Notes in Mathematical Statistics I

Dennis (Shaoxiong) Zheng

Abstract

These notes of the course STAT 30100 Mathematical Statistics I in Winter 2020 Quarter at The University of Chicago are written based on the lectures taught by Professor Mary Sara McPeek, the notes taken by Tom Hen (tomhen@uchicago.edu) in Winter 2016 Quarter and the two textbooks *Statistical Inference, 2nd edition* by Casella, Berger and *Course in Large Sample Theory, 1st edition* by Ferguson.

The notes are written for review purpose. I follow the way of formal concept analysis (FCA) when naming the section names, so that one may look at the table of contents, lists of definitions, theorems and examples, and then try to recall all the underlying details. If he can do so, he should be confident for the exam.

If you find any typo, please notify me at denniszheng@uchicago.edu. For more course notes, please visit my website.

Last update: March 10, 2020.

Contents

List of Definitions

List of Theorems

List of Examples

0 Preliminaries

We do not need $X_1, \cdots, X_n$ to be independent and identically distributed, in this section.

0.1 Convergence

0.1.1 Convergence in Probability

Definition 1 (Converges in Probability)**.** A sequence of random variables, $X_1, \cdots, X_n$ converges in probability to a random variable X if, for every $\epsilon > 0$

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \epsilon\right) = 0$$

or equivalently

$$\lim_{n\to\infty} P\left(|X_n - X| < \epsilon\right) = 1$$

denoted by

$$X_n \xrightarrow{\mathcal{P}} X$$

0.1.2 Almost Sure Convergence

Definition 2 (Converges Almost Surely)**.** A sequence of random variables, $X_1, \cdots, X_n$ converges in almost surely to a random variable X if, for every $\epsilon > 0$

$$P\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = 1$$

denoted by

$$X_n \overset{a.s.}{\to} X$$

The definition of convergence in probability is the limit of probability while here the definition is probability of an event: a inequality involving a limit. Recall a random variable is a real-valued function. Let S be aa sample space S containing elements s. Then $X_n(s)$ and $X(s)$ are functions defined on S. The definition here states that $X_n(s)$ converges to $X(s)$ for all $s \in S$ except for $s \in N \subset S$ but $P(N) = 0$.

0.1.3 Convergence in Distribution

Definition 3 (Converges in Distribution)**.** A sequence of random variables, $X_1, \cdots, X_n$ converges in distribution to a random variable X if, for every $\epsilon > 0$, at all points x where $F_X(x)$ is continuous,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

denoted by

$$X_n \xrightarrow{\mathcal{D}} X$$

Remark. It is the CDF that converge, not the random variables.

0.1.4 Relation: Convergence *a.s.* $\Rightarrow \mathcal{P} \Rightarrow \mathcal{D}$

Theorem 4 (Convergence *a.s.* $\Rightarrow \mathcal{P} \Rightarrow \mathcal{D}$)**.** *If the sequence of random variables $X_1, \cdots, X_n$ converges almost surely to a random variable X, the sequence also converges in probability to X.*

If the sequence of random variables $X_1, \cdots, X_n$ converges in probability to a random variable X, the sequence also converges in distribution to X.

0.2 Law of Large Numbers

0.2.1 Weak Law of Large Numbers $\bar{X}_n \xrightarrow{\mathcal{P}} \mu$

Theorem 5 (Weak Law of Large Numbers)**.** *Let $X_1, \cdots, X_n$ be iid r.v. with mean μ and variance $\sigma^2 < \infty$. Then for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| < \epsilon\right) = 1$$

i.e.

$$\bar{X}_n \xrightarrow{\mathcal{P}} \mu$$

Proof. By Chebychev's Inequality,

$$P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) = P\left(\left(\bar{X}_n - \mu\right)^2 \geq \epsilon^2\right) \leq \frac{\mathrm{E}\left(\bar{X}_n - \mu\right)^2}{\epsilon^2} = \frac{\mathrm{Var}\,\bar{X}_n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

□

0.2.2 Strong Law of Large Numbers $\bar{X}_n \xrightarrow{a.s.} \mu$

Theorem 6 (Strong Law of Large Numbers)**.** *Let $X_1, \cdots, X_n$ be iid r.v. with mean μ and variance $\sigma^2 < \infty$. Then for every $\epsilon > 0$,*

$$P\left(\lim_{n\to\infty} \left|\bar{X}_n - \mu\right| < \epsilon\right) = 1$$

i.e.

$$\bar{X}_n \xrightarrow{a.s.} \mu$$

Note that in both theorems we assume finite variance $\sigma^2 < \infty$, but in fact this is a stronger assumption than is needed. Both the weak and strong laws hold without this assumption. The only condition needed is $E|X_i| < \infty$.

0.3 Central Limit Theorem $\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \xrightarrow{\mathcal{D}} N(0,1)$

Theorem 7 (Central Limit Theorem)**.** *Let $X_1, \cdots, X_n$ be a sequence of i.i.d. random variables such that $E(X_i) = \mu$, $Var(X_i) = \sigma^2 > 0$. Let $G_n(x)$ denote the CDF of $\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma}$, then*

$$\lim_{n\to\infty} G_n(x) = \Phi(x)$$

or equivalently,

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \xrightarrow{\mathcal{D}} N(0,1)$$

0.4 Existence of Expectation

For any random variable X we can decompose it as

$$X = X^+ - X^-$$

where $X^+ = \max(X, 0)$ and $X^- = \min(-X, 0)$. As a result, $E(X) = E(X^+) - E(X^-)$ and $|X| = X^+ + X^-$.

- If $E(X^+) < \infty$ and $E(X^-) < \infty$, or equivalently, $E|X| < \infty$ then we say X is *integrable.*
- If at least one of $E(X^+)$ and $E(X^+)$ is finite, then we say $E(X) = E(X^+) - E(X^-)$ *exists.*
- If $E(X^+) = E(X^-) = \infty$, then we say $E(X)$ *doesn't exist.*

0.5 Asymptotic Theorems

0.5.1 Asymptotic Distribution $b_n(X_n - a_n) \xrightarrow{\mathcal{D}} X$

Definition 8 (Univariate asymptotic distribution)**.** Consider a sequence of real-valued r.v.'s $X_1, X_2, \cdots X_n$. If there are fixed sequences of real numbers $a_n \in \mathbb{R}$ and $b_n > 0$ such that $b_n(X_n - a_n) \xrightarrow{\mathcal{D}} X$ where X is a non-degenerate r.v., then we say we have found an asymptotic distribution of the sequence $\{X_n\}$.

Remark. Non-degenerate means that X does not obtain a single value w.p.1.

Remark. If you are asked to find an asymptotic distribution you need to provide the sequences $\{a_n\}$ and $\{b_n\}$ and the non-degenerate r.v. X, these together characterize the asymptotic distribution.

Definition 9 (Multivariate asymptotic distribution)**.** Consider a sequence of real-valued random vectors $\mathbf{X}_1, \mathbf{X}_2, \cdots \mathbf{X}_n \in \mathbb{R}^d$. If there are fixed sequences of real numbers $a_n \in \mathbb{R}^d$ and $b_n \in \mathbb{R}^d > 0$ and a random vector $\mathbf{Y} \in \mathbb{R}^d$ for which $Y_1, \ldots, Y_d$ are non-degenerate, such that

$$\begin{pmatrix} b_{n_1}(X_{n_1} - a_{n_1}) \\ \vdots \\ b_{n_d}(X_{n_d} - a_{n_d}) \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} Y_1 \\ \vdots \\ Y_d \end{pmatrix}$$

then we say we have found an asymptotic distribution of the sequence $\{\mathbf{X}_n\}$.

0.5.2 Slutsky's Theorem

Theorem 10 (Slutsky)**.**

1. Suppose $X_n \in \mathbb{R}^d$, $X_n \xrightarrow{\mathcal{D}} X$ and $f : \mathbb{R}^d \to \mathbb{R}^k$ is a function such that $\mathbb{P}(X \in \mathcal{C}(f)) = 1$, where $\mathcal{C}(f)$ is the continuity set of f. Then $f(X_n) \xrightarrow{\mathcal{D}} f(X)$.
2. Suppose $X_n \in \mathbb{R}^d$, $Y_n \in \mathbb{R}^d$, $X_n \xrightarrow{\mathcal{D}} X$ and $(X_n - Y_n) \xrightarrow{\mathcal{P}} 0$, then $Y_n \xrightarrow{\mathcal{D}} X$.
3. Suppose $X_n \in \mathbb{R}^d$, $Y_n \in \mathbb{R}^k$, $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{\mathcal{D}} c = \text{constant}$, then $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} X_n \\ c \end{pmatrix}$.

Corollary. *By combining 1 and 3 above we have $f(X_n, Y_n) \xrightarrow{\mathcal{D}} f(X_n, c)$, such as $X_n + Y_n \xrightarrow{\mathcal{D}} X + c$ $Y_n^\top X_n \xrightarrow{\mathcal{D}} c^\top X_n$ and $X_n / Y_n \xrightarrow{\mathcal{D}} X_n / c$ in one-dimensional case. This is very useful when X is random and $Y_n \xrightarrow[CLT]{\mathcal{D}} \mu_Y$.*

Proof. Ferguson Chapter 6. □

0.5.3 Cramer's Theorem (Delta-Method) $\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{\mathcal{D}} \dot{g}(\mu)X$

Theorem 11 (Cramer)**.** *Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be a mapping such that g is continuous in a neighborhood of $\mu \in \mathbb{R}^d$. Suppose $X_n \in \mathbb{R}^d$ is a sequence of r.v. such that $\sqrt{n}(X_n - \mu) \xrightarrow{\mathcal{D}} X$. Then*

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{\mathcal{D}} \dot{g}(\mu)X$$

where $\dot{g}$ is the Jacobian matrix of g.

Remark. $\sqrt{n}(X_n - \mu) \xrightarrow{\mathcal{D}} X$ implies $X_n \xrightarrow{\mathcal{D}} \mu$.

Proof. (See Ferguson Chapter 7 for more details) By Taylor expansion,

$$g(x) = g(\mu) + g'(\mu)(x - \mu) + \frac{1}{2} g''(w)(x - \mu)^2$$

for some w between x and μ. Then

$$\sqrt{n}\left(g\left(X_n\right)-g(\mu)\right)=g'(\mu)\underbrace{\sqrt{n}\left(X_n-\mu\right)}_{\xrightarrow{\mathcal{D}}X}+\frac{1}{2}g''(w)\underbrace{\sqrt{n}\left(X_n-\mu\right)}_{\xrightarrow{\mathcal{D}}X}\underbrace{\left(X_n-\mu\right)}_{\xrightarrow{\mathcal{D}}0}$$

$$\xrightarrow{\mathcal{D}} g'(\mu)X$$

□

Remark. What if $\dot{g}(\mu)=0$ such that RHS is degenerate? Since determining convergence in distribution to a constant does not constitute finding an asymptotic distribution, we consider a higher order approximation.

$$g(x)=g(\mu)+g'(\mu)(x-\mu)+\frac{1}{2}g''(w)(x-\mu)^2+\frac{1}{6}g^{(3)}(w)(x-\mu)^3$$

Then

$$n\left(g\left(X_n\right)-g(\mu)\right)=\sqrt{n}\underbrace{g'(\mu)}_{0}\sqrt{n}(X_n-\mu)+\frac{1}{2}g''(\mu)\left[\sqrt{n}(X_n-\mu)\right]^2+\frac{1}{6}g^{(3)}(w)\left[\sqrt{n}(X_n-\mu)\right]^2\underbrace{\left(X_n-\mu\right)}_{\xrightarrow{\mathcal{D}}0}$$

$$\xrightarrow{\mathcal{D}}\frac{1}{2}g''(\mu)X^2$$

Example 12 (Apply Cramer to Normal)**.** If

$$\sqrt{n}\left(X_n-\mu\right)\xrightarrow{\mathcal{D}}N(0,\Sigma)$$

then

$$\sqrt{n}\left(g\left(X_n\right)-g(\mu)\right)\xrightarrow{\mathcal{D}}N_k\left(0,\dot{g}(\mu)\Sigma\dot{g}(\mu)^\top\right)$$

Example 13 (Apply higher order Cramer)**.** Suppose $Y_1,\ldots,Y_n\overset{iid}{\sim}F$ where F is some univariate distribution with finite mean μ and finite variance σ^2. The CLT tells us

$$\sqrt{n}\left(\bar{Y}_n-\mu\right)\xrightarrow{\mathcal{D}}N\left(0,\sigma^2\right)$$

We are interested in the asymptotic distribution of $\bar{Y}_n^2$ or $\left(\bar{Y}_n\left(1-\bar{Y}_n\right)\right)$. By Cramer's Theorem, let $g(x)=x(1-x)$, then

$$\sqrt{n}\left(\bar{Y}_n\left(1-\bar{Y}_n\right)-\mu(1-\mu)\right)\xrightarrow{\mathcal{D}}N\left(0,g'(\mu)^2\sigma^2\right)=N\left(0,(1-2\mu)^2\sigma^2\right)$$

Note that when $\mu=\frac{1}{2}$, we have

$$\sqrt{n}\left(\bar{Y}_n\left(1-\bar{Y}_n\right)-\frac{1}{4}\right)\xrightarrow{\mathcal{D}}0$$

so the RHS distribution is degenerate. This implies that $\bar{Y}_n\left(1-\bar{Y}_n\right)$ converges to $1/4$ at a rate higher than $\frac{1}{\sqrt{n}}$. Then by taking 3rd order approximation, we have

$$n\left[\bar{Y}_n\left(1-\bar{Y}_n\right)-\frac{1}{4}\right]\xrightarrow{\mathcal{D}}\frac{1}{2}g''(\mu)\left[N\left(0,\sigma^2\right)\right]^2=-\sigma^2\chi_1^2$$

0.5.4 Relation between Slutsky's and Cramer's Theorems

Given a asymptotic distribution of X_n, when we want to find the asymptotic distribution $g(X_n)$, it seems that both theorems work. There is a subtle deference.

Example 14 (Slutsky v.s. Cramer)**.** Given $\sqrt{n}(Z_n - \theta) \xrightarrow{\mathcal{D}} Z$, what is the asymptotic distribution for $\frac{n}{n+1}X_n$? First, by Slutsky, let $X_n = \sqrt{n}(Z_n - \theta) \xrightarrow{\mathcal{D}} Z$, and $Y_n = \frac{n}{n+1} \to 1$ then

$$X_n Y_n = \sqrt{n}(\frac{n}{n+1}Z_n - \frac{n}{n+1}\theta) \xrightarrow{\mathcal{D}} Z \times 1$$

Instead, by Cramer's Theorem, let $g(x) = \frac{n}{n+1}x$, then $g'(x) = \frac{n}{n+1}$, thus

$$\sqrt{n}(\frac{n}{n+1}Z_n - \frac{n}{n+1}\theta) \xrightarrow{\mathcal{D}} \frac{n}{n+1}Z$$

which need one more step to apply Slutsky to show $\frac{n}{n+1}Z \xrightarrow{\mathcal{D}} Z$.

So in this example, Slutsky is more efficient.

Part I
Distributions

1 Families of Distributions

1.1 Exponential Families

1.1.1 Definition $f(x|\theta) = h(x) \cdot c(\theta) \exp\left(\sum_{i=1}^{k} w_i(\theta) t_i(x)\right)$

Definition 15 (k-parameter-exponential family). A family of PDFs or PMFs $\{f(x|\theta)\}$, where $\theta \in \Theta \subset \mathbb{R}^k$, is called a k-parameter-exponential family if for every $\theta \in \Theta$ one can represent $f(x|\theta)$ in the following form

$$f(x|\theta) = h(x) \cdot c(\theta) \exp\left(\sum_{i=1}^{k} w_i(\theta) t_i(x)\right)$$

where

- $h(x) \geq 0$ for all $x \in X$.
- $0 < c(\theta) = \left[\int h(x) \exp\left(\sum_{i=1}^{k} w_i(\theta) t_i(x)\right) dx\right]^{-1} < \infty$ for all $\theta \in \Theta$. Note $c : \mathbb{R}^k \to \mathbb{R}$ and the quantity $c(\theta)$ serves as normalizing constant.
- $t_i(x)$ and $w_i(\theta)$ are real valued functions for $1 \leq i \leq k$.

Remark. Let the support of f be $S := \{x \in X | f(x|\theta) > 0\}$. It can be seen that if $f(x|\theta)$ belongs to an exponential family then $S = \text{supp}h(x)$. Since $h(x)$ does not depend on θ, then S can not depend on θ.

Example 16 ($Bin(p)$ is from exponential family). If n is known and p is unknown, then $\Theta = (0, 1)$ and

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} = \overbrace{\binom{n}{x}}^{} \overbrace{(1-p)^n}^{n(x)} \exp(\overbrace{x}^{t_1(x)} \overbrace{\log \frac{p}{1-p}}^{\omega_i(p)})$$

So it belongs to an exponential family.

If p is known and n is unknown, then the support of f depends on the parameter n. So it does not belong to an exponential family.

1.1.2 Definition of Natural Parameter Space

Definition 17 (Natural Parameter Space). If $f(x|\theta)$ belongs to an exponential family then we can reparameterize it,

$$f(x|\eta) = h(x) c^*(\eta) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right)$$

The natural parameter space is

$$H := \left\{(\eta_1, \ldots, \eta_k) \in \mathbb{R}^k | 0 < \int h(x) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) dx < \infty\right\} \subseteq \mathbb{R}^k$$

The condition above is to ensure $c^*(\eta) = \left[\int h(x) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) dx\right]^{-1}$. Essentially, we have a mapping $\eta : \Theta \to H$.

Note that the natural parameter space is only defined up to affine transformations since one can write

$$f(x|\eta) = \left[\frac{h(x)}{\exp\left\{\sum_{i=1}^{k} \frac{b_i}{a_i} t_i(x)\right\}}\right] c^*(\eta) \exp\left(\sum_{i=1}^{k} (a_i\eta_i + b_i)\frac{t_i(x)}{a_i}\right)$$

Example 18 (Natural Parameter Space for $N(\mu, \sigma^2)$)**.** The Normal PDF

$$f\left(x|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right); \mu \in \mathbb{R}, \sigma^2 > 0, x \in \mathbb{R}$$

can be written as

$$f\left(x|\mu,\sigma^2\right) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\mu^2}{2\sigma^2}\right)}_{c*(\eta)} \exp(\underbrace{\frac{-x^2}{2}}_{t_1(x)}\underbrace{\left(\frac{1}{\sigma^2}\right)}_{\eta_1} + \underbrace{x}_{t_2(x)}\underbrace{\left(\frac{\mu}{\sigma^2}\right)}_{\eta_2}); \eta = \left(\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}\right), c^*(\eta) = e^{\frac{-n^2}{2\pi n}}\sqrt{\frac{\eta_1}{2\pi}}$$

Thus the natural parameterization is given by the transformation

$$\left(\mu,\sigma^2\right) \mapsto \left(\frac{1}{\sigma^2}, \frac{\mu}{\sigma^2}\right) = \eta\left(\mu,\sigma^2\right)$$

Theorem 19 (Relation Between $c^*(\eta)$ and $E[\mathbf{t}(X)]$)**.** *It can be shown that in the natural parameterization one has*

$$\frac{\partial \log c^*(\eta)}{d\eta} = -E[\mathbf{t}(X)]; \mathbf{t}(X) := (t_1(X), \ldots, t_k(X))$$

and

$$\frac{\partial^2 \log c^*(\eta)}{d\eta d\eta^\top} = -\operatorname{Cov}(\mathbf{t}(X), \mathbf{t}(X)) = -E\left[\mathbf{t}(X)(\mathbf{t}(X))^\top\right] - E[\mathbf{t}(X)]E[\mathbf{t}(X)]^\top$$

In the one dimensional case,

$$\frac{\partial^2 \log c^*(\eta)}{d\eta^2} = \operatorname{Var}(t(X))$$

1.2 Curved Exponential Families

1.2.1 Definition: $\dim(\Theta) = d < k$

Definition 20 (Curved Exponential Families)**.** Let $f(x|\theta) = h(x)c(\theta)\exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\right\}, \theta \in \Theta$ be a family of densities, then

- If $\dim(\Theta) = d < k$, the family is known as a curved exponential family.
- If $d = k$, the family is known as a full exponential family.
- If $d > k$, then the parameter Θ is non-identifiable. That is, there exists $\theta_1 \neq \theta_2$ such that $f\left(x|\theta_1\right) = f\left(x|\theta_2\right)$ for all x.

Example 21 ($N(\mu, \mu^2)$ is a curved exponential family)**.** The family $N(\mu, \mu^2)$ is a curved exponential family.

$$f(x|\mu) = \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(\frac{-(x-\mu)^2}{2\mu^2}\right) = \underbrace{\frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2}\right)}_{c(\theta)} \exp\left(-\frac{x^2}{2\mu^2} + \frac{x}{\mu}\right)$$

So $d = 1 < k = 2$. The natural parameters are

$$\eta_1 = \frac{1}{\mu}, \eta_2 = \frac{1}{\mu^2}, H = \left\{(\mu_1, \mu_2) \in \mathbb{R}^2 | \mu_1 \neq 0, \mu_2 = \mu_1^2\right\}$$

1.3 Location-Scale Families

1.3.1 Definition: $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$

Definition 22 (Location-Scale Families)**.** Let f: $\mathbb{R} \to \mathbb{R}$ be a PDF, the family

$$\mathcal{F} := \left\{ \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) | (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+ \right\}$$

indexed by the parameter (μ, σ) is called a location-scale family with standard PDF f, location parameter μ and scale parameter σ.

