

MATH357 - Statistics

Based on lectures from Winter 2025 by Prof. Abbas Khalili.
Notes by Louis Meunier

Contents

1 Review of Probability	2
2 Common Statistical Tools	6
2.1 Definition of Statistics	6
2.2 Properties of Normal and other Common Distributions	7
2.3 Order Statistics	10
2.4 Large Sample/Asymptotic Theory	12
3 Parametric Inference	14
3.1 Uniformly Minimum Variance Unbiased Estimators (UMVUE), Cramér-Rau Lower Bound (CRLB)	17
3.2 Sufficiency	23
3.3 Completeness	28
3.4 Existence of a UMVUE	32
4 System Parameter Estimation	34
4.1 Method of Moments	34
4.2 Maximum Likelihood Estimation (MLE)	36

§1 REVIEW OF PROBABILITY

↪ **Definition 1.1** (Measurable Space, Probability Space): We work with a set Ω = sample space = {outcomes}, and a σ -algebra \mathcal{F} , which is a collection of subsets of Ω containing Ω and closed under taking complements and countable unions. The tuple (Ω, \mathcal{F}) is called *measurable space*.

We call a nonnegative function $P : \mathcal{F} \rightarrow \mathbb{R}$ defined on a measurable space a *probability function* if $P(\Omega) = 1$ and if $\{E_n\} \subseteq \mathcal{F}$ a disjoint collection of subsets of Ω , then $P(\bigcup_{n \geq 1} E_n) = \sum_{n \geq 1} P(E_n)$. We call the tuple (Ω, \mathcal{F}, P) a *probability space*.

↪ **Definition 1.2** (Random Variables): Fix a probability space (Ω, \mathcal{F}, P) . A Borel-measurable function $X : \Omega \rightarrow \mathbb{R}$ (namely, $X^{-1}(B) \in \mathcal{F}$ for every $B \in \mathfrak{B}(\mathbb{R})$) is called a *random variable* on \mathcal{F} .

- *Probability distribution*: X induces a probability distribution on $\mathfrak{B}(\mathbb{R})$ given by $P(X \in B)$
- *Cumulative distribution function (CDF)*:

$$F_X(x) := P(X \leq x).$$

Note that $F(-\infty) = 0, F(+\infty) = 1$ and F right-continuous.

We say X *discrete* if there exists a countable set $S := \{x_1, x_2, \dots\} \subset \mathbb{R}$, called the *support* of X , such that $P(X \in S) = 1$. Putting $p_i := P(X = x_i)$, then $\{p_i : i \geq 1\}$ is called the *probability mass function* (PMF) of X , and the CDF of X is given by

$$P(X \leq x) = \sum_{i: x_i \leq x} p_i.$$

We say X *continuous* if there is a nonnegative function f , called the *probability distribution function* (PDF) of X such that $F(x) = \int_{-\infty}^x f(t) dt$ for every $x \in \mathbb{R}$. Then,

- $\forall B \in \mathfrak{B}(\mathbb{R}), P(X \in B) = \int_B f(t) dt$
- $F'(x) = f(x)$
- $\int_{-\infty}^{\infty} f(x) dx = 1$

If $X : \Omega \rightarrow \mathbb{R}$ a random variable and $g : \mathbb{R} \rightarrow \mathbb{R}$ a Borel-measurable function, then $Y := g(X) : \Omega \rightarrow \mathbb{R}$ also a random variable.

↪ **Definition 1.3** (Moments): Let X be a discrete/random variable with pmf/pdf f and support S . Then, if $\sum_{x \in S} |x| f(x) / \int_S |x| f(x) dx < \infty$, then we say the first moment/mean of X exists, and define

$$\mu_X = \mathbb{E}[X] = \begin{cases} \sum_{x \in S} x f(x) \\ \int_S x f(x) dx \end{cases}.$$

Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a Borel-measurable function. Then, we have

$$\mathbb{E}[g(X)] = \begin{cases} \sum_{x \in S} g(x) f(x) \\ \int_S g(x) f(x) dx \end{cases}.$$

Taking $g(x) = |x|^k$ gives the so-called “ k th absolute moments”, and $g(x) = x^k$ gives the ordinary “ k th moments”. Notice that $\mathbb{E}[\cdot]$ is linear in its argument.

For $k \geq 1$, if μ exists, define the central moments

$$\mu_k := \mathbb{E}[(X - \mu)^k],$$

where they exist.

↪ **Definition 1.4** (Moment Generating Function (mgf)): If X a r.v., the mgf of X is given by

$$M(t) := \mathbb{E}[e^{tX}],$$

if it exists for $t \in (-h, h)$, $h > 0$. Then, $M^{(n)}(0) = \mathbb{E}[X^n]$.

↪ **Definition 1.5** (Multiple Random Variable): $X = (X_1, \dots, X_n) : \Omega \rightarrow \mathbb{R}^n$ a random vector if $X^{-1}(I) \in \mathcal{F}$ for every $I \in \mathfrak{B}_{\mathbb{R}^n}$. (It suffices to check for “rectangles” $I = (-\infty, a_1] \times \dots \times (-\infty, a_n]$, as before.)

Let F be the CDF of X , and let $A \subseteq \{1, \dots, n\}$, enumerating A by $\{i_1, \dots, i_k\}$. Then, the CDF of the subvector $X_A = (X_{i_1}, \dots, X_{i_k})$ is given by

$$F_{X_A}(x_{i_1}, \dots, x_{i_k}) = \lim_{\substack{x_{i_j} \rightarrow \infty, \\ i_j \in \mathcal{I} \setminus A}} F(x_1, \dots, x_n).$$

In particular, the marginal distribution of X_i is given by

$$F_{X_i}(x) = \lim_{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n \rightarrow +\infty} F(x_1, \dots, x, \dots, x_n).$$

Let $g : \mathbb{R}^n \rightarrow \mathbb{R}$ measurable. Then,

$$\mathbb{E}[g(X_1, \dots, X_n)] = \begin{cases} \sum_{(x_1, \dots, x_n)} g(x_1, \dots, x_n) f(x_1, \dots, x_n) \\ \int \dots \int g(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1 \dots dx_n \end{cases}.$$

We have the notion of a joint mgf,

$$M(t_1, \dots, t_n) = \mathbb{E}\left[e^{\sum_{i=1}^n t_i X_i}\right],$$

if it exists for $0 < \left(\sum_{i=1}^n t_i^2\right)^{\frac{1}{2}} < h$ for some $h > 0$. Notice that $M(0, \dots, 0, t_i, 0, \dots, 0)$ is equal to the mgf of X_i .

↪ **Definition 1.6** (Conditional Probability): Let (X_1, \dots, X_n) a random vector. Let $\mathcal{I} = \{1, \dots, n\}$ and A, B disjoint subsets of \mathcal{I} with $k := |A|, h := |B|$. Write $X_A = (X_{i_1}, \dots, X_{i_k})^t$, similar for B . Then, the conditional probability of A given B is given by

$$f_{X_A|X_B}(x_a|x_b) := f_{X_A|X_B=x_B}(x_A) = \frac{f_{X_A, X_B}(x_a, x_b)}{f_{X_B}(x_b)},$$

provided the denominator is nonzero. Sometimes we have information about conditional probabilities but not the main probability function; we have that

$$f(x_1, \dots, x_n) = f(x_1)f(x_2|x_1)f(x_3|x_1, x_2)\dots f(x_n|x_1, \dots, x_{n-1}),$$

which follows from expanding the previous definition and observing the cancellation.

Let $X = (X_1, \dots, X_n) \sim F$. We say X_1, \dots, X_n (mutually) independent and write $\prod_{i=1}^n X_i$ if

$$F(x_1, \dots, x_n) = \prod_{i=1}^n F_{X_i}(x_i),$$

where F_{X_i} the marginal cdf of X_i . Equivalently,

$$\begin{aligned} \prod_{i=1}^n X_i &\Leftrightarrow f(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i) \\ &\Leftrightarrow P(X_1 \in B_1, \dots, X_n \in B_n) = \prod_{i=1}^n P(X_i \in B_i) \quad \forall B_i \in \mathfrak{B}_{\mathbb{R}} \\ &\Leftrightarrow M_X(t_1, \dots, t_n) = \prod_{i=1}^n M_{X_i}(t_i). \end{aligned}$$

If X, Y are two random variables with cdfs F_X, F_Y such that $F_X(z) = F_Y(z)$ for every z , we say X, Y identically distributed and write $X \stackrel{d}{=} Y$ (note that X need not equal Y pointwise). If X_1, \dots, X_n a collection of random variables that are independent and identically distributed with common cdf F , we write $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$.

Further, define the covariance, correlation of two random variables X, Y respectively:

$$\text{Cov}(X, Y) := \sigma_{X,Y} := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mu_X \mu_Y, \quad \rho_{X,Y} := \frac{\sigma_{XY}}{\sigma_X \sigma_Y},$$

if $\mathbb{E}[|X - \mathbb{E}[X]| |Y - \mathbb{E}[Y]|] < \infty$.

↪ **Theorem 1.1**: If X_1, \dots, X_n independent and $g_1, \dots, g_n : \mathbb{R} \rightarrow \mathbb{R}$ borel-measurable functions, then $g_1(X_1), \dots, g_n(X_n)$ also independent.

↪ **Definition 1.7** (Conditional Expectation): Let X, Y be random variables and $g : \mathbb{R} \rightarrow \mathbb{R}$ a borel-measurable function. We define the following notions:

$$\mathbb{E}[g(X)|Y = y] = \begin{cases} \sum_{x \in S_X} g(x)f(x|y) & \text{discrete} \\ \int_{S_X} g(x)f(x|y) dx & \text{cnts} \end{cases}.$$

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

↪ **Theorem 1.2**: If $\mathbb{E}[g(X)]$ exists, then $\mathbb{E}[g(X)] = \mathbb{E}[\mathbb{E}[g(X)|Y]]$, where the first nested \mathbb{E} is with respect to x , the second y .

↪ **Theorem 1.3**: If $\mathbb{E}[X^2] < \infty$, then $\text{Var}(X) = \text{Var}(\mathbb{E}[X|Y]) + \mathbb{E}[\text{Var}(X|Y)]$. In particular, $\text{Var}(X) \geq \text{Var}(\mathbb{E}[X|Y])$.

§2 COMMON STATISTICAL TOOLS

§2.1 Definition of Statistics

↪ **Definition 2.1** (Inference): We consider some population with some characteristic we wish to study. We can model this characteristic as a random variable $X \sim F$. In general, we don't have access to F , but wish to take samples from our population to make inferences about its properties.

(1) *Parametric inference*: in this setting, we assume we know the functional form of X up to some parameter, $\theta \in \Theta \subset \mathbb{R}^d$, where Θ our "parameter space". Namely, we know $X \sim F_\theta \in \mathcal{F} := \{F_\theta \mid \theta \in \Theta\}$.

(2) *Non-parametric inference*: in this setting we know nothing about F itself, except perhaps that F continuous, discrete, etc.

Other types exist. We'll focus on these two.

↪ **Definition 2.2** (Random Sample): Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$. Then X_1, \dots, X_n called a *random sample* of the population.

We also call X_i the "pre-experimental data" (to be observed) and x_i the "post-experimental data" (been observed).

