, we must have $\{p_0(x) = kp_1(x)\}$, hence whenever $\{p_0(x) \neq kp_1(x)\}$, $\phi = \phi^*$.
From the construction of ϕ, $E_{\theta_0}(\phi^*(x)) = E_{\theta_0}(\phi(x)) = \alpha$ unless $k = 0$, but $k = 0$ if and only if $E_{\theta_1}(\phi^*(x)) = 1$.

∎

Remark From the above proof we see that if $\{x : p_{\theta_1}(x) = kp_{\theta_0}(x)\}$ is of measure zero, most powerful test is unique.

Example Let $X_1, ..., X_n \sim N(\theta, 1)$. We are testing $H_0 : \theta = 0$ versus $H_1 : \theta = 1$ at level α.

$$\frac{p_{\theta=1}(x_1, ..., x_n)}{p_{\theta=0}(x_1, ..., x_n)} = \frac{\frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^n (x_i-1)^2}}{\frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^n (x_i)^2}} = e^{-\sum_{i=1}^n x_i - \frac{n}{2}}$$

Hence $\phi = 1$ if $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > k$ iff $\sum_{i=1}^n x_i - \frac{n}{2} > \log k$ iff $\sum_{i=1}^n x_i > k' := \log k + \frac{n}{2}$.
Hence

$$\phi(x) \begin{cases} 1 & \text{if } \sum x_i > k' \\ 0 & \text{if } \sum x_i < k' \end{cases}$$

In practise, α is given, how we find k is through the following:
where $\alpha = E_{\theta_0}\phi(x) = p_{\theta_0}(\sum_{i=1}^n x_i > k') \implies k' = \sqrt{n}Z_{1-\alpha}$.

Example $X_1, X_2 \sim Bernouli(\theta)$. Test $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta = \frac{2}{3}$ at level $\alpha = \frac{1}{2}$. We consider

Sample	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$p_{\theta_0}(x_1, x_2)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$p_{\theta_1}(x_1, x_2)$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{4}{9}$
$\frac{p_{\theta_1}}{p_{\theta_0}}$	$\frac{4}{9}$	$\frac{8}{9}$	$\frac{8}{9}$	$\frac{16}{9}$

Which suggests us to take $k = \frac{8}{9}$. Let

$$\phi(x_1, x_2) = \begin{cases} 1\ (x_1, x_2) = (1, 1) \\ 0\ (x_1, x_2) = (0, 0) \\ \text{randomized such that } E_{\theta_0}(\phi(x_1, x_2)) = \frac{1}{2}. \end{cases}$$

Corollary 17.1.1 *Let $\beta = \beta(\theta_1)$ denote the power of the MP test for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, at level $\alpha \in (0, 1)$. Then $\beta \geq \alpha$. Furthermore, $\beta > \alpha$ unless $p_{\theta_1} = p_{\theta_0}$.*

Proof Let ϕ be the MP test from part 1. of the NP Lemma. Let $\psi(x) = \alpha$ then $E_{\theta_0}(\psi) = \alpha$ is of level α, hence $\beta = E_{\theta_1}(\phi(x)) \geq E_{\theta_1}(\psi(x)) = \alpha$. Suppose $\beta = \alpha$, then ψ is a MP test, since $\psi \neq 0, 1$, $p_{\theta_1}(x) = k p_{\theta_0}(x)$ a.s. μ, and hence $k = 1$ hence $p_{\theta_1} = p_{\theta_0}$. ∎

Example Suppose $X_1, ..., X_n \sim N(\theta, 1)$. Testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ at level α. For $\theta_1 > 0$, we test $H_0 : \theta = 0$ versus $H_1' : \theta = \theta_1$ at level α. The MP test for this problem is

$$\phi(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^n x_i > \sqrt{n} Z_{1-\alpha} \\ 0 & \text{if } \sum_{i=1}^n x_i < \sqrt{n} Z_{1-\alpha} \end{cases}$$

Hence ϕ is uniformly MP for testing testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$.