Example 23 (Normal family is a location-scale family)**.** The Normal family $\left\{ \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) | \mu \in \mathbb{R}, \sigma > 0 \right\}$ where ϕ is the standard normal PDF is a location-scale family.

1.3.2 Characterization:

Theorem 24 (Characterization of location-scale families)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a PDF and let $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$. Then X is a real-valued r.v. with PDF $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ iff there exists a r.v. Z with pdf $f(z)$ such that $X = \sigma Z + \mu$.*

In order to show some family of densities $\mathcal{F}$ is a location-scale family, one has to find some PDF $g : \mathbb{R} \to \mathbb{R}$ such that $\frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right) \in \mathcal{F}$ for all (μ, σ), and such that for all $f \in \mathcal{F}$ can be written as $\frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right)$.

1.3.3 Location Families $\mathcal{F} := \{f(x-\mu) | \mu \in \mathbb{R}^n\}$

Generalizing the definition above, let f: $\mathbb{R}^n \to \mathbb{R}$ be a PDF, the family

$$\mathcal{F} := \{f(x-\mu) | \mu \in \mathbb{R}^n\}$$

indexed by μ is called a location family.

1.3.4 Scale Families $\mathcal{F} := \left\{ \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) | \sigma > 0 \right\}$

Univariate location families are

$$\mathcal{F} := \left\{ \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right) | \sigma > 0 \right\}$$

There are several approaches to generalize the definition to multivariate case. One would be let $Z \sim f$ and define other members of the family by

$$X = AZ$$

where A is a matrix from suitable transformation group. An example is the multivariate normal distribution $X = \Sigma Z + \mu$, where Σ

2 Dependence and Correlation

Consider a case of a bivariate random vector in $\mathbb{R}^2$ with PDF f.

2.1 Independence

Definition 25 (Independence)**.** Two random variables X, Y are independent if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \text{ for all } (x,y) \in \mathbb{R}^2$$

Remark. A simple criterion is that the support of $f_{X,Y}$ can be factorized to be the product of the support of X and the support of Y. If the two marginal supports are dependent, then they two random variables are dependent.

Theorem 26 (Equivalent definitions for independence)**.**

1. *X, Y are independent by the previous definition.*
2. *Conditional PDF = Unconditional PDF:*

$$f_{Y|X}(y|x) = f_Y(y) \text{ for all } (x,y) \in \text{Supp}(X,Y)$$

where $f(y|x) := \frac{f(x,y)}{f(x)}$, or

$$f_{X|Y}(x|y) = f_X(x) \text{ for all } (x,y) \in \text{Supp}(X,Y)$$

where $f(x|y) := \frac{f(x,y)}{f(y)}$.

3. *Factorization of PDF: There exist functions $g, h : \mathbb{R}^2 \to \mathbb{R}$ such that*

$$f_{X,Y}(x,y) = g(x)h(y)$$

Note g and h may not be marginal PDFs.

4. *Factorization of CDF to marginal CDFs:*

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \text{ for all } (x,y) \in \mathbb{R}^2$$

5. *Factorization of expectation to marginal expectations: If for all bounded function $g, h : \mathbb{R}^2 \to \mathbb{R}$,*

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

6. *Factorization of characteristic function to marginal characteristic functions*

$$\phi_{X,Y}(t_1,t_2) = \phi_X(t_1)\,\phi_Y(t_2) \text{ for all } (t_1,t_2) \in \mathbb{R}^2$$

Remark. All of the above definitions can be extended to the multivariate case.

2.2 Correlation

2.2.1 Definition $\rho(X,Y) := \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$

Definition 27 (Correlation)**.** Let X, Y be two r.v. with finite first and second moments. The correlation between X, Y is

$$\rho(X,Y) := \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

2.2.2 Range $|\rho(X,Y)| \leq 1$

Theorem 28 (Range of a Correlation)**.** *The range of a correlation is*

$$|\rho(X,Y)| \leq 1$$

The equality is attained iff there exists $a, b \in \mathbb{R}$ such that

$$\mathbb{P}(Y = aX + b) = 1$$

Furthermore, $a > 0 \Leftrightarrow \rho = 1$ and $a < 0 \Leftrightarrow \rho = -1$.

Remark. Lack of correlation does not imply independence, it is possible for X, Y to be dependent but $\rho(X,Y) = 0$. For instance, $X \sim N(0,1)$ and $Y = X^2$.

2.2.3 Correlation of Transformed Variables $\rho(h(X), g(X))$

Theorem 29 (Correlation of Transformed Variables $\rho(h(X), g(X))$)**.** *For functions $h, g : \mathbb{R} \to \mathbb{R}$, if both h and g are non-decreasing, or both are non-increasing, then $\rho(h(X), g(X)) \geq 0$.*

If one is non-decreasing and the other is non-increasing, then $\rho(h(X), g(X)) \leq 0$.

Proof. It suffices to show that

$$\mathbb{E}[h(X)g(X)] \geq \mathbb{E}[h(X)] \cdot \mathbb{E}[g(X)]$$

since

$$\text{sign}(\rho(h(X), g(X))) = \text{sign}(\text{Cov}(h(X), g(X))) = \mathbb{E}[h(X)g(X)] - \mathbb{E}[h(X)] \cdot \mathbb{E}[g(X)]$$

Suppose X_1, X_2 are independent copies of X. By assumption,

$$(g(X_1) - g(X_2))(h(X_1) - h(X_2)) \geq 0$$

then

$$\begin{aligned} 0 &\leq \mathbb{E}\left[(g(X_1) - g(X_2))(h(X_1) - h(X_2))\right] \\ &= \mathbb{E}\left[g(X_1)h(X_1) - g(X_1)h(X_2) - g(X_2)h(X_1) + g(X_2)h(X_2)\right] \\ &= 2\mathbb{E}[g(X)h(X)] - 2\mathbb{E}[g(X)]\mathbb{E}[h(X)] \\ &= 2(\mathbb{E}[h(X)g(X)] - \mathbb{E}[h(X)] \cdot \mathbb{E}[g(X)]) \end{aligned}$$

The proof for the second case is similar. □

Remark. Note that it is not necessarily true that if two r.v. X and Y have $\rho(X,Y) > 0$ then $\rho(h(X), h(Y)) \geq 0$ for any non-decreasing function g. An example is $U_1 \sim$ Uniform $(0,1)$ and

$$U_2 = \begin{cases} \frac{1}{2} - U_1 & 0 \leq U_1 \leq \frac{1}{2} \\ \frac{3}{2} - U_1 & \frac{1}{2} < U_1 \leq 1 \end{cases}$$

It can be shown that $U_2 \sim$ Uniform $(0,1)$ and $\rho(U_1, U_2) = \frac{1}{2} > 0$.
Let $g(x) = 1_{\left\{x \geq \frac{3}{4}\right\}}$ be a non-decreasing function, then

$$\mathbb{E}[g(U_1)g(U_2)] = 0, \mathbb{E}[g(U_1)] = \mathbb{E}[g(U_2)] = \frac{1}{4}$$

Thus $\rho(g(U_1), g(U_2)) < 0$.

2.2.4 Maximal Correlation $\rho_{\max}(X,Y) = \sup_{g,h}\rho(g(X),h(Y))$

Definition 30 (Maximal correlation)**.** The maximal correlation between two r.v. X, Y is defined as

$$\rho_{\max}(X,Y) = \sup_{g,h}\rho(g(X),h(Y))$$

where g, h are functions such that $\rho(g(X),h(Y))$ exists.
It can be shown that $\rho_{\max}(X,Y) = 0$ iff $X \perp\!\!\!\perp Y$.

2.3 Iterative Formulas

2.3.1 Conditional Variance: $\mathrm{Var}(X|Y) := \mathbb{E}\left[X^2|Y\right] - (\mathbb{E}[X|Y])^2$

2.3.2 Conditional Covariance: $\mathrm{Cov}(X,Y|Z) := \mathbb{E}[X\cdot Y|Z] - \mathbb{E}[X|Z]\cdot\mathbb{E}[Y|Z]$

2.3.3 Iterative Expectation: $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$

A.k.a. tower rule, law of total expectation.

Proof. When there is a joint density $f_{X,Y}(x,y)$,

$$\mathbb{E}[X] = \iint x f_{X,Y}(x,y)dxdy = \iint x f_{X|Y}(x|y)f_Y(y)dxdy = \mathbb{E}[\mathbb{E}[X|Y]]$$

□

Theorem 31 (MGF of random sum of r.v.)**.** *Suppose $\xi_1, \xi_2, \dots$ are i.i.d. real-valued r.v. with MGF $M_\xi(t)$ and suppose Y is a positive integer-valued r.v. with MGF $M_Y(t)$. Assume that $\{Y, \xi_1, \xi_2, \dots\}$ is an independent collection, then the MGF of $X := \sum_{i=1}^{Y}\xi_i$ is given by $M_X(t) = M_Y\left[\log\left(M_\xi(t)\right)\right]$.*

Proof. By iterative expectation

$$\begin{aligned}
M_X(t) &= \mathbb{E}\left[e^{tX}\right] \\
&= \mathbb{E}\left[e^{t\sum_{i=1}^{Y}\xi_i}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{t\sum_{i=1}^{Y}\xi_i|Y}\right]\right] \text{ by iterative expectation} \\
&= \mathbb{E}\left[\left[M_\xi(t)\right]^Y\right] \\
&= \mathbb{E}\left[e^{Y\log M_\xi(t)}\right] \\
&= M_Y\left[\log M_\xi(t)\right]
\end{aligned}$$

□

2.3.4 Iterative Variance: $\mathrm{Var}(X) = \mathrm{Var}(\mathbb{E}[X|Y|) + \mathbb{E}[\mathrm{Var}(X|Y)]$

Example 32 (Iterative Variance)**.** Suppose an insect lays Y eggs, each surviving with probability p. Let X be the number of surviving eggs and assume $Y \sim Poi(\lambda)$, $X|Y \sim Bin(Y,p)$. Then

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Var}(\mathbb{E}[X|Y|) + \mathbb{E}[\mathrm{Var}(X|Y) \\
&= \mathrm{Var}(Yp) + E[Yp(1-p)] \\
&= p^2\lambda + p(1-p)\lambda \\
&= p\lambda
\end{aligned}$$

By the previous theorem, we can further specify the distribution of X. Since $M_\xi(t) = pe^t + (1-p)$ and $M_Y(t) = e^{\lambda\left(e^t-1\right)}$, we have

$$M_X(t) = e^{\lambda\left(e^{\log\left(M_\xi(t)\right)}-1\right)} = e^{\lambda(M_\xi(t)-1)} = e^{\lambda p\left(e^t-1\right)}$$

which is the MGF of $Poi(\lambda p)$.

2.3.5 Iterative Covariance: $\mathrm{Cov}(X,Y) = \mathrm{Cov}(\mathbb{E}[X|Z],\mathbb{E}[Y|Z]) + \mathbb{E}[\mathrm{Cov}(X,Y|Z)]$

2.3.6 Simpson's paradox: $\mathrm{Cov}(X,Y|Z) \geq 0 \not\Longleftrightarrow \mathrm{Cov}(X,Y) \geq 0$

2.4 Mixture Distribution

2.4.1 Definition: $f(x)$ depends on other r.v.

Definition 33 (Mixture distribution)**.** A r.v. X is said to have a mixture distribution if its distribution depends on quantity that also has a distribution. An example is shown in Example Iterative Variance. More common examples are $f_{X|Y}(x|y)$.

Remark. For conditional distributions $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$, we require $f_Y(y) \neq 0$. There are alternative suitable definitions when X, Y are both discrete, the case where one is discrete and one is continuous is a little bit more sensitive.

2.4.2 Two Conditional Distributions (under conditions)⇒ Joint Distribution

Theorem 34 (Two Conditional Distribution (under conditions) ⇒ Joint Distribution)**.** *Given two functions $f_1(x|y) \geq 0$ and $f_2(y|x) \geq 0$, such that*

$$\int_{\mathbb{R}} f_1(x|y)dx = \int_{\mathbb{R}} f_2(y|x)dy = 1$$

and

$$\mathrm{Supp}\, f_1 = \mathrm{Supp}\, f_1 := \mathcal{S} = \mathcal{S}_x \times \mathcal{S}_y \subseteq \mathbb{R}^2$$

Then there exists a joint density $f(x,y)$ such that $f(x|y) = f_1(x|y)$ and $f(y|x) = f_2(y|x)$, iff

- *Functional Compatibility*
  There exist function $h : \mathcal{S}_x \to \mathbb{R}$ and $g : \mathcal{S}_y \to \mathbb{R}$ such that

$$\frac{f_1(x|y)}{f_2(y|x)} = h(x)g(y)\ \forall (x,y) \in \mathcal{S}$$

- *Finite Integral*

$$\int_{\mathcal{S}_x} \frac{f_1(x|y)}{f_2(y|x)}dx = \frac{1}{f(y)} < \infty\ \forall y \in \mathcal{S}$$

The first condition is motivated from

$$f(x,y) = f_1(x|y)f(x) = f_2(y|x)f(y) \Rightarrow \frac{f_1(x|y)}{f_2(y|x)} = \frac{f(x)}{f(y)}\ \forall (x,y) \in \mathcal{S}$$

For the second condition, we also require

$$\int_{\mathcal{S}_y} \frac{f_2(y|x)}{f_1(x|y)}\mathrm{d}y = \frac{1}{f(x)} < \infty\ \forall x \in \mathcal{S}$$

but this is implied by $\int_{\mathcal{S}_x} \frac{f_1(x|y)}{f_2(y|x)}dx = \frac{1}{f(y)} < \infty$ (See the paper Arnold (1989) JASA 84:152-156.) They also give conditions for uniqueness.

Example 35 (No joint density due to functional incompatibility)**.** Suppose $X|Y \sim \exp(y)$ and $Y|X \sim \exp(x)$, then there is no joint density $f(x,y)$ since the functional compatibility does not holds.

Why? Since $f(x|y) = \frac{1}{y}e^{\frac{-z}{y}}$ and $f(y|x) = \frac{1}{x}e^{-\frac{y}{x}}$, we have $\frac{f(x|y)}{f(y|x)} = \frac{x}{y}e^{-\frac{x}{y}+\frac{y}{x}}$. Suppose the functional compatibility holds, then $\frac{f(x|y)}{f(y|x)} = h(x)g(y)$ for all $(x,y) \in \mathbb{R}^+ \times \mathbb{R}^+$. However, plugging in $x = 1$ gives $g(y) = \frac{\frac{1}{y}e^{y-\frac{1}{y}}}{h(1)}$ and plugging in $y = 1$ gives $h(x) = \frac{xe^{\frac{1}{x}-x}}{g(1)}$ and thus

$$\begin{aligned}
\frac{x}{y}e^{-\frac{x}{y}+\frac{y}{z}} &= h(x)g(y) = \frac{\frac{x}{y}e^{y-x-\frac{1}{y}+\frac{1}{x}}}{g(1)h(1)} \\
\Rightarrow g(1)h(1) &= \exp\left(\frac{xy^2 - x^2y^2 - x + y + x^2 - y^2}{xy}\right) = \exp\left(\frac{(x-1)(y-1)(y-x)}{xy}\right)
\end{aligned}$$

But the RHS is not constant, which should be the case if the $\frac{f(1|1)}{f(1|1)} = h(1)g(1)$ holds.

Remark. The above method is a recipe to check functional compatibility, or check if a decomposition exists.

Example 36 (No joint density due to infinite integral)**.** eSuppose the $Y|X \sim N(x,1)$ and $X|Y \sim N(y,1)$, then there is no joint density since

$$\int_{\mathbb{R}} \frac{f(y|x)}{f(x|y)}\mathrm{d}x = \int_{\mathbb{R}} \frac{f(y|x)}{f(x|y)}\mathrm{d}y = \infty \ \forall x, y \in \mathbb{R}$$

2.4.3 Two Marginal Distributions (trivial)⇒Joint Distribution but NOT Unique.

Problem. Given densities $f_1(x), f_2(y)$, is there a joint density $f(x,y)$ such that f_1, f_2 are its marginals? If so, is it unique?

The answer is, $f(x,y)$ always exists, but is not unique.

Consider any function $h(x,y)$ such that $|h(x,y)| \leq 1$ and

$$\int_{\mathbb{R}} f_1(x)f_2(y)h(x,y)\mathrm{d}y = \int_{\mathbb{R}} f_1(x)f_2(y)h(x,y)\mathrm{d}x = 0\Big]$$

Then $f(x,y) = f_1(x)f_2(y)(1 + h(x,y))$ will be a joint density.

One such h can be constructed as: Let $S_{y_1}, S_{y_2} \in \sigma(Y)$ be disjoint sets which are assigned positive measure by $Y \sim f_2$ and let $S_{x_1}, S_{x_2} \in \sigma(X)$ be disjoint sets assigned positive measure by $X \sim f_1$. It is easy to confirm that

$$h(x,y) = \left(\mathbb{I}_{S_{y_1}}(y) - \mathbb{I}_{S_{v_2}}(y)\right)\left(\mathbb{I}_{S_{x_1}}(x) - \mathbb{I}_{S_{x_2}}(x)\right)$$

works.

3 Multivariate Normal Distribution (TBC)

3.1 Bivariate Normal Distributions

3.2 Distribution of Quadratic Forms

3.3 The t and F distribution

3.4 Pearson χ^2 Test

4 Order Statistics

4.1 Definitions and Distributions

4.1.1 Definitions of k-th order statistic, range, quantiles.

Definition 37 (Order statistics)**.** Given a sample from i.i.d. r.v. $X_1, \cdots, X_n$ we define the k-th order statistic for $1 \leq k \leq n$ to be the k-th largest value in the sample, denoted by $X_{(k)}$. Specifically, $X_{(1)} = \min X_i$ and $X_{(n)} = \max X_i$.

Definition 38 (Range, Median, Quantiles)**.** More definitions:

- Sample Range $R = X_{(n)} - X_{(1)}$
- Sample Median
$$M = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & n \text{ is odd} \\ \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & n \text{ is even} \end{cases}$$
- The $(100p)$-th percentile is a value above which there are approximately $(1-p)n$ of the observations
- Lower quantile is the 25-th percentile
- Upper quantile is the 75-th percentile
- Interquartile range is $UQ - LQ$.

4.1.2 Discrete: Joint PMF $n! \prod_{i \in I} \frac{1}{n_i!} p_i$

Theorem 39 (Joint distribution of order statistics (discrete))**.** *Suppose $X_1, \cdots, X_n \overset{iid}{\sim} X$ be discrete r.v. with PMF f_X supported on a sample space $\mathcal{X} = \{x_i\}_{i \in I}$. For all $i \in I$ let $p_i := f_X(x_i)$. Then the joint PMF of order statistics $\left(X_{(1)}, \ldots, X_{(n)}\right)$ is*

$$f_{\left(X_{(1)},\ldots,X_{(n)}\right)}\left(x_{k_1}, \ldots, x_{k_n}\right) = \begin{cases} n! \prod_{i \in I} \frac{1}{n_i!} p_i & x_{k_1} \leq \ldots \leq x_{k_n} \\ 0 & \textit{otherwise} \end{cases}$$

where $n_i := \sum_{j=1}^{n} \mathbb{I}_{\{k_j = i\}}$ is the number of time that x_i occurs in the sample.

Remark. When $|x_i|$ is finite for all i, this distribution is a multinomial distribution.

Proof. Let $\mathcal{S} \subseteq \mathcal{X}^n$ be all $v \in \mathcal{X}^n$ such that there exists a permutation $\sigma \in S_n$, s.t. $\sigma(v) = (x_{k_1}, \cdots, x_{k_m})$, then

$$\begin{aligned} f_{\left(X_{(1)},\ldots,X_{(n)}\right)}\left(x_{k_1}, \ldots, x_{k_n}\right) &= \sum_{v \in \mathcal{S}} f_{(X_1,\ldots,X_n)}(v) \mathbb{I}_{\left\{x_{k_1} \leq \ldots \leq x_{k_n}\right\}} \\ &= \sum_{v \in \mathcal{S}} \prod_{i=1}^{n} f_X\left(v_i\right) \mathbb{I}_{\left\{x_{k_1} \leq \ldots \leq x_{k_i}\right.} \\ &= \sum_{v \in S} \prod_{i=1}^{n} f_X\left(x_{k_i}\right) \mathbb{I}_{\left\{x_{k_1} \leq \ldots \leq x_{k_n}\right\}} \\ &= \sum_{v \in S} \prod_{j \in I} p_j^{n_j} \mathbb{I}_{\left\{x_{k_1} \leq \ldots \leq x_{k_n}\right\}} \\ &= |\mathcal{S}| \cdot \prod_{j \in I} p_j^{n_j} \mathbb{I}_{\left\{x_{k_1} \leq \ldots \leq x_{k_n}\right\}} \end{aligned}$$

It can be confirmed that $|\mathcal{S}| = \frac{n!}{\prod_{i \in I} n_i!}$ and thus we get the required result. □

4.1.3 Discrete: PDF $X_{(j)}$

Theorem 40 (PDF of $X_{(j)}$ (discrete)). *Let F_X be the corresponding CDF. Then for all $1 \leq j \leq n$ the CDF of $X_{(j)}$ is*

$$P_{X_{(j)}}(t) = \mathbb{P}\left(X_{(j)} \leq t\right) = \sum_{k=j}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) F_X(t)^k \left(1 - F_X(t)\right)^{n-k}$$

Proof. Let $Y = \Sigma_i \mathbb{I}\{x_i \leq t\}$, then $Y \sim Bin(n, F_X(t))$ and so

$$P(X_{(j)} \leq t) = P(Y \geq j)$$

which means "The j-th smallest number is less than t" $\Leftrightarrow$ "at least j number is less than t". The RHS can be easily calculate by the CDF of $Bin(n, F_X(t))$. □

4.1.4 Continuous: Joint PDF $n! \prod_{i=1}^{n} f_\theta(x_i) 1_{\{x_{k_1} \leq \ldots \leq x_{k_n}\}}(x_{k_1}, \ldots, x_{k_n})$

Theorem 41 (Joint distribution of order statistics (continuous)). *$X_1, \cdots, X_n \overset{iid}{\sim} X$ be discrete r.v. with PDF f_X. Then the joint PDF of order statistics $\left(X_{(1)}, \ldots, X_{(n)}\right)$ is*

$$\begin{aligned} f_{\left(X_{(1)}, \ldots, X_{(n)}\right)}(x_{k_1}, \ldots, x_{k_n}) &= \begin{cases} n! \Pi_{i=1}^{n} f(x_{k_i}) & x_{k_1} \leq \ldots \leq x_{k_n} \\ 0 & \textit{otherwise} \end{cases} \\ &= n! \prod_{i=1}^{n} f_\theta(x_i) 1_{\{x_{k_1} \leq \ldots \leq x_{k_n}\}}(x_{k_1}, \ldots, x_{k_n}) \end{aligned}$$

which is simply $n!$ times the joint distribution of $X_1, \cdots, X_n$, and a indicator variable.