↪ **Definition 2.3** (Statistics): Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$ where X_i a d -dimensional random vector. Let

$$T : \underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \dots \times \mathbb{R}^d}_{n\text{-fold}} \rightarrow \mathbb{R}^k$$

be a borel-measurable function. Then, $T(X_1, \dots, X_n)$ is called a *statistic*, provided it does not depend on any unknown.

⊗ **Example 2.1:** $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$ (the “sample mean”) and $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, (the “sample variance”) are both typical statistics.

§2.2 Properties of Normal and other Common Distributions

↪ **Theorem 2.1:** Let $x_1, \dots, x_n \in \mathbb{R}$, then

- (a) $\operatorname{argmin}_{\alpha \in \mathbb{R}} \left\{ \sum_{i=1}^n (x_i - \alpha)^2 \right\} = \bar{x}_n$;
- (b) $\sum_{i=1}^n (x_i - \bar{x}_n)^2 = \sum_{i=1}^n (x_i^2) - n\bar{x}_n^2$;
- (c) $\sum_{i=1}^n (x_i - \bar{x}_n) = 0$.

↪ **Theorem 2.2:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$, and $g : \mathbb{R} \rightarrow \mathbb{R}$ borel-measurable such that $\operatorname{Var}(g(X)) < \infty$. Then,

- (a) $\mathbb{E} \left[\sum_{i=1}^n g(X_i) \right] = n\mathbb{E}[g(X_1)]$;
- (b) $\operatorname{Var} \left(\sum_{i=1}^n g(X_i) \right) = n \operatorname{Var}(X_1)$.

↪ **Theorem 2.3:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$ with $\sigma^2 < \infty$, then

1. $\mathbb{E}[\bar{X}_n] = \mu$, $\operatorname{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$, $\mathbb{E}[S_n^2] = \sigma^2$.
2. If $M_{X_1}(t)$ exists in some neighborhood of 0, then $M_{\bar{X}_n}(t) = M_{X_1}\left(\frac{t}{n}\right)^n$, where it exists.

↪ **Theorem 2.4:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then

1. $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$;
2. \bar{X}_n, S_n^2 are independent;
3. $\frac{(n-1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi_{(n-1)}^2$.

Remark 2.1:

2. actually holds iff the underlying distribution is normal.

PROOF. We prove 2. We first write S_n^2 as a function of $(X_2 - \bar{X}_n, \dots, X_n - \bar{X}_n)$:

$$\begin{aligned}
S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n-1} \left\{ \sum_{i=2}^n (X_i - \bar{X}_n)^2 + (X_1 - \bar{X}_n)^2 \right\} \\
&= \frac{1}{n-1} \left\{ \sum_{i=2}^n (X_i - \bar{X}_n)^2 + \left(\sum_{i=2}^n (X_i - \bar{X}_n) \right)^2 \right\}.
\end{aligned}$$

Then, it suffices to show that \bar{X}_n and $(X_2 - \bar{X}_n, \dots, X_n - \bar{X}_n)$ are independent.

Consider now the transformation

$$\begin{cases} y_1 = \bar{x}_n \\ y_2 = x_2 - \bar{x}_n \\ \vdots \\ y_n = x_n - \bar{x}_n \end{cases} \Rightarrow \begin{cases} x_1 = y_1 - \sum_{i=2}^n y_i \\ x_2 = y_2 + y_1 \\ \vdots \\ x_n = y_n + y_1 \end{cases},$$

which gives Jacobian

$$|J| = \left| \begin{pmatrix} 1 & -1 & \dots & -1 \\ 1 & 1 & 0 & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 1 \end{pmatrix} \right| = n,$$

and so

$$\begin{aligned}
f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) &= |J| \cdot f_{X_1, \dots, X_n}(x_1(y_1, \dots, y_n), \dots, x_n(y_1, \dots, y_n)) \\
&= n \cdot \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i(y_1, \dots, y_n) - \mu)^2} \\
&\approx \underbrace{e^{-\frac{n(y_1 - \mu)^2}{2\sigma^2}}}_{\text{only } y_1} \cdot \underbrace{e^{-\frac{1}{2\sigma^2}\{(\sum_{i=2}^n y_i)^2 + \sum_{i=2}^n y_i^2\}}}_{\text{no } y_1 \text{ dependence}},
\end{aligned}$$

and hence as the PDFs split, we conclude Y_1 independent of Y_2, \dots, Y_n and so \bar{X}_n independent of $(X_2 - \bar{X}_n, \dots, X_n - \bar{X}_n)$ and so in particular of any Borel-measurable function of this vector such as S_n^2 , completing the proof.

For 3, note that

$$\begin{aligned}
V &:= \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n ((X_i - \bar{X}_n) - (\mu - \bar{X}_n))^2 \\
&= \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} + \frac{n(\bar{X}_n - \mu)^2}{\sigma^2} =: W_1 + W_2.
\end{aligned}$$

The first part, W_1 , of this summation is just $(n-1) \frac{S_n^2}{\sigma^2}$, a function of S_n^2 , and the second, W_2 , is a function of \bar{X}_n . By what we've just shown in the previous part, these two are independent. In addition, $V \sim \chi_{(n)}^2$ and

$$W_2 = \frac{n(\bar{X}_n - \mu)^2}{\sigma^2} = \left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \right)^2 \sim \chi_{(1)}^2,$$

since the inner random variable is a standard normal. Then, since W_1, W_2 independent, $M_V(t) = M_{W_1}(t)M_{W_2}(t)$, so for $t < \frac{1}{2}$,

$$M_{W_1}(t) = \frac{M_V(t)}{M_{W_2}(t)} = \frac{(1-2t)^{-\frac{n}{2}}}{(1-2t)^{-\frac{1}{2}}} = (1-2t)^{-\frac{(n-1)}{2}},$$

hence $W_1 \sim \chi^2_{(n-1)}$. ■

↪ **Proposition 2.1:** Let $X \sim t(\nu)$, the Student t -distribution i.e

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu} \cdot \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

then

- $\text{Var}(X) = \frac{\nu}{\nu-2}$ for $\nu > 2$
- If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{(\nu)}$ are independent random variables, then $T = \frac{Z}{\sqrt{V/\nu}} \sim t(\nu)$.

↪ **Theorem 2.5:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then,

$$T = \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim t(n-1).$$

Remark 2.2: By combining CLT and Slutsky's Theorem, T asymptotes to $\mathcal{N}(0,1)$, but this gives a general distribution. Note that for large n , $t(n-1)$ approximately normal too.

PROOF. Notice that

$$W_1 := \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \sim \mathcal{N}(0,1), \quad W_2 := \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

are independent, and

$$T = \frac{W_1}{\sqrt{W_2/(n-1)}}$$

so by the previous proposition $T \sim t(n-1)$. ■

↪ **Proposition 2.2:** Given $U \sim \chi^2_{(m)}, V \sim \chi^2_{(n)}$ independent, then $F = \frac{U/m}{V/n} \sim F(m,n)$. If $T \sim t(\nu)$, $T^2 \sim F(1, \nu)$.

↪ **Theorem 2.6:** Let $X_1, \dots, X_m \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$ be mutually independent random samples. Then,

$$F = \frac{S_m^2/\sigma_1^2}{S_n^2/\sigma_2^2} \sim F(m-1, n-1).$$

PROOF. We have that $U = \frac{(m-1)S_m^2}{\sigma_1^2} \sim \chi_{(m-1)}^2$ and $V = \frac{(n-1)S_n^2}{\sigma_2^2}$ are independent so by the previous proposition

$$F = \frac{U/(m-1)}{V/(n-1)} \sim F(m-1, n-1).$$

■

§2.3 Order Statistics

↪ **Definition 2.4:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$. Then, the *order statistics* are

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)},$$

where $X_{(i)}$ the i th largest of X_1, \dots, X_n .

↪ **Definition 2.5** (Related Functions of Order Statistics): The *sample range* is defined

$$R_n := X_{(n)} - X_{(1)}.$$

The *sample median* is defined

$$M(X_1, \dots, X_n) := \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ odd} \\ \frac{X_{(\frac{n}{2})} + X_{(\frac{n+1}{2})}}{2} & \text{if } n \text{ even.} \end{cases}$$

↪ **Theorem 2.7** (Distribution of Max, Min): Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F, f$.

(Discrete)

$$(a) P(X_{(1)} = x) = [1 - F(x^-)]^n - [1 - F(x)]^n$$

$$(b) P(X_{(n)} = y) = [F(y)]^n - [F(y^-)]^n$$

(Continuous)

$$(c) F_{X_{(1)}}(x) = P(X_{(1)} \leq x) = 1 - [1 - F(x)]^n, \quad f_{X_{(1)}}(x) = n \cdot f(x)[1 - F(x)]^{n-1}$$

$$(d) F_{X_{(n)}}(y) = [F(y)]^n, \quad f_{X_{(n)}}(y) = n \cdot f(y)[F(y)]^{n-1}$$

PROOF. (a) Notice

$$P(X_{(1)} = x) = P(X_{(1)} \leq x) - P(X_{(1)} < x).$$

We have

$$\begin{aligned}
P(X_{(1)} \leq x) &= 1 - P(X_{(1)} > x) \\
&= 1 - P(X_1 > x, X_2 > x, \dots, X_n > x) \\
&= 1 - P(X_1 > x)P(X_2 > x) \cdots P(X_n > x) \\
&= 1 - [1 - F(x)]^n,
\end{aligned}$$

and similarly

$$P(X_{(1)} < x) = 1 - P(X_{(1)} \geq x) = 1 - [1 - F(x^-)]^n,$$

where $F(x^-) = \lim_{z \rightarrow x^-} F(z)$. So in all,

$$P(X_{(1)} = x) = [1 - F(x^-)]^n - [1 - F(x)]^n.$$

(b) is very similar. For (c), we have

$$\begin{aligned}
P(X_{(1)} \leq x) &= 1 - P(X_{(1)} > x) \\
&= 1 - P(X_1 > x, \dots, X_n > x) \\
&= 1 - [1 - F(x)]^n.
\end{aligned}$$

(d) is similar. ■

↪ **Theorem 2.8** (Distribution of j th Order Statistics): Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F, f$.

(Discrete) Suppose the X_i 's take values in $S_x = \{x_1, x_2, \dots\}$ and put $p_i = P(X_i)$. Then,

$$F_{X_{(j)}}(x_i) = P(X_{(j)}(x_i) \leq x_i) = \sum_{k=j}^n \binom{n}{k} p_i^k (1 - p_i)^{n-k},$$

where $P_i = P(X_i \leq x_i) = \sum_{\ell=1}^i p_\ell$.

(Continuous)

$$F_{X_{(j)}}(x) = \sum_{k=j}^n \binom{n}{k} F^k(x) [1 - F(x)]^{n-k},$$

so

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f(x) [F(x)]^{j-1} [1 - F(x)]^{n-j}.$$

PROOF. For discrete, we have

$$P(X_{(j)}(x_i) \leq x_i) = P(\text{at least } j \text{ out of } X_1, \dots, X_n \leq x_i).$$

Then,

$$P(\text{at least } j \text{ out of } X_1, \dots, X_n \leq x_i) = \sum_{k=j}^n \binom{n}{k} p_i^k (1 - p_i)^{n-k}.$$

Continuous is similar. ■

§2.4 Large Sample/Asymptotic Theory

↪ **Definition 2.6** (Convergence in Probability): We say $T_n = T(X_1, \dots, X_n)$ converges in probability to θ $T_n \xrightarrow{P} \theta$ as $n \rightarrow \infty$ if for any $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|T_n - \theta| > \varepsilon) = 0.$$

↪ **Definition 2.7** (Convergence in Distribution): Find a positive sequence $\{r_n\}$ with $r_n \rightarrow \infty$ such that

$$r_n(T_n - \theta) \xrightarrow{d} T,$$

where T a random variable.

