MATH 557 - Mathematical Statistics II

Pr. Masoud Asgharian

Course notes by
Léo Raymond-Belzile
Leo.Raymond-Belzile@mail.mcgill.ca

The current version is that of November 2, 2015
Winter 2013, McGill University

Please signal the author by email if you find any typo.

These notes have not been revised and should be read carefully.

Contents

Section 1
Preliminaries

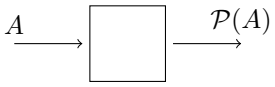
Let the indicator function be

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if otherwise} \end{cases}$$

where $A \in \mathcal{F} = \mathcal{P}(\Omega)$

where $(\Omega, \mathcal{F}, \mathsf{P})$ and the random variable X is a mapping $(\Omega, \mathcal{F}, \mathsf{P}) \to (\mathbb{R}, \mathbb{B})$.

Figure 1: Probability as a transformation



Recall the Kolmogorov's axioms, σ-field closed under countable unions, complementation rules. We also require $\mathsf{P}(\emptyset) = 0$ and similarly, $\mathsf{P}(\Omega) = 1$. If $A_i \cap A_j = \emptyset$, then $\mathsf{P}\left(A^{\complement}\right) = 1 - \mathsf{P}(A)$. If all A_n are disjoint, then $\mathsf{P}\left(\bigcup_{i=1}^{\infty} A_n\right) = \sum_{i=1}^{\infty} \mathsf{P}(A_n)$ Note thereafter we assume that $\mathcal{F}$ is a $\sigma-$field, not necessarily $\mathcal{P}(\Omega)$.

Definition 1.1 (Simple function)

We mean by **simple function**

$$\varphi(\omega) = \sum_{i=1}^{K} a_i \mathbf{1}_{A_i}(\omega), \quad A_i \in \mathcal{F}$$

and $a_i \in \mathbb{R}$, for $i = 1, \ldots, K$.

Definition 1.2 (Borel functions)

$f : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathbb{B})$ is called a Borel function if $f^{-1}(A) \in \mathcal{F}$ for any $A \in \mathbb{B}$ where $\mathbb{B}$ is the smallest σ- field generated by open sets of $\mathbb{R}$.

Proposition 1.3

Let f be a non-negative Borel function on $(\Omega, \mathcal{F})$. Then there exists a sequence of simple functions $\{\varphi_n\}$ satisfying $0 \leq \varphi_1 \leq \varphi_2 \leq \cdots \leq f$ and $\lim_{n\to\infty} \varphi_n = f$ (almost everywhere or pointwise).

Using Lebesgue rather than Riemann integral allows interesting examples which are mixture of continuous and discrete random variables, for example

$$f_X(x) = p\delta_0(x) + (1-p)U$$

where $U \sim \mathcal{U}(0,1)$ and where

$$\delta_0(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise .} \end{cases}$$

1.1 Integration

Let $\varphi \geq 0$ be a simple function, then we define

$$\int \varphi d\nu = \sum_{i=1}^{K} a_i \nu(A_i)$$

where $\varphi(\omega) = \sum_{i=1}^{K} a_i \mathbf{1}_{A_i}(\omega)$ and ν is a measure on $(\Omega, \mathcal{F})$. [1]

Recall that $\mathcal{F}$ is a $\sigma-$field if and only if

(i) $\emptyset \in \mathcal{F}$,

(ii) $A \in \mathcal{F} \Rightarrow A^{\complement} \in \mathcal{F}$

(iii) $\{A_i\}_{i=1}^{\infty} \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Recall that a set function ν defined on $\mathcal{F}$ is called a σ-additive measure if and only if

(i) $0 \leq \nu(A) \leq \infty \; \forall \; A \in \mathcal{F}$

(ii) $\nu(\emptyset) = 0$

(iii) If $A_i \in \mathcal{F}, i = 1, \ldots,$ and $A_i \cap A_j = \emptyset, \; \forall \; i \neq j$, then

$$\nu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \nu(A_i)$$

Proposition 1.4

Let $f \geq 0$ be a Borel function. Then

$$\int f \, \mathrm{d}\nu = \sup\left\{\int \varphi \, \mathrm{d}\nu : \varphi \in S_f\right\}$$

where

$$S_f = \left\{\varphi = \sum_{i=1}^{K} a_i \mathbf{1}_{A_i}\right\}$$

[1] In the case of step functions, this is the area under the curve.

for some k and $A_1, \ldots, A_K \in \mathcal{F}, \varphi \leq f$

Theorem 1.5

Let $\{f_i\}_{i=1}^{\infty}$ be a sequence of Borel functions on $(\Omega, \mathcal{F}, \nu)$. Assume f_n, f, g are measurable. Then

(i) **Fatou's lemma**: If $f_n \geq 0$, we have

$$\int \liminf_{n\to\infty} f_n \, \mathrm{d}\nu \leq \liminf_{n\to\infty} \int f_n \, \mathrm{d}\nu \tag{1.1}$$

where $\liminf_{n\to\infty} f_n(x) = \sup_n \inf_{k\geq n} f_n(x)$.