4.1.5 Continuous: PDF of $X_{(j)}$ e.g. $F_{(1)}(x) = 1 - [1 - F(x)]^n, f_{(1)}(x) = n[1 - F(x)]^{n-1} f(x)$

Theorem 42 (PDF of $X_{(j)}$ (continuous)). *The CDF of $X_{(j)}$ is*

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

Proof. (See textbook for details) Suppose the i-th variable is $X_{(j)}$, and the realization is x. Then there are exactly $j-1$ variables smaller than x, and $n-j$ variables greater than x. This becomes a combinatorial problem. That is, from the remaining $n-1$ variables, we select $j-1$ to put on the left side of x, and the remaining go to the right hand side. So there are C_{n-1}^{j-1} number of combinations. For each combination, the probability is $[F_X(x)]^{j-1} [1 - F_X(x)]^{n-j}$. Note that there are n candidates of variables to be $X_{(j)}$, so there is a multiplier n.

In summary,

$$f_{X_{(j)}}(x) = \underbrace{\sum_{i=1}^{n}}_{n} \underbrace{P(X_i = x)}_{f_X(x)} \underbrace{P(j-1 \text{ variables from } n-1 \text{ variables are less than } x)}_{C_{n-1}^{j-1} [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j}}$$

□

Example 43 (PDF of $X_{(1)}$ and $X_{(n)}$ (continuous)). In particular, it is easy to show, by differentiation, that

$$F_{(1)}(x) = 1 - [1 - F(x)]^n, f_{(1)}(x) = n[1 - F(x)]^{n-1} f(x)$$

and

$$F_{(n)}(x) = [F(x)]^n, f_{(n)}(x) = n[F(x)]^{n-1} f(x)$$

4.1.6 Continuous: Joint PDF of $(X_{(i)}, X_{(j)})$ e.g. $(X_{(i)}, X_{(j)})$ and thus range, median

Theorem 44 (Joint PDF of $\left(X_{(i)}, X_{(j)}\right)$ (continuous))**.** *Given $1 \leq i < j \leq n$, the joint PDF of $(X_{(i)}, X_{(j)})$ is*

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

Proof. For the second part, suppose there are two cutoffs u, v on the real line $\mathbb{R}$, and $u < v$. Denote the numbers of counts of variables in the three intervals as U, V and $n - U - V$. Then $U = \Sigma_{k=1}^{n} \mathbb{I}\left\{x_i \leq u\right\}$, $V = \Sigma_{k=1}^{n} \mathbb{I}\left\{u \leq x_i \leq v\right\}$. Following this notation we have

$$\begin{aligned}
F_{\left(x_{(t)}, x_{(t)}\right)}(u,v) &= \mathbb{P}\left(X_{(i)} \leq u, X_{(j)} \leq v\right) \\
&\equiv \mathbb{P}(U \geq i, U + V \geq j, V < j) + \mathbb{P}(U \geq i, U + V \geq j, U \geq j) \\
&= \mathbb{P}(i \leq U \leq j-1, j - U \leq V \leq n - U) + \mathbb{P}(U \geq j) \\
&= \sum_{k=1}^{j-1} \sum_{l=j-k}^{n-k} \mathbb{P}(U = k, V = l) + \mathbb{P}(U \geq j)
\end{aligned}$$

Note (U, V) follow a multinomial distribution, so

$$\mathbb{P}(U = k, V = l) = \left(\begin{array}{c} n \\ k, l, n-k, l \end{array} \right) F_X(u)^k \left(F_X(v) - F_X(u)\right)^l \left(1 - F_X(v)\right)^{n-k-l}$$

And additionally U follows a binomial distribution and so we have

$$\mathbb{P}(U \geq j) = \sum_{m=j}^{n} \left(\begin{array}{c} n \\ m \end{array} \right) F(u)^m (1 - F(u))^{n-m}$$

By differentiating and simplifying these two quantities, the result follows. □

Example 45 (Joint distribution of $\left(X_{(1)}, X_{(n)}\right)$)**.** From the theorem, we have

$$F_{X(1),X(n))}\left(x_1, x_2\right) = F_X^n\left(x_2\right) - \left[F_X\left(x_2\right) - F_X\left(\min\left\{x_1, x_2\right\}\right)\right]^n$$

from

$$\begin{aligned}
F_{X(1),X(n))}\left(x_1, x_2\right) &= P(X_{(n)} < x_2, X_{(1)} < \min(x_1, x_2)) \\
&= P(X_{(n)} < x_2) - P(X_{(n)} < x_2, X_{(1)} > \min(x_1, x_2)) \\
&= F_X^n(x_2) - P(\min(x_1, x_2) < X_i < x_2)
\end{aligned}$$

Example 46 (Joint distribution of range and median)**.** Recall that $R = X_{(n)} - X_{(1)}$ and let $V = \frac{X_{(1)} + X_{(n)}}{2}$. By applying this theorem we can find the joint distribution of (R, V). Note that

$$\left(X_{(1)}, X_{(n)}\right) = (V - \frac{R}{2}, V + \frac{R}{2})$$

Let $g : \mathbb{R}^2 \to \mathbb{R}^2$ be a function such that $g(r, v) = (v - \frac{r}{2}, v + \frac{r}{2})$. Thus $g(R, V) = \left(X_{(1)}, X_{(n)}\right)$. By changing of variables,

$$\begin{aligned}
f_{(R,V)}(r,v) &= \left|\det \mathcal{J}_g(r,v)\right| f_{\left(X_{(1)}, X_{(n)}\right)}\left(v - \frac{r}{2}, v + \frac{r}{2}\right) \\
&= \underbrace{\det \left(\begin{array}{cc} 1 & -\frac{1}{2} \\ 1 & +\frac{1}{2} \end{array} \right)}_{1} f_{\left(X_{(1)}, X_{(n)}\right)}\left(v - \frac{r}{2}, v + \frac{r}{2}\right) \\
&= n(n-1)\left(F_X\left(v + \frac{r}{2}\right) - F_X\left(v - \frac{r}{2}\right)\right)^{n-2} f_X\left(v + \frac{r}{2}\right) f_X\left(v - \frac{r}{2}\right)
\end{aligned}$$

4.2 Asymptotic Properties

4.2.1 For Uniform Sample Quantiles $\sqrt{n}(U_{(\lceil np_1 \rceil)} - p_1) \xrightarrow{\mathcal{D}} N(0, p_1(1-p_1))$

Theorem 47 (Asymptotic distribution of uniform sample quantiles)**.** *Let $U_1, \ldots, U_n \overset{iid}{\sim} U(0,1)$ and let $U_{(1)} \leq \ldots \leq U_{(n)}$ be the corresponding order-statistics. Suppose $n \to \infty$ and let k_1, k_2 be functions of n such that $k_1(n) \overset{n\to\infty}{\to} \infty$ and $k_2(n) \overset{n\to\infty}{\to} \infty$. Furthermore, suppose that $\sqrt{n}\left(\frac{k_1(n)}{n} - p_1\right) \overset{n\to\infty}{\longrightarrow} 0$ and $\sqrt{n}\left(\frac{k_2(n)}{n} - p_2\right) \overset{n\to\infty}{\longrightarrow} 0$ where $0 < p_1 < p_2 < 1$, such as the ceiling function $\lceil np_1 \rceil$. Then*

$$\sqrt{n}\begin{pmatrix} U_{(k_1(n))} - p_1 \\ U_{(k_2(n))} - p_2 \end{pmatrix} \xrightarrow{\mathcal{D}} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} p_1(1-p_1) & p_1(1-p_2) \\ p_1(1-p_2) & p_2(1-p_2) \end{pmatrix}\right)$$

Proof. The proof appears in Ferguson chapter 13 and relies on the following lemma. □

Lemma. *Let $Y_1, \ldots, Y_{n+1} \overset{iid}{\sim} Exp(1)$ bet $n+1$ waiting times and let $S_j = \sum_{i=1}^{j} Y_i$ for $i = 1, 2, \cdots, n+1$ be the cumulative waiting time. Then*

$$\left(\frac{S_1}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}}\right) \sim \left(U_{(1)}, \ldots, U_{(n)}\right)$$

In particular,

$$\left(\frac{S_{k_1}}{S_{n+1}}, \ldots, \frac{S_{k_2}}{S_{n+1}}\right) \sim \left(U_{(k_1)}, \ldots, U_{(k_2)}\right)$$

Proof. The joint density of $Y := (Y_1, \ldots, Y_{n+1})$ is

$$f_Y(y_1, \ldots, y_{n+1}) = \begin{cases} \exp\left(-\sum_{i=1}^{n+1} y_i\right) & y_i > 0 \forall 1 \leq i \leq n+1 \\ 0 & \text{otherwise} \end{cases}$$

Consider a transformations $(Y_1, \ldots, Y_{n+1}) \to \left(\frac{S_1}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}}, S_{n+1}\right) := (Z_1, \ldots, Z_n, S_{n+1})$. The Jacobian determinant can be shown to be S_{n+1}^n, (easier to prove by considering an intermediate $(S_1, \ldots, S_{n+1})$) Thus,

$$f(z_1, \ldots, z_n, S_{n+1})(z_1, \ldots z_n, s_{n+1}) = s_{n+1}^n \exp(-s_{n+1}) \mathbb{I}_{\{0<z_1<\ldots<z_n<1\}}(z_1, \ldots, z_n)$$

This joint density can be factored into a function of s_{n+1} and a function of $(z_1, \ldots, z_n)$. As a result, we can conclude $(z_1, \ldots, z_n) \perp\!\!\!\perp S_{n+1}$. Furthermore, since $S_{n+1} \sim \Gamma(n+1, 1)$, its density is

$$f_{S_{n+1}}(s_{n+1}) = \begin{cases} \frac{s_{n+1}^n e^{-s_{n+1}}}{n!} & s_{n+1} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Hence from the factorization $f_{Z,S_{n+1}} = f_Z(z) f_{S_{n+1}}(s_{n+1})$ we have

$$f_Z(z) = \begin{cases} n! & 0 < z_1 < \ldots < z_n < 1 \\ 0 & \text{otherwise} \end{cases}$$

which is the joint density of the order statistics of n i.i.d. $U(0,1)$ r.v. □

Theorem 48 (Independency of $U_{(k)}$ and $1 - U_{(k)}$)**.** *(Ferguson Chapter 15) For fixed values of $0 < p_1 < \cdots < p_n < 1$ and fixed k,*

(a)

$$n\left(U_{(1)}, \ldots, U_{(k)}\right) \xrightarrow{\mathcal{D}} (S_1, \ldots, S_k)$$

(b) *The three vectors,*
(1) lower k order statistics

$$n\left(U_{(1)}, \ldots, U_{(k)}\right)$$

(2) upper k (flipped) order statistics

$$n\left(1-U_{(n)},\ldots,1-U_{(n-k+1)}\right)$$

(3) k quantiles

$$\sqrt{n}\left(U_{(np_1)}-p_1,\ldots,U_{(np_k)}-p_k\right)$$

are asymptotically independent, with distribution of (1) and (2) vectors as in (a), and (3) in previous theorem.

Example 49 (Asymptotic distributions of uniform range and midrange)**.** Since the range $R_n = U_{(n)} - U_{(1)}$ we have

$$n(1-R_n) = n(1-U_{(n)}) + nU_{(1)} \xrightarrow{\mathcal{D}} Y_1 + Y_2 \sim \Gamma(2,1)$$

where by the above theorem $Y_1, Y_2 \overset{iid}{\sim} Exp(1)$.
For the midrange $M_n = \frac{1}{2}\left(U_{(n)} - U_{(1)}\right)$ we have

$$n\left(M_n - \frac{1}{2}\right) = \frac{1}{2}\left(nU_{(1)} - n\left(1-U_{(n)}\right)\right) \xrightarrow{\mathcal{D}} \frac{1}{2}\left(Y_1 - Y_2\right)$$

The RHS has the Laplace (double exponential) distribution with density $f(z) = e^{-2|z|}$.

4.2.2 For General Sample Quantiles $\sqrt{n}(X_{(\lceil np_1\rceil)} - x_{p_1}) \xrightarrow{\mathcal{D}} N(0, \frac{p_1(1-p_1)}{f^2(x_{p_1})})$

Theorem 50 (Asymptotic distribution of general sample quantiles)**.** *Let $X_1, \cdots, X_n \overset{iid}{\sim} F$ with density $f(x)$ which is continuous and positive in a neighborhood of x_{p_1}, x_{p_2} where $x_{p_i} = F^{-1}(p_i), 0 < p_i < 1$ and $0 < p_1 < p_2 < 1$ (thus F is strictly monotone in these neighborhoods). Then*

$$\sqrt{n}\begin{pmatrix} X_{(\lceil np_1\rceil)} - x_{p_1} \\ X_{(\lceil np_2\rceil)} - x_{p_2} \end{pmatrix} \xrightarrow{\mathcal{D}} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{p_1(1-p_1)}{f^2(x_{p_1})} & \frac{p_1(1-p_1)}{f(x_{p_1})f(x_{p_2})} \\ \frac{p_1(1-p_1)}{f(x_{p_1})f(x_{p_2})} & \frac{p_2(1-p_2)}{f^2(x_{p_2})} \end{pmatrix}\right)$$

Proof. Note $X_{(\lceil np_i\rceil)} = F^{-1}\left(U_{(\lceil np_i\rceil)}\right)$, then we can define a transformation $g: \mathbb{R}^2 \to \mathbb{R}^2$ such that $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \xrightarrow{g} \begin{pmatrix} F^{-1}(y_1) \\ F^{-1}(y_2) \end{pmatrix}$, then

$$\begin{pmatrix} U_{(\lceil np_1\rceil)} \\ U_{(\lceil np_2\rceil)} \end{pmatrix} \overset{g}{\mapsto} \begin{pmatrix} X_{(\lceil np_1\rceil)} \\ X_{(\lceil np_2\rceil)} \end{pmatrix}, \text{ and } \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \overset{g}{\longleftrightarrow} \begin{pmatrix} x_{p_1} \\ x_{p_2} \end{pmatrix}$$

The Jacobian matrix of this transformation is

$$\dot{g}(y_1, y_2) = \begin{bmatrix} \frac{1}{f(F^{-1}(y_1))} & 0 \\ 0 & \frac{1}{f(F^{-1}(y_2))} \end{bmatrix} \Longrightarrow \dot{g}(p_1, p_2) = \begin{bmatrix} \frac{1}{f(x_{p_1})} & 0 \\ 0 & \frac{1}{f(x_{p_2})} \end{bmatrix}$$

Thus, apply Cramer's Theorem to the previous theorem, we have

$$\begin{aligned}
\sqrt{n}\begin{pmatrix} X_{(\lceil np_1\rceil)} - x_{p_1} \\ X_{(\lceil np_2\rceil)} - x_{p_2} \end{pmatrix} &= \sqrt{n}\left(g\begin{pmatrix} U_{(\lceil np_1\rceil)} \\ U_{(\lceil np_2\rceil)} \end{pmatrix} - g\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}\right) \\
&\xrightarrow{\mathcal{D}} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \dot{g}(p_1,p_2)\begin{pmatrix} p_1(1-p_1) & p_1(1-p_2) \\ p_1(1-p_2) & p_2(1-p_2) \end{pmatrix}\dot{g}(p_1,p_2)^\top\right) \\
&= \begin{pmatrix} \frac{p_1(1-p_1)}{f^2(x_{p_1})} & \frac{p_1(1-p_1)}{f(x_{p_1})f(x_{p_2})} \\ \frac{p_1(1-p_1)}{f(x_{p_1})f(x_{p_2})} & \frac{p_2(1-p_2)}{f^2(x_{p_2})} \end{pmatrix}
\end{aligned}$$

□

Remark. This theorem only gives asymptotic distribution for $X_{(\lceil np_i \rceil)}$ where p_i is fixed. For instance, the median $X_{(n/2)}$ and the lower quantile $X_{(n/4)}$. It does not specify the asymptotic distribution of $X_{(i)}$ for $i \in 1, 2, \cdots, n$, since if $\lceil np_i \rceil = i$ then p_i is not fixed. As an alternative, we can use Continuous: PDF of $X_{(j)}$ e.g. $F_{(1)}(x) = 1 - [1 - F(x)]^n, f_{(1)}(x) = n[1 - F(x)]^{n-1} f(x)$ and infer what happens to their CDFs as $n \to \infty$. An example is given below. In particular, for $X_{(n)}$, called extreme order statistic, its asymptotic distribution, if exists, have special properties.

Example 51 (Infer asymptotic distribution of Cauchy interquartile range)**.** For the Cauchy distribution $C(\mu, \sigma)$ with density

$$f(x) = \frac{1}{\pi\sigma} \frac{1}{1 + [(x-\mu)/\sigma]^2}$$

It has median μ, first quartile $x_{1/4} = \mu - \sigma$ and third quartile $x_{3/4} = \mu + \sigma$. For the sample median,

$$\sqrt{n}(m_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\pi^2\sigma^2}{4}\right)$$

To find the asymptotic distribution of the semi-interquartile range $\left(X_{(n/4)} - X_{(3n/4)}\right)/2$, we first find the asymptotic joint distribution of $X_{(n/4)}$ and $X_{(3n/4)}$. By the theorem,

$$\sqrt{n} \begin{bmatrix} X_{(n/4)} - (\mu - \sigma) \\ X_{(3n/4)} - (\mu + \sigma) \end{bmatrix} \xrightarrow{\mathcal{L}} N\left(0, \pi^2\sigma^2 \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}\right)$$

Hence

$$\sqrt{n}\left[\frac{X_{(3n/4)} - X_{(n/4)}}{2} - \sigma\right] \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \pi^2\sigma^2/4\right)$$

Example 52 (Infer asymptotic distribution of $X_{(1)}$)**.** Suppose $f(x) = \theta x^{-2}, 0 < \theta \leq x < \infty$. Then the CDF of $X_{(1)}$ is $F_{(1)}(t) = 1 - \left[\frac{\theta}{t}\right]^n$. Thus,

$$\begin{aligned} P(n(X_{(1)} - \theta) \leq t) &= P(X_{(1)} \leq \frac{t}{n} + \theta) \\ &= 1 - \left[\frac{\frac{t}{n} + \theta}{\theta}\right]^{-n} \\ &\to 1 - e^{-x/\theta} \end{aligned}$$

So

$$n(X_{(1)} - \theta) \xrightarrow{\mathcal{D}} Exp(\theta)$$

4.3 Extremal Distributions

4.3.1 Definition of Extremal Distributions

Definition 53 (Extremal distributions)**.** A CDF $G(x)$ is said to be extremal if it is non-degenerate and if there is some continuous distribution F and sequences $\{a_n\}, \{b_n\}, b_n > 0$ such that

$$[F(b_n x + a_n)]^n \xrightarrow{n\to\infty} G(x), \text{ pointwise.}$$

Theorem 54 (Extremal distributions have three types)**.** *All extremal distributions are one of the three following types up to change of location and scale.*

1. $G_{1,r}(x) = e^{-x^{-r}} \mathbb{I}_{\{x>0\}}$ for some $r > 0$
2. $G_{2,r}(x) = \begin{cases} e^{-(-x)^r} & x < 0 \\ 1 & x \geq 0 \end{cases}$ for some $r > 0$
3. $G_3(x) = e^{-e^{-x}}$

Remark. If $Y \sim Exp(1)$ then for $r > 0$, $Y^{-\frac{1}{r}} \sim G_{1,r}, -Y^{\frac{1}{r}} \sim G_{2,r}$ and $-\log(Y) \sim G_3$.

4.3.2 Extreme Order Statistic $X_{(n)}$

Previously we found the CDF and PDF of $X_{(n)}$ but they are too complicated. Now we are interested in the asymptotic distribution of $X_{(n)}$. Asymptotic distributions of $X_{(n)}$ may not exist. If it exists, it must be one of the above three types, up to location and scale. See Ferguson Chapter 14 for details.