↪ **Theorem 2.9** (Slutsky's): Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} a$ for some $a \in \mathbb{R}$. Then,

$$X_n + Y_n \xrightarrow{d} X + a$$

$$X_n Y_n \xrightarrow{d} aX,$$

and if $a \neq 0$,

$$\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{a}.$$

↪ **Theorem 2.10** (Continuous Mapping Theorem (CMT)): Suppose $X_n \xrightarrow{P} X$ and g is continuous on the set C such that $P(X \in C) = 1$. Then,

$$g(X_n) \xrightarrow{P} g(X).$$

⊗ **Example 2.2:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$ with $\mu = \mathbb{E}[X_i]$, $\sigma^2 = \text{Var}(X_i) < \infty$. Then,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{d} \mathcal{N}(0, 1),$$

since we may rewrite

$$\frac{\sqrt{n}(\bar{X}_n - \mu)/\sigma}{S_n/\sigma}.$$

The numerator $\xrightarrow{d} \mathcal{N}(0, 1)$ by CLT. $S_n^2 \xrightarrow{P} \sigma^2$, so the denominator goes to 1 in probability.

↪ **Definition 2.8** (Big O , Little o Notation): Let $\{a_n\}, \{b_n\} \subseteq \mathbb{R}$ real sequences.

- We say $a_n = O(b_n)$ if $\exists 0 < c \in \mathbb{R}$ and $N \in \mathbb{N}$ such that $|\frac{a_n}{b_n}| \leq c$ for every $n \geq N$.
- We say $a_n = o(b_n)$ if $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 0$.

↪ **Definition 2.9** (Big O_p , Little o_p Notation): Let $\{X_n\}, \{Y_n\}$ sequences of random variables.

- We say $X_n = O_p(1)$ if $\forall \varepsilon > 0$ there is a $N_\varepsilon \in \mathbb{N}$ and $C_\varepsilon \in \mathbb{R}$ such that

$$P(|X_n| > C_\varepsilon) < \varepsilon$$

for every $n > N_\varepsilon$.

- We say $X_n = O_p(Y_n)$ if $X_n/Y_n = O_p(1)$.
- We say $X_n = o_p(1)$ if $X_n \xrightarrow{P} 0$.
- We say $X_n = o_p(Y_n)$ if $X_n/Y_n = o_p(1)$.

↪ **Proposition 2.3**: If $X_n \xrightarrow{d} X$, then $X_n = O_p(1)$.

↪ **Proposition 2.4** (The Delta Method (First Order)): Let $\sqrt{n}(X_n - \mu) \xrightarrow{d} V$ and g a real-valued function such that g' exists at $x = \mu$ and $g'(\mu) \neq 0$. Then,

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} g'(\mu)V.$$

In particular, if $V \sim \mathcal{N}(0, \sigma^2)$ then

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, g'(\mu)^2 \sigma^2).$$

PROOF. Taylor expanding the LHS,

$$\sqrt{n}\{g(X_n) - g(\mu)\} = g'(\mu)\sqrt{n}(X_n - \mu) + o_p(1) \rightarrow g'(\mu)V.$$

■

↪ **Proposition 2.5** (The Delta Method (Second Order)): Suppose $\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and $g'(\mu) = 0$ but $g''(\mu) \neq 0$. Then,

$$n\{g(X_n) - g(\mu)\} \xrightarrow{d} \sigma^2 \frac{g''(\mu)}{2} \cdot \chi_{(1)}^2.$$

PROOF.

$$g(X_n) - g(\mu) = \frac{g''(\mu)}{2}(X_n - \mu)^2 + o_p(1),$$

so

$$n(g(X_n) - g(\mu)) = \sigma^2 \frac{g''(\mu)}{2} \left[\frac{\sqrt{n}(X_n - \mu)}{\sigma} \right]^2 + o_p(1).$$

The bracketed term converges in distribution to $\mathcal{N}(0, 1)$ and the $o_p(1)$ term converges in probability to zero, so the proposition follows by applying Slutsky's Theorem. ■

⊗ **Example 2.3:** $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ by CLT. Letting $g(x) = x^2$, and assuming $\mu \neq 0$, then

$$\sqrt{n}(\bar{X}_n^2 - \mu^2) \rightarrow \mathcal{N}(0, 4\mu^2\sigma^2),$$

by the first-order delta method.

↪ **Proposition 2.6:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$, and denote the ECDF $F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_i \leq x)$. Then,

1. $\mathbb{E}[F_n(x)] = F(x)$;
2. $\text{Var}(F_n(x)) = \frac{1}{n}F(x)(1 - F(x))$;
3. $nF_n(x) = \sum_{i=1}^n \mathbb{1}(X_i \leq x) \sim \text{Bin}(n, F(x))$;
4. $\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{d} \mathcal{N}(0, 1)$.
5. $F_n(x) \xrightarrow{P} F(x)$.
6. $P(|F_n(x) - F(x)| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$, by Hoeffding's Inequality.
7. $\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = \|F_n - F\|_\infty \xrightarrow{\text{a.s.}} 0$, by the Glivenko-Cantelli Theorem.
8. $P(\|F_n - F\|_\infty > \varepsilon) \leq C\varepsilon e^{-2n\varepsilon^2}$ for some constant C (Dvoretzky-Kiefer-Wolfowitz Theorem).

Remark 2.3: The constant in 8. was shown to be 2 by Massart.

§3 PARAMETRIC INFERENCE

↪ **Definition 3.1** (Point Estimator): Let X_1, \dots, X_n a random sample. A *point estimator* $\hat{\theta} := \hat{\theta}(X_1, \dots, X_n)$ is an estimator of a parameter θ if it is a statistic.

⊗ **Example 3.1:** Let X be a random variable denoting whether a randomly selected electronic chip is operational or not, i.e. $X = \begin{cases} 1 & \text{operational} \\ 0 & \text{else} \end{cases}$, supposing $X \sim \text{Ber}(\theta)$, then $0 < \theta < 1$ is the probability a randomly selected chip is operational. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$. Then,

$$\mathcal{F} = \{\text{Ber}(\theta) : 0 < \theta < 1\}, \quad \Theta = (0, 1).$$

Then, possible estimators are $\bar{X}_n, \frac{X_1 + X_2}{2}$, just X_2 , etc.

↪ **Definition 3.2** (Bias): An estimator $\hat{\theta}_n$ is an *unbiased* estimator of θ if

$$\mathbb{E}_\theta[\hat{\theta}_n] = \theta, \quad \forall \theta \in \Theta,$$

where the expected value is taken with respect to the distribution of $\hat{\theta}_n$ (and thus depends on the distribution F_θ).

Generally, the *bias* of an estimator $\hat{\theta}_n$ is defined

$$\text{Bias}(\hat{\theta}_n) := \mathbb{E}_\theta[\hat{\theta}_n] - \theta, \quad \theta \in \Theta.$$

If $\hat{\theta}_n$ unbiased, then $\text{Bias}(\hat{\theta}_n) = 0$.

⊗ **Example 3.2:** For instance, recalling the previous example,

$$\mathbb{E}_\theta[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta[X_i] = \frac{1}{n} n\theta = \theta,$$

so \bar{X}_n unbiased. Also,

$$\mathbb{E}_\theta[X_1] = \theta,$$

so just X_1 also unbiased, as is $\frac{X_1 + X_2}{2}$.

⊗ **Example 3.3:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F_\theta$, $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$, $\mu = \mathbb{E}[X_i]$, $\sigma^2 = \text{Var}(X_i)$. Then, $\hat{\mu}_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ an unbiased estimator of μ . Let $\hat{\sigma}_n^2 = S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, then recalling $\mathbb{E}[\hat{\sigma}_n^2] = \sigma^2$, this is an unbiased estimator of σ^2 . However, changing the constant term, to get

$$\hat{\sigma}_n^{*2} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

is biased, with

$$\mathbb{E}_\theta[\hat{\sigma}_n^{*2}] = \frac{n-1}{n} \sigma^2,$$

so

$$\text{Bias}(\hat{\sigma}_n^{*2}) = -\frac{\sigma^2}{n} < 0,$$

i.e. $\hat{\sigma}_n^{*2}$ underestimates the true parameter on average. Of course, in the limit it becomes 0.

⊗ **Example 3.4:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{U}(0, \theta)$, $\theta > 0$, $\Theta = (0, \infty)$. Recall $\mathbb{E}_\theta[X_i] = \frac{\theta}{2}$. Consider

$$\hat{\theta}_{n,1} := 2X_3, \quad \hat{\theta}_{n,2} := 2\bar{X}_n, \quad \hat{\theta}_{n,3} := X_{(n)}.$$

Then, $\mathbb{E}[\hat{\theta}_{n,i}] = \theta$ for $i = 1, 2$ and $\frac{n}{n+1}\theta$ for $i = 3$. Hence, we can scale the last one, $\hat{\theta}_{n,4} := \frac{n+1}{n}\hat{\theta}_{n,3}$, to get an unbiased estimator.

↪ **Definition 3.3** (Mean-Squared Error): The *Mean-Squared Error* (MSE) of an estimator is defined

$$\begin{aligned} \text{MSE}_\theta(\hat{\theta}_n) &:= \mathbb{E}_\theta[(\hat{\theta}_n - \theta)^2] \\ &= \mathbb{E}_\theta\left[\left((\hat{\theta}_n - \mathbb{E}_\theta[\hat{\theta}_n]) + (\mathbb{E}_\theta[\hat{\theta}_n] - \theta)\right)^2\right] \\ &= \text{Var}_\theta(\hat{\theta}_n) + [\text{Bias}(\hat{\theta}_n)]^2. \end{aligned}$$

Remark that if $\mathbb{E}_\theta[\hat{\theta}_n] = \theta$, i.e. $\hat{\theta}_n$ unbiased, then $\text{MSE}_\theta(\hat{\theta}_n) = \text{Var}_\theta(\hat{\theta}_n)$.

↪ **Definition 3.4** (Consistency): We say an estimator $\hat{\theta}_n$ of θ is *consistent* if $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \rightarrow \infty$.

Remark 3.1: There are many ways of establishing consistency; by direct definition of convergence in probability, the WLLN (maybe continuous mapping theorem), or checking if $\mathbb{E}_\theta[\hat{\theta}_n] \rightarrow \theta$ (if this happens we say $\hat{\theta}_n$ “asymptotically unbiased”) and $\text{Var}_\theta(\hat{\theta}_n) \rightarrow 0$ as $n \rightarrow \infty$, for in this case by Chebyshev’s Inequality we have consistency.

⊗ **Example 3.5:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F_\theta$.