(ii) **Dominated convergence theorem** If $\lim_{n\to\infty} f_n = f$ a.e. for some Borel measurable function g such that $\int g \, \mathrm{d}\nu < \infty$, then $\int \lim_{n\to\infty} f_n \, \mathrm{d}\nu = \lim_{n\to\infty} \int f_n \, \mathrm{d}\nu$.

Note

By **almost everywhere**, we mean

$$\nu(x : \lim_{n\to\infty} f_n(x) \neq f(x)) = 0$$

(iii) If $0 \leq f_1 \leq f_2 \leq \ldots$, and $\lim_{n\to\infty} f_n = f$ almost everywhere, then

$$\lim_{n\to\infty} f_n \, \mathrm{d}\nu = \lim_{n\to\infty} \int f_n \, \mathrm{d}\nu.$$

Note

For general, not necessarily non-negative functions, we define $\int f \, \mathrm{d}\nu = \int f_+ f \, \mathrm{d}\nu - \int f_- \, \mathrm{d}\nu$ where $f \coloneqq f_+ - f_-$ and where $f_+ = f \vee 0$ ($\vee$ is the maximum) and $-f_- = f \wedge 0$ ($\wedge$ here is the minimum).

1.2 Interchanging integral and differential operators

Let $(\Omega, \mathcal{F}, \nu)$ be a measurable space and for any fixed $\theta \in \mathbb{R}$, $f(\omega, \theta)$ be a Borel function on Ω. Suppose that $\partial f(\omega, \theta)/\partial\theta$ exists a.e. for $\theta \in (a, b) \subseteq \mathbb{R}$ and $|\partial f(\omega, \theta)/\partial\theta| \leq g(\omega)$ a.e. where g is an integrable function on Ω, *i.e.* $\int g \, \mathrm{d}\nu < \infty$. Then, for each $\theta \in (a, b)$,

$$\frac{\partial f(\omega, \theta)}{\partial\theta} \in \mathcal{L}^1,$$

that is

$$\int \left| \frac{\partial f(\omega, \theta)}{\partial\theta} \right| \mathrm{d}\nu(\omega) < \infty$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int f(\omega;\theta)\,\mathrm{d}\nu(\omega) = \int \frac{\partial f(\omega,\theta)}{\partial\theta}\,\mathrm{d}\nu(\omega)$$

Proposition 1.6 (Change of variable)
Let f be $(\Omega,\mathcal{F},\nu) - (\Lambda,\mathcal{S})$ measurable, that is $f^{-1}(A) \in \mathcal{F}$, $\forall\ A \in \mathcal{S}$ and g be a Borel function on $(\Lambda,\mathcal{S})$. Then

$$\int_\Omega g \circ f\,\mathrm{d}\nu = \int_\Lambda g\,\mathrm{d}(\nu \circ f^{-1})$$

1.3 Fubini's theorem

Definition 1.7
A measurable function ν is called σ-finite if there exists $\{A_i\}_{i=1}^\infty$ such that $\int_{i=1}^\infty A_i = \Omega$ and $\nu(A_i) < \infty$ for $i = 1, 2, \ldots$.

Theorem 1.8 (Fubini's theorem)
Let ν_i be a σ- finite measure on $(\Omega_i,\mathcal{F}_i)$ for $i = 1, 2$ and let f be a Borel measurable function on $(\Omega_1,\mathcal{F}_1) \times (\Omega_2,\mathcal{F}_2)$ whose integral with respect to $(\nu_1 \times \nu_2)$ exist. Then

1. $$g(\omega_2) = \int_{\Omega_1} f(\omega_1,\omega_2)\,\mathrm{d}\nu_1$$

   exists almost everywhere with respect to ν_2.

2. $$\begin{aligned}\int_{\Omega_1\times\Omega_2} f(\omega_1,\omega_2)\,\mathrm{d}(\nu_1\times\nu_2) &= \int_{\Omega_2}\left[\int_{\Omega_1} f(\omega_1,\omega_2)\,\mathrm{d}\nu_1\right]\mathrm{d}\nu_2 \\ &= \int_{\Omega_1}\left[\int_{\Omega_2} f(\omega_1,\omega_2)\,\mathrm{d}\nu_2\right]\mathrm{d}\nu_1\end{aligned}$$

Note
The product σ-field is generated by $\{A \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$. In general, the product σ-fied is defined in such a way that the coordinate maps remain measurable.

