MATH357 - Statistics

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

Contents

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

§1 Review of Probability

↪**Definition 1.1** (Measurable Space, Probability Space): We work with a set Ω = sample space = {outcomes}, and a σ-algebra $\mathcal{F}$, which is a collection of subsets of Ω containing Ω and closed under taking complements and countable unions. The tuple $(\Omega, \mathcal{F})$ 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 Ω, 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 $S := \{x_1, x_2, ...\} \subset \mathbb{R}$, 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 "kth absolute moments", and $g(x) = x^k$ gives the ordinary "kth moments". Notice that $\mathbb{E}[\cdot]$ linear in its argument.

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

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

where they exist.

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

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

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 $(X_1, \dots, X_n)$ a random vector. Let $\mathcal{I} = \{1, \dots, n\}$ and A, B disjoint subsets of $\mathcal{I}$ with $k := |A|, h := |B|$. Write $X_A = \left(X_{i_1}, \dots, X_{i_k}\right)^t$, 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

↪**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 Θ our "parameter space". Namely, we know $X \sim F_\theta \in \mathcal{F} := \{F_\theta \mid \theta \in \Theta\}$.

(2) *Non-parametric inference:* in this setting we know 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 $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. 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 $\left(X_2 - \overline{X}_n, \dots, X_n - \overline{X}_n\right)$ are independent.

Consider now the transformation

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

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

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

and hence as the PDFs split, we conclude Y_1 independent of $Y_2, \dots, Y_n$ and so $\overline{X}_n$ independent of $\left(X_2 - \overline{X}_n, \dots, X_n - \overline{X}_n\right)$ 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 $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. 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 $X_1, \ldots, X_m \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$ 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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$. Then, the *order statistics* are

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

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

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

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

The *sample median* is defined

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

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

(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

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

and similarly

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

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

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

(d) is similar. ■

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

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

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

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

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

Then,

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

Continuous is similar. ■

§2.4 Large Sample/Asymptotic Theory

↪**Definition 2.6** (Convergence in Probability): We say $T_n = T(X_1, \ldots, X_n)$ converges *in probability* to θ $T_n \xrightarrow{P} \theta$ 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 $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ with $\mu = \mathbb{E}[X_i], \sigma^2 = \mathrm{Var}(X_i) < \infty$. 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 $\mathcal{O}$, Little o Notation): Let $\{a_n\}, \{b_n\} \subseteq \mathbb{R}$ real sequences.

- We say $a_n = \mathcal{O}(b_n)$ 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 $\mathcal{O}_p$, Little o_p Notation): Let $\{X_n\}, \{Y_n\}$ sequences of random variables.

- We say $X_n = \mathcal{O}_p(1)$ 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 $X_n = \mathcal{O}_p(Y_n)$ if $X_n/Y_n = \mathcal{O}_p(1)$.
- 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 $X_n = \mathcal{O}_p(1)$.

↪**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 $X_1, ..., X_n \overset{\text{iid}}{\sim} F$, 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. $\operatorname{Var}(F_n(x)) = \frac{1}{n} F(x)(1 - F(x))$;
3. $nF_n(x) = \sum_{i=1}^n \mathbb{1}(X_i \leq x) \sim \operatorname{Bin}(n, F(x))$;
4. $\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)(1-F(x))}} \xrightarrow{d} \mathcal{N}(0,1)$.
5. $F_n(x) \xrightarrow{P} F(x)$.
6. $P(|F_n(x) - F(x)| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$, by Hoeffding's Inequality.
7. $\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = \|F_n - F\|_\infty \xrightarrow{\text{a.s.}} 0$, by the Glivenko-Cantelli Theorem.
8. $P(\|F_n - F\|_\infty > \varepsilon) \leq C\varepsilon e^{-2n\varepsilon^2}$ for some constant C (Dvoretzky-Kiefer-Wolfowitz Theorem).

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

§3 Parametric Inference

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

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

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

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 θ 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_θ).

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 $X_1, ..., X_n \overset{\text{iid}}{\sim} F_\theta$, $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$, $\mu = \mathbb{E}[X_i], \sigma^2\ \text{Var}(X_i)$. Then, $\hat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ an unbiased estimator of μ. 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 σ^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 $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{U}(0, \theta)$, $\theta > 0$, $\Theta = (0, \infty)$. 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

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

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

↪**Definition 3.4** (Consistency): We say an estimator $\hat{\theta}_n$ of θ is *consistent* if $\hat{\theta}_n \overset{P}{\to} \theta$ 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 $\mathrm{Var}_\theta\left(\hat{\theta}_n\right) \to 0$ 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 θ, and both are consistent. To see the second one, we have that for any $\varepsilon > 0$,

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

   We have too that

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

   Also

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

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

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

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

↪**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 θ;

(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 θ.