1. $\hat{\mu}_n := \bar{X}_n \xrightarrow{P} \mu$ by WLLN, and $S_n^2 \xrightarrow{P} \sigma^2$ similarly.
2. If $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{U}(0, \theta)$, then $\mathbb{E}[X_i] = \frac{\theta}{2}$. Note that $\hat{\theta}_{n,1} = 2\bar{X}_n$ and $\hat{\theta}_{n,2} = \frac{n+1}{n}X_{(n)}$ are both unbiased estimators of θ , and both are consistent. To see the second one, we have that for any $\varepsilon > 0$,

$$\begin{aligned} P(|X_{(n)} - \theta| > \varepsilon) &= P(\theta - X_{(n)} > \varepsilon) \\ &= P(X_{(n)} < \theta - \varepsilon) \\ &= \left(\frac{\theta - \varepsilon}{\theta}\right)^n \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

We have too that

$$\text{MSE}_\theta(\hat{\theta}_{n,1}) = \text{Var}_\theta(\hat{\theta}_{n,1}) = 4\text{Var}_\theta(\bar{X}_n) = \frac{4}{n} \text{Var}(X_i) = \frac{4}{n} \frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Also

$$\begin{aligned} \text{MSE}_\theta(\hat{\theta}_{n,2}) &= \text{Var}_\theta(\hat{\theta}_{n,2}) = \left(\frac{n+1}{n}\right)^2 \text{Var}(X_{(n)}) \\ &= \dots = \frac{\theta^2}{n(n+2)} = \frac{\theta^2}{3n} \cdot \frac{3}{n+2} \leq \text{MSE}_\theta(\hat{\theta}_{n,1}) \quad \forall n \geq 1. \end{aligned}$$

We will focus on the class of unbiased estimators of a real-valued parameter, $\tau(\theta)$, $\tau : \Theta \rightarrow \mathbb{R}$.

§3.1 Uniformly Minimum Variance Unbiased Estimators (UMVUE), Cramér-Rau Lower Bound (CRLB)

↪ **Definition 3.5** (UMVUE): Let $\mathbf{X} = (X_1, \dots, X_n)^t$ be a random variable with a joint pdf/pmf given by

$$p_\theta(\mathbf{x}) = p_\theta(x_1, \dots, x_n),$$

where θ some parameter in $\Theta \subseteq \mathbb{R}^d$. An estimator $T(\mathbf{X})$ of a real valued parameter $\tau(\theta) : \Theta \rightarrow \mathbb{R}$ is said to be a UMVUE of $\tau(\theta)$ if

1. $\mathbb{E}_\theta[T(\mathbf{X})] = \tau(\theta)$ for every $\theta \in \Theta$;
2. for any other unbiased estimator $T^*(\mathbf{X})$ of $\tau(\theta)$, we have

$$\text{Var}_\theta(T(\mathbf{X})) \leq \text{Var}_\theta(T^*(\mathbf{X})), \quad \forall \theta \in \Theta.$$

↪ **Proposition 3.1** (Cramér-Rau Lower Bound): We define in the case $d = 1$ ($\Theta \subseteq \mathbb{R}$) for convenience. Assume that

(1) the family $\{p_\theta : \theta \in \Theta\}$ has a common support $S = \{x \in \mathbb{R}^n : p_\theta(x) > 0\}$ that does not depend on θ ;

(2) for $x \in S, \theta \in \Theta, \frac{d}{d\theta} \log p_\theta(x) < \infty$;

(3) for any statistic $h(x)$ with $\mathbb{E}_\theta[|h(x)|] < \infty$ for every $\theta \in \Theta$, we have

$$\frac{d}{d\theta} \int_S h(x) p_\theta(x) dx = \int_S h(x) \frac{d}{d\theta} p_\theta(x) dx,$$

whenever the right-hand side is finite.

Let $T(X)$ be such that $\text{Var}_\theta(T(X)) < \infty$ and $\mathbb{E}_\theta[T(X)] = \tau(\theta)$ for every $\theta \in \Theta$. Then if $0 < \mathbb{E}_\theta \left[\left(\frac{d}{d\theta} \log(p_\theta(x)) \right)^2 \right] < \infty$ for every $\theta \in \Theta$, then the Cramér-Rao Lower Bound (CRLB) holds:

$$\text{Var}_\theta(T(X)) \geq \frac{[\tau'(\theta)]^2}{\mathbb{E}_\theta \left[\left(\frac{d}{d\theta} \log p_\theta(x) \right)^2 \right]}, \quad \forall \theta \in \Theta.$$

Remark 3.2: The quantity

$$I(\theta) := \mathbb{E}_\theta \left[\left(\frac{d}{d\theta} \log(p_\theta(x)) \right)^2 \right]$$

is called the *Fisher information* contained in X about θ .

PROOF. Note that $\tau(\theta) = \mathbb{E}_\theta[T(X)]$ implies

$$\begin{aligned} \tau'(\theta) &= \frac{d}{d\theta} \mathbb{E}[T(X)] \\ &= \frac{d}{d\theta} \left[\int_S T(x) p_\theta(x) dx \right] \\ \text{by ass. 2, 3} \quad &= \int_S T(x) \frac{d}{d\theta} p_\theta(x) dx \\ &= \int_S T(x) \frac{d}{d\theta} [\log p_\theta(x)] p_\theta(x) dx \\ &= \mathbb{E}_\theta \left[T(X) \frac{d}{d\theta} \log p_\theta(X) \right], \quad \forall \theta \in \Theta. \quad (\text{I}) \end{aligned}$$

On the other hand, by (3) with $h \equiv 1$, then

$$\begin{aligned}
0 &= \int_S \frac{d}{d\theta} p_\theta(x) dx = \int_S \left[\frac{d}{d\theta} \log p_\theta(x) \right] p_\theta(x) dx \quad \forall \theta \in \Theta \\
&\Rightarrow \mathbb{E}_\theta \left[\frac{d}{d\theta} \log p_\theta(X) \right] = 0. \quad (\text{II})
\end{aligned}$$

Combining (I) and (II),

$$\tau'(\theta) = \text{Cov}_\theta \left(T(X), \frac{d}{d\theta} \log p_\theta(x) \right),$$

since $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, but the second of these terms vanishes by (II). Thus,

$$[\tau'(\theta)]^2 = \text{Cov}_\theta^2 \left(T(X), \frac{d}{d\theta} \log p_\theta(X) \right).$$

By Cauchy-Schwarz, we find

$$\begin{aligned}
[\tau'(\theta)]^2 &\leq \text{Var}_\theta(T(X)) \text{Var}_\theta \left(\frac{d}{d\theta} \log p_\theta(X) \right) \\
&\leq \text{Var}_\theta(T(X)) \mathbb{E}_\theta \left\{ \left[\frac{d}{d\theta} \log p_\theta(X) \right]^2 \right\},
\end{aligned}$$

the last line following by the Bartlett Identity. ■

Remark 3.3: If $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f_\theta$, then $p_\theta(x) = \prod_{i=1}^n f(x_i; \theta)$, and

$$\begin{aligned}
I(\theta) &= \mathbb{E}_\theta \left\{ \left[\frac{d}{d\theta} \log p_\theta(X) \right]^2 \right\} = \mathbb{E}_\theta \left\{ \left[\sum_{i=1}^n \frac{d}{d\theta} \log f(X_i; \theta) \right]^2 \right\} \\
&= \underbrace{n \mathbb{E}_\theta \left\{ \left(\frac{d}{d\theta} \log f(X_1; \theta) \right)^2 \right\}}_{=I_1(\theta)},
\end{aligned}$$

so the CRLB in this case reads

$$\text{Var}_\theta(T(X)) \geq \frac{[\tau'(\theta)]^2}{nI_1(\theta)},$$

and moreover if $\tau(\theta) = \theta$ itself,

$$\text{Var}_\theta(T(X)) \geq \frac{1}{nI_1(\theta)}.$$

⊗ **Example 3.6:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$, so $f(x; \theta) = \theta^x (1 - \theta)^{1-x}$ for $x = 0, 1$. Then,

$$\log(f(x; \theta)) = x \log(\theta) + (1 - x) \log(1 - \theta)$$

so

$$\frac{d}{d\theta} \log(f(x; \theta)) = \frac{x}{\theta} - \frac{1-x}{1-\theta},$$

so the Fisher information in one X_1 is given

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

For any unbiased estimator of $\tau(\theta) = \theta$, the CRLB gives

$$\text{Var}_\theta(T(\mathbf{X})) \geq \frac{1}{nI_1(\theta)} = \frac{\theta(1-\theta)}{n}.$$

Recall our estimator $\hat{\theta}_n = \bar{X}_n$. We have that $\text{Var}_\theta(\bar{X}_n) = \frac{1}{n} \text{Var}_\theta(X_1) = \frac{\theta(1-\theta)}{n}$.

Remark 3.4: If p_θ additionally twice differentiable in θ and $\mathbb{E}_\theta \left\{ \frac{d}{d\theta} \log p_\theta(\mathbf{X}) \right\}$ is also differentiable under the \mathbb{E}_θ ,

$$\frac{d}{d\theta} \log p_\theta(\mathbf{X}) = \int \frac{d}{d\theta} \left\{ \left[\frac{d}{d\theta} \log p_\theta(x) \right] p_\theta(x) \right\} dx.$$

In particular, this implies $\int p_\theta''(x) dx = 0$. Then,

$$I(\theta) = \mathbb{E}_\theta \left\{ \left[\frac{d}{d\theta} \log p_\theta(\mathbf{X}) \right]^2 \right\} = -\mathbb{E}_\theta \left\{ \frac{d^2}{d\theta^2} p_\theta(\mathbf{X}) \right\},$$

making it easier to compute $I(\theta)$. This follows from the fact that

$$\frac{d^2}{d\theta^2} \log p_\theta(x) = \frac{p_\theta''(x)}{p_\theta(x)} - \left[\frac{d}{d\theta} \log p_\theta(x) \right]^2,$$

and so taking the expected value of both sides cancels the inner-most term by the differentiability condition of p_θ ;

$$\begin{aligned} \mathbb{E} \left[\frac{d^2}{d\theta^2} \log p_\theta(x) \right] &= \mathbb{E} \left[\frac{p_\theta''(x)}{p_\theta(x)} \right] - \mathbb{E} \left[\left[\frac{d}{d\theta} \log p_\theta(x) \right]^2 \right] \\ &= \int \cancel{p_\theta''(x)} dx - I(\theta). \end{aligned}$$

⊗ **Example 3.7:** Returning to the previous example, remark that

$$\frac{d^2}{d\theta^2} \log(f(x; \theta)) = -\frac{x}{\theta^2} - \frac{x-1}{(1-\theta)^2},$$

and so

$$\mathbb{E} \left[\frac{d^2}{d\theta^2} \log f(x; \theta) \right] = \frac{1}{\theta} + \frac{1}{1-\theta}$$

so $I_1(\theta) = \frac{1}{\theta(1-\theta)}$ as we found before.

Remark 3.5: The CRLB is *not* a sharp bound, in the sense that the UMVUE for a particular parameter may be strictly larger than the CRLB.

⊗ **Example 3.8:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \theta^2)$. Then, $\hat{\mu}_n$ the UMVUE for μ . If μ known, then $\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ is the UMVUE for σ^2 . If μ is unknown, then $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ would be the UMVUE for σ^2 .

However, if $X_i \stackrel{\text{iid}}{\sim} \exp(\beta)$, with $f(x; \beta) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$ for $x > 0$, S_n^2 is not the UMVUE for $\text{Var}_\beta(X_i) = \beta^2$.