Part II
Parametric Inference

5 Statistics

5.0.1 Definition: A form of data reduction or data summary. A RV.

Definition 55 (Statistic)**.** A statistic $T(\mathbf{X})$ is a form of data reduction or data summary. It is is a random variable (or vector) $\mathbf{Y} = T(\mathbf{X}_1, \cdots, \mathbf{X}_n)$. It is a function whose domain includes the sample space of the random vector $(\mathbf{X}_1, \cdots, \mathbf{X}_n)$. It may be real-valued or vector-valued. More formally, it may be defined as a mapping from one experiment to another experiment:

$$(\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta}) \to (\mathcal{T}, \mathcal{C}, \{Q_\theta\}_{\theta\in\Theta})$$

Example 56 (Statistics)**.** Sum, average, min, max, 2nd small, 2nd large, range, midrange.

Remark. Let $\mathcal{T} = \{t : t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\mathbf{x})$. Define the partition sets induced by $T(\mathbf{x})$ as $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$, i.e. partition $\mathcal{X}$ into parts $A_1, A_2, \cdots$, and for all $\mathbf{x}$ in the same part A_t, $T(\mathbf{x})$ has the same value t.

5.1 Sufficient Statistics

5.1.1 Definition: conditional distribution $f(\mathbf{x}|T(\mathbf{x}))$ is free of θ

A sufficient statistics captures all the information about θ in the sample. Experiment 1 who knows only $T(\mathbf{X}) = T(\mathbf{x})$, has just as much information about θ as does Experiment 2 who knows the entire sample $\mathbf{X} = \mathbf{x}$.

Definition 57 (Sufficient statistic)**.** $T(\mathbf{X})$ is a sufficient statistic for θ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ, i.e. $f(\mathbf{x}|T(\mathbf{x}))$ is free of θ.

Remark. Since $\{\mathbf{X} = \mathbf{x}\}$ is a subset of $\{T(\mathbf{X}) = T(\mathbf{x})\}$, we have

$$P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) = \frac{P_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = T(\mathbf{x}))}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))} = \frac{P_\theta(\mathbf{X} = \mathbf{x})}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))}$$

Thus, is equivalent to check whether the *ratio of the densities* of $\mathbf{X}$ and $T(\mathbf{X})$ is free of θ.

Example 58 (Binomial sufficient statistics)**.** By finding the conditional distribution, we see $T(\mathbf{X}) = X_1 + \cdots + X_n$ is a sufficient statistic for θ. The interpretation is: the total number of 1s in the Bernoulli sample contains all the information about θ that is in the data. Knowing other features of the data provides no additional information.

Example 59 (Normal sufficient statistics for μ when σ^2 is known)**.** If σ^2, is known $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ.

5.1.2 Characterization: Exists a factorization $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}),\ \forall x, \theta$

Sometimes it is hard to verify whether $T(\mathbf{X})$ is a sufficient statistic using the above definition if we do know now the distribution of it. Alternatively, we can use the following characterization theorem.

Theorem 60 (Fisher-Neyman Factorization)**.** *A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(\mathbf{x})$ such that*

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}),\ \forall x \in \mathcal{X}, \theta \in \Theta$$

Proof. ($\Rightarrow$) By the definition of sufficiency, $P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$ is free of θ,

$$\begin{aligned} f(\mathbf{x}|\theta) &= P_\theta(\mathbf{X} = \mathbf{x}) \\ &= P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\ &= P_\theta(T(\mathbf{X}) = T(\mathbf{x}))P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) \\ &= g(T(\mathbf{x})|\theta)h(\mathbf{x}) \end{aligned}$$

($\Leftarrow$) Let $q(T(\mathbf{x})|\theta)$ be the PMF of $T(\mathbf{X})$. Define $A_{T(\mathbf{x})} = \{\mathbf{y} : T(\mathbf{y}) = T(\mathbf{x})\}$ for some fixed $\mathbf{x}$. Note that

$$q(T(\mathbf{x})|\theta) = \Sigma_{A_{T(\mathbf{x})}} g(T(\mathbf{y})|\theta)h(\mathbf{y})$$

We have

$$\begin{aligned} \frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{q(T(\mathbf{x})|\theta)} \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\Sigma_{A_{T(\mathbf{x})}} g(T(\mathbf{y})|\theta)h(\mathbf{y})} \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{g(T(\mathbf{x})|\theta)\Sigma_{A_{T(\mathbf{x})}} h(\mathbf{y})} \\ &= \frac{h(\mathbf{x})}{\Sigma_{A_{T(\mathbf{x})}h(\mathbf{y})}} \end{aligned}$$

Thus the ratio does not depend on θ. □

Remark. It may happen that $h(\mathbf{x}) = 1$.

Remark. If $\{\mathbf{x} : f(\mathbf{x}|\theta) = 0\}$ depends on θ, better use indicator function . See the example below.

Example 61 (Uniform sufficient statistics)**.** Let $X_i \overset{iid}{\sim} U(0, \theta)$ then

$$f(x_1, \ldots, x_n|\theta) = \frac{1}{\theta^n}\mathbb{I}_{\{0<x_{(1)}<x_{(n)}<\theta\}} = \underbrace{\frac{1}{\theta^n}\mathbb{I}_{\{X_{(n)}<\theta\}}}_{g\left(x_{(n)},\theta\right)}\underbrace{\mathbb{I}_{\{0<x_{(1)}\}}}_{h(x)}$$

which implies that $X_{(n)}$ is a sufficient statistic for θ.

Example 62 (Normal sufficient statistics)**.** When both μ and σ^2 are unknown, it can be seen that the joint density

$$f\left(\mathbf{x}|\mu, \sigma^2\right) = g\left(T_1(\mathbf{x}), T_2(\mathbf{x})|\mu, \sigma^2\right) h(\mathbf{x})$$

where $h(\mathbf{x}) = 1$ and

$$\begin{aligned} g(\mathbf{t}|\theta) &= g\left(t_1, t_2|\mu, \sigma^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\left(n\left(t_1 - \mu\right)^2 + (n-1)t_2\right) / \left(2\sigma^2\right)\right) \end{aligned}$$

So $T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = \left(\bar{X}, S^2\right)$ is a sufficient statistic for (μ, σ^2).

Remark. This may not holds for other distribution. If one calculates only $\bar{X}$ and S^2 and totally ignores the rest of the data, he is placing strong faith in the normal model assumption.

Remark. When both μ and σ^2 are unknown, we need S^2 in addition to $\bar{X}$ even if we are only interested in μ.

Example 63 (Exponential family sufficient statistics)**.** Given a k-parameter exponential family with density

$$f(x|\boldsymbol{\theta}) = h(x)\underbrace{c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right)}_{g(T(\mathbf{x})|\boldsymbol{\theta})}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d), d \leq k$, then

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_j), \ldots, \sum_{i=1}^{n} t_k(X_j) \right)$$

is a sufficient statistic for $\boldsymbol{\theta}$.

5.1.3 Existence: always exists, e.g. whole sample, order statistics.

There always exists a sufficient statistic - the whole sample $T(\mathbf{X}) = \mathbf{X}$. By Fisher-Neyman Factorization Theorem, simply let $g(T(\mathbf{x})|\theta) = f(\mathbf{x}|\theta)$ and $h(\mathbf{x}) = 1$, then $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$. In addition, the order statistic is also a sufficient statistic.

Example 64 (Order sufficient statistics)**.** If we are unable to specify any more information about the PDF f_θ (e.g. nonparametric estimation), then the order statistic $T(\mathbf{X}) = (X_{(1)}, \cdots, X_{(n)})$ is sufficient for θ, which is easy to see from the density

$$f\left(x_{(1)}, \ldots, X_{(n)}\right)(x_{i_1}, \ldots, x_{i_n}) = n! \prod_{j=1}^{n} f_\theta\left(x_{i_j}\right) 1_{\left\{x_{i_1} \leq \ldots \leq x_{i_n}\right\}}(x_{i_1}, \ldots, x_{i_n})$$

We can also verify this by checking the conditional distribution

$$\begin{aligned} f_{\left(X_1, \ldots, X_n | X_{(1)}, \ldots, X_{(n)}\right)}(x_1, \ldots, x_n) &= \frac{f_{(X_1, \ldots, X_n)}(x_1, \ldots, x_n)}{f_{\left(X_{(1)}, \ldots, X_{(n)}\right)}(x_{i_1}, \ldots, x_{i_n})} \\ &= \frac{\prod_{j=1}^{n} f_\theta\left(x_{i_j}\right)}{n! \prod_{j=1}^{n} f_\theta\left(x_{i_j}\right) \mathbb{I}_{\left\{x_{i_1} \leq \ldots \leq x_{i_n}\right\}}(x_{i_1}, \ldots, x_{i_n})} \\ &= \frac{1}{n!} \end{aligned}$$

The value $\frac{1}{n!}$ is very intuitive by thinking of permutation.

5.1.4 Non-uniqueness: Any one-to-one function of a stuff. stat. is a stuff. stat.

It is always true that the complete sample $\mathbf{X}$ is a sufficient statistic if we let $g(\mathbf{x}|\theta) = f(\mathbf{x}|\theta)$ and $h(\mathbf{x}) = 1$. In general, sufficient statistic is not unique.

Proposition 65. *Any one-to-one function of a sufficient statistic is a sufficient statistic.*

Proof. Define a one-to-one function $T^*(\mathbf{x}) = r(T(\mathbf{x}))$ with inverse r^{-1}, then

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}) = \underbrace{g\left(r^{-1}\left(T^*(\mathbf{x})\right)|\theta\right)}_{g^*(T^*(\mathbf{x}))} h(\mathbf{x})$$

□

5.1.5 Sufficient Principle: If $T(\mathbf{x}) = T(\mathbf{y})$, then same inference of θ

If $T(\mathbf{X})$ is a sufficient statistics for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value $T(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference about θ (point estimate, confidence interval, hypothesis testing) should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

5.2 Minimal sufficient statistic:

Since sufficient statistics are not unique, we can measure their extent of data reduction.

5.2.1 Definition: a suff. stat. which is a function of any other suff. stat.

Definition 66 (Minimal Sufficient Statistic)**.** A sufficient statistic $T(\mathbf{X})$ is called a minimal sufficient statistic if, for any other sufficient statistic $U(\mathbf{X})$

$$T(\mathbf{x}) \text{ is a function of } U(\mathbf{x})$$

That is, $U(\mathbf{x}) = U(\mathbf{y}) \Rightarrow T(\mathbf{x}) = T(\mathbf{y})$.

Remark. Recall the understanding of T as a partition in Definition of Statistic. The above implies that T provides the *coarsest* possible partition, or greatest reduction of $\mathcal{X}$ while still being sufficient. One example of "coarser partition" is, for sample points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$, and two statistic $T(\mathbf{X})$ and $U(\mathbf{X})$,

$$U(\mathbf{x}_1) = U(\mathbf{x}_2) = u_1, U(\mathbf{x}_3) = U(\mathbf{x}_4) = u_2 \neq u_1$$

and

$$T(\mathbf{x}_i) = t \ \forall i = 1, 2, 3, 4.$$

5.2.2 Existence: under weak conditions

Minimal sufficient statistics can exist under weak conditions (Theory of Point of estimation page 37). In this course, MSS always exists.

5.2.3 Characterization: If $T(\mathbf{x}) = T(\mathbf{y}) \Leftrightarrow \frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)}$ is free of θ and $\Theta_{\mathbf{x}} = \Theta_{\mathbf{y}}$

Using the above definition is impractical, but we can use the characterization theorem below.

Theorem 67 (Lehmann-Scheffe for MSS)**.** *For two sample points* $\mathbf{x}$ *and* $\mathbf{y}$*, if*

$$T(\mathbf{x}) = T(\mathbf{y}) \Leftrightarrow \begin{cases} \Theta_{\mathbf{x}} = \Theta_{\mathbf{y}} \\ \frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} \text{ is a constant as a function of } \theta \end{cases}$$

then $T(\mathbf{X})$ *is a minimal sufficient statistic for* θ*, where* $\Theta_{\mathbf{x}} := \{\theta | f(\mathbf{x}|\theta) > 0\}$.

Proof. We first prove sufficiency. For a sample point $\mathbf{x}$, consider another sample point $\mathbf{x}_t$ which is in the same partition as $\mathbf{x}$, i.e. $T(\mathbf{x}) = T(\mathbf{x}_t) = t$, then by assumption ($\Rightarrow$)

$$f(\mathbf{x}|\theta) = \underbrace{f(\mathbf{x}_t|\theta)}_{g(t|\theta)} \underbrace{\frac{f(\mathbf{x}|\theta)}{f(\mathbf{x}_t|\theta)}}_{h(\mathbf{x})}$$

Then we prove minimality, by the following lemma. □

Lemma. *Assume* $\Theta_{\mathbf{x}} \neq \emptyset \ \forall \boldsymbol{x} \in \mathcal{X}$*. If* T *is sufficient and* $T(\mathbf{x}) = T(\mathbf{y})$ *then* $\Theta_{\mathbf{x}} = \Theta_{\mathbf{y}}$

Proof. For any other sufficient statistic $U(\mathbf{X})$ and two sample points $\mathbf{x}$ and $\mathbf{y}$ such that $U(\mathbf{x}) = U(\mathbf{y})$, then by the lemma,

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g^*(U(\mathbf{x})|\theta)h^*(\mathbf{x})}{g^*(U(\mathbf{y})|\theta)h^*(\mathbf{y})} = \frac{h^*(\mathbf{x})}{h^*(\mathbf{y})}$$

Since the ratio does not depend on θ, the assumption ($\Leftarrow$) implies that $T(\mathbf{x}) = T(\mathbf{y})$. Thus, $U(\mathbf{x}) = U(\mathbf{y}) \Rightarrow T(\mathbf{x}) = T(\mathbf{y})$ and hence T is a function of U. □

Example 68 (Normal minimal sufficient statistics)**.** We can check the ratio of the densities of two sample points

$$\begin{aligned}\frac{f\left(\mathbf{x}|\mu,\sigma^2\right)}{f\left(\mathbf{y}|\mu,\sigma^2\right)} &= \frac{\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\left[n(\bar{x}-\mu)^2+(n-1)s_{\mathbf{x}}^2\right]/\left(2\sigma^2\right)\right)}{\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\left[n(\bar{y}-\mu)^2+(n-1)s_{\mathbf{y}}^2\right]/\left(2\sigma^2\right)\right)} \\ &= \exp\left(\left[-n\left(\bar{x}^2-\bar{y}^2\right)+2n\mu(\bar{x}-\bar{y})-(n-1)\left(s_{\mathbf{x}}^2-s_{\mathbf{y}}^2\right)\right]/\left(2\sigma^2\right)\right)\end{aligned}$$

The ratio will be constant as a function of μ and σ^2 if and only if $\bar{x}-\bar{y}$ and $s_{\mathbf{x}}^2-s_{\mathbf{y}}^2$. Thus, $(\bar{X},S^2)$ is a minimal sufficient statistic for (μ,σ^2).

Example 69 ($U(\theta,\theta+1)$ minimal sufficient statistics)**.** Suppose $X_i \overset{iid}{\sim} U(\theta,\theta+1)$ then the joint PDF of $\mathbf{X}$ can be written as

$$f(\mathbf{x}|\theta)=\begin{cases}1 & \max_i x_i-1<\theta<\min_i x_i\\ 0 & \text{otherwise}\end{cases}$$

For two sample points $\mathbf{x}$ and $\mathbf{y}$,

$$\Theta_{\mathbf{x}}=\Theta_{\mathbf{y}} \Leftrightarrow \max_i x_i=\max_i y_i, \min_i x_i=\min_i y_i$$

And the ratio equals 1 if the above holds. Thus, $T(\mathbf{X})=\left(X_{(1)},X_{(n)}\right)$ is a minimal sufficient statistics for θ.

Remark. In the above example, $dim(\mathcal{T})=2>dim(\Theta)=1$. Actually there is no definite relation.

5.2.4 Non-uniqueness: any one-to-one function of a MSS is a MSS.

By the definition, it is easy to see any one-to-one function of a minimal sufficient statistics is a minimal sufficient statistics. One-to-one maintains the *coarsest* possible reduction.

In the above examples, $U(\mathbf{X})=\left(\Sigma_{i=1}^n X_i,\Sigma_{i=1}^n X_i^2\right)$ is a minimal sufficient statistics for (μ,σ^2), and $H(\mathbf{X})=\left(X_{(n)}-X_{(1)},\left(X_{(n)}+X_{(1)}\right)/2\right)$ (i.e. range and midrange) is a minimal sufficient statistics for θ.

5.3 Ancillary Statistics

5.3.1 Definition: Its distribution is free of θ

Definition 70 (Ancillary Statistics)**.** $S(\mathbf{X})$ is ancillary statistics for θ if its distribution does not depend on θ.

An ancillary statistic alone contains no information about θ. A common ancillary statistic for μ in normal distribution is $\frac{(n-1)S^2}{\sigma^2}$, if σ^2 is known.

Example 71 (Location family ancillary statistics)**.** For a location family, suppose $X_i \overset{iid}{\sim} F(x-\theta)$. Let $Z_i = X_i - \theta \sim F(x)$. Then $X_{(n)} - X_{(1)} = Z_{(n)} - Z_{(1)}$ which is free of θ. Thus in Example: $U(\theta, \theta+1)$ minimal sufficient statistics, $R = X_{(n)} - X_{(1)}$ is an ancillary statistic. More generally, any function of $(Y_1, \cdots, Y_{n-1})$ where $Y_i = X_{(n)} - X_{(i)}$ does not depend on θ.

Example 72 (Scale family ancillary statistics)**.** Similarly, Suppose $X_i \overset{iid}{\sim} F(x/\theta)$ and define $Z_i = \theta X_i \sim F(x)$. Then any function of the $n-1$ values $\left(\frac{X_1}{X_n}, \ldots, \frac{X_{n-1}}{X_n}\right)$, such as

$$\frac{X_1 + \cdots + X_n}{X_n} = \frac{X_1}{X_n} + \cdots + \frac{X_{n-1}}{X_n} + 1$$

and

$$\left(\frac{X_{(1)}}{X_{(n)}}, \ldots, \frac{X_{(n-1)}}{X_{(n)}}\right)$$

will be an ancillary statistic for θ, since the joint CDF of $\left(\frac{X_1}{X_n}, \ldots, \frac{X_{n-1}}{X_n}\right)$ is the same as that of $\left(\frac{Z_1}{Z_n}, \ldots, \frac{Z_{n-1}}{Z_n}\right)$

$$\begin{aligned} F(y_1, \ldots, y_{n-1}|\theta) &= P_\theta\left(X_1/X_n \le y_1, \ldots, X_{n-1}/X_n \le y_{n-1}\right) \\ &= P_\theta\left(\theta Z_1/(\theta Z_n) \le y_1, \ldots, \theta Z_{n-1}/(\theta Z_n) \le y_{n-1}\right) \\ &= P_\theta\left(Z_1/Z_n \le y_1, \ldots, Z_{n-1}/Z_n \le y_{n-1}\right) \end{aligned}$$

and the last probability does not depend on θ.

Remark. This implies that if $Z_1, Z_2 \overset{iid}{\sim} N(0,1) \Rightarrow \frac{Z_1}{Z_2} \sim Cauchy(0,1)$, then $X_1, X_2 \overset{iid}{\sim} N(0, \sigma^2) \Rightarrow \frac{X_1}{X_2} \sim Cauchy(0,1)$.

5.3.2 Existence (always)

Ancillary statistics always exist. A trivial case is $T =$ constant.

5.3.3 Non-uniqueness: any function of it is also ancillary.

Any function of an ancillary statistic is also ancillary.

5.3.4 Relation: minimal sufficient $\perp\!\!\!\perp$ or $\not\perp\!\!\!\perp$ ancillary statistics

In most cases, a minimal sufficient statistic is independent of any ancillary statistic.

In Example: $U(\theta, \theta+1)$ minimal sufficient statistics, $R = X_{(n)} - X_{(1)}$ is a ancillary statistic and is a part of a minimal sufficient statistic $T = (R, \frac{X_{(n)} + X_{(1)}}{2})$, and thus they are not independent. It alone does not provides any information about θ, but its value increased out knowledge about the *precision* of an estimate of θ (if R is close to 1 we can be very certain about the location of θ and if it's close to 0 we are very uncertain).

When will the independence hold? This brings the definition of complete statistics.

5.4 Complete Statistics

5.4.1 Definition: $\mathbb{E}_{P_\theta}[g(T)] = 0\ \forall\theta \in \Theta \Longrightarrow P_\theta(g(T) = 0) = 1\ \forall\theta \in \Theta$

Definition 73 (Complete Statistic)**.** A statistic $T : (\mathcal{X}, \mathcal{A}) \to (\mathcal{T}, \mathcal{C})$ is said to be complete if for *any* measurable function g (where the only unknown variable is T),

$$\mathbb{E}_{P_\theta}[g(T)] = 0\ \forall\theta \in \Theta \Longrightarrow P_\theta(g(T) = 0) = 1\ \forall\theta \in \Theta$$

The RHS means $g(T)$ must be degenerate w.p.1 at 0.

Remark. Typically, the way to prove completeness is to show $g(t) = 0$ for all t and all θ by using some reasoning of the summation (discrete T) or integral (continuous T) on the LHS. Below are two examples showing that sufficient statistics for Bernoulli and Uniform distribution are complete.

