# **MATH357** - Statistics

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

## Contents

## §1 Review of Probability

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

We call a nonnegative function $P : \mathcal{F} \to \mathbb{R}$ 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 $\Omega$, then $P\left(\bigcup_{n\geq 1} E_n\right) = \sum_{n\geq 1} P(E_n)$. We call the tuple $(\Omega, \mathcal{F}, P)$ a *probability space*.

↪**Definition 1.2** (Random Variables): Fix a probability space $(\Omega, \mathcal{F}, P)$. A Borel-measurable function $X : \Omega \to \mathbb{R}$ (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 $S := \{x_1, x_2, \ldots\} \subset \mathbb{R}$, 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 $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ for every $x \in \mathbb{R}$. Then,

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

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

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

$$\mu_X=\mathbb{E}[X]=\begin{cases}\sum_{x\in S}xf(x)\\ \int_S xf(x)\,\mathrm{d}x\end{cases}.$$

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

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

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 $\mathbb{E}[\cdot]$ linear in its argument.

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

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

where they exist.

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

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

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

↪**Definition 1.5** (Multiple Random Variable): $X = (X_1, ..., X_n) : \Omega \to \mathbb{R}^n$ 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" $I = (-\infty, a_1] \times \cdots \times (-\infty, a_n]$, as before.)

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

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

In particular, the marginal distribution of $X_i$ is given by

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

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

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

We have the notion of a joint mgf,

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

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

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

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

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

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

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

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

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

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

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

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 $X \overset{d}{=} Y$ (note that $X$ need not equal $Y$ pointwise). If $X_1, ..., X_n$ a collection of random variables that are independent and identically distributed with common cdf $F$, we write $X_1, ..., X_n \overset{\text{iid}}{\sim} F$.

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

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

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

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

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

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

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

↪**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 $\mathbb{E}\left[X^2\right] < \infty$, then $\mathrm{Var}(X) = \mathrm{Var}(\mathbb{E}[X|Y]) + \mathbb{E}[\mathrm{Var}(X|Y)]$. In particular, $\mathrm{Var}(X) \geq \mathrm{Var}(\mathbb{E}[X|Y])$.

# §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 $\Theta$ 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 noting 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 $X_1, ..., X_n \overset{\text{iid}}{\sim} F$. Then $X_1, ..., X_n$ 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 $X_1, ..., X_n \overset{\text{iid}}{\sim} F$ where $X_i$ a $d$-dimensional random vector. Let

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

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

⊛ **Example 2.1**: $\overline{X_n} := \frac{1}{n}\sum_{i=1}^n X_i$ (the "sample mean") and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X_n}\right)^2$, (the "sample variance") are both typical statistics.

## §2.2 Properties of Normal and other Common Distributions

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

(a) $\operatorname{argmin}_{\alpha\in\mathbb{R}}\left\{\sum_{i=1}^n (x_i - \alpha)^2\right\} = \overline{x_n}$;

(b) $\sum_{i=1}^n (x_i - \overline{x_n})^2 = \sum_{i=1}^n \left(x_i^2\right) - n\overline{x_n}^2$;

(c) $\sum_{i=1}^n (x_i - \overline{x_n}) = 0$.

↪**Theorem 2.2**: Let $X_1, ..., X_n \overset{\text{iid}}{\sim} F$, and $g : \mathbb{R} \to \mathbb{R}$ 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 $X_1, ..., X_n \overset{\text{iid}}{\sim} F$ with $\sigma^2 < \infty$, then

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

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

1. $\overline{X_n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$;
2. $\overline{X_n}, S_n^2$ are independent;
3. $\frac{(n-1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^n \left(X_i - \overline{X_n}\right)^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

**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 $\left(X_2 - \overline{X}_n, ..., X_n - \overline{X}_n\right)$:

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

Then, it suffices to show that $\overline{X}_n$ and $\left(X_2 - \overline{X}_n, \ldots, X_n - \overline{X}_n\right)$ are independent.

Consider now the transformation

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

which gives Jacobian

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

and so

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

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

For 3, note that

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

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 $\overline{X}_n$. By what we've just shown in the previous part, these two are independent. In addition, $V \sim \chi^2_{(n)}$ and

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

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

- $\mathrm{Var}(X) = \frac{\nu}{\nu-2}$ 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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then,

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

**Remark 2.2**: By combing 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

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

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 $X_1, ..., X_m \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_1, ..., Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$ 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 $U = \frac{(m-1)S_m^2}{\sigma_1^2} \sim \chi^2_{(m-1)}$ 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 $X_1, ..., X_n \overset{\text{iid}}{\sim} F$. Then, the *order statistics* are

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

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

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

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

The *sample median* is defined

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

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

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

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

Proof. (a) Notice

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

We have

$$\begin{aligned} P\big(X_{(1)} \le x\big) &= 1 - P\big(X_{(1)} > x\big) \\ &= 1 - P(X_1 > x, X_2 > x, \ldots, X_n > x) \\ &= 1 - P(X_1 > x)P(x_2 > x)\cdots P(X_n > x) \\ &= 1 - [1 - F(x)]^n, \end{aligned}$$

and similarly

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

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

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

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

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

(d) is similar. ■

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

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

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

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

*(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

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

Then,

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

Continuous is similar. ■

## §2.4 Large Sample/Asymptotic Theory

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

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

↪**Definition 2.7** (Convergence in Distribution): Find a positive sequence $\{r_n\}$ with $r_n \to \infty$ 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 $X_1, ..., X_n \overset{\text{iid}}{\sim} F$ with $\mu = \mathbb{E}[X_i], \sigma^2 = \text{Var}(X_i) < \infty$. Then,

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

since we may rewrite

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

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 $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$.

↪**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

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

Proof. Taylor expanding the LHS,

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

■

↪**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,

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

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

$$\sqrt{n}\left(\overline{X}_n^2 - \mu^2\right) \to \mathcal{N}(0, 4\mu^2\sigma^2),$$

by the first-order delta method.

↪**Proposition 2.6**: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$, 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 Hoeffing'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 $X_1, \ldots, X_n$ a random sample. A *point estimator* $\hat{\theta} := \hat{\theta}(X_1, \ldots, X_n)$ is an estimator of a parameter $\theta$ 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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Ber}(\theta)$. Then,

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

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

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

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

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

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

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

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

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

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

so $\overline{X}_n$ 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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$, $\theta = \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix}$, $\mu = \mathbb{E}[X_i], \sigma^2\ \text{Var}(X_i)$. Then, $\hat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ an unbiased estimator of $\mu$. Let $\hat{\sigma}_n^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$, then recalling $\mathbb{E}\left[\hat{\sigma}_n^2\right] = \sigma^2$, this is an unbiased estimator of $\sigma^2$. However, changing the constant term, to get

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

is biased, with

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

so

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

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

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

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

Then, $\mathbb{E}\left[\hat{\theta}_{n,i}\right] = \theta$ 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

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

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

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

**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 $\mathbb{E}_\theta\left[\hat{\theta}_n\right] \to \theta$ (if this happens we say $\hat{\theta}_n$ "asymptotically unbiased") and $\text{Var}_\theta\left(\hat{\theta}_n\right) \to 0$ as $n \to \infty$, for in this case by Chebyshev's Inequality we have consistency.