↪ **Theorem 3.1** (Attaining the CRLB): Suppose $\mathbf{X} = (X_1, \dots, X_n) \sim p_\theta$. Let $T(\mathbf{X})$ be unbiased for $\tau(\theta)$. Then, $T(\mathbf{X})$ attains the CRLB if and only if

$$a(\theta)\{T(\mathbf{x}) - \tau(\theta)\} = \frac{d}{d\theta} \log p(\mathbf{x}; \theta),$$

for some function $a(\theta)$, for every $\theta \in \Theta$ and \mathbf{x} in the support of p .

PROOF. In the proof of the CRLB, the only inequality arose from using Cauchy-Schwarz with bounding the covariance of $T(\mathbf{X})$ and $\frac{d}{d\theta} \log p_\theta(\mathbf{X})$. Equality in this inequality holds if and only if the terms are linearly dependent, namely if there is some function $a(\theta)$ and $b(\theta)$ such that $a(\theta)T(\mathbf{x}) + b(\theta) = \frac{d}{d\theta} \log p_\theta(\mathbf{x})$.

On the other hand,

$$\mathbb{E}_\theta\{a(\theta)T(\mathbf{X}) + b(\theta)\} = \mathbb{E}_\theta\left\{\frac{d}{d\theta} \log p_\theta(\mathbf{x})\right\} = 0 \Rightarrow b(\theta) = -\mathbb{E}_\theta\{a(\theta)T(\mathbf{X})\} = -a(\theta)\tau(\theta),$$

so combining these two gives the desired linear relation. ■

⊗ **Example 3.9** (Exponential family): $X_i \stackrel{\text{iid}}{\sim} f(x; \theta) = h(x)c(\theta) \exp\{\omega(\theta)T_1(x)\}$, where h a nonnegative function of only x and c a nonnegative function of only θ , with the support of f being independent of θ . Then

$$p_\theta(\mathbf{x}) = \prod_{i=1}^n f(x_i; \theta) = \left[\prod_{i=1}^n h(x_i) \right] (c(\theta))^n \exp\left(\omega(\theta) \sum_{i=1}^n T_1(x_i) \right).$$

Taking the log:

$$\begin{aligned} \frac{d}{d\theta} \log p_\theta(\mathbf{x}) &= n \frac{c'(\theta)}{c(\theta)} + \omega'(\theta) \sum_{i=1}^n T_1(x_i) \\ &= \omega'(\theta) \left\{ \sum_{i=1}^n T_1(x_i) - \frac{nc'(\theta)}{c(\theta)\omega'(\theta)} \right\}. \end{aligned}$$

Let

$$\tau(\theta) = -\frac{c'(\theta)}{c(\theta)\omega'(\theta)}.$$

Then, since

$$\mathbb{E}_\theta \left[\frac{d}{d\theta} \log p_\theta(\mathbf{x}) \right] = 0,$$

then

$$\mathbb{E}_\theta \left[\sum_{i=1}^n T_1(X_i) \right] = n\tau(\theta),$$

so

$$T(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n T_1(X_i)$$

is a UMVUE for $\tau(\theta)$ by the previous theorem.

⊗ **Example 3.10:** Let $X_i \stackrel{\text{iid}}{\sim} \text{Poisson}(\theta)$ so

$$f(x; \theta) = \frac{e^{-\theta}}{x!} \theta^x = \frac{e^{-\theta}}{x!} e^{x \log(\theta)},$$

with support $x \in \{0, 1, \dots\}$. Then, we notice that with

$$h(x) = \frac{1}{x!}, c(\theta) = e^{-\theta}, \omega(\theta) = \log(\theta), T_1(x) = x,$$

that X_i in the exponential family. Then, according to the previous example,

$$\tau(\theta) = -\frac{-e^{-\theta}}{e^{-\theta} \frac{1}{\theta}} = \theta,$$

has UMVUE

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

⊗ **Example 3.11:** Recall we found, for $X_i \stackrel{\text{iid}}{\sim} \mathcal{U}(0, \theta)$, that $\hat{\theta}_n := \frac{n+1}{n} X_{(n)}$ was an unbiased estimator but cannot obtain the CRLB since the regularity conditions are not satisfied (namely, the support of the pdfs depends on the parameter). Moreover, we found

$$\mathbb{E}_{\theta} \left\{ \frac{n+1}{n} X_{(n)} \right\} = \theta, \text{Var}_{\theta} \left\{ \frac{n+1}{n} X_{(n)} \right\} = \frac{\theta^2}{n(n+2)}.$$

If we temporarily ignore that we cannot apply CRLB, we would find

$$\text{CRLB} = \frac{1}{n I_1(\theta)} = \frac{\theta^2}{n},$$

so our estimator actually has a “better” variance. We’ll see later that this estimator actually the UMVUE.

§3.2 Sufficiency

We can’t always find unbiased estimators; here we look for other ways for comparing different estimators.

⊗ **Example 3.12:** Let $X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and consider the following estimators of σ^2 :

$$S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

$$S_2^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

$$S_3^2 = \frac{1}{n+1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

One verifies these have respective means, variances

	S_1^2	S_2^2	S_3^2
\mathbb{E}	$\frac{n-1}{n} \sigma^2$	σ^2	$\frac{n-1}{n+1} \sigma^2$
Var	$\frac{2(n-1)\sigma^4}{n^2}$	$\frac{2\sigma^4}{n-1}$	$\frac{2(n-1)\sigma^4}{(n+1)^2}$

. We notice then that

$$\text{MSE}(S_3^2) < \text{MSE}(S_2^2) < \text{MSE}(S_1^2),$$

so despite the fact that S_2^2 is unbiased, it does not minimize the MSE.

↪ **Definition 3.6** (Sufficiency): Suppose $\mathbf{X} = (X_1, \dots, X_n)$ has joint pdf (pmf) $p(\mathbf{x}; \theta)$ for $\theta \in \Theta$. A statistic $T(\mathbf{X}) : \mathbb{R}^n \supseteq \mathbf{X} \rightarrow S_T \subseteq \mathbb{R}^k, k \leq n$, is *sufficient* for θ or the parametric family $\{p_\theta : \theta \in \Theta\}$ if the conditional distribution of (X_1, \dots, X_n) given $T(\mathbf{X}) = t$ for any $\theta \in \Theta$ and $t \in S_T$ in the support such that $P_\theta(t \in S_T) = 1$, does not depend on θ . Namely,

$$f_{\mathbf{X}|T(\mathbf{X})=t}(x_1, \dots, x_n),$$

does *not* depend on θ .

⊗ **Example 3.13:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$. Let $T(\mathbf{X}) = \sum_{i=1}^n X_i$. We know that then $T(\mathbf{X}) \sim \text{Bin}(n, \theta)$. We claim T sufficient; we have

$$f_\theta(x_1, \dots, x_n | T(\mathbf{X}) = t) = \begin{cases} \frac{1}{\binom{n}{t}} & \text{if } \sum_{i=1}^n x_i = t \\ 0 & \text{else} \end{cases},$$

which is independent of θ so indeed sufficient.

Remark 3.6: A sufficient statistic induces a partitioning of the sample space $X \subseteq \mathbb{R}^n$; namely,

$$X = \bigcup_{t \in S_T} \Pi_t,$$

such that

$$\Pi_t = \{x = (x_1, \dots, x_n) \in X \mid T(x) = t\},$$

and S_T the support of T .

⊗ **Example 3.14:** Return to the Bernoulli example from before, and consider specifically the case when $n = 2$, so $T(X) = X_1 + X_2$ is a sufficient statistic as we showed. Then, the sample space is given by

$$X = \{(0, 0), (0, 1), (1, 0), (1, 1)\},$$

and T has support

$$T(x) = x_1 + x_2 \in \{0, 1, 2\} =: S_T.$$

This induces the partitioning

$$X = \Pi_0 \sqcup \Pi_1 \sqcup \Pi_2 = \{(0, 0)\} \sqcup \{(0, 1), (1, 0)\} \sqcup \{(1, 1)\}.$$

↪ **Theorem 3.2** (Neyman-Fisher Factorization Theorem): Let $X = (X_1, \dots, X_n)^t$ be a random vector with a joint pdf/pmf $p_\theta(x) = p(x; \theta)$. A statistic $T(X)$ is sufficient for θ if and only if there exist functions $g(\cdot; \theta)$ and $h(\cdot)$ such that

$$p_\theta(x) = h(x) \cdot g(\theta, T(x)),$$

for every $\theta \in \Theta$ and $x \in X$.

Note that g depends on x *only* through $T(x)$, and h does *not* depend on θ .

PROOF. We prove in the discrete case.

Note that

$$f_{X|T(X)=t_x}(x) = \frac{P_\theta(X_1 = x_1, \dots, X_n = x_n, T(X) = t_x)}{P_\theta(T(X) = t_x)},$$

for every x such that $T(x) = t_x$, and 0 otherwise;

$$= \frac{P_\theta(X_1 = x_1, \dots, X_n = x_n)}{\sum_{y=(y_1, \dots, y_n): T(y)=t_x} P(X_1 = y_1, \dots, X_n = y_n)}.$$

If $T(X)$ a sufficient statistic for θ , then the above ratio, by definition, does not depend on θ ; hence, putting $h(x)$ to be the ratio above, it is independent of θ (is only a function of the data), and if we take g to be the denominator of the ratio above, then g depends on the data only through T . Hence, we can write $p_\theta(x) = h(x) \cdot g(t_x; \theta)$.

Conversely, suppose $p_\theta(x) = g(T(x); \theta)h(x)$. Then,

$$f_{X|T(X)=t_x}(x; \theta) = \frac{g(t_x; \theta)h(x)}{\sum_{y: T(y)=t_x} g(T(y); \theta)h(y)} = \frac{h(x)}{\sum_{y: T(y)=t_x} h(y)},$$

which depends only on x and hence $T(X)$ a sufficient statistic. ■

⊗ **Example 3.15:** Let again $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$ so

$$p_\theta(x_1, \dots, x_n) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i} \prod_{i=1}^n \mathbb{1}\{x_i \in \{0, 1\}\}.$$

for $x_i = 0, 1$.

One notices that the LHS (not the product) can be written as a function of θ and $\sum_{i=1}^n x_i$ only, and the remaining term is independent of θ . Hence by the previous theorem $T(X) = \sum_{i=1}^n X_i$ a sufficient statistic for θ .

⊗ **Example 3.16:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{U}(0, \theta)$, so $f(x; \theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ 0 & \text{else} \end{cases}$. Then

$$\begin{aligned} p_\theta(x) &= \prod_{i=1}^n \frac{1}{\theta} \mathbb{1}(0 < x_i < \theta) \\ &= \underbrace{\frac{1}{\theta^n} \mathbb{1}(0 < x_{(n)} < \theta)}_{=: g(T(x; \theta))} \underbrace{\mathbb{1}(0 < x_{(1)} < \theta)}_{=: h(x)}, \end{aligned}$$

so $X_{(n)}$ is a sufficient statistic for θ .

Remark 3.7: If T is a sufficient statistic for θ and $T(X) = \Phi(T^*(X))$ where Φ is a measurable function and T^* another statistic, then T^* is also a sufficient statistic.

⊗ **Example 3.17:** In the exponential family, we claim $T(X_1, \dots, X_n) = \sum_{i=1}^n T_1(X_i)$.