Example 74 (Binomial complete sufficient statistic)**.** Suppose T has a $Bin(n, p)$ distribution, $0 < p < 1$. Let g be a function such that $E_p[g(T)] = 0$, Then

$$\begin{aligned} 0 = E_p[g(T)] &= \sum_{t=0}^{n} g(t) \left(\begin{array}{c} n \\ t \end{array} \right) p^t (1-p)^{n-t} \\ &= (1-p)^n \sum_{t=0}^{n} g(t) \left(\begin{array}{c} n \\ t \end{array} \right) \left(\frac{p}{1-p} \right)^t \end{aligned}$$

If we require the above equality holds for all p, since $(1-p)^n > 0$, then it must be that

$$0 = \sum_{t=0}^{n} g(t) \left(\begin{array}{c} n \\ t \end{array} \right) \underbrace{\left(\frac{p}{1-p} \right)^t}_{r} = \sum_{t=0}^{n} g(t) \left(\begin{array}{c} n \\ t \end{array} \right) r^t$$

So this is a polynomial of degree n in r with coefficients $g(t) \left(\begin{array}{c} n \\ t \end{array} \right)$. If we require the above equality holds for all r, then *each* coefficient must 0, which implies $g(t) = 0$ for $t = 0, 1, 2, \cdots, n$. This yields that $P_p(g(T) = 0) = 1$ for all p, the desired conclusion. Hence T is a complete statistic.

Example 75 (Uniform complete sufficient statistic)**.** As shown in Example: Uniform sufficient statistics, a uniform sufficient statistic for $X_i \overset{iid}{\sim} U(0, \theta)$ is $T(\mathbf{X}) = X_{(n)}$, and its PDF is

$$f(t|\theta) = \begin{cases} nt^{n-1}\theta^{-n} & 0 < t < \theta \\ 0 & \text{otherwise} \end{cases}$$

Then $\mathbb{E}_{P_\theta}[g(T)] = 0\ \forall\theta \in \Theta$ implies that $\mathbb{E}_{P_\theta}[g(T)]$ is constant as a function of θ, and thus its derivative w.r.t. θ is 0. Thus,

$$\begin{aligned} 0 = \frac{d}{d\theta} \mathrm{E}_\theta g(T) &= \frac{d}{d\theta} \int_0^\theta g(t) n t^{n-1} \theta^{-n} dt \\ &= \left(\theta^{-n}\right) \frac{d}{d\theta} \int_0^\theta g(t) n t^{n-1} dt + \left(\frac{d}{d\theta} \theta^{-n} \right) \int_0^\theta g(t) n t^{n-1} dt \quad \text{(product rule)} \\ &= \theta^{-n} g(\theta) n \theta^{n-1} + \left(\frac{d}{d\theta} \theta^{-n} \right) \theta^n \mathrm{E}_\theta g(T) \\ &= \theta^{-1} g(\theta) n \end{aligned}$$

Thus $g(\theta) = 0$ for all $\theta > 0$. Hence, $g(t) = 0$ for all t since $0 < t < \theta$. Therefore, $P_p(g(T) = 0) = 1$.

Example 76 (Location exponential complete sufficient statistic)**.** Suppose $f(x|\theta) = e^{-(x-\theta)}, \theta < x < \infty, -\infty < \theta < \infty$. Given a sample $\mathbf{x}$, the parameter space is $\Theta_x = \left\{\theta : \theta > x_{(1)}\right\}$. Note that $\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = 1$ is free of θ. Hence, $T(\mathbf{X}) = X_{(1)}$ is a minimal sufficient statistics, and it is easy to show that $f(t|\theta) = ne^{-(t-\theta)n}, t > \theta$. Now we prove it is complete by mimicking the proof above.

$$
\begin{aligned}
0 = \frac{d}{d\theta}\mathrm{E}_\theta g(T) &= \frac{d}{d\theta}\int_\theta^\infty g(t)ne^{-(t-\theta)n}dt \\
&= \left(e^{n\theta}\right)\frac{d}{d\theta}\int_\theta^\infty g(t)ne^{-tn}dt + \left(\frac{d}{d\theta}e^{n\theta}\right)\int_\theta^\infty g(t)ne^{-tn}dt \qquad \text{(product rule)} \\
&= e^{n\theta}\left(-g(\theta)ne^{-\theta n}\right) + e^{n\theta}\int_\theta^\infty g(t)ne^{-tn}dt \\
&= -g(\theta)n + \mathrm{E}_\theta g(T)
\end{aligned}
$$

Thus $g(\theta) = 0$ for all $\theta \in \mathbb{R}$. Hence, $g(t) = 0$ for all t. Therefore, $P_p(g(T) = 0) = 1$.

Example 77 (Exponential family complete sufficient statistic)**.** Given a k-parameter *full* exponential family with density

$$
f(x|\boldsymbol{\theta}) = h(x)\underbrace{c(\boldsymbol{\theta})\exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right)}_{g(T(\mathbf{x})|\boldsymbol{\theta})}
$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$, d=k, then

$$
T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_j), \ldots, \sum_{i=1}^{n} t_k(X_j)\right)
$$

is a complete (and sufficient statistic) for $\boldsymbol{\theta}$.

For instance, for $N(\mu, \sigma^2)$, $t_1(x) = x^2$ and $t_2(x) = x$. Thus a complete statistic is $T(X) = \left(\sum X, \sum X^2\right)$. A one-to-one function of this is $T_2(X) = \left(\bar{X}, \sum X^2 - n\bar{X}^2\right) = \left(\bar{X}, S^2\right)$. So e

Remark. Another way of saying "full exponential family" is that the parameter space Θ contains an open set in $\mathbb{R}^k$. For instance, for the distribution $N(\theta, \theta^2)$, the parameter space (θ, θ^2), consisting of only the points on a parabola, does not contain a two-dimensional open set.

5.4.2 Existence

It always exists for exponential family. For other distributions, need to prove case by case.

5.4.3 Non-Uniqueness: any function of it is also complete

If a complete statistic T exists, any function of T is also a complete statistics.

Proof. Suppose T is a complete statistic, by definition □

$$
\mathbb{E}_{P_\theta}[g(T)] = 0 \ \forall\theta \in \Theta \Longrightarrow P_\theta(g(T) = 0) = 1 \ \forall\theta \in \Theta
$$

Consider a statistic $\tilde{T}$ to be a function of T, i.e. $\tilde{T} = h(T)$, then $g(\tilde{T}) = g(h(T)) = g \circ h(T)$. Thus, for any function g,

$$
\mathbb{E}_{P_\theta}[g(\tilde{T})] = 0 \Leftrightarrow \mathbb{E}_{P_\theta}[g \circ h(T)] = 0 \Longrightarrow P_\theta(g \circ h(T) = 0) = 1 \Leftrightarrow P_\theta(g(\tilde{T}) = 0) = 1
$$

Hence $\tilde{T}$ is also a complete statistic.

5.4.4 Relation: complete and sufficient ⇒ minimal

Theorem 78 (Complete and sufficient ⇒ minimal)**.** *If a minimal sufficient statistic exists, then any complete and sufficient statistic is also a minimal sufficient statistic.*

Remark. In Casella&Berger the condition of sufficiency is inaccurately omitted. We need this condition to ensure $D(T)$ depends only on T. Otherwise, it becomes $D(T, \tilde{T})$.

Proof. Let T be a complete and sufficient statistic and let $\tilde{T}$ be a minimal sufficient statistic. It suffices to prove T is a function of $\tilde{T}$.

Since $\tilde{T}$ is an MSS, it is a function of T. We can write $\tilde{T} = g(T)$. Then define a function D on $\mathcal{T}$ by

$$D(T) = T - E[T|\tilde{T} = g(T)]$$

By sufficiency of $\tilde{T}$, $E[T|\tilde{T}]$ does not depend on θ

By the law of total expectation, $E[D(T)] = 0$ for all θ. Thus, by the completeness of T,

$$P(T - E[T|\tilde{T}] = 0) = 1 \ \forall \ \theta$$

Define a function $h(\tilde{T}) = E[T|\tilde{T}]$, then the above implies $P(T = h(\tilde{T})) = 1$ for all θ. Thus, T is a function of $\tilde{T}$. □

5.4.5 Relation: complete and (minimal) sufficient ⫫ ancillary

Theorem 79 (Basu's)**.** *If $T(\mathbf{X})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic.*

Proof. We give the proof for discrete distributions. Let $T(\mathbf{X})$ be a complete and minimal sufficient statistic, and $S(\mathbf{X})$ be a ancillary statistic. Then it suffices to show that

$$P(S(\mathbf{X}) = s|T(\mathbf{X}) = t) = P(S(\mathbf{X}) = s), \ \forall t \in \mathcal{T}$$

Note that we can write $P(S(\mathbf{X}) = s)$ in two ways. First, use the conditional probability,

$$P(S(\mathbf{X}) = s) = \sum_{t \in \mathcal{T}} P(S(\mathbf{X}) = s|T(\mathbf{X}) = t) P_\theta(T(\mathbf{X}) = t)$$

where $P(S(\mathbf{X}) = s|T(\mathbf{X}) = t) = P(\mathbf{X} \in \{\mathbf{x} : S(\mathbf{x}) = s\}|T(\mathbf{X}) = t)$ does not depend on θ by the sufficiency of $T(\mathbf{X})$.

Second, since $\sum_{t \in \mathcal{T}} P_\theta(T(\mathbf{X}) = t) = 1$, we can write

$$P(S(\mathbf{X}) = s) = P(S(\mathbf{X}) = s) \sum_{t \in \mathcal{T}} P_\theta(T(\mathbf{X}) = t)$$

If we define a statistic

$$g(t) = P(S(\mathbf{X}) = s|T(\mathbf{X}) = t) - P(S(\mathbf{X}) = s)$$

then subtracting the above two equations gives

$$0 = \sum_{t \in \mathcal{T}} g(t) P_\theta(T(\mathbf{X}) = t) = \mathrm{E}_\theta g(T), \ \forall \theta$$

By the definition of complete statistic, we have $g(t) = 0$ for all t, as desired. □

Basus's Theorem gives a way to deduce the independence of two statistic without ever finding their joint distribution.

Example 80 (Using Basus's Theorem to find expectation)**.** For $X_i \overset{iid}{\sim} Exp(\theta)$, we are interested in computing the expected value of

$$g(\mathbf{X}) = \frac{X_n}{X_1 + \cdots + X_n}$$

First note that exponential distributions form a scale parameter family.

By Example: Scale family ancillary statistics, $g(\mathbf{X})$ is a ancillary statistic.

Then note that exponential distributions also form an exponential family with $t(x) = x$. Thus by Example: Exponential family sufficient statistics and Example: Exponential family complete sufficient statistic, the statistic

$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

is a complete and sufficient statistic (also minimal, which could easily be verified) for θ.
Thus, $g(\mathbf{X}) \perp\!\!\!\perp T(\mathbf{X})$ and hence $E_\theta[g(\mathbf{X})T(\mathbf{X})] = E_\theta[g(\mathbf{X})]E_\theta[T(\mathbf{X})]$. It follows that

$$\begin{aligned} E_\theta[g(\mathbf{X})] &= \frac{E_\theta[g(\mathbf{X})T(\mathbf{X})]}{E_\theta[T(\mathbf{X})]} \\ &= \frac{E_\theta(X_n)}{\sum E_\theta(X_i)} \\ &= \frac{1}{n} \end{aligned}$$

Example 81 (Using Basus's Theorem to show $\bar{X} \perp\!\!\!\perp S^2$ in Normal)**.** Suppose $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, then if we fixed σ^2, by Example: Normal sufficient statistics for μ when σ^2 is known, $\bar{X}$ is a sufficient statistic for μ. Furthermore, by Example: Exponential family complete sufficient statistic, $\bar{X}$ is complete since we can write $t(x) = \frac{1}{n}x$.

Note that $\frac{n-1}{\sigma^2}S^2 \chi^2_{n-1}$ which does not depend on μ. Thus, it is ancillary for μ.

By Basu's Theorem, $\bar{X} \perp\!\!\!\perp \frac{n-1}{\sigma^2}S^2$ for any μ and fixed σ^2.

Since σ^2 is arbitrary, we have for any μ and σ^2.

6 Likelihood Based Inference

6.1 Likelihood Function and Properties

6.1.1 Definition: $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$

Definition 82 (Likelihood Function)**.** Let $f(\mathbf{x}|\theta)$ bee the joint density of a sample $\mathbf{x}$ (might not be IID). Then given we observe $\mathbf{X} = \mathbf{x}$, the likelihood function of θ is defined by $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$ as a function of θ.

6.1.2 The Likelihood Principle: same inference regarding θ if $L(\theta|\mathbf{x}) = c(x, y)L(\theta|\mathbf{y})$

Likelihood Principle: Any inference regarding θ based on samples $\mathbf{X} = \mathbf{x}$ and $\mathbf{X} = \mathbf{y}$ should be the same if $L(\theta|\mathbf{x}) \propto L(\theta|\mathbf{y})$. That is, if there exists a constant $c(x_1, x_2) \neq 0$ independent of θ such that $L(\theta|\mathbf{x}) = c(x, y)L(\theta|\mathbf{y})$ for all $\theta \in \Theta$.

Example 83 (Likelihood Principle)**.** Suppose there are two independent samples $\mathbf{x}$ and $\mathbf{y}$ from an exponential family. Suppose both samples have the same MSS, then $L(\theta|\mathbf{x}) = c(x, y)L(\theta|\mathbf{y})$.

The likelihood principle implies that the ratio $\frac{L(\theta_1|x)}{L(\theta_2|x)}$ is an appropriate measure of the plausibility of θ_1 and θ_2.

6.1.3 Property: Moments $\mathbb{E}_\theta[\nabla_\theta \ell] = 0, \mathbb{E}_\theta\left[\nabla^2_\theta \ell\right] + \mathbb{E}_\theta\left[\nabla_\theta \ell \left(\nabla_\theta \ell\right)^\top\right] = 0$

Proposition 84. *Under mild regularity condition (Dominated Convergence Theorem to interchange the order of differentiation and integration), the log-likelihood derivatives satisfy the moment identities*

- $\mathbb{E}_\theta[\nabla_\theta \ell] = 0$ where $\ell(\theta|x) = \log L(\theta|x)$
- $\mathbb{E}_\theta\left[\nabla^2_\theta \ell\right] = -\operatorname{Var}_\theta\left(\nabla_\theta \ell\right)$ where $\operatorname{Var}_\theta\left(\nabla_\theta \ell\right) = \mathbb{E}_\theta\left[\nabla_\theta \ell \left(\nabla_\theta \ell\right)^\top\right]$

where the i, j-th entry of the $k \times k$ matrix $\nabla^2 l$ is $\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}$.

Note that for exponential families, the regularity condition always holds.

Proof. Consider the equality $\int_{\mathbb{R}} f(x|\theta)dx = 1$, assume we can interchange the order of differentiation and integration when differentiating the LHS w.r.t. θ, then

$$\begin{aligned}
0 &= \frac{\partial}{\partial \theta_i} 1 \\
&= \frac{\partial}{\partial \theta_i} \int f(x|\theta)dx \\
&= \int \frac{\partial}{\partial \theta_i} f(x|\theta)dx \\
&= \int \frac{\frac{\partial f(x|\theta)}{\partial \theta_i}}{f(x|\theta)} f(x|\theta)dx \\
&= \int \frac{\partial \log f(x|\theta)}{\partial \theta_i} f(x|\theta)dx \\
&= \mathbb{E}_\theta\left[\frac{\partial \ell}{\partial \theta_i}\right]
\end{aligned}$$

Likewise, for the equality $\frac{\partial^2}{\partial \theta_i \partial \theta_j} \int f(x|\theta)dx = 0$, by interchanging we get

$$
\begin{aligned}
0 &= \int \frac{\partial^2}{\partial\theta_i \partial\theta_j} f(x|\theta)dx \\
&= \int \left(\frac{\frac{\partial^2 f(x|\theta)}{\partial\theta_i\partial\theta_j} f(x\theta) - \frac{\partial f}{\partial\theta_i}\frac{\partial f}{\partial\theta_j}}{f(x|\theta)^2} + \frac{\frac{\partial f(x|\theta)}{\partial\theta_i}}{f(x|\theta)} \cdot \frac{\partial f(x|\theta)}{\partial\theta_j} \right) f(x|\theta)dx \\
&= \int \left(\left[\nabla_\theta^2 \ell\right]_{ij} + \left(\nabla_\theta \ell\right)_{ij} \left(\nabla_\theta \ell\right)_j \right) f(x|\theta)dx \\
&= \mathbb{E}_\theta \left[\left[\nabla_\theta^2 \ell\right]_{ij} \right] + \mathbb{E}_\theta \left[\nabla_\theta \ell_i \cdot \nabla_\theta \ell_j \right]
\end{aligned}
$$

The above holds for all i, j and thus

$$\mathbb{E}_\theta \left[\nabla_\theta^2 \ell\right] = -\mathbb{E}_\theta \left[\nabla_\theta \ell \left(\nabla_\theta \ell\right)^\top \right]$$

□

Remark. Unless the support of $f(\mathbf{x}|\theta)$ is the same for all values of θ or at least for all values in an open neighborhood of the true parameter value we should *not* generally expect interchangeability of the differentiation and integration. In which case these equalities will not hold. ? $U(0, \theta)$.

6.1.4 Fisher Information $I(\theta) := \mathbb{E}_\theta \left[\nabla_\theta \ell \left(\nabla_\theta \ell\right)^\top \right]$

Definition 85 (Fisher Information)**.** The quantity $I(\theta) := \mathbb{E}_\theta \left[\nabla_\theta \ell \left(\nabla_\theta \ell\right)^\top \right]$ is called the Fisher-Information Matrix.

Under *regularity* conditions the moment equalities imply that

$$I(\theta) = \mathrm{Var}_\theta \left(\nabla_\theta \ell\right) = -\mathbb{E}_\theta \left[\nabla_\theta^2 \ell\right]$$

This quantity is used in Theorem Cramer-Rao Lower-Bound. As the information gets bigger, we have more information about θ, and we have a smaller lower bound on the variance of the unbiased estimator.

Theorem 86 (Fisher Information for i.i.d sample)**.** *If the data is an i.i.d. sample, then the sample quantities equal the sum of the quantity of single observations*

$$\ell(\theta|\mathbf{x}) = \sum_{i=1}^n \tilde{\ell}\left(\theta|x_i\right)$$

Proof. It can be seen that □

$$
\begin{aligned}
\ell(\theta|x) &= \log \left(\prod_{i=1}^n f\left(x_i|\theta\right) \right) \\
&= \sum_{i=1}^n \log f\left(x_i|\theta\right) \\
&= \sum_{i=1}^n \tilde{\ell}\left(\theta|x_i\right)
\end{aligned}
$$

where $\tilde{\ell}\left(\theta|x_i\right)$ is the log-likelihood for a single observation X_i.

The score function is

$$\begin{aligned} S(\theta|\mathbf{x}) &= \nabla_\theta \ell(\theta|x) \\ &= \nabla_\theta \left(\sum_{i=1}^{n} \tilde{\ell}\left(\theta|x_i\right) \right) \\ &= \sum_{i=1}^{n} \nabla_\theta \tilde{\ell}\left(\theta|x_i\right) \\ &= \sum_{i=1}^{n} \tilde{S}\left(\theta|x_i\right) \end{aligned}$$

And

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta \left[\nabla_\theta \ell \left(\nabla_\theta \ell \right)^\top \right] \\ &= \mathbb{E}_\theta \left[\left(\sum_{i=1}^{n} \nabla_\theta \tilde{\ell}\left(\theta|x_i\right) \right) \left(\sum_{j=1}^{n} \nabla_\theta \tilde{\ell}\left(\theta|x_j\right)^\top \right) \right] \\ &= \mathbb{E}_\theta \left[\sum_{i=1}^{n} \sum_{j=1}^{n} \nabla_\theta \tilde{\ell}\left(\theta|x_i\right) \left(\nabla_\theta \tilde{\ell}\left(\theta|x_j\right) \right)^\top \right] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}_\theta \left[\nabla_\theta \tilde{\ell}\left(\theta|x_i\right) \left(\nabla_\theta \tilde{\ell}\left(\theta|x_j\right) \right)^\top \right] \end{aligned}$$

By independence, for $i \neq j$,

$$\mathbb{E}_\theta \left[\nabla_\theta \tilde{\ell}\left(\theta|x_i\right) \left(\nabla_\theta \tilde{\ell}\left(\theta|x_j\right) \right)^\top \right] = \overbrace{\mathbb{E}_\theta \left[\nabla_\theta \tilde{\ell}\left(\theta|x_i\right) \right]}^{=0} \cdot \overbrace{\mathbb{E}_\theta \left[\nabla_\theta \tilde{\ell}\left(\theta|x_j\right) \right]}^{=0} = 0$$

Hence

$$I(\theta) = \sum_{i=1}^{n} \mathbb{E}_\theta \left[\left(\nabla \tilde{\ell}\left(\theta|x_i\right) \right) \left(\nabla_\theta \tilde{\ell}\left(\theta|x_i\right) \right)^\top \right] := nI_1(\theta)$$

where $I_1(\theta)$ is the Fisher-Information matrix for a single observation.