⊛ **Example 3.5**: Let $X_1, ..., X_n \overset{\text{iid}}{\sim} F_\theta$.

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

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

We have too that

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

Also

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

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

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

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

$$p_\theta(\boldsymbol{x}) = p_\theta(x_1, ..., x_n),$$

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

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

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

↪**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 $S = \{\boldsymbol{x} \in \mathbb{R}^n : p_\theta(\boldsymbol{x}) > 0\}$ that does not depend on $\theta$;

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

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

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

whenever the right-hand side is finite.

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

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

**Remark 3.2**: The quantity

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

is called the *Fisher information* contained in $\boldsymbol{X}$ about $\theta$.

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

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

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

$$0 = \int_S \frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_S \left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(\boldsymbol{x})\right] p_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \forall\,\theta\in\Theta$$

$$\Rightarrow \mathbb{E}_\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(\boldsymbol{X})\right] = 0. \qquad \text{(II)}$$

Combining (I) and (II),

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

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

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

By Cauchy-Schwarz, we find

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

the last line following by the Bartlett Identity. ■

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

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

so the CRLB in this case reads

$$\mathrm{Var}_\theta(\mathrm{T}(\boldsymbol{X})) \ge \frac{\left[\tau'(\theta)\right]^2}{nI_1(\theta)},$$

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

$$\mathrm{Var}_\theta(\mathrm{T}(\boldsymbol{X})) \ge \frac{1}{nI_1(\theta)}.$$

⊛ **Example 3.6**: Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, 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

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

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

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

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

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

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

In particular, this implies $\int p''_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$. Then,

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

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

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

and so taking the expected value of both sides cancels the inner-most term by the differentiability condition of $p_\theta$;

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

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

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

and so

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

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 $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \theta^2)$. Then, $\hat{\mu}_n$ the UMVUE for $\mu$. If $\mu$ known, then $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ is the UMVUE for $\sigma^2$. If $\mu$ is unknown, then $\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$ would be the UMVUE for $\sigma^2$.

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

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

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

for some function $a(\theta)$, for every $\theta \in \Theta$ and $\boldsymbol{x}$ 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 $\mathrm{T}(\boldsymbol{X})$ and $\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(\boldsymbol{X})$. 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 $a(\theta)T(\boldsymbol{x}) + b(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(\boldsymbol{x})$.

On the other hand,

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

so combining these two gives the desired linear relation. ■

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

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

Taking the log:

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

Let

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

Then, since

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

then

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

so

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

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

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

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

with support $x \in \{0, 1, ...\}$. 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

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

⊛ **Example 3.11**: Recall we found, for $X_i \overset{\text{iid}}{\sim} \mathcal{U}(0, \theta)$, 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

$$\text{CRLB} = \frac{1}{nI_1(\theta)} = \frac{\theta^2}{n},$$

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 $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, and consider the following estimators of $\sigma^2$:

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

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

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

One verifies these have respective means, variances

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

. We notice then that

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

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

↪**Definition 3.6** (Sufficiency): Suppose $\boldsymbol{X} = (X_1, ..., X_n)$ has joint pdf (pmf) $p(\boldsymbol{x}; \theta)$ for $\theta \in \Theta$. A statistic $T(\boldsymbol{X}) : \mathbb{R}^n \supseteq X \to S_T \subseteq \mathbb{R}^k$, $k \leq n$, is *sufficient* for $\theta$ or the parametric family $\{p_\theta : \theta \in \Theta\}$ if the conditional distribution of $(X_1, ..., X_n)$ given $T(\boldsymbol{X}) = t$ 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 $\theta$. Namely,

$$f_{\boldsymbol{X}|T(\boldsymbol{X})=t}(x_1, ..., x_n),$$

does *not* depend on $\theta$.

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

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

which is independent of $\theta$ so indeed sufficient.

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

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

such that

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

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 $T(\boldsymbol{X}) = X_1 + X_2$ is a sufficient statistic as we showed. Then, the sample space is given by

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

and $T$ has support

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

This induces the partitioning

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

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

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

for every $\theta \in \Theta$ and $\boldsymbol{x} \in \mathrm{X}$.

Note that $g$ depends on $\boldsymbol{x}$ *only* through $T(\boldsymbol{x})$, and $h$ does *not* depend on $\theta$.

Proof. We prove in the discrete case.

Note that

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

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

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

If $T(\boldsymbol{X})$ a sufficient statistic for $\theta$, then the above ratio, by definition, does not depend on $\theta$; hence, putting $h(\boldsymbol{x})$ to be the ratio above, it is independent of $\theta$ (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 $p_\theta(\boldsymbol{x}) = h(\boldsymbol{x}) \cdot g(t_x; \theta)$.

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

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

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

⊛ **Example 3.15**: Let again $X_1, ..., X_n \overset{\text{iid}}{\sim} \text{Ber}(\theta)$ so

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

for $x_i = 0, 1$.

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

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

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

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

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

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

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

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

is a sufficient statistic for $\theta$, as is $\left(\overline{X}_n, S_n^2\right)$.

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

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

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

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 $\theta$.

↪**Definition 3.7** (Minimal Sufficient Statistic): A statistic $T(\boldsymbol{X})$ is a *minimal sufficient statistic* for $\theta$ iff

- $T(\boldsymbol{X})$ is sufficient;
- For any other sufficient statistic $T^*(\boldsymbol{X})$ of $\theta$, $T(\boldsymbol{X})$ is a function of $T^*(\boldsymbol{X})$, i.e.

  $$T(\boldsymbol{X}) = \varphi(T^*(\boldsymbol{X})),$$

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

**Remark 3.9**: If $T(\boldsymbol{X})$ minimally sufficient and induces a partitioning

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

and $T^*(\boldsymbol{X})$ any sufficient statistic that induces a partitioning

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

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 $T(\boldsymbol{X})$ is the *coarsest* possible partition of X.

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

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

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

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

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

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

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

## §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 $\mathbb{E}_\theta[g(\boldsymbol{X})] < \infty$, then $\mathbb{E}_\theta[g(\boldsymbol{X})] = 0$ for all $\theta \in \Theta$ implies $P_\theta(g(X) = 0) = 1$.