1.4 Radon-Nikodym derivative

We say λ is **absolutely continuous** with respect to ν if $\nu(A) = 0$ implies $\lambda(A) = 0\ \forall\ A \in \mathcal{F}$ (where both ν and λ are defined on $(\Omega,\mathcal{F})$). We denote this by $\lambda << \nu$.

Theorem 1.9 (Radon-Nikodym)

Let ν and λ be two measures on $(\Omega, \mathcal{F})$ such that ν is σ- finite. If $\lambda << \nu$, then there exists a Borel function $f \geq 0$ on Ω such that

$$\lambda(A) = \int_A f \, \mathrm{d}\nu$$

for all $A \in \mathcal{F}$. [2] Furthermore, f is unique a.e. ν that is if $\lambda(A) = \int_A g \, \mathrm{d}\nu, \ \forall \ A \in \mathcal{F}$ then $f = g$ almost everywhere ν.

Proposition 1.10 (Properties of the Radon-Nikodym derivative)

- **Cancellation rule:** If $\lambda << \nu$ and $h \geq 0$, then $\int h \, \mathrm{d}\lambda = \int h \frac{\mathrm{d}\lambda}{\mathrm{d}\nu} \, \mathrm{d}\nu$.
- If $\lambda_i << \nu$ for $i = 1, 2$, then so is $\lambda_1 + \lambda_2$ and

$$\frac{\mathrm{d}(\lambda_1 + \lambda_2)}{\mathrm{d}\nu} = \frac{\mathrm{d}\lambda_1}{\mathrm{d}\nu} + \frac{\mathrm{d}\lambda_2}{\mathrm{d}\nu} \quad \text{a.e.}[\nu].$$

- **Chain rule:** If λ is a σ-finite measure and $\tau << \lambda << \nu$, then

$$\frac{\mathrm{d}\tau}{\mathrm{d}\nu} = \frac{\mathrm{d}\tau}{\mathrm{d}\lambda} \cdot \frac{\mathrm{d}\lambda}{\mathrm{d}\nu} \quad \text{a.e.}[\nu].$$

  In particular, if $\lambda << \nu$ and $\nu << \lambda$ (in which case we say λ and ν are equivalent), then

$$\frac{d\lambda}{\mathrm{d}\nu} = \left(\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} \right)^{-1} \quad \text{a.e.}[\nu]. \text{ or } [\lambda]$$

- Let $(\Omega_i, \mathcal{F}_i, \nu_i)$ be a measure space, $i = 1, 2$ and ν_i be σ-finite for $i = 1, 2$. Let λ_i be a σ-finite measure on $(\Omega_i, \mathcal{F}_i)$ and $\lambda_i << \nu_i$ then $\lambda_1 \times \lambda_2 << \nu_1 \times \nu_2$

$$\frac{\mathrm{d}(\lambda_1 \times \lambda_2)}{\mathrm{d}(\nu_1 \times \nu_2)}(\omega_1, \omega_2) = \frac{\mathrm{d}\lambda_1}{\mathrm{d}\nu_1}(\omega_1) \cdot \frac{\mathrm{d}\lambda_2}{\mathrm{d}\nu_2}(\omega_2) \quad \text{a.e.}[\nu_1 \times \nu_2].$$

1.5 Asymptotic theory and modes of convergence

Definition 1.11 (Convergence)

Let $X, X_1, X_2, \ldots$ be random vectors (functions) on a common probability space. [3]

1. **Convergence in distribution:** We say $X_n \xrightarrow{d} X$ (or weakly; in this case we use

[2] Densities are thus derivatives with respect to the Lebesgue measure.
[3] Recall that a probability space is a measurable space such that $\mathsf{P}(\Omega) = 1$.

$F_n \xrightarrow{w} F$) if and only if

$$\lim_{n\to\infty} F_n(x) = F(x), \quad \forall\, x \in C_F.$$

where $C_f = \{\text{continuity points of } F\}$ and where $F, F_1, F_2, \ldots$ are respectively CDF of $X, X_1, X_2, \ldots$. When we use "weakly", we may use $\mathsf{P}_n \xrightarrow{w} \mathsf{P}$ instead of $F_n \xrightarrow{w} F$, although authors have been also using $X_n \xrightarrow{w} X$.