Example 87 (No Moments Function Identities for $U(0,\theta)$)**.** Suppose $X_i \sim U(0,\theta)$, then

$$\begin{aligned} L(\theta|x) &= \frac{1}{\theta} \mathbb{I}_{[0,\theta]}\left(x_i\right) \\ \implies \ell(\theta|x) &= -\log(\theta) \mathbb{I}_{[0,\theta]}\left(x_i\right) \\ \implies \frac{\partial \ell(\theta|x)}{\partial \theta} &= -\frac{1}{\theta} \mathbb{I}_{[0,\theta]}(x) \\ \implies \mathbb{E}_\theta \left[\frac{\partial \ell(\theta|x)}{\partial \theta} \right] &= -\frac{1}{\theta} \neq 0 \end{aligned}$$

So the regularity conditions don't hold (in particular this *always* happens when the support of X_i depends on θ). Furthermore,

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta\left[\left(\frac{\partial \ell(\theta|x)}{\partial \theta}\right)^2\right] \\ &= \mathbb{E}_\theta\left[\left(-\frac{1}{\theta}\mathbb{I}_{[0,\theta]}(x)\right)^2\right] \\ &= \frac{1}{\theta^2}\mathbb{E}_\theta\left[\mathbb{I}_{[0,\theta]}(x)\right] \\ &= \frac{1}{\theta^2} \end{aligned}$$

$$\begin{aligned} \mathrm{Var}_\theta\left(\frac{\partial \ell(\theta|x)}{\partial \theta}\right) &= \mathbb{E}_\theta\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] - \left(\mathbb{E}_\theta\left[\frac{\partial \ell(\theta|x)}{\partial \theta}\right]\right)^2 \\ &= \frac{1}{\theta^2} - \left(-\frac{1}{\theta}\right)^2 \\ &= 0 \\ &\neq I(\theta) \end{aligned}$$

And lastly it can also be seen that

$$\frac{\partial^2 \ell(\theta|x)}{\partial \theta^2} = \frac{1}{\theta^2}\mathbb{I}_{[0,\theta]}(x) \Longrightarrow -\mathbb{E}_\theta\left[\frac{\partial^2 \ell(\theta|x)}{\partial \theta^2}\right] = -\frac{1}{\theta^2} \neq I(\theta)$$

So in this case none of the equalities hold.

6.1.5 Identifiability of a Parameter θ: $f_{\theta_1}(x) \equiv f_{\theta_2}(x)\ \forall x \Longrightarrow \theta_1 = \theta_2$

Definition 88 (Identifiable)**.** A parameter θ is said to be identifiable for the family $\{f_\theta(x), \theta \in \Theta\}$ if

$$f_{\theta_1}(x) \equiv f_{\theta_2}(x)\ \forall x \Longrightarrow \theta_1 = \theta_2$$

Note the converse is always true.

6.1.6 KL Information $K(\theta_0|\theta) = \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X|\theta_0)}{f(X|\theta)}\right)\right]$

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} f(x|\theta)$ and the true value of θ is θ_0, then

$$\begin{aligned} \frac{1}{n}\left[\ell_n(\theta_0|x) - \ell_n(\theta|x)\right] &= \frac{1}{n}\left(\sum_{i=1}^n \log f(x_i|\theta_0) - \sum_{i=1}^n \log f(x_i|\theta)\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^n \log\left(\frac{f(x_i|\theta_0)}{f(x_i|\theta)}\right)\right) \\ &\underset{\text{SLLN}}{\overset{a.s.}{\to}} \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X|\theta_0)}{f(X|\theta)}\right)\right] \\ &= \int \log\left(\frac{f(x|\theta_0)}{f(x|\theta)}\right) f(x|\theta_0)\, dx \\ &=: K(\theta_0|\theta) \text{ if the integral converges} \end{aligned}$$

where the convergence is contingent on the RHS integral being well defined. In the case where everything is well defined the quantity $K(\theta_0|\theta)$ is known as the Kullback-Leibler information in favor of θ_0 against θ, when θ_0 is true.

6.1.7 Shannon-Kolmogorov Information Inequality $K(\theta_0|\theta) \geq 0$, equal iff $f_\theta = f_{\theta_0}$

If $K(\theta_0|\theta)$ exists, then $K(\theta_0|\theta) \geq 0$ with equality iff $f(x|\theta_0) = f(x|\theta)$, $a.s.$. Assuming identifiability, we have $\theta = \theta_0$.

Proof. Jensen's inequality says that for any convex function g,

$$E[g(X)] \geq g(E(X))$$

with equality iff $P(X = E(X)) = 1$
Since $-\log x$ is a convex function, by Jensen's inequality we have

$$\begin{aligned}
K(\theta_0|\theta) &= \int -\log\left(\frac{f(x|\theta)}{f(x|\theta_0)}\right) f(x|\theta_0)\, dx \\
&= E_{\theta_0}\left[-\log\left(\frac{f(x|\theta)}{f(x|\theta_0)}\right)\right] \\
&\geq -\log\left(E_{\theta_0}\left[\frac{f(x|\theta)}{f(x|\theta_0)}\right]\right) \text{ by Jensen} \\
&= -\log\left(\int\left(\frac{f(x|\theta)}{f(x|\theta_0)}\right) f(x|\theta_0)\, dx\right) \\
&= -\log\left(\int f(x|\theta) dx\right) \\
&= 0
\end{aligned}$$

with equality iff $P\left(\frac{f(x|\theta)}{f(x|\theta_0)} = E_{\theta_0}\left[\frac{f(x|\theta)}{f(x|\theta_0)}\right]\right) = 1$.
Since

$$E_{\theta_0}\left[\frac{f(x|\theta)}{f(x|\theta_0)}\right] = \int\left(\frac{f(x|\theta)}{f(x|\theta_0)}\right) f(x|\theta_0)\, dx = 1$$

the equality condition is

$$\frac{f(x|\theta)}{f(x|\theta_0)} \overset{a.s.}{=} 1$$

If the condition holds, then

$$\frac{1}{n}\left[\ell_n(\theta_0|x) - \ell_n(\theta|x)\right] \overset{\text{a.s}}{\to} K(\theta_0|\theta) = 0$$

If θ is identifiable then the equality condition is

$$\frac{f(x|\theta)}{f(x|\theta_0)} \overset{a.s.}{=} 1 \Leftrightarrow \theta_0 \overset{a.s.}{=} \theta$$

□

So when one has identifiability eventually the likelihood function will be larger at θ_0 than at any other θ. In general though, we need additional conditions in order to show the MLE actually converges to θ_0 (Ferguson Ch.17).

6.2 Maximum Likelihood Estimator

6.2.1 Estimator and Estimate

Definition 89 (Point Estimator)**.** A point estimator is any function $W(X_1, \ldots, X_n)$ of a sample. Any statistic is a point estimator.

Remark. An estimator is a function of the sample while is a function of $X_1, \ldots, X_n$, while an *estimate* is the realized value of an estimator, which is a function of $x_1, \ldots, x_n$.

6.2.2 Definition of Maximum Likelihood Estimator $\hat{\theta}_{\text{MLE}} = \text{argmax}_{\theta \in \Theta} L(\theta|x)$

Definition 90 (Maximum Likelihood Estimator)**.** The maximum likelihood estimator for θ is defined by

$$\hat{\theta}_{\text{MLE}} = \text{argmax}_{\theta \in \Theta} L(\theta|x)$$

6.2.3 Existence: may not exist. If it exists, must be a function of a suff. stat.

The MLE $\hat{\theta}_{\text{MLE}}$ may not exists. See Example: Mixture Normal-Binomial: No MLE.

Note that by the characterization of sufficient statistics, $f(x|\theta) = g(T(x)|\theta)h(x)$, thus

$$\frac{\partial \log f(x|\theta)}{\partial \theta} = \frac{\partial \log g(T(x)|\theta)}{\partial \theta} = 0$$

The solution to the normal equation must involves X only through the quantity $T(x)$. So $\hat{\theta}_{\text{MLE}}$ must be a function depends only on a sufficient statistic $T(X)$. This provides a simple criteria to identify whether a statistic is an MLE or not.

6.2.4 Uniqueness: unique if L is concave

If exists it may not be unique. It is unique if L is concave.

6.2.5 Definitions of log-likelihood, score function and normal equation.

Definition 91 (Log-Likelihood)**.** If the likelihood is differentiable, we can take log and obtain the log-likelihood $\ell(\theta|x) = \log L(\theta|x) = \log f(x|\theta)$.

Definition 92 (Score Function)**.** The score function is the gradient of log-likelihood

$$S(\theta|x) = \nabla_\theta \ell(\theta|x)$$

Definition 93 (Normal Equations)**.** The normal equations are defined by the set of equations

$$S(\theta|x) = \underbrace{0}_{k \times 1}$$

Any solution to these equations is a stationary point of the likelihood.

6.2.6 Computation

In order to find maxima of the likelihood one needs to consider all stationary points in the interior of the parameter space *and* all points on the *boundary* of the parameter space. Sometimes boundary points can be excluded using algebraic results such as inequalities. Alternatively it's possible to use numerical search procedures. Finding MLEs is hard and sometimes it may not even be possible to find even a local maxima let alone global maxima of the likelihood.

Example 94 (Univariate Normal given $\sigma^2 = 1$ MLE)**.** Consider $X_i \overset{iid}{\sim} N(\theta, 1)$, then

$$L(\theta|x) = C \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right)$$

where C is a constant not depending on θ. Thus

$$\ell(\theta|x) = \log C - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2$$

and

$$S(\theta|x) = \sum_{i=1}^{n}(x_i - \theta)$$

So $\hat{\theta}_n = \bar{X}$ is a stationary point.
We can check the second derivative

$$\frac{\partial S(\theta|X)}{\partial \theta} = -n < 0$$

which implies L is concave. So $\hat{\theta}_n = \bar{X}$ is a global maximum.
For the boundary (why?),

$$\lim_{\theta\to\infty}\ell(\theta|x) = \lim_{\theta\to-\infty}\ell(\theta|x) = -\infty$$

Thus $\hat{\theta}_n = \bar{X}$ is indeed the MLE.
It can also be seen that

$$\sum_{i=1}^{n}(x_i-\theta)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\theta)^2 \geq \sum_{i=1}^{n}(x_i-\bar{x})^2$$

The quality holds iff $\theta = \bar{x}$.

Example 95 (Multivariate Normal MLE)**.** Consider $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} N_p(\mu, \Sigma)$ where Σ is p.d. and $n > \frac{1}{2}p + \frac{3}{2}$. It can be shown that the MLE is

$$\hat{\mu} = \bar{X} \text{ and } \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^\top$$

Example 96 (Uniform MLE)**.** Consider $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} U(0,\theta)$, then

$$L(\theta|\mathbf{x}) = \frac{1}{\theta^n}\mathbb{I}_{\{x_{(n)}\leq\theta\}}\mathbb{I}_{\{x_{(1)}\geq 0\}}$$

Note that when $\theta < x_{(n)}$, we have $L(\theta|\mathbf{x}) = 0$ and when $\theta > x_{(n)}$, we have

$$L(\theta|x) = \frac{1}{\theta^n} < \frac{1}{x_{(n)}^n} = L\left(x_{(n)}|x\right)$$

So $X_{(n)}$ is the MLE.

Example 97 (Mixture Normal-Binomial: No MLE)**.** Consider a sample $\mathbf{x}$ from the density

$$f(x|\theta) = \alpha\frac{1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right) + (1-\alpha)\frac{1}{\sigma_2}\phi\left(\frac{x-\mu_2}{\sigma_2}\right)$$

where $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)^\top$ is unknown, ϕ is the density for $N(0,1)$ and $\alpha \in \left(0, \frac{1}{2}\right)$ is known. It can be interpreted as choosing one from two normal distributions according to a Bernoulli experiment with probability α.

It turns out there is no MLE in this case. The likelihood is

$$L(\theta|x) = \prod_{i=1}^{n}\left(\frac{\alpha}{\sigma_1}\exp\left(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right) + \frac{(1-\alpha)}{\sigma_2}\exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right)\right)$$

It drifts off to infinity as you approach certain points on the boundary of the parameter space (likelihood is finite in the interior of the parameter space). For instance, if we set $\mu_1 = x_1$, then as $\sigma_1 \to 0$, it diverges to $+\infty$. Thus, there is no global maxima.

Some alternative approaches to tackle cases like this include

- Using discretization (Cox).
- Using largest local maximum (may be hard to find it numerically)

6.2.7 Property: Invariance. unique $\hat{\theta}_{MLE} \Rightarrow$ unique $\hat{\eta}_{MLE} = \tau(\hat{\theta}_{MLE})$

Definition 98 (Induced Likelihood)**.** Let $\eta = \tau(\theta)$ be some function (may not be one-to-one) of the parameter θ and define the induced likelihood of η as

Fact.

$$L^*(\eta|x) = \sup\{L(\theta|x)|\theta \ s.t \ \tau(\theta) = \eta\}$$

The value $\hat{\eta}$ that maximizes $L^*(\eta|x)$ is called the MLE of η.

Theorem 99 (Invariance property of MLE)**.** *If $\hat{\theta}$ is the unique MLE of θ then $\tau(\hat{\theta})$ is the unique MLE of $\eta = \tau(\theta)$, for any function τ. It also holds in multivariate case.*

Example 100 (Invariance property of MLE)**.** Suppose $X_i \overset{iid}{\sim} Ber(p)$, then clearly $\hat{p}_{MLE} = \bar{X}$. The natural parameter is $\theta = \ln \frac{p}{1-p}$. Then we have $\hat{\theta}_{MLE} = \ln \frac{\bar{X}}{1-\bar{X}}$.

6.2.8 Property: Asymptotic Normality $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{D}} N\left(0, I_1^{-1}(\theta_0)\right)$

Theorem 101 (Asymptotic Normality of MLE (1-d))**.** *Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$ where $\theta \in \Theta \subseteq \mathbb{R}$. Let θ_0 be the true value of θ. Suppose that*

1. Θ is an open set.
2. $\frac{\partial^3 \log f(x|\theta)}{\partial \theta^3}$ exists and is dominated by an integrable function, i.e. $\frac{\partial^3 \log f(x|\theta)}{\partial \theta^3} \leq H(x)$ for all x and θ, where $E_\theta[H(x)] \leq M$ for all θ.
3. $0 < I_1(\theta) := \mathbb{E}_\theta\left[\left(\frac{\partial \log f(x|\theta)}{\partial \theta}\right)^2\right] < \infty$
4. The support $S_\theta = \{x|f(x|\theta) > 0\}$ does not depend on θ for all θ.
5. There is identifiability of θ_0 for $f(x|\theta)$: $f_{\theta_1}(x) \equiv f_{\theta_2}(x) \ \forall x \Longrightarrow \theta_1 = \theta_2$

Then there exists a strongly consistent sequence $\hat{\theta}_n$, i.e. $\hat{\theta}_n \xrightarrow{\text{a.s}} \theta_0$, of roots of the normal equation $S_n(\theta) = \sum_1^n \frac{\partial \log f(x|\theta)}{\partial \theta} = 0$ (i.e. stationary points of $l_n(\theta|\mathbf{x})$) such that

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{D}} N\left(0, I_1^{-1}(\theta_0)\right)$$

Remark. All this theorem claims is that, under the conditions stated, there exists a consistent sequence of roots of the likelihood equation that is asymptotically normal, with $[I_1^{-1}(\theta_0)/n]$ as its variance. However, it does not claim that *all* sequences of roots (e.g. MLE) will be consistent and asymptotically normal. But if there is a *unique* root of the likelihood equation for every n (i.e. MLE), as in many applications, then this sequence of roots will be consistent and asymptotically normal.

Proof. Sketch of the proof

1. The existence of a sequence of roots $\left\{\hat{\theta}_n\right\}$ of $S_n(\theta|x)$.
2. This sequence $\left\{\hat{\theta}_n\right\}$ is strongly consistent: $\hat{\theta}_n \overset{\text{a.s}}{\to} \theta_0$

3. This sequence $\left\{\hat{\theta}_n\right\}$ is asymptotically normal.

4. The asymptotic variance of $\left\{\hat{\theta}_n\right\}$ is $\frac{1}{nI_1(\theta_0)} = \frac{1}{I_n(\theta_0)}$

For (1), note

$$\frac{1}{n}\ell_n(\theta|x) = \frac{1}{n}\sum_{i=1}^{n}\log f\left(x_i|\theta\right) \underset{SLLN}{\overset{a.s.}{\to}} E_{\theta_0}\left[\log f(x|\theta)\right] := Z(\theta)$$

and

$$\frac{1}{n}\ell_n\left(\theta_0|x\right) \underset{SLLN}{\overset{a.s.}{\to}} Z\left(\theta_0\right)$$

Recall K-L information, for any $\theta \neq \theta_0$

$$Z(\theta_0) - Z\left(\theta\right) = K\left(\theta_0|\theta\right) > 0$$

So we have for any $\epsilon > 0$

$$Z\left(\theta_0\right) - Z\left(\theta_0 \pm \varepsilon\right) > 0$$

Thus, except for a set of measure zero (or, the below holds w.p.1) there exists $N > 0$ such that for all $n \geq N$ we have

$$\max\left\{\ell_n\left(\theta_0 + \varepsilon|x\right), \ell_n\left(\theta_0 - \varepsilon|x\right)\right\} < \ell_n\left(\theta_0|x\right)$$

This implies that for all $\epsilon > 0$, $n \geq N$, there exists $\hat{\theta}_n \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$ that maximizes $\ell_n\left(\theta_0|x\right)$, which is the root of $S_n\left(\hat{\theta}_n|x\right) = 0$.

For (2), since $|\hat{\theta}_n - \theta_0| < \varepsilon$ for all $\varepsilon > 0$ and $n \geq N$, we can conclude that

$$\hat{\theta}_n \overset{\text{a.s}}{\to} \theta_0$$

For (3), Taylor expansion of $S_n\left(\theta|x\right)$ around $\theta = \theta_0$ we have that

$$\underbrace{S_n\left(\hat{\theta}_n|x\right)}_{0} - S_n\left(\theta_0|x\right) = S_n'\left(\theta_0|x\right)\left(\hat{\theta}_n - \theta_0\right) + \frac{1}{2}S_n''\left(\theta^*\right)\left(\hat{\theta}_n - \theta_0\right)^2$$

where $\theta^* = \lambda\theta_0 + (1-\lambda)\hat{\theta}_n$ for some $\lambda \in (0,1)$. Thus, dividing both side by $\sqrt{n}$

$$-\underbrace{\sqrt{n}\frac{S_n\left(\theta_0|x\right)}{n}}_{\underset{CLT}{\overset{D}{\to}} N(0,I_1(\theta_0))} = \left[\underbrace{\frac{S_n'\left(\theta_0|x\right)}{n}}_{\underset{SLLN}{\overset{a.s.}{\to}} -I_1(\theta_0)} + \frac{1}{2}\underbrace{\frac{S_n''(\theta^*|x)}{n}}_{\leq M}\underbrace{\left(\hat{\theta}_n - \theta_0\right)}_{\overset{\text{a.s}}{\to} 0}\right]\sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$$

Because

$$\frac{\sqrt{n}S_n\left(\theta_0|x\right)}{n} = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\ell\left(x_i|\theta_0\right)}{\partial\theta} - 0\right] \underset{CLT}{\overset{D}{\to}} N\left(0, \text{Var}\left(\frac{\partial\ell\left(x|\theta_0\right)}{\partial\theta}\right)\right) = N\left(0, I_1\left(\theta_0\right)\right)$$

and

$$\frac{S_n'\left(\theta_0|x\right)}{n} \underset{SLLN}{\overset{a.s.}{\to}} \mathbb{E}_{\theta_0}\left[\frac{\partial^2\log f\left(x|\theta_0\right)}{\partial\theta^2}\right] = -I_1\left(\theta_0\right)$$

and the third derivative is dominated by an integrable function

$$\left|\frac{S_n''\left(\theta^*\right)}{n}\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3\log f\left(x_i|\theta^*\right)}{\partial\theta^3}\right| \leq \frac{1}{n}\sum_{i=1}^{n}H\left(x_i\right) \overset{\text{a.s.}}{\to} \mathbb{E}_{\theta_0}[H(x)] \leq M$$

Combining these three into the equation we have

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{D}} N\left(0, \frac{I_1(\theta_0)}{I_1(\theta_0)^2}\right) = N\left(0, \frac{1}{I_1(\theta_0)}\right)$$

□

Example 102 (Asymptotic Normality of Bernoulli MLE)**.** For Bernoulli distribution, the Fisher information is

$$I_1(\theta) = \mathbb{E}_\theta\left[\frac{-X}{\theta^2} - \frac{(1-X)}{(1-\theta)^2}\right] = \frac{1}{\theta(1-\theta)}$$

Thus by the above theorem,

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathcal{D}} N\left(0, \theta(1-\theta)\right)$$

Example 103 (Non-Asymptotic Normality of Uniform MLE)**.** For $U(0,\theta)$, the support depends on θ so the above theorem does not apply. Instead, we can using the asymptotic of extreme order statistics to find the non-standard asymptotic of the MLE: $X_{(n)}$:

$$\frac{n\left(X_{(n)} - \theta\right)}{\theta} \xrightarrow{\mathcal{D}} -\xi; \xi \sim Exp(1)$$

Remark. $X_{(n)}$ converges to θ at a much faster rate than standard theory, and is not asymptotically normal.

6.3 Uniform Minimum Variance Unbiased Estimator (UMVUE)

6.3.1 Definition: $\mathbf{Var}_\theta(W^*) \leq \mathbf{Var}_\theta(W)$ and $E_\theta(W^*) = \theta$

Definition 104 (UMVUE)**.** An estimator W^* is said to be the UMVUE for θ if

- $E_\theta(W^*) = \theta$ (unbiased)
- $\mathrm{Var}_\theta(W^*) \leq \mathrm{Var}_\theta(W)$ for all other unbiased estimator W of θ.