⊗ **Example 3.18:** Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $\theta = (\mu, \sigma^2)$ both unknown. Using the factorization theorem, we can see that

$$T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right)$$

is a sufficient statistic for θ , as is (\bar{X}_n, S_n^2) .

Remark 3.8: This does *not* imply that say $\sum_{i=1}^n X_i$ sufficient for μ ! Namely T is a sufficient statistic for the 2-dimensional parameter θ . We cannot simply separate the dependence.

⊗ **Example 3.19:** Recall the Bernoulli example once again. We claim that

$$T_m^*(X) = \left(\sum_{i=1}^m X_i, \sum_{i=m+1}^n X_i \right), \quad 1 \leq m \leq n-1$$

is also sufficient for $0 < \theta < 1$. Clearly this is no different then just using the one-dimensional statistic $\sum_{i=1}^n X_i$; we'd like to formalize how to differentiate such statistics. Namely, $\sum_{i=1}^n X_i$ is called a *minimal* sufficient statistic for θ .

↪ **Definition 3.7** (Minimal Sufficient Statistic): A statistic $T(X)$ is a *minimal sufficient statistic* for θ iff

- $T(X)$ is sufficient;
- For any other sufficient statistic $T^*(X)$ of θ , $T(X)$ is a function of $T^*(X)$, i.e.

$$T(X) = \varphi(T^*(X)),$$

where $\varphi(\cdot)$ some measurable function, or equivalently, $\forall x, y \in X \subseteq \mathbb{R}^n$, if $T^*(x) = T^*(y)$ then $T(x) = T(y)$.

Remark 3.9: If $T(X)$ minimally sufficient and induces a partitioning

$$X = \bigcup_{t \in S_T} \Pi_t, \quad \Pi_t := \{x \in X : T(x) = t\}$$

and $T^*(X)$ any sufficient statistic that induces a partitioning

$$X = \bigcup_{t^* \in S_{T^*}} \Pi_{t^*}^*, \quad \Pi_{t^*}^* := \{x \in X : T^*(x) = t^*\},$$

then we find that $\forall t^* \in S_{T^*}$, there is some $t \in S_T$ such that $\Pi_{t^*}^* \subseteq \Pi_t$; namely, the partition induced by $T(X)$ is the *coarsest* possible partition of X .

↪ **Theorem 3.3** (Lehmann-Scheffé): For a parametric family $p_\theta(\cdot)$ (the joint pdf/pmf of X), suppose a statistic $T(X) = T(X_1, \dots, X_n)$ is such that for every $x, y \in X \subseteq \mathbb{R}^n$ $T(x) = T(y) \Leftrightarrow \frac{p_\theta(x)}{p_\theta(y)}$ does not depend on θ . Then, $T(X)$ is a minimal sufficient statistic for θ .

⊗ **Example 3.20:** Suppose $X_i \stackrel{\text{iid}}{\sim} \mathcal{U}(0, \theta)$, then $p_\theta(\mathbf{x}) = \frac{1}{\theta^n} \mathbb{1}\{x_{(n)} < \theta\} \mathbb{1}\{x_{(1)} > 0\}$; then $T(\mathbf{X}) = X_{(n)}$ is a sufficient statistic for θ . For any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we find

$$\frac{p_\theta(\mathbf{x})}{p_\theta(\mathbf{y})} = \frac{\mathbb{1}\{x_{(n)} < \theta\} \mathbb{1}\{x_{(1)} > 0\}}{\mathbb{1}\{y_{(n)} < \theta\} \mathbb{1}\{y_{(1)} > 0\}},$$

which does not depend on θ iff $x_{(n)} = y_{(n)}$ iff $T(\mathbf{x}) = T(\mathbf{y})$ and therefore by the previous theorem $T(\mathbf{X})$ is a minimally sufficient statistic.

⊗ **Example 3.21:** If $X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $\theta = (\mu, \sigma^2)$, it can be shown that

$$T(\mathbf{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right)$$

is a minimal sufficient statistic for θ . Any one-to-one function of a minimally sufficient statistic also minimally sufficient, hence this implies (\bar{X}_n, S_n^2) is also minimally sufficient for θ .

§3.3 Completeness

↪ **Definition 3.8** (Completeness): Let X be a random variable with a pmf/pdf belonging to a parametric family $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. This family is said to be *complete* if for any measurable function g with $\mathbb{E}_\theta[g(X)] < \infty$, then $\mathbb{E}_\theta[g(X)] = 0$ for all $\theta \in \Theta$ implies $P_\theta(g(X) = 0) = 1$.

A statistic $T(\mathbf{X}) = T(X_1, \dots, X_n)$ is said to be *complete* if the family of its distributions is complete.

Remark 3.10: Complete and sufficient \Rightarrow minimal, but minimally sufficient may not be complete, as we'll see.

⊗ **Example 3.22:** Let $X_i \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$, then note $T(\mathbf{X}) = \sum_{i=1}^n X_i \sim \text{Bin}(n, \theta)$. Let g a measurable function. Then,

$$\begin{aligned} 0 = \mathbb{E}_\theta[g(\mathbf{X})] &\Rightarrow 0 = \sum_{t=0}^n g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} \\ &= \cancel{(1-\theta)^n} \sum_{t=0}^n g(t) \binom{n}{t} \left(\frac{\overbrace{\theta}^{=: \eta}}{1-\theta} \right)^t \\ &= \sum_{t=0}^n g(t) \binom{n}{t} \eta^t. \end{aligned}$$

Then, this is just a polynomial in η , which, being equal to zero implies all the coefficients $g(t) \binom{n}{t} = 0$ for every t and hence $g(t) = 0$. Hence, $T(\mathbf{X})$ is a complete statistic.

⊗ **Example 3.23:** If $X \sim \mathcal{N}(0, \theta)$, the family is not complete. For instance with $g(x) := x$, $\mathbb{E}_\theta(X) = 0$ but $g(x)$ is not identically zero. On the other hand, $T(\mathbf{X}) = X^2$ is a complete statistic. To see this, we know $\frac{X^2}{\theta} \sim \chi_{(1)}^2$, so

$$\begin{aligned} \mathbb{E}_\theta(g(T)) = 0 &\Rightarrow 0 = \int_0^\infty g(t) f_T(t; \theta) dt \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\theta}} g(t) t^{-\frac{1}{2}} e^{-\frac{t}{2\theta}} dt \\ &= \mathcal{L} \left\{ g(t) t^{-\frac{1}{2}} \frac{1}{\sqrt{2\pi\theta}} \right\}. \end{aligned}$$

By uniqueness of the Laplace transform, it must be that $g(t) t^{-\frac{1}{2}} \equiv 0$ hence $g(t) = 0$ and thus $T(\mathbf{X}) = X^2$ is a complete statistic.

⊗ **Example 3.24:** In the exponential family, $\sum_{i=1}^n T_1(X_i)$ is a complete statistic.

Note that an unbiased estimator of a parameter of interest may not even exist. For instance,

⊗ **Example 3.25:** If $X \sim \text{Bin}(n, \theta)$, let $\tau(\theta) = \frac{1}{\theta}$. If $\delta(X)$ is an unbiased estimator of $\tau(\theta)$, we must have $\mathbb{E}_\theta[\delta(X)] = \frac{1}{\theta}$ i.e.

$$\sum_{x=0}^n \delta(x) \binom{n}{x} \theta^x (1-\theta)^{n-x} = \frac{1}{\theta}.$$

As $\theta \rightarrow 0$, the left-hand side will just be $\delta(0)$, while the right-hand side will diverge to ∞ , so no such estimator exists.

↪ **Theorem 3.4** (Rao-Blackwell): Let $U(X)$ be an unbiased estimator of $\tau(\theta)$ and let $T(X)$ be a sufficient statistic for the parametric family. Set

$$\delta(t) = \mathbb{E}_\theta[U(X) | T(X) = t], \quad t \in S_T.$$

Then,

- $\delta(T(X))$ is a statistic, i.e. only depends on X ;
- $\mathbb{E}_\theta[\delta(T(X))] = \tau(\theta)$;
- $\text{Var}_\theta(\delta(T(X))) \leq \text{Var}_\theta[U(X)]$.

PROOF.

- $\delta(T(X)) = \mathbb{E}_\theta[U(X)|T(X)]$ is a random variable in its own right, and is a statistic because $T(X)$ is sufficient, hence conditioning on $T(X)$ will result in no reliance on θ .
- $\mathbb{E}_\theta[\delta(T(X))] = \mathbb{E}_\theta[\mathbb{E}_\theta[U(X)|T(X)]] = \mathbb{E}_\theta[U(X)] = \tau(\theta)$ (using the law of total expectation), since $U(X)$ is an unbiased estimator of $\tau(\theta)$.
- Using the law of total variance, we find

$$\begin{aligned} \text{Var}_\theta(U(X)) &= \text{Var}_\theta(\underbrace{\mathbb{E}_\theta[U(X)|T(X)]}_{=\delta(T(X))}) + \mathbb{E}_\theta[\text{Var}_\theta(U(X)|T(X))] \\ &= \text{Var}_\theta[\delta(T(X))] + \mathbb{E}_\theta[\underbrace{\text{Var}_\theta(U(X)|T(X))}_{\geq 0}] \\ &\geq \text{Var}_\theta[\delta(T(X))]. \end{aligned}$$

■

Remark 3.11: This theorem gives a systematic manner of improving unbiased estimators, by taking an unbiased estimator and a sufficient statistic, and “Rao-Blackwell-izing”, leading to a uniform improvement in variance.

↪ **Theorem 3.5** (Lehmann-Scheffé: Uniqueness): Let $T(X)$ be a complete sufficient statistic. Let $U(X) = h(T(X))$, for a measurable function h , an unbiased estimator of $\tau(\theta)$ such that $\mathbb{E}_\theta[U(X)^2] < \infty$. Then, $U(X)$ is the unique unbiased estimator of $\tau(\theta)$ with the smallest variance in the class of unbiased estimators of $\tau(\theta)$.

PROOF. By the Rao-Blackwell Theorem, it suffices to restrict attention to unbiased estimators that are only functions of $T(X)$; for any other such unbiased statistic, applying Rao-Blackwell to it results in a new statistic with smaller variance.

Now, let $V(X) = h^*(T(X))$ be any other unbiased estimator of $\tau(\theta)$. Then,

$$\mathbb{E}_\theta[V(X)] = \mathbb{E}_\theta[U(X)] = \tau(\theta)$$

hence

$$\mathbb{E}_\theta[V(\mathbf{X}) - U(\mathbf{X})] = \mathbb{E}_\theta[h^*(T(\mathbf{X})) - h(T(\mathbf{X}))] = 0.$$

Let $g(T(\mathbf{X})) = h^*(T(\mathbf{X})) - h(T(\mathbf{X}))$; then, since $T(\mathbf{X})$ complete, it must be that $P(g = 0) = 1$ i.e.

$$P(h(T(\mathbf{X})) = h^*(T(\mathbf{X}))) = 1,$$

so $U(\mathbf{X}), V(\mathbf{X})$ are almost surely identical, hence we indeed have uniqueness. ■

Remark 3.12: This, combined with the Rao-Blackwell theorem, provides a method for obtaining the UMVUE for $\tau(\theta)$ starting with a complete sufficient statistic and an unbiased statistic.