18 Monotone Likelihood Ratio

Definition Suppose Θ is an interval. We say that $\{p_\theta : \theta \in \Theta\}$ has the monotone likelihood ratio property in a statistic $T(X)$ if for all $\theta_1 < \theta_2$, $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}$ is a nondecreasing function of $T(X)$.

Example $p_\theta(x) = e^{\eta(\theta)T(x) - B(\theta)} h(x)$, $\theta \in (a, b)$, η non decreasing. Then

$$\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} = e^{(\eta(\theta_2) - \eta(\theta_1))T(x)} e^{-B(\theta_2) + B(\theta_1)} := g(T(X)).$$

where

$$g(t) = e^{(\eta(\theta_2) - \eta(\theta_1))t} e^{-B(\theta_2) + B(\theta_1)}.$$

Example $X_1, ..., X_n \sim Uniform(0, \theta)$. $p_\theta(x) = \frac{1}{\theta^n} I(X_{(n)} < \theta)$. Let $T = X_{(n)}$, $\theta_1 < \theta_2$. If $0 < T < \theta_1$, $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} = (\frac{\theta_1}{\theta_2})^n$. If $\theta_1 \leq T \leq \theta_2$, $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} = \infty$. If $\theta_2 \leq T$. $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} = 0/0$ (set to be ∞).

Theorem 18.1 *Let $\{p_\theta(\cdot); \theta \in \Theta\}$ be MLR on $T(X)$ such that $p_{\theta_1} \neq p_{\theta_2}$ if $\theta_1 \neq \theta_2$ and Θ is an interval.*

1. *For testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ at level $\alpha \in (0, 1)$. There exists a UMP test ϕ of the form*

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

and $E_{\theta_0}\phi(x) = \alpha$.

2. *The power function $\beta(\theta) = E_\theta\phi$ is strictly increasing on the set $\{\theta : 0 < \beta(\theta) < 1\}$, i.e., if $\theta_1 < \theta_2 \in \Theta$ such that $\beta(\theta_1), \beta(\theta_2) \in (0,1)$ then $\beta(\theta_1) < \beta(\theta_2)$.*

3. *For all $\theta' \in \Theta$, the test of part 1. is UMP for testing $H_0 : \theta \leq \theta'$ vs $H_1 : \theta > \theta'$ at level $\alpha' = \beta(\theta')$.*

4. *For any $\theta < \theta_0$, ϕ minimizes $\beta(\theta)$ amongst all test ψ satisfying $E_{\theta_0}\psi(x) = \alpha$.*

Proof 1. Let $f(c) = p_{\theta_0}(T(x) > c)$, $f(\infty) = 0$, $f(-\infty) = 1$. There exists $c_0 \in [-\infty, \infty]$ such that $f(c_0-) \geq \alpha \geq f(c_0)$. Let

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > c_0 \\ \frac{\alpha - f(c_0)}{f(c_0-) - f(c_0)} & \text{if } T(x) = c_0 \\ 0 & \text{if } T(x) < c_0 \end{cases}$$

Fix $\theta_1 > \theta_0$, we need to show ϕ is MP for $\theta = \theta_0$ vs $\theta = \theta_1$. Let $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = g_{\theta_0,\theta_1}(T(x))$ where $g_{\theta_0,\theta_1}(\cdot)$ is non decreasing.
Set $K = g_{\theta_2,\theta_1}(c_0)$. If $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} > K$ iff $g_{\theta_2,\theta_1}(T(x)) > g_{\theta_2,\theta_1}(c_0) \implies T(x) > c_0 \implies \phi = 1$.
If $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < k \implies \phi = 0$. Hence ϕ is of NP form $\implies$ ϕ is MP for $\theta = \theta_0$ vs $\theta = \theta_1$ at level α. Hence ϕ is UMP for $\theta = \theta_0$ vs $\theta > \theta_0$ at level α.(Since we only consider $\theta \in \Theta_1$ when talking about UMP).
Now we need to show $\sup_{\theta \leq \theta_0} E_\theta\phi \leq \alpha$ such tht ϕ is UMP for $\theta \leq \theta_0$ versus $\theta > \theta_0$.
Fix $\theta_0' \leq \theta_0$. Consider the test problem of $\theta = \theta_0'$ versus $\theta = \theta_0$ at level $\beta(\theta_0') = E_{\theta_0'}\phi$. Repeating the proof as above, ϕ is MP for this problem hence $\beta(\theta_0') = E_{\theta_0'}\phi \leq E_{\theta_0}\phi = \alpha$ (because size $\leq$ power, in this case we set the alternative to be θ_0 hence this time it is the power). So ϕ is level α UMP test for $\theta \leq \theta_0$ versus $\theta > \theta_0$.