In d-dimensional case, the second condition is

$$\mathrm{Cov}_\theta\left(W^*(X)\right) \preceq \mathrm{Cov}_\theta(W(X))$$

which is equivalent to

$$\mathrm{Var}_\theta\left(c^\top W^*(X)\right) \leq \mathrm{Var}_\theta\left(c^\top W(X)\right) \ \forall\, c \in \mathbb{R}^n$$

6.3.2 Characterization: a statistic uncorrelated with any unbiased estimator of zero

Theorem 105 (Necessary and Sufficient Condition for UMVUE)**.** *Given an experiment $\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta}\big)$ such that $\Theta \subseteq \mathbb{R}^n$. Let $W(X)$ be a 1-dimensional statistics such that $\mathbb{E}_\theta[W(X)] = \tau(\theta)$. Then W is an UMVUE of $\tau(\theta)$ iff it is uncorrelated with any unbiased estimator of zero.*

That is W is UMVUE of $\tau(\theta)$ iff

$$\mathrm{Cov}_\theta(W(X), T(X)) = 0$$

For any T (unbiased estimator of zero) such that

$$\mathbb{E}_\theta[T(X)] = 0 \text{ and } \mathrm{Var}_\theta(T) < \infty \text{ for all } \theta \in \Theta$$

Proof. ($\Rightarrow$ by contradiction) Let $\mathcal{U}_0$ be a set of unbiased estimators of zero, $\mathcal{U}_0 = \{U | \mathbb{E}_\theta[U] = 0, \mathrm{Var}_\theta(U) < \infty \forall\theta\}$. Suppose $U_0 \in \mathcal{U}_0$ such that

$$\mathrm{Cov}_{\theta_1}(W, U_0) \neq 0$$

for some $\theta = \theta_1$. Consider $W_1 = W - \beta U_0$ where $\beta = \frac{\mathrm{Cov}_{\theta_1}(W,U_0)}{\mathrm{Var}_{\theta_1}(U_0)}$. We are doing projection of W into space spanned by U_0.

Note that $\mathbb{E}_\theta[W_1] = \mathbb{E}_\theta[W]$, so W_1 is unbiased. Furthermore

$$\begin{aligned}
\mathrm{Var}_{\theta_1}(W_1) &= \mathrm{Var}_{\theta_1}(W) + \beta^2 \mathrm{Var}_{\theta_1}(U_0) - 2\beta \mathrm{Cov}_{\theta_1}(W, U_0) \\
&= \mathrm{Var}_{\theta_1}(W) - \frac{\mathrm{Cov}^2_{\theta_1}(W, U_0)}{\mathrm{Var}_{\theta_1}(U_0)} \\
&< \mathrm{Var}_{\theta_1}(W)
\end{aligned}$$

which is a contradiction to W being UMVUE. □

Remark. This is a recipe to construct a uniformly *better* unbiased estimator of $\tau(\theta)$ given an unbiased estimator W.

Proof. ($\Leftarrow$) Assume W is uncorrelated with every $U_0 \in \mathcal{U}_0$ and let

$$\mathcal{W} = \{W | \mathbb{E}_\theta[W] = \tau(\theta), \mathrm{Var}_\theta(W) < \infty \forall\theta \in \Theta\}$$

Clearly $W \in \mathcal{W}$. For any $W_1 \in \mathcal{W}$ we have $W - W_1 \in \mathcal{U}_0$. Then by assumption that $\mathrm{Cov}_\theta(W, W - W_1) = 0$ we have

$$\begin{aligned}
&\mathbb{E}_\theta\left[W\left(W-W_1\right)\right]=0\\
&\quad\implies \mathbb{E}_\theta\left[W^2\right]=\mathbb{E}_\theta\left[WW_1\right]\\
&\qquad\qquad\leq \sqrt{\mathbb{E}_\theta\left[W^2\right]}\sqrt{\mathbb{E}_\theta\left[W_1^2\right]}\\
&\implies \sqrt{\mathbb{E}_\theta\left[W^2\right]}\leq\sqrt{\mathbb{E}_\theta\left[W_1^2\right]}\\
&\quad\implies \mathrm{Var}_\theta(W)\leq \mathrm{Var}_\theta\left(W_1\right)\ \text{ since } \mathbb{E}_\theta[W]=\mathbb{E}_\theta[W_1]=\tau(\theta)
\end{aligned}$$

Since this is true for any unbiased W_1 we have the W is UMVUE. □

6.3.3 Uniqueness. Must be a function of complete sufficient stat.

Theorem 106 (Uniqueness of UMVUE)**.** *If a UMVUE exists it is unique (up to a.s equality).*

Proof. Suppose W_1, W_2 are two different UMVUE of θ. Then $W_1 - W_2$ is an unbiased estimator of zero and so by the previous theorem W_1, W_2 are both uncorrelated with $W_1 - W_2$ which implies that

$$\mathbb{E}_\theta\left[W_1\left(W_1-W_2\right)\right]=0=\mathbb{E}_\theta\left[W_2\left(W_1-W_2\right)\right]$$

$$\begin{aligned}
&\implies \mathbb{E}_\theta\left[\left(W_1-W_2\right)^2\right]=0\\
&\quad\implies \mathbb{P}\left(W_1=W_2\right)=1
\end{aligned}$$

□

6.4 Cramer-Rao Lower-Bound (Information Inequality)

6.4.1 Definition: $\mathrm{Cov}_\theta(\mathbf{W}(X)) \succeq \mathcal{J}_\tau(\theta)[I(\theta)]^{-1}\mathcal{J}_\tau(\theta)^\top, \mathrm{Var}_\theta\left(W(X)\right) \geq \frac{\tau'(\theta)^2}{I(\theta)}$

Given a unbiased 1-d estimator $W(X)$ of $\tau(\theta)$, we are interested in a lower bound of $Var\left(W(X)\right)$. For multi-dimensional case, we want to find a matrix M such that $\mathrm{Cov}(W(X)) \succeq M$.

Theorem 107 (Cramer-Rao Lower-Bound)**.** *Given an experiment $\left(\mathcal{X}, \mathcal{A}, \{f_\theta\}_{\theta\in\Theta}\right)$, where $\mathcal{X} \subset \mathbb{R}^d, \Theta$ is an open set in $\mathbb{R}^k$, f_θ is a density of a continuous r.v. X_1. Suppose*

1. *The Fisher-Information $I(\theta)$ exists and is nonsingular for all θ.*
2. *Let $\mathbf{W}(X)$ be an r-dimentionsl statistic such that $\tau(\theta) := \mathbb{E}_\theta[\mathbf{W}(X)]$ exists for all θ.*
3. *$\nabla_\theta f(x|\theta)$ exists and that ∇ can be passed under the integral sign in $\int_{\mathbb{R}} f(x|\theta)dx$ and $\int_{\mathbb{R}} W(x)f(x|\theta)dx$ (regularity condition).*

Then the following inequality holds, known as the information inequality

$$\mathrm{Cov}_\theta(\mathbf{W}(X)) \succeq \mathcal{J}_\tau(\theta)[I(\theta)]^{-1}\mathcal{J}_\tau(\theta)^\top, \forall\theta \in \Theta \subset \mathbb{R}^k$$

where $\mathcal{J}_\tau(\theta)$ is the $r \times k$ Jacobian matrix of $\tau(\theta)$.
Equivalently one has

$$\mathrm{Var}_\theta\left(\mathbf{c}^\top\mathbf{W}(X)\right) \geq \mathbf{c}^\top\left\{\mathcal{J}_\tau(\theta)[I(\theta)]^{-1}\mathcal{J}_\tau(\theta)^\top\right\}\mathbf{c}, \forall\mathbf{c} \in \mathbb{R}^d$$

Note $E\left[\mathbf{c}^\top\mathbf{W}(X)\right] = \mathbf{c}^\top\tau(\theta)$.
For one-dimensional case, when $k = r = 1$. The inequality is

$$\mathrm{Var}_\theta\left(W(X)\right) \geq \frac{\tau'(\theta)^2}{I(\theta)}$$

Note if $\mathbf{W}(X)$ is an unbiased estimator of θ, then the inequality becomes

$$\mathrm{Cov}_\theta(\mathbf{W}(X)) \succeq [I(\theta)]^{-1}$$

And further if X_i's are i.i.d., the inequality becomes

$$\mathrm{Cov}_\theta(\mathbf{W}(X)) \succeq [I(\theta)]^{-1} = \frac{1}{n}\left[I_1(\theta)\right]^{-1}$$

Proof. (1-d case) By the regularity conditions, $\mathbb{E}_\theta[S(\theta|X)] = 0$, and thus

$$\begin{aligned}
\mathrm{Cov}(S(\theta|X), W(X)) &= \mathbb{E}_\theta[S(\theta|X)W(X)] \\
&= \int_{\mathbb{R}} W(x)\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)f(x|\theta)dx \\
&= \int_{\mathbb{R}} W(x)\frac{\partial}{\partial\theta}f(x|\theta)dx \\
&= \frac{\partial}{\partial\theta}\int_{\mathbb{R}} W(x)f(x|\theta)dx \text{ by regularity condition} \\
&= \frac{\partial}{\partial\theta}\mathbb{E}_\theta[W(X)] \\
&= \frac{\partial}{\partial\theta}\tau(\theta) \\
&= \dot{\tau}(\theta)
\end{aligned}$$

Thus

$$\begin{aligned}[\dot{\tau}(\theta)]^2 &= \mathrm{Cov}^2(S(\theta|X), W(X)) \\ &\leq \mathrm{Var}_\theta(S(\theta|X))\,\mathrm{Var}_\theta(W(X)) \text{ by C.S.} \\ &= I(\theta)\,\mathrm{Var}_\theta(W(X))\end{aligned}$$

□

Example 108 (Unbiased estimator whose variance is smaller than the CR LB)**.** When the support of f_θ depends on θ, such as $X_i \overset{iid}{\sim} U(0,\theta)$, then $I(\theta) = \frac{n}{\theta^2}$. So if the CR LB applies, we wound have $\mathrm{Var}_\theta(W(X)) \geq \frac{n}{\theta^2}$ for any unbiased estimator W. Now for the MLE $X_{(n)}$, it can be shown $\mathbb{E}_\theta\left[X_{(n)}\right] = \frac{n}{n+1}\theta$ and so $W := \frac{n+1}{n}X_{(n)}$ is unbiased. Further more it can be shown

$$\mathrm{Var}_\theta(W(X)) = \frac{\theta^2}{n(n+2)} \ll \frac{\theta^2}{n^2}$$

So there is an unbiased estimator that smaller than the CR LB (since the regularity condition does not hold here).

6.4.2 Attainment of the CR LB: $S(\theta|X) = \frac{I(\theta)}{\tau'(\theta)}(W(X) - \tau(\theta))$

Note the CR LB may not be attained. That is, the variance of unbiased estimator can be strictly larger than the CR LB. Now we try to answer under what conditions the CR LB is attained.

Theorem 109 (Sufficient and Necessary Condition for attaining the CR LB)**.** *($k = r = 1$) Recall the lower bound is*

$$\mathrm{Var}_\theta\left(W(X)\right) \geq \frac{\tau'(\theta)^2}{I(\theta)}$$

An unbiased estimator $W(X)$ of $\tau(\theta) \in \mathbb{R}$ attains the CR LB iff $S(\theta|X)$ is a linear function of $W(X)$, i.e.,

$$S(\theta|X) = a(\theta)(W(X) - \tau(\theta))$$

More specifically, if $\tau'(\theta) \neq 0$ for all θ, then $S(\theta|X) = \frac{I(\theta)}{\tau'(\theta)}(W(X) - \tau(\theta))$.
(Multi-parameter) Recall the lower bound is

$$\mathrm{Cov}_\theta(\mathbf{W}(X)) \succeq \mathcal{J}_\tau(\theta)[I(\theta)]^{-1}\mathcal{J}_\tau(\theta)^\top, \forall\theta \in \Theta \subset \mathbb{R}^k$$

where $\mathcal{J}_\tau(\theta)$ is the $r \times k$ Jacobian matrix of $\tau(\theta)$. The equality condition is

$$\mathbf{W}(X) = \tau(\theta) + \mathcal{J}_\tau(\theta)I(\theta)^{-1}S(\theta|X)$$

Proof. ($k = r = 1$) Recall we use C.S. inequality to prove CR LB. The equality holds iff

$$\mathrm{Cov}^2_\theta(S(\theta|X), W(X)) = \mathrm{Var}_\theta(S(\theta|X))\,\mathrm{Var}_\theta(W(X))$$

$$\implies \rho^2(S(\theta|X), W(X)) = \frac{\mathrm{Cov}^2_\theta(S(\theta|X), W(X))}{\mathrm{Var}_\theta(S(\theta|X))\,\mathrm{Var}_\theta(W(X))} = 1$$

Thus the LB is attained when $S(\theta|X)$ is a linear function of $W(X)$.

$$\rho^2(S(\theta|X), W(X)) = 1$$

$$\Leftrightarrow S(\theta|X) = a(\theta)(W(X) - \tau(\theta))$$

for some function $a(\theta)$. Note $S(\theta|X)$ must be mean 0 so we have $\tau(\theta)$.
By differentiating the equality

$$\frac{\partial}{\partial\theta}\log f(x|\theta) = S(\theta|X) = a(\theta)(W(X) - \tau(\theta))$$

we get

$$\frac{\partial^2}{\partial\theta^2}\log f(x|\theta) = \frac{\partial a(\theta)}{\partial\theta}\left[W(X) - \tau(\theta)\right] - a(\theta)\cdot\tau'(\theta)$$

And by taking expectation on both sides since $\mathbb{E}_\theta[W(X)] = \tau(\theta)$ we get

$$I(\theta) = \mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right] = \mathbb{E}_\theta\left[a(\theta)\cdot\tau'(\theta)\right] = a(\theta)\cdot\tau'(\theta)$$

By assumption $\tau'(\theta) \neq 0$ then

$$a(\theta) = \frac{I(\theta)}{\tau'(\theta)}$$

Plugging this back gives

$$S(\theta|X) = \frac{I(\theta)}{\tau'(\theta)}(W(X) - \tau(\theta))$$

So this gives a more specific form of the linear Relation. □

6.4.3 Relation to UMVUE: Unbiased, attains CR LB ⇒ UMVUE

Theorem 110 (Sufficient Condition for UMVUE)**.** *If $W(X)$ is an unbiased estimator of θ that obtains the CR LB, then W is UMVUE.*

Remark. The converse does not hold. The variance of a UMVUE can be strictly larger than the CR LB. See Example: UMVUE may not attain CR LB.

Remark. By the uniqueness of $UMVUE$, there can only be one unbiased estimator that attains the CR LB.

Example 111 (Poisson UMVUE $\bar{X}$)**.** Given $X_i \overset{iid}{\sim} Poi(\lambda)$, it can be shown $\mathbb{E}_\lambda[\bar{X}] = \lambda$ and $\mathbb{E}_\lambda\left[S^2\right] = \lambda$. So we can ask which is better of if either is UMVUE. We can also consider other unbiased estimators which are convex combinations of the form $W(X) = \alpha\bar{X} + (1-\alpha)S^2$ for some $\alpha \in (0,1)$. It can be shown that $S(\lambda|X_1) = \frac{X_1}{\lambda} - 1$ and hence $I(\lambda) = \frac{n}{\lambda}$. So the CR LB is $\frac{\lambda}{n}$ for any unbiased estimator of λ. Note that $\mathrm{Var}_\theta(\bar{X}) = \frac{\lambda}{n}$. So $\bar{X}$ is UMVUE.

6.4.4 Relation to Exponential Family

Theorem 112 (Relation to exponential family: $f(x|\theta) = h(x)c(\theta)\exp\{\xi(\theta)W(X)\}$.)**.** *Given an experiment $\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta}\big)$ such that $\Theta \subseteq \mathbb{R}^n$. Let $W(X)$ be an unbiased estimator of $\tau(\theta)$. Suppose the conditions of CR theorem are satisfied, then the CR LB is attained by $W(X)$ iff $f_\theta(x)$ can be written as a 1-parameter exponential family with "natural" sufficient statistic $W(X)$. That is $f(x|\theta) = h(x)c(\theta)\exp\{\xi(\theta)W(X)\}$.*

If $k = r > 1$, then this theorem becomes: The CR LB is attained by $\mathbf{W}(X)$ iff $f_\theta(x)$ can be written as a k-parameter exponential family with natural sufficient statistic $\mathbf{W}(X)$. That is $f(x|\theta) = h(x)c(\theta)\exp\{\mathbf{W}(X)^\top\boldsymbol{\xi}(\theta)\}$.

Remark. The theorem implies that the CR LB cannot be attained if the distribution is not from the exponential family distribution. In general, if the regularity conditions are revised, this is not true. See V. M. Joshi's paper: On The Attainment Of The Cramer-rao Lower Bound.

Remark. This theorem provides a method to calculate $\mathrm{Var}(W(X))$, i.e. $\mathrm{Var}(W(X)) = \text{CR LB} = \frac{\left[\tau'(\theta)\right]^2}{I(\theta)}$.

Proof. **(⇒)**

Suppose CR LB is attained by $W(X)$ then by Theorem we must have

$$\frac{\partial}{\partial\theta}\log f(x|\theta) = S(\theta|X) = a(\theta)(W(X) - \tau(\theta))$$

By integrating both sides on some interval $[\theta_0, \theta]$ we have

$$\log f(x|\theta) - \log f(x|\theta_0) = W(X)\int_{\theta_0}^{\theta} a(t)dt - \int_{\theta_0}^{\theta} a(t)\tau(t)dt$$

This implies that

$$f(x|\theta) = \overbrace{f(x|\theta_0)}^{h(x)} \overbrace{\exp\left(-\int_{\theta_0}^{\theta} a(\theta)\tau(\theta)d\theta\right)}^{c(\theta)} \exp\left\{W(X)\overbrace{\int_{\theta_0}^{\theta} a(\theta)d\theta}^{\xi(\theta)}\right\}$$

Thus it is a 1-parameter exponential family.
($\Leftarrow$)
Suppose $W(X)$ is the natural sufficient statistic for the exponential family,

$$f(x|\theta) = h(x)c(\theta)\exp\{\xi(\theta)W(X)\}$$

By taking the logarithm of both sides and differentiating w.r.t. θ we get

$$S(\theta|X) = \frac{\partial}{\partial\theta}\log c(\theta) + \frac{\partial}{\partial\theta}\xi(\theta)W(X)$$

Since

$$0 = E_\theta\left[S(\theta|X)\right] = \frac{\partial}{\partial\theta}\log c(\theta) + \frac{\partial}{\partial\theta}\xi(\theta)\tau(\theta)$$

We have $\frac{\partial}{\partial\theta}\log c(\theta) = -\frac{\partial}{\partial\theta}\xi(\theta)\tau(\theta)$, then

$$S(\theta|X) = \frac{\partial\xi(\theta)}{\partial\theta}(W(X) - \tau(\theta))$$

So $S(\theta|X)$ is a linear function of $W(X)$. Thus, the CR LB is attained by the Theorem: 109. □

6.4.5 Relation to MLE

Theorem 113 (Relation between to MLE: monotonic $\tau(\theta)$ $\Rightarrow W(X) = \hat{\tau}_{MLE}(\theta)$). *Given an experiment $\mathcal{X}, \mathcal{A}, \{P_\theta\}_{\theta\in\Theta}\big)$ such that $\Theta \subseteq \mathbb{R}^n$. Let $W(X)$ be an unbiased estimator of $\tau(\theta)$. Suppose the conditions of CR theorem are satisfied. If $\tau(\theta)$ is strictly monotone in θ then $W(X)$ is the MLE of $\tau(\theta)$.*

Proof. When $\tau(\theta)$ is strictly monotone it is a one-to-one function on Θ and so we can give an equivalent parameterization of the log-likelihood in terms of $\eta = \tau(\theta)$. That is,

$$\ell^*(\eta|x) = \ell\left(\tau^{-1}(\eta)|x\right)$$

We can find the MLE of η by differentiation of ℓ^*

$$\begin{aligned}\frac{\partial}{\partial\eta}\ell^*(\eta|x) &= \frac{\partial}{\partial\tau}\ell\left(\tau^{-1}(\eta)|x\right)\\ &= \frac{\partial}{\partial\theta}\ell(\theta|x)\Big|_{\theta=\tau^{-1}(\eta)} \cdot \frac{\partial\tau^{-1}(\eta)}{\partial\eta}\\ &= a\left(\tau^{-1}(\eta)\right)\left[W(X) - \eta\right]\frac{\partial\tau^{-1}(\eta)}{\partial\eta}\end{aligned}$$

So $W(X)$ is a root of $\frac{\partial}{\partial\eta}\ell^*(\eta|x) = 0$ and is thus a critical point. To see that $W(X)$ is actually the MLE we can use the fact that $a(\theta) = \frac{I(\theta)}{\tau'(\theta)}$ and get $a\left(\tau^{-1}(\eta)\right) = \frac{I\left(\tau^{-1}(\eta)\right)}{\tau'(\tau^{-1}(\eta))}$. Plugging this back gives

$$a\left(\tau^{-1}(\eta)\right)\frac{\partial\tau^{-1}(\eta)}{\partial\eta} = \frac{I\left(\tau^{-1}(\eta)\right)}{\frac{\partial\tau(\theta)}{\partial\theta}\Big|_{\theta=\tau^{-1}(\eta)}} \cdot \frac{1}{\frac{\partial\tau(\theta)}{\partial\theta}\Big|_{\theta=\tau^{-1}(\eta)}}$$
$$= \frac{I\left(\tau^{-1}(\eta)\right)}{\left(\frac{\partial\tau(\theta)}{\partial\theta}\Big|_{\theta=\tau^{-1}(\eta)}\right)^2} > 0$$

This proves $\frac{\partial}{\partial\eta}\ell^*(\eta|x) > 0$ for $\eta < W(X)$, and $\frac{\partial}{\partial\eta}\ell^*(\eta|x) < 0$ for $\eta > W(X)$ and so $W(X)$ is a maximum of $\ell^*(\eta|x)$ and thus is the MLE of $\eta = \tau(\theta)$. □

6.4.6 Problem: Only want to estimate θ_i

Problem 114 (Estimating one parameter in multi-parameters)**.** For $N(\mu, \sigma^2)$, we want to estimate σ^2. The score function is

$$S(\mu, \sigma^2|X) = -\frac{n}{2\sigma^2} + \frac{\sum(X_i-\mu)^2}{2\sigma^4}$$
$$= \underbrace{\frac{n}{2\sigma^2}}_{a(\sigma^2)}\left(\underbrace{\frac{1}{n}\sum(X_i-\mu)^2}_{W_\mu(X)} - \underbrace{\sigma^2}_{\tau(\sigma^2)=\sigma^2}\right)$$

which shows the linear Relation between $S(\mu, \sigma^2|X)$ and $W_\mu(X)$.