⊗ **Example 3.26:** Let $X_i \stackrel{\text{iid}}{\sim} \text{Ber}(\theta), i = 1, \dots, n$ and $\hat{\theta}_n = \bar{X}_n$. This is unbiased, and $\sum_{i=1}^n X_i$ is a complete and sufficient statistic. Hence, $\hat{\theta}_n$ is a unbiased estimator that is a function of a complete and sufficient statistic and thus is the UMVUE for θ by the Lehmann-Scheffé Theorem.

⊗ **Example 3.27:** Let $X_i \stackrel{\text{iid}}{\sim} \text{Pos}(\theta), i = 1, \dots, n$ and $\hat{\theta}_n = \bar{X}_n$. This is unbiased, and again $\sum_{i=1}^n X_i$ is a complete sufficient statistic hence $\hat{\theta}_n$ is the UMVUE of θ .

Suppose now $\tau(\theta) = P_\theta(X = 0) = e^{-\theta}$; can we obtain a UMVUE for this (function of) a parameter? Define

$$U(X_1) = \mathbb{1}\{X_1 = 0\},$$

which will be unbiased for $\tau(\theta)$. We already have a complete and sufficient statistic. Applying now the Rao-Blackwell theorem, we obtain

$$\delta(t) = \mathbb{E}_\theta \left[U(X_1) \mid \sum_{j=1}^n X_j = t \right].$$

One verifies that

$$\left(X_i \mid \sum_{j=1}^n X_j = t \right) \sim \text{Bin} \left(t, \frac{1}{n} \right),$$

therefore

$$\delta(t) = P_\theta(X_1 = 0 \mid T(\mathbf{X}) = t) = \left(1 - \frac{1}{n} \right)^t.$$

So, $\delta(T(\mathbf{X})) = \left(1 - \frac{1}{n} \right)^{\sum_{i=1}^n X_i}$ is the UMVUE of $e^{-\theta}$. Remark that

$$\delta(T(\mathbf{X})) = \left(1 - \frac{1}{n} \right)^{n\bar{X}_n} \approx e^{-\bar{X}_n} \text{ for large } n.$$

⊗ **Example 3.28:** Let $X_i \stackrel{\text{iid}}{\sim} \text{Ber}(\theta), i = 1, \dots, n$, and suppose $\tau(\theta) = \text{Var}(X_i) = \theta(1 - \theta)$. Recall the UMVUE for θ is $\hat{\theta}_n$. Note that

$$T(\mathbf{X}) = \sum_{i=1}^n X_i \sim \text{Bin}(n, \theta),$$

is complete and sufficient. We know $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = U(\mathbf{X})$ is unbiased for $\tau(\theta)$. We may write

$$\begin{aligned} U(\mathbf{X}) &= \frac{1}{n-1} \left[\sum_{i=1}^n X_i^2 - n\bar{X}_n^2 \right] \\ \text{since } X_i \in \{0, 1\} \quad &= \frac{1}{n-1} \left[\sum_{i=1}^n X_i - n\bar{X}_n^2 \right] \\ &= \frac{1}{n-1} \left(T(\mathbf{X}) - \frac{T^2(\mathbf{X})}{n} \right) \\ &= \frac{n}{n-1} \bar{X}_n (1 - \bar{X}_n) \end{aligned}$$

Hence, $U(\mathbf{X})$ a function of $T(\mathbf{X})$, a complete sufficient statistic, and $U(\mathbf{X})$ is unbiased, so we conclude $U(\mathbf{X})$ the UMVUE for $\tau(\theta)$.

§3.4 Existence of a UMVUE

↪ **Definition 3.9** (Unbiased Estimators of Zero): An estimator $\delta(\mathbf{X})$ satisfying $\mathbb{E}_\theta[\delta(\mathbf{X})] = 0$ is called an *unbiased estimator of zero*.

↪ **Theorem 3.6:** An estimator $U(\mathbf{X})$ of $\tau(\theta) = \mathbb{E}_\theta[U(\mathbf{X})]$ is the best unbiased estimator iff $U(\mathbf{X})$ is uncorellated with all unbiased estimators of zero, i.e.

$$\text{Cov}_\theta(U(\mathbf{X}), \delta(\mathbf{X})) = \mathbb{E}_\theta[U(\mathbf{X})\delta(\mathbf{X})] = 0$$

for every $\delta(\mathbf{X})$ such that $\mathbb{E}_\theta[\delta(\mathbf{X})] = 0$.

PROOF. (Necessity) Let $U(\mathbf{X})$ be a UMVUE of $\tau(\theta)$ and $\delta(\mathbf{X})$ any unbiased estimator of zero. Then $U^*(\mathbf{X}) = U(\mathbf{X}) + a\delta(\mathbf{X})$ for some nonzero $a \in \mathbb{R}$ is also an unbiased estimator $\tau(\theta)$;

$$\mathbb{E}_\theta[U^*(\mathbf{X})] = \mathbb{E}_\theta[U(\mathbf{X})] + a\mathbb{E}_\theta[\delta(\mathbf{X})] = \mathbb{E}_\theta[U(\mathbf{X})] = \tau(\theta).$$

Now,

$$\text{Var}_\theta[U^*(\mathbf{X})] = \text{Var}_\theta[U(\mathbf{X})] + a^2\text{Var}_\theta[\delta(\mathbf{X})] + 2a\text{Cov}_\theta[U(\mathbf{X}), \delta(\mathbf{X})].$$

If this covariance term is non-zero for some θ_0 , then we may choose some a such that

$$a^2\text{Var}_{\theta_0}[\delta(\mathbf{X})] + 2a\text{Cov}_{\theta_0}[U(\mathbf{X}), \delta(\mathbf{X})] < 0$$

i.e.

$$a \in \left\{ \begin{pmatrix} 0, -2 \frac{\text{Cov}_{\theta_0}(U(X), \delta(X))}{\text{Var}_{\theta_0}(\delta(X))} \\ -2 \frac{\text{Cov}_{\theta_0}(U(X), \delta(X))}{\text{Var}_{\theta_0}(\delta(X))}, 0 \end{pmatrix} \right\}'$$

which ever makes sense. Hence,

$$\text{Var}_{\theta_0}[U^*(X)] < \text{Var}_{\theta_0}(U(X)),$$

a contradiction to the minimality of the variance of $U(X)$ hence the covariance term must be zero.

(Sufficiency) Suppose that $\mathbb{E}_{\theta}[U(X), \delta(X)] = 0$ for every θ . Let $U'(X)$ be any arbitrary unbiased estimator, then since $U'(X) = U(X) + (U'(X) - U(X))$, then since $(U'(X) - U(X))$ an unbiased estimator of zero, we find

$$\begin{aligned} \text{Var}_{\theta}[U'(X)] &= \text{Var}_{\theta}[U(X)] + \text{Var}_{\theta}[(U'(X) - U(X))] + \underbrace{2\text{Cov}_{\theta}[U(X), U'(X) - U(X)]}_{=0 \text{ by assumption}} \\ &\geq \text{Var}_{\theta}[U(X)], \end{aligned}$$

for every θ . ■

Remark 3.13: This theorem can be used to investigate the existence of a UMVUE of $\tau(\theta)$, or to determine that an estimator is *not* a UMVUE.

⊗ **Example 3.29:** Let $X \sim \text{unif}\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$ for $\theta \in \mathbb{R}$. Let $\delta(X)$ be an unbiased estimator of zero. Then,

$$0 = \mathbb{E}_\theta[\delta(X)] = \int_{\theta - \frac{1}{2}}^{\theta + \frac{1}{2}} \delta(x) dx, \quad \forall \theta \in \mathbb{R}.$$

Hence, it must be that $\delta\left(\theta + \frac{1}{2}\right) - \delta\left(\theta - \frac{1}{2}\right) = 0$ (taking the derivative of the above with respect to θ) or moreover $\delta(x) = \delta(x + 1)$ for every $x \in \mathbb{R}$. Letting now $U(X)$ be a UVMUE of $\tau(\theta)$, then by the previous theorem it must be that $\text{Cov}_\theta(U(X), \delta(X)) = 0$ for any $\theta \in \mathbb{R}$, i.e.

$$0 = \mathbb{E}_\theta[U(X)\delta(X)].$$

Hence, $U(X)\delta(X)$ also an unbiased estimator of zero so also has the property that $U(x)\delta(x) = U(x + 1)\delta(x + 1)$. δ also unbiased for zero so $\delta(x) = \delta(x + 1)$, so it must be that

$$U(x) = U(x + 1), \quad \forall x \in \mathbb{R}.$$

But also, $U(X)$ is unbiased for $\tau(\theta)$, so

$$\mathbb{E}_\theta[U(X)] = \int_{\theta - \frac{1}{2}}^{\theta + \frac{1}{2}} U(x) dx = \tau(\theta) \Rightarrow \tau'(\theta) = U\left(\theta + \frac{1}{2}\right) - U\left(\theta - \frac{1}{2}\right).$$

But since $U\left(\theta + \frac{1}{2}\right) = U\left(\theta - \frac{1}{2}\right)$ by the remarks above, it follows that $\tau'(\theta) = 0$ so $\tau(\theta)$ is a constant, for some $c \in \mathbb{R}$. We conclude, thus, that there is no UMVUE for any non-constant function $\tau(\theta)$.

§4 SYSTEM PARAMETER ESTIMATION

This chapter is devoted to systematic manners of deriving estimators for particular statistical models.

§4.1 Method of Moments

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f_\theta$ with $\theta = (\theta_1, \dots, \theta_d) \in \Theta \subseteq \mathbb{R}^d$ such that $\mathbb{E}_\theta[|X_i|^d] < \infty$. Let $\mu_j(\theta) = \mathbb{E}_\theta[X_1^j]$ for $j = 1, \dots, d$, the non-central moments. Also define

$$m_j(\mathbf{X}) := \frac{1}{n} \sum_{i=1}^n X_i^j,$$

the *non-central sample moments*. Note that $\mathbb{E}_\theta[m_j(\mathbf{X})] = \mu_j(\theta)$ and by the iid assumption, WLLN implies $m_j(\mathbf{X}) \xrightarrow{P} \mu_j(\theta)$.

Typically, $\mu_j(\theta) = h_j(\theta_1, \dots, \theta_d)$ for some real-valued function $h_j(\cdot)$ for each $j = 1, \dots, d$. The Method of Moments (MM) gives estimates of $\theta_1, \dots, \theta_d$ by solving the following system of equations:

$$m_j(\mathbf{X}) = \mu_j(\theta) = h_j(\theta_1, \dots, \theta_d), \quad j = 1, \dots, d,$$

and solving for each θ_j as a function of the data. In general, this yields

$$\hat{\theta}_j(\mathbf{X}) = g_j(m_1(\mathbf{X}), \dots, m_d(\mathbf{X})), \quad j = 1, \dots, d.$$

These $\hat{\theta}_1, \dots, \hat{\theta}_d$ are the so-called *MM estimators* of $\theta_1, \dots, \theta_d$. In general, these may 1) have no solutions, 2) have a unique solution, 3) have multiple solutions.

⊗ **Example 4.1:** Let $X_i \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$. Then $\mu_1(\theta) = \theta$ and $m_1(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i$. Setting $\mu_1 = m_1$ gives that $\hat{\theta}_n = \bar{X}_n$.