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

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

the last line following by the Bartlett Identity. ■

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

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

so the CRLB in this case reads

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

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

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

⊛ **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

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

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

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

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_θ;

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

⊛ **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 μ. If μ known, then $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$ is the UMVUE for σ^2. If μ is unknown, then $\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$ would be the UMVUE for σ^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 θ, with the support of f being independent of θ. 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 σ^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
$\mathbb{E}$	$\frac{n-1}{n}\sigma^2$	σ^2	$\frac{n-1}{n+1}\sigma^2$
Var	$\frac{2(n-1)\sigma^4}{n^2}$	$\frac{2\sigma^4}{n-1}$	$\frac{2(n-1)}{(n+1)^2}\sigma^4$

. We notice then that

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

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 θ 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 θ. Namely,

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

does *not* depend on θ.

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

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

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

such that

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

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

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

and T has support

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

This induces the partitioning

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

↪**Theorem 3.2** (Neyman-Fisher Factorization Theorem): Let $\boldsymbol{X} = (X_1, \dots, X_n)^t$ 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 θ 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 $\boldsymbol{x} \in \mathcal{X}$.

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

PROOF. We prove in the discrete case.

Note that

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

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

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

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

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

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

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 θ and $\sum_{i=1}^n x_i$ only, and the remaining term is independent of θ. Hence by the previous theorem $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$ a sufficient statistic for θ.

⊛ **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

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

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

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

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

⊛ **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 θ, 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 μ! Namely T is a sufficient statistic for the 2-dimensional parameter θ. We cannot simply separate the dependence.

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

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

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 θ.

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

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

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

  where $\varphi(\cdot)$ some measurable function, or equivalently, $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X} \subseteq \mathbb{R}^n$, 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

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

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

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

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 $\mathcal{X}$.

↪**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 $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X} \subseteq \mathbb{R}^n$ $T(\boldsymbol{x}) = T(\boldsymbol{y}) \Leftrightarrow \frac{p_\theta(\boldsymbol{x})}{p_\theta(\boldsymbol{y})}$ does not depend on θ. Then, $T(\boldsymbol{X})$ is a minimal sufficient statistic for θ.

⊛ **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 θ. 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 θ 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 θ. 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 θ.

§3.3 Completeness

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

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

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

⊛ **Example 3.22**: Let $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta)$, 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 η, which, being equal to zero implies all the coefficients $g(t)\binom{n}{t} = 0$ for every t and hence $g(t) = 0$. Hence, $T(\boldsymbol{X})$ 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 ∞, so no such estimator exists.

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

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

Then,

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

PROOF.

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

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

■

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

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

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

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 $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta), i = 1, ..., n$ 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 θ by the Lehmann-Scheffé Theorem.

⊛ **Example 3.27**: Let $X_i \overset{\text{iid}}{\sim} \text{Pos}(\theta), i = 1, ..., n$ 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 θ.

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

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

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

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

One verifies that

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

therefore

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

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 $X_i \overset{\text{iid}}{\sim} \text{Ber}(\theta), i = 1, \ldots, n$, and suppose $\tau(\theta) = \text{Var}(X_i) = \theta(1-\theta)$. Recall the UMVUE for θ 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 estimaor 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, then we may choose some a such that

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

so

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

which ever makes sense. Hence,

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

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 θ. 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 θ. ■

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

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

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

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 θ) or moreover $\delta(x) = \delta(x+1)$ for every $x \in \mathbb{R}$. Letting now $U(X)$ be a UVMUE of $\tau(\theta)$, then by the previous theorem it must be that $\text{Cov}_\theta(U(X), \delta(X)) = 0$ for any $\theta \in \mathbb{R}$, i.e.

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

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

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

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 Parameter Estimation

§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 $\mathbb{E}_\theta[\![X_i|^d]\!] < \infty$. Let $\mu_j(\theta) = \mathbb{E}_\theta\left[X_1^j\right]$ for $j = 1, ..., d$, the non-central moments. Also define

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

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 θ_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 then the MM estimators of $\theta_1, ..., \theta_d$.

⊛ **Example 4.1**:

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 $\hat{\theta}_n = \overline{X}_n$.
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

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

This yields

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