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

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

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

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

Then, this is just a polynomial in $\eta$, which, being equal to zero implies all the coefficients $g(t)\binom{n}{t} = 0$ for every $t$ and hence $g(t) = 0$. Hence, $T(\boldsymbol{X})$ 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, $T(X) = X^2$ is a complete statistic. To see this, we know $\frac{X^2}{\theta} \sim \chi^2_{(1)}$, so

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

By uniqueness of the Laplace transform, it must be that $g(t)t^{-\frac{1}{2}} \equiv 0$ hence $g(t) = 0$ and thus $T(X) = X^2$ 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 $\theta \to 0$, the left-hand side will just be $\delta(0)$, while the right-hand side will diverge to $\infty$, so no such estimator exists.

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

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

Then,

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

Proof.

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

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

■

**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 $T(\boldsymbol{X})$ be a complete sufficient statistic. Let $U(\boldsymbol{X}) = h(T(\boldsymbol{X}))$, for a measurable function $h$, an unbiased estimator of $\tau(\theta)$ such that $\mathbb{E}_\theta\left[U(\boldsymbol{X})^2\right] < \infty$. Then, $U(\boldsymbol{X})$ 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 $T(\boldsymbol{X})$; for any other such unbiased statistic, applying Rao-Blackwell to it results in a new statistic with smaller variance.

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

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

hence

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

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

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

so $U(\boldsymbol{X}), V(\boldsymbol{X})$ 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 $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, $i = 1, \ldots, n$ and $\hat{\theta}_n = \overline{X}_n$. 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 $\theta$ by the Lehmann-Scheffé Theorem.

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

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

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

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

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

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

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

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

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

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

## §3.4 Existence of a UMVUE

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

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

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

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

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

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

Now,

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

If this covariance term is non-zero for some $\theta_0$, then we may choose some $a$ such that

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

i.e.

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

which ever makes sense. Hence,

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

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

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

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

for every $\theta$. ■

**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,

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

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 $\theta$) 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 $U(x)\delta(x) = U(x+1)\delta(x+1)$. $\delta$ also unbiased for zero so $\delta(x) = \delta(x+1)$, so it must be that

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

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

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

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 Systematic Parameter Estimation

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

### §4.1 Method of Moments

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

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

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

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

$$m_j(\boldsymbol{X}) = \mu_j(\theta) = h_j(\theta_1, ..., \theta_d), \qquad j = 1, ..., d,$$

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

$$\hat{\theta}_j(\boldsymbol{X}) = g_j(m_1(\boldsymbol{X}), ..., m_d(\boldsymbol{X})), \qquad j = 1, ..., d.$$

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

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

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

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

which gives a system of equations

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

This yields

$$\hat{\mu}_n = \overline{X}_n, \qquad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2.$$

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

$$\frac{1}{n}\sum_{i=1}^n X_i^2 = \frac{\widehat{\theta_n}^2}{3},$$

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 $X_i \overset{\text{iid}}{\sim} \text{Geo}(p)$, so $f(x;p) = p(1-p)^{x-1}$ with $x = 1, 2, \ldots$. Then, $\mathbb{E}_p[X_i] = \frac{1}{p}$ and $m_1(\boldsymbol{X}) = \overline{X}_n$, so

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

It it an unbiased estimator?

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

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

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

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

is given by

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

Note that $\boldsymbol{x}$ is fixed in this definition; $L_n$ a function of $\theta$.

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

**Remark 4.1**: If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta$, 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 $L_n(\theta) = c \cdot p_\theta(\boldsymbol{x})$ for some constant $c > 0$, a proportionality constant. It is not a pdf; $\theta$ varies, and $\boldsymbol{x}$ is fixed.

**Remark 4.3**: If $T(\boldsymbol{X})$ a sufficient statistic for $\theta$, 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 $\boldsymbol{x}$, the likelihood contains all the information in the data about $\theta$". In addition, two likelihood functions contain the same information about $\theta$ 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 $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$ for $i = 1, ..., 20$ 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 a "maximum likelihood estimation point of view", both likelihoods contains the same information about $\theta$.

**Remark 4.5**: In the discrete case, $L_n(\theta)$ is the probability of observing $\boldsymbol{x}$, given distribution with parameter $\theta$; 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 $\theta_1$ rather than $\theta_2$. A "similar" interpretation can be made in the continuous case.

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

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

⊛ **Example 4.6**: Let $X_i \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$, 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,

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

Moreover, since

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

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

**Remark 4.6**:

1. MLE of $\theta$ may or may not exist over $\Theta$, when $\Theta$ is open. It always exists over the closure of $\Theta$.
2. If $\Theta$ is finite, the 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 $\Theta$.
3. If $L_n(\theta)$ or $\ell_n(\theta)$ is differentiable on $\Theta^\circ$, then possible candidates for the MLE are values of $\theta \in \Theta^\circ$ that satisfy the so-called "likelihood equations" or "score equations",

$$\frac{\mathrm{d}\ell_n(\theta)}{\mathrm{d}\theta} = 0. \qquad \circledast$$

   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 $X_i \overset{\text{iid}}{\sim} \mathcal{U}(0,\theta)$ 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 $\theta$ is $\hat{\theta}_n(x_1, ..., x_n) = x_{(n)}$.

⊛ **Example 4.8**: Let $X_i \overset{\text{iid}}{\sim} \mathcal{U}\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$, 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 $\theta$, for instance $\frac{x_{(1)}+x_{(n)}}{2}$ (the midpoint). In short, the MLE is *not* unique in this case.

⊛ **Example 4.9**: Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, 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 equatiosn 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

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

and so the second gives

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

Then, we find

$$\left.\frac{\partial^2 \ell_n(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}_n} = -\begin{pmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{pmatrix} < 0,$$

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

⊛ **Example 4.10**: Let $X_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$ 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

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

The likelihood equations:

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

This gives

$$\begin{cases} 0 = \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \log\left(\hat{\beta}\right) - \frac{1}{n}\sum_{i=1}^n \log(x_i) \\ \hat{\beta} = \frac{\overline{x}_n}{\hat{\alpha}} \end{cases},$$

which gives $\hat{\beta}$ as a function of $\hat{\alpha}$. Plugging this expression into the first, we find

$$\log(\hat{\alpha}) - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \frac{1}{n}\sum_{i=1}^n \log(x_i) - \log(\overline{x_n}) = 0,$$

which does not have a nice closed form. So, we must resort to numerical methods to approximate $\hat{\theta}_n$.