2. Fix $\theta' \leq \theta''$ assume $\beta(\theta'), \beta(\theta'') \in (0,1)$. Need to show $\beta(\theta') \leq \beta(\theta'')$. Consider the problem of testing $\theta = \theta'$ vs $\theta = \theta''$ at level $\beta(\theta')$, ϕ is MP for this problem $\implies$ $\beta(\theta') < \beta(\theta'')$ because of again size $\leq$power.

3. Repeat the proof. See [TSH 3.4.1].

4. Fix $\theta' < \theta_0$, need to show ϕ minimizes $E_{\theta'}\tilde{\phi}$ for all tests with $E_{\theta_0}\tilde{\phi} = \alpha$, that is, if and only if

$$\{1 - \phi \text{ maximises } 1 - E_{\theta'}\tilde{\phi} \text{ subject to } 1 - E_{\theta_0}\tilde{\phi} = 1 - \alpha\} \text{ iff}$$
$$\{\psi = 1 - \phi \text{ maximises } E_{\theta'}\tilde{\psi} \text{ subject to } E_{\theta_0}\tilde{\psi} = 1 - \alpha\}$$

i.e., ϕ is MP for $\theta = \theta_0$ versus $\theta = \theta'$ at level $1 - \alpha$, where

$$\psi = \begin{cases} 1 \text{ if } T(x) < c_0 \\ \gamma \text{ if } T(x) = c_0 \\ 0 \text{ if } T(x) > c_0 \end{cases}$$

∎

Example $X_1, ..., X_n \sim U(0, \theta)$. Test $H_0 : \theta = 1$ versus $H_1 : \theta > 1$ at level α. By the theorem, a UMP test is given by

$$\phi = \begin{cases} 1 \text{ if } X_{(n)} > K \\ 0 \text{ if } X_{(n)} < K. \end{cases}$$

and $\alpha = E_{\theta=1}\phi = p_{\theta=1}(X_{(n)} > K) = 1 - p_{\theta=1}(X_{(n)} \leq K) = 1 - K^n$. Hence $K = (1-\alpha)^{\frac{1}{n}}$.

Example (CauchyLocation Model). Let X have the density $p_\theta(x) = \frac{1}{1-(x-\theta)^2}$. We find two points at which the MLR condition fails. For any fixed $\theta > 0$,

$$\frac{p_\theta(x)}{p_0(x)} = \frac{1+x^2}{1+(x-\theta)^2} \to 1 \text{ as } x \to \infty, x \to -\infty.$$

But $\frac{p_\theta(x)}{p_0(x)} = \frac{1}{1+\theta^2} < 1$ which is strictly less than 1, hence as $T(x) = x$, it is not MLR.

18.1 Strategies for finding UMPs

Note that the existence of UMP is not guaranteened. (The following steps imitates the proof of theorem of MLR above).

1. Reduce the composite alternative to a simple alternative if H_1 is composite, for $\theta_1 \in \Theta_1$ and test H_0 against $H_1 : \theta = \theta_1$ and hoping that it doesn't depend on θ_1.
2. collapse the composite null to simple null.
3. Apple NP lemma: Find the MP LRT test for simple vs simple case/ use MLR trick.

19 Least Favourable Distribution

Skipped first since not included in the exam.

References

[1] Tony Sit *STAT5010 Advanced Statistical Inferences 2018-2019*

[2] Shao Jun *Mathematical Statistic*

[3] Robert W. Keener *Theoretical Statistics*