⊗ **Example 4.2:** Let $X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$. Then,

$$\begin{cases} m_1(\mathbf{X}) = \bar{X}_n \\ m_2(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases} \quad \begin{cases} \mu_1(\theta) = \mu \\ \mu_2(\theta) = \sigma^2 + \mu^2 \end{cases}$$

which gives a system of equations

$$\begin{cases} \bar{X}_n = \mu \\ \frac{1}{n} \sum_{i=1}^n X_i^2 = \sigma^2 + \mu^2 \end{cases}$$

This yields

$$\hat{\mu}_n = \bar{X}_n, \quad \hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

⊗ **Example 4.3:** Let $X_i \stackrel{\text{iid}}{\sim} \mathcal{U}(-\theta, \theta)$. Then, $\mathbb{E}_\theta[X_i] = 0$, so we need to move onto to the second moment. We have $\mathbb{E}_\theta[X_i^2] = \text{Var}_\theta[X_i] = \frac{\theta^2}{3}$. $m_2(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i^2$, so we have system of equations

$$\frac{1}{n} \sum_{i=1}^n X_i^2 = \frac{\hat{\theta}_n^2}{3},$$

which has solution

$$\hat{\theta}_n = \sqrt{\frac{3}{n} \sum_{i=1}^n X_i^2}.$$

Note that we have positive and negative roots, but ignore the negative one since $\theta > 0$.

⊗ **Example 4.4:** Let $X_i \stackrel{\text{iid}}{\sim} \text{Geo}(p)$, so $f(x; p) = p(1-p)^{x-1}$ with $x = 1, 2, \dots$. Then, $\mathbb{E}_p[X_i] = \frac{1}{p}$ and $m_1(\mathbf{X}) = \bar{X}_n$, so

$$\hat{p}_n = \frac{1}{\bar{X}_n}.$$

It it an unbiased estimator?

$$\mathbb{E}\left[\frac{1}{\bar{X}_n}\right] = \sum_{x_1=1} \cdots \sum_{x_n=1} \frac{1}{\frac{1}{n} \sum_{i=1}^n x_i} p^n (1-p)^{x_1+\cdots+x_n-n}.$$

Suppose now $n = 1$ so $X \sim \text{Geo}(p)$. Assume $T(X)$ is an unbiased estimator of p . Then,

$$p = \sum_{x=1}^{\infty} T(x) f(x; p) = \sum_{x=1}^{\infty} T(x) p(1-p)^{x-1}$$

so it must be $T(1) = 1, T(x) = 0$ for $x \geq 2$, since these are two polynomials in p . This is an “unreasonable” estimator.

§4.2 Maximum Likelihood Estimation (MLE)

Let $\mathbf{X} = (X_1, \dots, X_n)^t$ have a joint pdf/pmf $p_{\theta}(\mathbf{x})$ for $\theta \in \Theta \subseteq \mathbb{R}^d$ and $\mathbf{x} \in \mathbf{X} \subseteq \mathbb{R}^n$.

↪ **Definition 4.1** (Likelihood Function): Having observed (post-experimental data) $\mathbf{x} = (x_1, \dots, x_n) \in \mathbf{X}$, the *likelihood function*

$$L_n : \Theta \rightarrow [0, \infty),$$

is given by

$$L_n(\theta; \mathbf{x}) = L_n(\theta) := p_{\theta}(\mathbf{x}).$$

Note that \mathbf{x} is fixed in this definition; L_n a function of θ .

The *log-likelihood* is then defined $\ell_n(\theta) := \log(L_n(\theta))$.

Remark 4.1: If $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f_{\theta}$, then

$$L_n(\theta) = \prod_{i=1}^n f(x_i; \theta),$$

so

$$\ell_n(\theta) = \sum_{i=1}^n \log(f(x_i; \theta)).$$

Remark 4.2: Some texts write $L_n(\theta) = c \cdot p_{\theta}(\mathbf{x})$ for some constant $c > 0$, a proportionality constant. It is not a pdf; θ varies, and \mathbf{x} is fixed.

Remark 4.3: If $T(X)$ a sufficient statistic for θ , it contains all the necessary information needed to compute the likelihood function (by the factorization theorem).

Remark 4.4: The *likelihood principle* states “in light of the data x , the likelihood contains all the information in the data about θ ”. In addition, two likelihood functions contain the same information about θ if they are proportional to each other.

⊗ **Example 4.5:** Consider the following two experiments; Exp. 1: a coin was tossed 20 times and 8 heads observed. Let $X_i \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$ for $i = 1, \dots, 20$ and $Y = \sum_{i=1}^{20} X_i \sim \text{Bin}(20, \theta)$. Then

$$L_1(\theta) = P(Y = 8) = \binom{20}{8} \theta^8 (1 - \theta)^{12} \propto \theta^8 (1 - \theta)^{12}.$$

In Exp. 2., we toss 20 coins until 8 heads are observed. So, this is a negative binomial distribution, and we find

$$L_2(\theta) = \binom{19}{7} \theta^7 (1 - \theta)^{12} \propto \theta^8 (1 - \theta)^{12},$$

so $L_1(\theta), L_2(\theta)$ are both proportional to $\theta^8 (1 - \theta)^{12}$.

From a “maximum likelihood estimation point of view”, both likelihoods contains the same information about θ .

Remark 4.5: In the discrete case, $L_n(\theta)$ is the probability of observing x , given distribution with parameter θ ; in particular, if $L_n(\theta_1) > L_n(\theta_2)$, this means we were more likely to observe our data if the parameter value was θ_1 rather than θ_2 . A “similar” interpretation can be made in the continuous case.

↔ **Definition 4.2** (Maximum Likelihood Estimation): Given $x = (x_1, \dots, x_n)$, $\hat{\theta}_n = \hat{\theta}(x_1, \dots, x_n)$ is called a *maximum likelihood estimate* (MLE) of θ if it maximizes $L_n(\theta)$ or equivalently $\ell_n(\theta)$. I.e., $\hat{\theta}_n(x) = \arg\max_{\theta \in \Theta} L_n(\theta)$.

If $\hat{\theta}_n$ exists and $\hat{\theta}_n : X \rightarrow \Theta$ is measurable, then $\hat{\theta}(X_1, \dots, X_n)$ is called the *maximum likelihood estimator* of θ .

⊗ **Example 4.6:** Let $X_i \stackrel{\text{iid}}{\sim} \text{Poisson}(\theta)$, then

$$L_n(\theta) = \prod_{i=1}^n f(x_i; \theta) = \prod_{i=1}^n e^{-\theta} \frac{\theta^{x_i}}{x_i!},$$

then

$$\ell_n(\theta) = - \sum_{i=1}^n \ln(x_i!) - n\theta + n\bar{x}_n \ln(\theta).$$

Then,

$$\frac{d\ell_n(\theta)}{d\theta} = -n + \frac{n\bar{x}_n}{\theta} = 0 \Rightarrow \bar{x}_n = \theta \Rightarrow \hat{\theta}_n = \bar{x}_n.$$

Moreover, since

$$\frac{d^2(\ell_n(\theta))}{d\theta^2} = -\frac{n\bar{x}_n}{\theta^2} < 0,$$

it follows that $\hat{\theta}_n = \bar{x}_n$ is the maximum likelihood estimate of θ .

Remark 4.6:

1. MLE of θ may or may not exist over Θ , when Θ is open. It always exists over the closure of Θ .
2. If Θ is finite, then certainly $\Theta = \overline{\Theta}$ and the MLE always exists and can be computed by comparing the values of $L_n(\theta)$ (or $\ell_n(\theta)$) over Θ .
3. If $L_n(\theta)$ or $\ell_n(\theta)$ is differentiable on Θ° , then possible candidates for the MLE are values of $\theta \in \Theta^\circ$ that satisfy the so-called “likelihood equations” or “score equations”,

$$\frac{d\ell_n(\theta)}{d\theta} = 0. \quad \otimes$$

If $\ell_n(\theta)$ not differentiable (or in particular not everywhere differentiable), then extrema may occur at non-differentiability or even discontinuity points of $\ell_n(\theta)$. So, its crucial to analyze the entire likelihood function to find its maximum.

⊗ **Example 4.7:** Let $X_i \stackrel{\text{iid}}{\sim} \mathcal{U}(0, \theta)$ so $f(x; \theta) = \frac{1}{\theta} \mathbb{1}\{0 < x < \theta\}$. Then,

$$L_n(\theta) = \prod_{i=1}^n \frac{1}{\theta} \mathbb{1}\{0 < x_i < \theta\} = \frac{1}{\theta^n} \mathbb{1}\{x_{(1)} > 0\} \cdot \mathbb{1}\{\theta > x_{(n)}\}.$$

Then, $L_n(\theta)$ is strictly decreasing on $(x_{(n)}, \infty)$ and equal to zero on $(0, x_{(n)})$. Hence, the MLE of θ is $\hat{\theta}_n(x_1, \dots, x_n) = x_{(n)}$.

⊗ **Example 4.8:** Let $X_i \stackrel{\text{iid}}{\sim} \mathcal{U}\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$, then

$$L_n(\theta) = \prod_{i=1}^n \mathbb{1}\left\{\theta - \frac{1}{2} < x_i < \theta + \frac{1}{2}\right\} = \mathbb{1}\left\{x_{(n)} - \frac{1}{2} < \theta < x_{(1)} + \frac{1}{2}\right\}.$$

So, any choice of $\hat{\theta}_n \in \left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right]$ is an MLE of θ , for instance $\frac{x_{(1)} + x_{(n)}}{2}$ (the midpoint). In short, the MLE is *not* unique in this case.

⊗ **Example 4.9:** Let $X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$. Then

$$L_n(\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right),$$

so its more convenient to consider

$$\ell_n(\theta) = -\frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 + \text{const. indep. of } \theta.$$

Then, the likelihood equations give

$$\begin{cases} \frac{\partial \ell_n(\theta)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \\ \frac{\partial \ell_n(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0 \end{cases}.$$

The first equation gives

$$\hat{\mu}_n = \bar{x}_n,$$

and so the second gives

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2.$$

Then, we find

$$\frac{\partial^2 \ell_n(\theta)}{\partial \theta^2} = -\begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} < 0,$$

a negative-definite matrix, hence $\hat{\theta}_n = \left(\bar{x}_n, \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2\right)$ is the MLE of $\theta = (\mu, \sigma^2)$.

⊗ **Example 4.10:** Let $X_i \stackrel{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$ with $\theta = (\alpha, \beta)$, with pdf $f(x; \theta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right)$ for $x > 0$. Then

$$L_n(\theta) = [\Gamma(\alpha)\beta^\alpha]^{-n} \left(\prod_{i=1}^n x_i \right)^{\alpha-1} \exp\left(-\sum_{i=1}^n \frac{x_i}{\beta}\right),$$

so

$$\ell_n(\theta) = -n \log(\Gamma(\alpha)) - n\alpha \log(\beta) + (\alpha - 1) \sum_{i=1}^n \log(x_i) + \text{const indep of } \theta.$$

The likelihood equations:

$$\frac{\partial \ell_n(\theta)}{\partial \theta} = 0 \Rightarrow \begin{cases} \frac{\partial \ell_n(\theta)}{\partial \alpha} = -n \log(\beta) - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^n \log(x_i) = 0 \\ \frac{\partial \ell_n(\theta)}{\partial \beta} = -\frac{n\alpha}{\beta} + \sum_{i=1}^n \frac{x_i}{\beta^2} = 0. \end{cases}$$