⊛ **Example 4.11** (Newton-Raphson): One way to numerically approximate MLEs (and more generally approximate roots of functions) such as in the previous example, is to approximate via linear functions. For instance, suppose we are interested in solving

$$\frac{\partial \ell_n(\theta)}{\partial \theta} = \ell_n'(\theta) = 0.$$

The Newton-Raphson starts with some initial guess $\theta^{(0)}$, and is then defined inductively. Given $\theta^{(t)}$, an approximation of $\theta$, the $t+1$-st iteration performs the following approximation to obtain $\theta^{(t+1)}$, by Taylor exanding,

$$\ell_n'\big(\theta^{(t)}\big) + \ell_n''\big(\theta^{(t)}\big)\big[\theta^{(t+1)} - \theta^{(t)}\big] = 0,$$

implying

$$\theta^{(t+1)} = \theta^{(t)} - \big[\ell_n''\big(\theta^{(t)}\big)\big]^{-1}\ell_n\big(\theta^{(t)}\big),$$

where in the general case $\ell_n'(\theta)$ a $d \times 1$ vector and $\ell_n''(\theta)$ a $d \times d$ matrix. In general, this procedure need not converge to the true value; typically, one stops after some "proximity standard" is met, e.g. if for some fixed allowance $\varepsilon > 0$, one may chooes to stop once $\big\|\theta^{(t+1)} - \theta^{(t)}\big\| < \varepsilon$.

⊛ **Example 4.12**: Let $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$ for $0 < \theta < 1$. Then,

$$\begin{aligned} L_n(\theta) &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ \Rightarrow \ell_n(\theta) &= n\bar{x}_n \log(\theta) + n(1-\bar{x}_n)\log(1-\theta) \\ \Rightarrow \frac{\mathrm{d}\ell_n(\theta)}{\mathrm{d}\theta} &= \frac{n\bar{x}_n}{\theta} - \frac{n(1-\bar{x}_n)}{1-\theta} = 0 \Rightarrow \hat{\theta}_n = \bar{x}_n, \end{aligned}$$

while also,

$$\frac{\mathrm{d}^2\ell_n(\theta)}{\mathrm{d}\theta^2} = -\frac{n\bar{x}_n}{\theta^2} - \frac{n(1-\bar{x}_n)}{1-\theta} < 0,$$

so $\hat{\theta}_n = \bar{x}_n$ is the unique maximizer of $\ell_n(\theta)$ when $0 < \bar{x}_n < 1$.

If $\bar{x}_n = 0$, then $L_n(\theta) = (1-\theta)^n$ is strictly decreasing in $\theta$, with unique maximizer at 0; but 0 is not in our parameter space. Similarly if $\bar{x}_n = 1$, then $L_n(\theta) = \theta^n$ is maximized at $\theta = 1$ which is again not in our parameter space. Combining these facts, the MLE indeed $\hat{\theta}_n = \bar{x}_n$.

When $\theta \in (0,1)$, the probability of $\bar{x}_n = 0$ or $\bar{x}_n = 1$ goes to zero as $n \to \infty$, exponentially.

### 4.2.1 Properties of MLE

↪**Theorem 4.1** (Invariance Property): If $\hat{\theta}_n$ the MLE of $\theta$, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\widehat{\tau(\theta)} = \tau\big(\hat{\theta}\big)$.

↪**Theorem 4.2** (Large Sample Behaviour): Under the regularity conditions from the CRLB theorem, then

- $\hat{\theta}_n$ is a consistent estimator of the parameter of interest;
- $\hat{\theta}_n$, properly cscaled and centeralized, is asymptotically normal.

## §4.3 Bayesian Estimation

Let $\boldsymbol{X} = (X_1, ..., X_n) \sim p_\theta(\cdot)$ be data distributed according to some parametrically indexed joint pdf. In Bayesian inference, the parameter $\theta$ is also treated as a random variable, with a pdf/pmf $\pi(\theta)$, called the *prior distribution* of $\theta$. Then, for post-experimental (observed) data $\boldsymbol{x} = (x_1, ..., x_n)$, then we write

$$p_\theta(x_1, ..., x_n) = p_\theta(\boldsymbol{x}) = p(\boldsymbol{x}|\theta),$$

i.e. treated as a conditional distribution of $X|\theta$.

By Baye's theorem, where $p_X(\boldsymbol{x})$ the marginal pdf/pmf of $\boldsymbol{X}$,

$$\pi(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)\pi(\theta)}{p_X(\boldsymbol{x})} = \frac{p_\theta(\boldsymbol{x})\pi(\theta)}{\int_\Theta p_\theta(\boldsymbol{x})\pi(\theta)\,\mathrm{d}\theta},$$

where $\Theta$ the entire parameter space (i.e., support of $\pi$). Hence, the so-called *posterior distribution*, $\pi(\theta|\boldsymbol{x})$, is proportional

$$\pi(\theta|\boldsymbol{x}) \propto p_\theta(\boldsymbol{x})\pi(\theta).$$

$\pi(\theta)$ is purely based on our "prior" belief/knowledge of $\theta$; $\pi(\theta|\boldsymbol{x})$ reflects the "updated" knowledge about $\theta$ given some data $\boldsymbol{x}$.

Recall that $\mathrm{Var}_\pi(\theta) \geq \mathbb{E}_\theta[\mathrm{Var}(\theta|\boldsymbol{X})]$; so, the prior variance of $\theta$ is at least as big as the expected posterior variance.

↪**Definition 4.3** (Loss Function): Given data $\boldsymbol{X} = (X_1, ..., X_n)$, a *loss function* $L(\delta(\boldsymbol{X}), \theta)$ is a measure of loss ("penatly") when $\theta$ is estimated by some function $\delta(\boldsymbol{X})$; for instance, $L(\delta(\boldsymbol{X}), \theta) = (\delta(\boldsymbol{X}) - \theta)^2$.

↪**Definition 4.4** (Baye's Risk): Given a loss function $L$, *Baye's Risk* of $\delta(\boldsymbol{X})$ is the function

$$R(\delta) := \mathbb{E}_\pi\left\{\mathbb{E}_{\boldsymbol{X}|\theta}[L(\delta(\boldsymbol{X}), \theta)]\right\}.$$

I.e., heuristically, the first nested expected value averages the loss of the estimator $\delta(\boldsymbol{X})$ over all data $\boldsymbol{X}$ given parameter $\theta$, then the second averages over all $\theta$'s.

↪**Definition 4.5** (Baye's Estimator): The *Baye's estimator* is defined

$$\hat{\delta}(\boldsymbol{X}) := \text{argmin}_{\delta \in D} R(\delta),$$

where $D$ the collection of all possible estimators; i.e. the estimator that minimizes Baye's Risk.