- If μ is known, then by the Theorem: Sufficient and Necessary Condition for attaining the CR LB, the CR LB for estimating σ^2 is attained by $W_\mu(X)$. Thus, by Theorem: Sufficient Condition for UMVUE, $W_\mu(X)$ is the UMVUE. Note that $W_\mu(X)$ is also the unique unbiased estimator that attains the CR LB when μ is known.
- If μ is unknown, we cannot $W_\mu(X)$. Is there any unbiased estimator of σ^2 that can attain the CR LB?

In general, suppose we are only interested in estimating 1 component of $\boldsymbol{\theta}$, say θ_i. Then whether the other components are known or unknown can affect our CR LB on unbiased estimators of θ_i. Suppose $W_i(X)$ (1-d) is an unbiased estimator of θ_i, based on i.i.d. data.

- Case 1: Other components of $\boldsymbol{\theta}$ are known. Then it is the case $k = r = 1$ and $\tau'(\theta)^2 = 1$. The CR LB is

$$\mathrm{Var}_\theta\left(W_i(X)\right) \geq \frac{\tau'(\theta)^2}{I(\theta_i)}$$
$$= \frac{1}{I(\theta_i)}$$
$$= \frac{1}{[I(\boldsymbol{\theta})]_{ii}} \text{ since both are } (\nabla_{\theta_i}\ell)^2$$
$$= \frac{1}{n\,[I_1(\boldsymbol{\theta})]_{ii}} \text{ if i.i.d.}$$

- Case 2: Other components of $\boldsymbol{\theta}$ are unknown. Then $\underbrace{\mathcal{J}_\tau(\boldsymbol{\theta})}_{1\times k} = \mathbf{e}_i^\top$. The CR LB is

$$\mathrm{Cov}_\theta(W_i(X)) \succeq \mathcal{J}_\tau(\boldsymbol{\theta})[I(\boldsymbol{\theta})]^{-1}\mathcal{J}_\tau(\boldsymbol{\theta})^\top$$
$$= \left[I(\boldsymbol{\theta})^{-1}\right]_{ii}$$
$$= \frac{\left[I_1(\boldsymbol{\theta})^{-1}\right]_{ii}}{n} \text{ if i.i.d}$$

In fact, we can show

$$\frac{1}{[I_1(\boldsymbol{\theta})]_{ii}} \leq \left[I_1(\boldsymbol{\theta})^{-1}\right]_{ii}$$

with equality iff the off-diagonal elements with index i are 0 (no need $I(\boldsymbol{\theta})$ to be diagonal, can be proved by the formula $A^{-1} = \text{adj}(A)/\det(A)$).

In this example, since

$$I_1(\boldsymbol{\theta}) = I_1(\mu, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{n\sigma^4} \end{bmatrix}$$

the CR LB on unbiased estimator of σ^2 remains the same whether or not μ is known.

This implies that if there exists an unbiased estimator Y of σ^2 that achieves the CR LB when μ is unknown, it also achieves the CR LB when μ is known. But we have shown $W_\mu(X) = \frac{1}{n}\sum(X_i - \mu)^2$ is the only unbiased estimator of σ^2 that archives the CR LB when μ is known, so it must be that $Y = W_\mu(X)$, i.e. Y must depend on μ. However, when μ is unknown, it cannot be valid estimator. Therefore, there is no unbiased estimator of σ^2 that archives the CR LB when μ is unknown. But UMVUE can exists. See Example: UMVUE may not attain CR LB.

6.4.7 Relation: E(unbiased estimator|sufficient stat) has lower variance

Theorem 115 (Rao-Blackwell)**.** *Let $W(X)$ be an unbiased estimator of $\tau(\theta)$ and $T(X)$ is a sufficient statistic for θ. Then*

$$\phi(T) = \mathbb{E}_\theta[W|T]$$

is a uniformly better unbiased estimate of $\tau(\theta)$ than $W(X)$. That is,

$$\text{Var}_\theta(W) \geq \text{Var}_\theta(\phi(T)) \text{ for all } \theta$$

When W is an r-dimentional estimator, the theorem claims

$$\text{Cov}_\theta(W) \succeq \text{Cov}_\theta(\phi(T)) \text{ for all } \theta$$

Remark. Note we need the sufficiency of T to ensure $\phi(T)$ is an valid estimator for $\tau(\theta)$, i.e. it is free of θ. Actually, the variance decreases if we condition on anything (as shown in the proof), but the resulting quantity may depends on θ, and hence is not an estimator.

Proof. By the law of total expectation of total variance,

$$\begin{aligned} E_\theta(\phi(T)) &= \mathbb{E}_\theta\left\{\mathbb{E}_\theta[W|T]\right\} \\ &= \mathbb{E}_\theta[W] \\ &= \tau(\theta) \end{aligned}$$

and

$$\begin{aligned} \text{Var}_\theta(W)) &= \text{Var}_\theta(E[W|T]) + E_\theta\left[\text{Var}_\theta(W|T)\right] \\ &= \text{Var}_\theta(\phi(T)) + E_\theta\left[\text{Var}_\theta(W|T)\right] \\ &\geq \text{Var}_\theta(\phi(T)) \text{ for all } \theta \end{aligned}$$

It holds in multivariate case since

$$\text{Var}_\theta(W|T) \succeq 0 \Rightarrow E_\theta\left[\text{Var}_\theta(W|T)\right] \succeq 0$$

□

6.4.8 Relation: E(unbiased estimator|complete suff stat) = UMVUE

Theorem 116 (Lehmann–ScheffÃ© for UMVUE)**.** *Let $T(X)$ be a **complete** sufficient statistic for θ. If $\phi(T)$ depends only on T, then $\phi(T)$ is the UMVUE for $E[\phi(T)]$.*

Equivalently, in this class we say that, if $W(X)$ is an unbiased estimator of $\tau(\theta)$ and . Then

$$\phi(T) = \mathbb{E}_\theta[W|T]$$

is a UMVUE of $\tau(\theta)$. Note $\phi(T)$ depends only on T since T is sufficient.

Proof. Given $W(X)$, let $U(X)$ be an unbiased estimator of $\tau(\theta)$. Define

$$\tilde{\phi}(T) := \mathbb{E}_\theta[U|T]$$

which is also an unbiased estimator of $\tau(\theta)$. By Theorem: Rao-Blackwell,

$$\mathrm{Var}_\theta(U) \geq \mathrm{Var}_\theta(\tilde{\phi}(T)) \text{ for all } \theta$$

Note that

$$\mathbb{E}_\theta[\underbrace{\tilde{\phi}(T) - \phi(T)}_{g(T)}] = \tau(\theta) - \tau(\theta) = 0 \text{ for all } \theta$$

By completeness of T, we have

$$\tilde{\phi}(T) \overset{a.s.}{=} \phi(T)$$

which implies

$$\mathrm{Var}_\theta(U) \geq \mathrm{Var}_\theta(\tilde{\phi}(T)) = \mathrm{Var}_\theta(\phi(T)$$

This is true for all θ and for any unbiased estimator $U(X)$. So $\phi(T)$ is UMVUE. □

This theorem provides a recipe to find a UMVUE of $\tau(\theta)$ when a complete sufficient statistic is known.

Step 1. Find a unbiased estimator $W(X)$ of $\tau(\theta)$

Step 2. Take $\phi(T) = \mathbb{E}_\theta[W|T]$.

Remark. Since the UMVUE is unique, we conclude that the UMVUE must depends only on a complete sufficient statistic for θ. This provides a simple criteria to identify whether a statistic is UMVUE or not.

Example 117 (UMVUE may not attain CR LB)**.** For $N(\mu, \sigma^2)$, when both μ and σ^2 are unknown and we want to estimate σ^2. We know that $\left(\bar{X}, S^2\right)$ is a complete sufficient statistic for (μ, σ^2) and S^2 is unbiased for σ^2. Thus, by the above theorem, $\mathbb{E}_\theta\left[S^2 | \left(\bar{X}, S^2\right)\right] = S^2$ is a UMVUE of σ^2. Note that there is a UMVUE for σ^2 even though there is no unbiased estimator for σ^2 that attains the CR LB, as shown in Problem: Estimating one parameter in multi-parameters. This gives an example that the variance of a UMVUE can be strictly larger than the CR LB.

6.5 Evaluation of Estimators

6.5.1 Mean Squared Error

Definition 118 (Mean Squared Error)**.** Suppose W is an estimator of θ. Then

$$\begin{aligned}\mathrm{MSE}_\theta(W(X)) &= \mathbb{E}_{P_\theta}\left[\|W(X)-\theta\|^2\right] \\ &= Var_\theta(W) + Bias_\theta^2(W)\end{aligned}$$

where

$$Bias_\theta(W) = E_\theta(W) - \theta$$

Definition 119 (Globally Optimal Estimator)**.** An estimator $W(X)$ of θ is said to be globally optimal w.r.t. MSE-risk if for any other estimator $T(X)$ of θ one has

$$\mathrm{MSE}_\theta(W(X)) \leq \mathrm{MSE}_\theta(T(X))$$

for all θ.

Remark. There is generally almost never a globally optimal estimator among all possible estimators. Consider the estimator $W = c$ regardless of the data. Then if it happens that $\theta = c$, we get $MSE(W) = 0$. So for W_{opt} to exist, we must have

$$0 \leq \mathrm{MSE}_{\theta=c}(W_{opt}) \leq \mathrm{MSE}_{\theta=c}(W) = 0 \Longrightarrow \mathrm{MSE}_{\theta=c}(W_{opt}) = 0$$

which is generally not possible in any non-trivial case.

So instead of seeking globally optimal estimators among all estimators we could limit ourselves to smaller classes of estimators. One such class of interest is the class of all unbiased estimators of θ. Another possible class is the class of all linear unbiased estimators.

6.5.2 Asymptotic Efficiency $\sqrt{n}(\tilde{\theta}-\theta) \xrightarrow{\mathcal{D}} N_k(0, I_1(\theta)^{-1})$

Definition 120 (Asymptotic Efficiency)**.** Let $X_1, \cdots, X_n$ be i.i.d. R.V.'s with distribution depending on k parameters $\theta \in \Theta \subset \mathbb{R}^k$. A sequence of estimates $\left\{\tilde{\theta}_n(X)\right\}$ of θ such that

$$\sqrt{n}(\tilde{\theta}-\theta) \xrightarrow{\mathcal{D}} N_k(0, \Sigma)$$

for all θ is said to be asymptotically efficient if

$$\Sigma = I_1(\theta)^{-1} \text{ for all } \theta \in \Theta$$

One could think of this as the CR LB is attained in the limit.

From Theorem: Asymptotic Normality of MLE (1-d), MLE is asymptotically efficient.

	Frequentist	Bayesian
Probabilities are objective or subjective	Probabilities are objective properties of the real world. It refers to limiting relative frequencies.	Probability describes degree of belief, not limiting frequency. One might say "The probability that Shakespeare wrote this play is 0.35".
Parameters are fixed or random	Parameters are fixed unknown constant. Because they are not fluctuating, no useful probability statements can be made about parameters.	Parameters are random variables with distributions. One can make probability statements about them. $P(\mu \in (-0.5, 0.5)) = 0.9$
Statistics	Statistical procedures should be designed to have well-defined long-run properties.	One can say$P(\mu \in (-0.5, 0.5)|x_1, \cdots, x_n) = 0.95$

Table 1: Comparison of frequentist and Bayesian point of view

7 A Short Intro to Bayesian Statistics

7.1 Comparison of Frequentist and Bayesian Point of View

7.2 Bayesian Estimation

7.2.1 Posterior distribution

Definition 121 (Posterior distribution on θ)**.** Start with our frequentist notion of experiment $(\mathcal{X}, \mathcal{A}, \{f(x|\theta), \theta \in \Theta\})$. Note here $f(x|\theta)$ is a new notation, a conditional probability.

Add a new element to this prior prior distribution on θ, call it $\pi(\theta)$. It incorporate prior knowledge of θ.

Collect a sample $\mathbf{X} = \mathbf{x}$. The likelihood is

$$L(\theta; \mathbf{x}) = f(\mathbf{x}|\theta)$$

Posterior distribution on θ is

$$\pi(\theta|\mathbf{x}) = \frac{\pi(\theta) f(\mathbf{x}|\theta)}{\int \pi(t) f(\mathbf{x}|t) \mathrm{d}t}$$

Note that the dominator is a function of $\mathbf{x}$, not θ. It's a normalizing constant. Sometimes one can avoid calculating it, so that

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta) f(\mathbf{x}|\theta)$$

Example 122 (Beta-Binomial)**.** Suppose $X \sim Bin(n, p)$. A 2-parameter family of priors for p is $Beta(\alpha, \beta)$, with density

$$f(p|\alpha, \beta) = \frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

and mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, then

$$\begin{aligned} \pi(\theta|x) &\propto \pi(\theta) f(x|\theta) \\ &= \begin{pmatrix} n \\ x \end{pmatrix} \frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{x+\alpha-1}(1-p)^{n-x+\beta-1} \\ &\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1} \end{aligned}$$

Thus, the posterior distribution must be $Beta(x + \alpha, n - x + \beta)$.

7.2.2 Estimator

One natural choice os point estimator is posterior mean. In the above example, $E(p|x) = \frac{x+\alpha}{n+\alpha+\beta}$. Compare to the frequentist estimator $\frac{x}{n}$, which is MLE, UMVUE, we see

$$E(p|x) = \frac{x+\alpha}{n+\alpha+\beta} = \frac{\alpha+\beta}{\alpha+\beta+n} \underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{prior mean}} + \frac{n}{\alpha+\beta+n} \underbrace{\frac{x}{n}}_{\text{MLE}}$$

is a weighted average or prior mean and standard estimator. Note the effect of prior $Beta(\alpha, \beta)$ is same as if there are $\alpha + \beta$ trials occurred with α successes.

If n is fixed, $\alpha \to \infty, \beta \to \infty$, then $E(p|x) \to \frac{\alpha}{\alpha+\beta}$, i.e. prior information overwhelms data. In this case, $Beta(\alpha, \beta)$ tends to a point mass at $\frac{\alpha}{\alpha+\beta}$. On the other hand, if α, β are fixed and $n \to \infty$ then $E(p|x) \to \frac{x}{n}$, i.e. data overwhelms prior.

7.3 Choice of Priors

7.3.1 Conjugate Prior Family

Definition 123 (Conjugate prior family)**.** Let $\mathcal{F} = \{f(x|\theta), \theta = \Theta\}$, if family Π of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the family Π for all $f \in \mathcal{F}$, all points in Π, and all $x \in \mathcal{X}$.

From the above section we see Beta is a conjugate family for Binomial (n known, p unknown).

Theorem 124 (Conjugate prior for exponential family)**.** *Exponential families have natural conjugate prior distributions. Note*

$$f(\mathbf{x}|\theta) = c(\theta)^n \left[\Pi h(x_i)\right] \exp\left\{\sum_{j=1}^{k}\left[w_j(\theta)\sum_{i=1}^{n} t_i(x_j)\right]\right\}$$

If the prior is specified as

$$\pi(\theta) = c(\theta)^\eta \exp\left\{\sum_{j=1}^{k} w_j(\theta)\gamma_j\right\}$$

Then the posterior density is

$$\pi(\theta|\mathbf{x}) \propto c(\theta)^{n+\eta} \exp\left\{\sum_{j=1}^{k}\left[w_j(\theta)\left(\gamma_j + \sum_{i=1}^{n} t_i(x_j)\right)\right]\right\}$$

7.3.2 Noninformative Priors

Noninformative priors has little impact on the posterior distribution. For instance, normal distribution with a large variance, or uniform distribution with a large range.

Noninformative priors can be used if

- don't have prior info to incorporate.
- want to avoid controversy about your conclusions.

7.3.3 Proper vs Improper Priors

A prior is improper if it does not depend on data and integrates to 1.

An improper prior is generally used to refer to a prior that integrates to ∞.

Example 125 (Improper prior of Normal μ)**.** $X_i \sim N(\mu, \sigma^2)$ where σ^2 is known. To estimate μ, consider prior $\pi(\mu) \propto 1$ on $\mathbb{R}$. Obviously this is an improper prior. In this case the posterior for μ is $N(\bar{X}, \frac{\sigma^2}{n})$ which is a proper posterior (but it's not always the case).

Example 126 (Improper prior of Binomial p)**.** $Beta(0,0)$ is a improper prior. Whether posterior is proper also depends on data. The posterior distribution in this example is $Beta(x, n-x)$, which is proper if $1 \leq x \leq n-1$, and improper if $x = 0$ or $x = n$.

7.3.4 Empirical Bayes

Prior depends on data.

7.3.5 Jefferys' Invariance Principle and Jefferys' Prior

Consider a 1-dim parameter. Consider any one-to-one transformation of θ, $\phi = h(\theta)$. Jefferys' invariance principle is that the rule for determining the prior on θ should give an equivalent result if applied to the transformed parameter ϕ. By change of variables,

$$\tilde{\pi}(\phi) = \pi(h^{-1}(\phi))|h'(\theta)|^{-1}_{\theta=h^{-1}(\phi)}$$

Definition 127 (Jeffreys' prior)**.** Jeffreys' prior is

$$\pi(\theta) \propto [I(\theta)]^{1/2}$$

Theorem 128 (Jeffrey's invariance prior)**.** *If we choose Jeffrey's prior on θ, then under Jefferys' invariance principle, the prior of a new parameterization $\phi = h(\theta)$ satisfies*

$$\tilde{\pi}(\phi) \propto \left[\tilde{I}(\phi)\right]^{1/2}$$

Proof. Note □

$$\tilde{I}(\phi) = E\left[\left(\frac{\partial \tilde{\ell}(\phi|x)}{\partial \phi}\right)^2\right]$$

$$\tilde{\ell}(\phi, x) = \ell(h^{-1}(\phi), x)$$

$$\frac{\partial \tilde{\ell}(\phi|x)}{\partial \phi} = \frac{\partial \ell(\theta, x)}{\partial \theta}|_{\theta=h^{-1}(\phi)} \frac{\partial h^{-1}(\phi)}{\partial \phi}$$

Thus,

$$\begin{aligned}\tilde{I}(\phi) &= E\left[\left(\frac{\partial \ell(\theta, x)}{\partial \theta}|_{\theta=h^{-1}(\phi)}\right)^2\right]\left|\frac{\partial h^{-1}(\phi)}{\partial \phi}\right|^2 \\ &= I(h^{-1}(\phi))\,|h'(\theta)|^2_{\theta=h^{-1}(\phi)}\end{aligned}$$

which is the same result you would get if you started with $\pi(\theta) \propto [I(\theta)]^{1/2}$ and transformed variables to find $\tilde{\pi}(\phi) = \pi(h^{-1}(\phi))|h'(\theta)|^{-1}_{\theta=h^{-1}(\phi)}$.

Example 129 (Jeffrey's prior for binomial)**.** Suppose $X \sim Bin(n, p)$, since $I(p) = \frac{n}{p(1-p)}$, then Jeffreys' prior $\pi(p) \propto [p(1-p)]^{-1/2}$ which is $Beta(\frac{1}{2}, \frac{1}{2})$ (convex, bell-shape). Note the prior is inversely proportional to the s.d. of the data. Intuitively, the data has the least effect on the posterior when true $\theta = \frac{1}{2}$ and has greatest effect near the extremes. Jeffreys' prior puts more mass near the extremes where the data has the strongest effect.

Jefferys' prior works well for single-parameter model but not for multi-parameter models. Extension to $\theta \in \Theta \subset \mathbb{R}^k$ is $\pi(\theta) \propto \sqrt{\det(I(\theta))}$.

Intuitively, the Jefferys prior in multidimensional tends to put too much mass out at a far distance, whereas we might normally want to shrink our estimates towards 0 in a multidimensional case.

7.3.6 Reference Prior (?)

Formalize idea of "noninformative" prior, reference prior maximizes K-L divergence between the posterior and the prior in order to make/allow the data to have the maximum possible effect on the posterior estimates.

Theorem 130 (Relation between Jeffreys' prior and reference prior)**.** *For one-dimensional θ, Jeffreys' prior and reference prior are equivalent. For multi-dimensional parameters, they differ.*