In the continuous case, we may write, where $\Theta$ the parameter space and X the support of $\delta$,

$$\begin{aligned} R(\delta) &= \int_{\Theta} \left[ \int_{\mathrm{X}} L(\delta(\boldsymbol{x}), \theta) p_{\theta}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right] \pi(\theta) \, \mathrm{d}\theta \\ &= \int_{\mathrm{X}} \left[ \int_{\Theta} L(\delta(\boldsymbol{x}), \theta) \pi(\theta | \boldsymbol{x}) \, \mathrm{d}\theta \right] p_{\boldsymbol{X}}(\boldsymbol{x}) \, \mathrm{d}x. \end{aligned}$$

The outside integral is independent of $\theta$, so it suffices to minimize the inner (bracketed) integral, hence

$$\hat{\delta}(\boldsymbol{X}) = \text{argmin}_{\delta \in D} \left\{ \int_{\Theta} L(\delta(\boldsymbol{X}); \theta) \pi(\theta | \boldsymbol{x}) \, \mathrm{d}\theta \right\}.$$

This expression is called the *posterior expected loss*. For instance if $L(\delta, \theta) = (\delta - \theta)^2$, then

$$\hat{\delta}(\boldsymbol{X}) = \text{argmin}_{\delta \in D} \left\{ \int_{\Theta} (\delta(\boldsymbol{X}) - \theta)^2 \pi(\theta | \boldsymbol{x}) \, \mathrm{d}\theta \right\}.$$

Recalling that the minimizer of $\mathbb{E}\left[ (X - a)^2 \right]$ is $a = \mathbb{E}[X]$, we readily find that

$$\hat{\delta}(\boldsymbol{X}) = \mathbb{E}_{\theta | \boldsymbol{X} = \boldsymbol{x}}[\theta | \boldsymbol{X} = \boldsymbol{x}],$$

called the *posterior mean*.

Similarly, if we take the absolute value loss function $L(\delta, \theta) = |\delta - \theta|$, then we'd find $\hat{\delta}(\boldsymbol{X}) =$ *posterior median*.

⊛ **Example 4.13**: Let $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$ and assume a Beta prior for $\theta$, namely $\theta \sim \pi(\theta) = \text{Beta}(\alpha, \beta)$, where $\alpha, \beta$ are so-called "hyperparameters" (namely, they are known), so

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

We aim to find the Baye's estimator of $\theta$ under the square loss. We have

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}) &\propto p_\theta(\boldsymbol{x})\pi(\theta) \\
&= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{n\bar{x}_n+\alpha-1}(1-\theta)^{n-n\bar{x}_n+\beta-1},
\end{aligned}$$

so in particular one observes $\theta|\boldsymbol{X} = \boldsymbol{x} \sim \text{Beta}(n\bar{x}_n + \alpha, n - n\bar{x}_n + \beta)$. Thus, using the known mean of a Beta distribution,

$$\begin{aligned}
\hat{\delta}(\boldsymbol{X}) &= \mathbb{E}_{\theta|\boldsymbol{X}}[\theta|\boldsymbol{X}] \\
&= \frac{n\overline{X}_n + \alpha}{n\overline{X}_n + \alpha + n - n\overline{X}_n + \beta} \\
&= \frac{n\overline{X}_n + \alpha}{n + \alpha + \beta} \\
&= \frac{n}{n+\alpha+\beta}\overline{X}_n + \frac{\alpha+\beta}{n+\alpha+\beta}\frac{\alpha}{\alpha+\beta},
\end{aligned}$$

where we notice this a convex combination of $\overline{X}_n$, the MLE, and $\frac{\alpha}{\alpha+\beta}$, the prior mean.

## §4.4 Large Sample Properties of MLE

Let $\mathcal{F} = \left\{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\right\}$ and $X \sim f_{\theta_0}$ for some $\theta_0 \in \Theta$. Throughout, we'll assume the following regularity conditions about the distribution:

- **R0**: $\Theta$ is either open, or contains an open set such $N$ such that $\theta_0$ an interior point of $N$
- **R1**: The pdf/pmf $f_\theta$ has a common support X for all $\theta \in N$ and is identifiable in $\theta$ for every $x \in \text{X}$. That is, for every $\theta_1, \theta_2 \in N, f(x;\theta_1) = f(x;\theta_2)$ for every $x \in \text{X}$ iff $\theta_1 = \theta_2$
- **R2**: $f_\theta$ is thrice differentiable in $\theta$ for almost every $x \in \text{X}$
- **R3**: There exists functions $M_i(x)$ for $i = 1, 2, 3$ (possibly depending on $\theta_0$) such that for every $\theta \in N$,

$$\left|\frac{\partial f(x;\theta)}{\partial \theta_i}\right| < M_1(x), \qquad \left|\frac{\partial^2 f(x;\theta)}{\partial \theta_i \partial \theta_j}\right| < M_2(x), \qquad \left|\frac{\partial^3 f(x;\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k}\right| < M_3(x)$$

  for every $x \in \text{X}$, such that the integral of each $M_i$ over X is finite
- **R4**: for all $\theta \in N$, $I_1(\theta) > 0$ is a positive definite matrix, as defined below

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_{\theta_0}$. Let $\hat{\theta}_n(\boldsymbol{X}) = \text{argmax}_{\theta\in\Theta} L_n(\theta)$. Assume we obtained the MLE by solving the likelihood equations $\frac{\partial \ell_n(\theta)}{\partial \theta} = 0$. Under R0 - R4, we find

$$\hat{\theta}_n \xrightarrow{P} \theta, \qquad \sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{d} \mathcal{N}_d\big(0, I_1^{-1}(\theta_0)\big),$$

where $I_1(\theta)$ the *Fisher information matrix* given by

$$I_1(\theta) = \mathbb{E}_\theta\left\{\left[\frac{\partial \log(f(x;\theta))}{\partial\theta}\right]\cdot\left[\frac{\partial \log(f(x;\theta))}{\partial\theta}\right]^t\right\}.$$

Before proceeding we need some tools.

$$\mathbb{E}_{\theta_0}\left\{\log\frac{f(\boldsymbol{X};\theta_0)}{f(\boldsymbol{X};\theta)}\right\}$$

is called the *Kullback-Leibler* (KL) distance between $f(\boldsymbol{x};\theta)$ and $f(\boldsymbol{x};\theta_0)$.

↪**Proposition 4.1**: The Kullback-Leibler distance is strictly positive for $\theta \neq \theta_0$ and equal for $\theta = \theta_0$.

Proof. We may write, by Jensen's inequality

$$= -\mathbb{E}_{\theta_0}\left\{\log\frac{f(\boldsymbol{X};\theta)}{f(\boldsymbol{X};\theta_0)}\right\} \geq -\log\mathbb{E}_{\theta_0}\left\{\frac{f(\boldsymbol{X};\theta)}{f(\boldsymbol{X};\theta_0)}\right\} = -\log\int\frac{f(\boldsymbol{x};\theta)}{\cancel{f(\boldsymbol{x};\theta_0)}}\cancel{f(\boldsymbol{x};\theta_0)}\,\mathrm{d}\boldsymbol{x} = -\log 1 = 0.$$

■

↪**Lemma 4.1**: $P(\ell_n(\theta) < \ell_n(\theta_0)) \to 1$ for every $\theta \neq \theta_0$.

Proof.

$$\frac{1}{n}[\ell_n(\theta) - \ell_n(\theta_0)] = \frac{1}{n}\sum_{i=1}^n \log\left(\frac{f(x_i;\theta)}{f(x_i;\theta_0)}\right) \xrightarrow{\text{a.s.}} \mathbb{E}_{\theta_0}\left[\log\left(\frac{f(X_1;\theta)}{f(X_1;\theta_0)}\right)\right] < 0,$$

using the strong law of large numbers and the properties of the KL distance. ■

↪**Theorem 4.3**: Under the regularity conditions,

1. $\mathbb{E}_\theta\left[\frac{\partial \log(f(X;\theta))}{\partial\theta}\right] = 0$ for every $\theta \in \Theta$;
2. $\mathbb{E}_\theta\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right] = -\mathbb{E}_\theta\left[\left[\frac{\partial \log f(X;\theta)}{\partial\theta}\right]\cdot\left[\frac{\partial \log f(X;\theta)}{\partial\theta}\right]^t\right] = -I(\theta)$ for every $\theta \in \Theta$;
3. for $d = 1$ i.e. $\Theta \subseteq \mathbb{R}$, $\mathbb{E}_{\theta_0}\left[\frac{\partial \log f(X;\theta)}{\partial\theta}\right]$ is a strictly decreasing function of $\theta$ in a small neighborhood of $\theta_0$.

These first two are the so-called *Bartlett Identities*.

Proof. 1., 2., were already proven in the discussion following the CRLB theorem, Prop. 3.1. For 3., note $\mathbb{E}_{\theta_0}\left[\frac{\partial \log f(X;\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right] = 0$. Using the regularity conditions,

$$\frac{\partial}{\partial\theta}\left(\mathbb{E}_{\theta_0}\left[\frac{\partial \log f(X;\theta)}{\partial\theta}\right]\right)\Big|_{\theta=\theta_0} \mathbb{E}_{\theta_0}\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\Big|_{\theta=\theta_0}\right] = -I(\theta_0) < 0,$$

since $I(\theta_0)$ a positive definite matrix. Hence, since $\mathbb{E}_{\theta_0}\left[\frac{\partial \log f(X;\theta)}{\partial \theta}\right]$, as a function of $\theta$, is decreasing at $\theta = \theta_0$, so strictly decreasing in some neighborhood of $\theta_0$. ■

↪**Theorem 4.4**: Under the regularity conditions, there exists a sequence $\hat{\theta}_n = \hat{\theta}_n(\boldsymbol{X})$ such that
1. $\ell_n'\left(\hat{\theta}_n\right) = 0$;
2. $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$.

Proof. By the previous theorem, for a sufficiently small $\varepsilon > 0$, by SLLN

$$\begin{cases} \frac{1}{n}\ell_n'(\theta_0 - \varepsilon) \\ \frac{1}{n}\ell_n'(\theta_0 + \varepsilon) \end{cases} \xrightarrow{\text{a.s.}} \begin{cases} \mathbb{E}_{\theta_0}\left[\left.\frac{\partial \log f(X;\theta)}{\partial \theta}\right|_{\theta=\theta_0-\varepsilon}\right] > 0 \\ \mathbb{E}_{\theta_0}\left[\left.\frac{\partial \log f(X;\theta)}{\partial \theta}\right|_{\theta=\theta_0+\varepsilon}\right] < 0 \end{cases}.$$

Therefore for large $n$, $\ell_n'(\theta_0 + \varepsilon) < 0 < \ell_n'(\theta_0 - \varepsilon)$. For large $n$ we had by the lemma as well that $\ell_n(\theta_0 + \varepsilon), \ell_n(\theta_0 - \varepsilon) < \ell_n(\theta_0)$ a.s., thus by the intermediate value theorem there is some $\hat{\theta}_n \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ such that $\ell_n'\left(\hat{\theta}_n\right) = 0$. Since $\varepsilon$ arbitrary, we also get $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$. ■

**Remark 4.7**: This result gives asymptotic existence of a sequence of "consistent" roots $\hat{\theta}_n$ of $\ell_n'(\theta) = 0$. For a given set of roots of $\ell_n'(\theta) = 0$, its consistency must be verified individually, unless it is unique, in which case it is consistent.

↪**Theorem 4.5** (Asymptotic Normality): Under the regularity conditions,

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

Proof. We have that $\ell_n'\left(\hat{\theta}_n\right) = 0$ Then,

$$0 = \ell_n'\left(\hat{\theta}_n\right) = \ell_n'(\theta_0) + \ell_n''(\theta_0)\left(\hat{\theta}_n - \theta_0\right) + \frac{\ell_n'''\left(\tilde{\theta}_n\right)\left(\hat{\theta}_n - \theta_0\right)^2}{2},$$

where $\tilde{\theta}_n$ is between $\theta_0$ and $\hat{\theta}_n$. Hence,

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \frac{\ell_n'(\theta_0)/\sqrt{n}}{-\frac{1}{n}\ell_n''(\theta_0) - \frac{1}{2n}\ell_n'''\left(\tilde{\theta}_n\right)\left(\hat{\theta}_n - \theta_0\right)}. \qquad \star$$

Now, by CLT and the Bartlett identities,

$$\frac{\ell_n'(\theta_0)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, I_1(\theta_0)).$$

By WLLN and Bartlett,

$$-\frac{1}{n}\ell_n''(\theta_0) \xrightarrow{P} -\mathbb{E}_{\theta_0}\left[\left.\frac{\partial^2 f(X_1;\theta)}{\partial \theta^2}\right|_{\theta=\theta_0}\right] = I_1(\theta_0).$$

Finally, by R3,

$$\left|\frac{1}{n}\ell_n'''\left(\tilde{\theta}_n\right)\right| = \frac{1}{n}\left|\sum_{i=1}^{n}\frac{\partial^3 \log f(X_i;\theta)}{\partial\theta^3}\bigg|_{\theta=\tilde{\theta}_n}\right|$$
$$\leq \frac{1}{n}\sum_{i=1}^{n} M_3(X_i) \xrightarrow{P} \mathbb{E}_{\theta_0}[M_3(X_i)],$$

so in particular

$$\frac{1}{n}\ell_n'''\left(\tilde{\theta}_n\right) = o_p(1).$$

Thus, combining all these convergences via Slutsky's theorem in $\star$, we find

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

■

**Remark 4.8**: The MLE is Fisher-Efficient as its asymptotic variance approaches the CRLB.

# §5 Confidence Intervals

## §5.1 Interpretations

A standard approach to representing uncertainty in point estimation is to report a "confidence interval" for a parameter of interest.

Let $\boldsymbol{X} = (X_1, ..., X_n) \overset{\text{iid}}{\sim} f_\theta$ be our "data" and $\boldsymbol{x} = (x_1, ..., x_n)^t$ be our "observed data".

↪**Definition 5.1** (Interval Estimator/Confidence Interval): Let $L(\boldsymbol{X}), U(\boldsymbol{X})$ be two statistics such that $L(\boldsymbol{x}) < U(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathrm{X}$. A random interval $(L(\boldsymbol{X}), U(\boldsymbol{X}))$ is called an *interval estimator/confidence interval (CI)* with confidence level $1-\alpha$ with $0 < \alpha < 1$ if

$$P(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})) = 1 - \alpha.$$

The *post-experimental confidence interval* is given $(L(\boldsymbol{x}), U(\boldsymbol{x}))$ for given data $\boldsymbol{x}$.

It is *wrong* to say that $(L(\boldsymbol{x}), U(\boldsymbol{x}))$ captures $\theta$ with probability $1-\alpha$; this interval either includes $\theta$ or not (basically, it captures $\theta$ with probability 0 or 1). How do we then interpret $(L(\boldsymbol{x}), U(\boldsymbol{x}))$, for a given $\alpha$? If we were to repeat our experiment (i.e. collect data under the same conditions) and compute similar confidence intervals for $\theta$, we expect $100 \times (1-\alpha)\%$ of those (post-experimental) intervals to capture $\theta$.

## §5.2 Construction of CI's

↪**Definition 5.2** (Pivotal Quantity (PQ)): A random function $Q(\boldsymbol{X}, \theta)$ is called a *pivotal quantity (PQ)* if its distribution does not depend on $\theta$, and $Q$ is only a function of $\boldsymbol{X}$ and $\theta$ (i.e. of no other unknown parameter).

Once/if we have a PQ, we proceed as follows to obtain a CI with confidence $1-\alpha$:

1. find constants $c_1, c_2$ such that $P(c_1 \leq Q(\boldsymbol{X};\theta) \leq c_2) = 1-\alpha$;
2. having $c_1, c_2$, solve the inequality from 1. with respect to $\theta$ to get something of the form $P(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})) = 1-\alpha$.

When $Q$ is monotone with respect to $\theta$, then inverting the inequality in 1. is easier. Otherwise, the resulting interval could be the union of several intervals. Further, for a parameter family, there may not exist a PQ, or there may exist many PQs. In this second case, we choose a PQ based of a sufficient statistic.

⊛ **Example 5.1**: Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), i = 1, ..., n$, where $\sigma$ known. We seek a confidence interval for $\mu$. Recall the UMVUE for $\mu$ is $\overline{X}_n$. Then, a PQ is given by

$$Q(\boldsymbol{X};\mu) = \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \sim \mathcal{N}(0,1).$$

Next, we seek $a, b$ such that

$$P\left(a \leq \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \leq b\right) = 1-\alpha.$$

Suppose we know $a, b$. Solving the inequality for $\mu$, we find

$$P\left(\overline{X}_n - \frac{b\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n - \frac{a\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Thus, our $100 \times (1-\alpha)\%$ CI for $\mu$ is

$$\left(\overline{X}_n - \frac{b\sigma}{\sqrt{n}}, \overline{X}_n - \frac{a\sigma}{\sqrt{n}}\right).$$

What are $a, b$ then? We find that the length of this interval is $\ell(\boldsymbol{X}; a, b) = \frac{(b-a)\sigma}{\sqrt{n}}$; we'd like to minimize this length (or in general the expected length, since in general this length is random). Suppose $b = b(a)$ (which it will be from our restriction above). Then,

$$\frac{\mathrm{d}}{\mathrm{d}a}\ell(\boldsymbol{X}; a, b) = \left(\frac{\mathrm{d}b}{\mathrm{d}a} - 1\right)\frac{\sigma}{\sqrt{n}} = 0 \Rightarrow \frac{\mathrm{d}b}{\mathrm{d}a} = 1 \Rightarrow b(a) = a + c,$$

for a constant $c$. Putting $\Phi, \varphi$ to be the CDF, PDF respectively of the standard norm, we know $\Phi(b) - \Phi(a) = 1-\alpha$. Taking the derivative, we find

$$\varphi(b)\frac{\mathrm{d}b}{\mathrm{d}a} - \varphi(a) = 0 \Rightarrow \frac{\mathrm{d}b}{\mathrm{d}a} = \frac{\varphi(a)}{\varphi(b)},$$

and thus all together, $\varphi(a) = \varphi(b)$. Thus, by symmetry of $\varphi$, it must be that $a = \pm b$. We take then $a = -z_{\alpha/2}, b = z_{\alpha/2}$ such that $P(Z \geq z_{\alpha/2}) = \alpha/2, Z \sim \mathcal{N}(0,1)$ so our CI becomes

$$\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

⊛ **Example 5.2**: In the same setup as the previous, but with $\sigma^2$ unknown, a PQ is given by

$$Q(\boldsymbol{X};\mu) = \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n} \sim t(n-1).$$

Following similar work to the previous, we find

$$\left(\overline{X}_n - t_{(n-1,\alpha/2)}\frac{S_n}{\sqrt{n}}, \overline{X}_n + t_{(n-1,\alpha/2)}\frac{S_n}{\sqrt{n}}\right)$$

to be the shortest CI for $\mu$ with unknown $\sigma$, where $t_{(n-1,\alpha/2)}$ the analgous quantile of the appropriate $t$ distribution.

⊛ **Example 5.3**: In the same setup, with both $\left(\mu, \sigma^2\right)$ unknown,

$$Q\left(\boldsymbol{X};\sigma^2\right) = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

is a PQ for $\sigma^2$ now. This distribution is no longer symmetric as in the previous two cases; we choose now

$$P\left(\chi^2_{(n-1,\alpha/2)} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_{(n-1,1-\alpha/2)}\right) = 1-\alpha,$$

where

$$P\left(Z \leq \chi^2_{(n-1,\alpha/2)}\right) = \frac{\alpha}{2} = P\left(Z \geq \chi^2_{(n-1,1-\alpha/2)}\right), \qquad Z \sim \chi^2_{(n-1)}.$$

This ends up with confidence interval

$$\left(\frac{(n-1)S_n^2}{\chi^2_{(n-1,1-\alpha/2)}}, \frac{(n-1)S_n^2}{\chi^2_{(n-1,\alpha/2)}}\right).$$

What would be the confidence interval with $\mu$ known?

⊛ **Example 5.4**: If $X_i$ an iid sample from a population with unknown mean $\mu$ and known variance $\sigma^2$ with $\mathbb{E}\left[X^4\right] < \infty$, by CLT $\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$. For large $n$, then, this gives an "approximate" PQ for the unknown family, so the previous analysis can be applied to find an "approximate" confidence interval for $\mu$. Similarly if $\sigma$ unknown, $\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n} \xrightarrow{d} \mathcal{N}(0,1)$ from which again we can use the confidence interval for when $X_i$ normal to find an "approximate" interval in this general case.

⊛ **Example 5.5**: Suppose we have two independent iid samples $X_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu_1, \sigma_1^2\right), i = 1, \ldots, m$ and $Y_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu_2, \sigma_2^2\right)$ with $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and we seek a CI for the difference $\mu_1 - \mu_2$. Let

$$S^2_{\text{pooled}} = \frac{1}{m+n-2}\left\{\sum_{i=1}^{m}\left(X_i - \overline{X}_m\right)^2 + \sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2\right\}.$$

Then, $\frac{(m+n-2)S^2_{\text{pooled}}}{\sigma^2} \sim \chi^2_{m+n-2}$. Under these conditions, we have that

$$\frac{\overline{X}_m - \overline{Y}_n - (\mu_1 - \mu_2)}{S_{\text{pooled}}\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2).$$

This can then be used to approximate the confidence interval. If the distributions are not known, we can use CLT to approximate as in the previous cases with one sample.

⊛ **Example 5.6**: Let $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$ and consider the point estimator $\hat{\theta}_n = \overline{X}_n$. This is consistent by WLLN, and so by CLT and Slutsky's,

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

Using this as a PQ, this results in a two-sided CI

$$\left(\hat{\theta}_n - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}_n\left(1-\hat{\theta}_n\right)}{n}}, \hat{\theta}_n + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}_n\left(1-\hat{\theta}_n\right)}{n}}\right).$$

⊛ **Example 5.7**: If $X_i, Y_j, i = 1, \ldots, m, j = 1, \ldots, n$ are two iid, independent samples from two Bernoulli distributions with parameters $\theta_1, \theta_2$, then as $m, n \to \infty$ (with $m/n \to \rho$),

$$\frac{\left(\hat{\theta}_1 - \theta_1\right) - \left(\hat{\theta}_2 - \theta_2\right)}{\sqrt{\hat{\theta}_1\left(1-\hat{\theta}_1\right)/m + \hat{\theta}_2\left(1-\hat{\theta}_2\right)/n}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\hat{\theta}_1 = \overline{X}_m, \hat{\theta}_2 = \overline{Y}_n$.

⊛ **Example 5.8**: Know This! Suppose $X_i$ are iid from a parametric model $f(\cdot, \theta)$ with $\theta$ unknown and $\hat{\theta}_n$ be the MLE of $\theta$. Assuming the regularity conditions R1-R4, recall

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

In practice, $I_1(\theta)$ is estimated either with $I_1\left(\hat{\theta}_n\right)$ or the so-called "empirical Fisher", given by

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\log(f(x_i;\theta))\Big|_{\theta=\hat{\theta}_n}\right)^2.$$

Then, this gives approximate CI given by

$$\left(\hat{\theta}_n - z_{\alpha/2}\cdot\sqrt{\frac{1}{n}\cdot[\widehat{I_1(\theta)}]^{-1}}, \hat{\theta}_n + z_{\alpha/2}\cdot\sqrt{\frac{1}{n}\cdot[\widehat{I_1(\theta)}]^{-1}}\right).$$

## §5.3 Hypothesis Testing

Consider a partitioning of the parameter space $\Theta = \Theta_0 \cup \Theta_1$. Rather than esimating $\theta$, the goal is to decided, based on the data, whether the unknown $\theta$ lies in $\Theta_0$ or $\Theta_1$.

↪**Definition 5.3** (Hypotheses): For a parametric family $\mathcal{F} = \{f(\cdot;\theta) \mid \theta \in \Theta \subset \mathbb{R}\}$, set

$$\mathcal{H}_0 : \theta \in \Theta_0 \qquad \mathcal{H}_1 : \theta \in \Theta_1,$$

such that $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \varnothing$.

↪**Definition 5.4** (Test): A statistical procedure that is used to decide whether to reject $\mathcal{H}_0$ in favour of the alternative $\mathcal{H}_1$ or not reject the null hypothesis $\mathcal{H}_0$ is called a *statistical test procedure* or simply a *test*.

A test defines a partition of the sample space X into two regions, $\mathcal{R} \sqcup \mathcal{R}^c$. The hypotheses $\mathcal{H}_0$ is then reject in favour of $\mathcal{H}_1$ depending on where the data $X_1, \ldots, X_n$ or a suitably chosen statistic $T(X_1, \ldots, X_n)$ falls into a so-called "rejection region", $\mathcal{R}$, of X. Formally, we may write the test as

$$\phi(T(\boldsymbol{X})) = \begin{cases} 1(\text{reject } \mathcal{H}_0) & \text{if } T(\boldsymbol{X}) \in \mathcal{R} \\ 0 & \text{if } T(\boldsymbol{X}) \in \mathcal{R}^c \end{cases}.$$

↪**Definition 5.5** (Types of Error): *Type I error* is made if $\mathcal{H}_0$ is rejected when $\mathcal{H}_0$ is true. *Type II error* is made if $\mathcal{H}_0$ is not rejected when $\mathcal{H}_1$ is true.

↪**Definition 5.6**: Given a statistical test $\phi$ with a rejection region $\mathcal{R}$,the *power function* of the test is defined as

$$\pi(\theta) = \mathbb{E}_\theta[\phi(T(\boldsymbol{X}))] = P_\theta\big(\text{rejecting } \mathcal{H}_0\big) = P_\theta(T(\boldsymbol{X}) \in \mathcal{R}).$$

Then,

$$\alpha(\phi) := P(\text{type I error}) = P_{\mathcal{H}_0}(T(\boldsymbol{X}) \in \mathcal{R})$$
$$\beta(\phi) := P(\text{type II error}) = P_{\mathcal{H}_1}(T(\boldsymbol{X}) \in \mathcal{R}^c).$$

↪**Definition 5.7** (Size): The *size* of a statistical test is defined

$$\overline{\alpha} = \sup_{\theta \in \mathcal{H}_0} \pi(\theta) = \sup_{\theta \in \mathcal{H}_0} [P_\theta[T(\boldsymbol{X}) \in \mathcal{R}]].$$