2. **Convergence in probability** We say $X_n \xrightarrow{\mathsf{P}} X$ if and only if

$$\lim_{n\to\infty} \mathsf{P}\left(\|X_n - X\| > \varepsilon\right) = 0, \quad \forall\, \varepsilon > 0.$$

3. **Convergence in $\mathbf{r^{th}}$ moment:** We say $X_n \xrightarrow{L_r}$ if and only if

$$\lim_{n\to\infty} \mathsf{E}\|X_n - X\|^r = 0, \quad r > 0.$$

4. **Almost-sure convergence:** We say $X_n \xrightarrow{\text{a.s.}} X$ if and only if

$$\mathsf{P}\left(\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right) = 1.$$

Theorem 1.12 (Polya's theorem)
Let

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}},$$

F_n is the CDF of Z_n and F is the CDF of $Z \sim \mathcal{N}(0,1)$. If $F_n \xrightarrow{w} F$ and F is continuous on $\mathbb{R}^k$, then

$$\lim_{n\to\infty} \|F_n - F\|_\infty = 0$$

where

$$\|F_n - F\|_\infty = \sup_{\boldsymbol{x}\in\mathbb{R}^k} |F_n(x) - F(x)|$$

that is the convergence is uniform.

A useful result for almost sure convergence is

$$X_n \xrightarrow{\text{a.s.}} X \quad \Leftrightarrow \quad \lim_{n\to\infty} \mathsf{P}\left(\bigcup_{m=n}^{\infty} \{\|X_m - X\| > \varepsilon\}\right) = 0.$$

Theorem 1.13

1. $X_n \xrightarrow{\text{a.s.}} X \Rightarrow X_n \xrightarrow{\mathsf{P}} X \Rightarrow X_n \xrightarrow{w} X$. Furthermore, if $\mathsf{P}(X = c) = 1$ for some constant c, then $X_n \xrightarrow{w} X \Rightarrow X_n \xrightarrow{\mathsf{P}} X$.
2. $X_n \xrightarrow{L_r} X \Rightarrow X_n \xrightarrow{\mathsf{P}} X$ for $r > 0$
3. **Skorohod's theorem:** If $X_n \xrightarrow{w} X$, then there are random vectors $Y, Y_1, \ldots$ defined on a common probability space such that $\mathsf{P}_Y = \mathsf{P}_X, \mathsf{P}_{Y_n} = \mathsf{P}_{X_n}, n = 1, 2, \ldots$ and $Y_n \xrightarrow{\text{a.s.}} Y$.
4. If, for any $\varepsilon > 0$, $\sum_{n=1}^{\infty} \mathsf{P}(\|X_n - X\| \geq \varepsilon) < \infty$, then $X_n \xrightarrow{\text{a.s.}} X$.
5. If $X_n \xrightarrow{\mathsf{P}} X$, then there is a subsequence $\{X_{n_j}, j = 1, 2, \ldots\}$ such that $X_{n_j} \xrightarrow{\text{a.s.}} X$ as $j \to \infty$.
6. Suppose $X_n \xrightarrow{d} X$, where $X_n \in \mathbb{R}^k$. Then, for any $r > 0$

$$\lim_{n\to\infty} \mathsf{E}\|X_n\|^r = \mathsf{E}\|X\|^r < \infty$$

if and only if $\{\|X_n\|^r\}$ is **uniformly integrable** in the sense[4] that

$$\lim_{t\to\infty} \sup_n \mathsf{E}\left(\|X_n\|^r \mathbf{1}_{\|X_n\|>t}\right) = 0$$

Remark

An easy sufficient condition for uniform integrability is

$$\sup_n \mathsf{E}\|X_n\|^{r+\delta} < \infty$$

for some $\delta > 0$.

Lemma 1.14 (Borel-Cantelli)

Let A_n be a sequence of events in a probability space and $\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m$.

(a) If $\sum_{n=1}^{\infty} \mathsf{P}(A_n) < \infty$, then $\mathsf{P}(\limsup_{n\to\infty} A_n) = 0$.

(b) If $\{A_n\}$ is a sequence of pairwise independent events and $\sum_{n=1}^{\infty} \mathsf{P}(A_n) = \infty$, then $\mathsf{P}(\limsup_{n\to\infty} A_n) = 1$.

[4] For some norm (not metric), one can take the absolute value or Euclidian, or $\mathcal{L}^1$ norm

Notation ("O" and "o")

We say that

$$a_n = O(b_n) \quad \Leftrightarrow \quad \left|\frac{a_n}{b_n}\right| \leq c, n \geq N$$

for some N, where $\{a_n\}$ and $\{b_n\}$ are two sequences of real numbers.

We say that

$$a_n = o(b_n) \quad \Leftrightarrow \quad \frac{a_n}{b_n} \to 0 \text{ as } n \to \infty.$$

Likewise, we can define O, o almost surely and O_p, o_p. Suppose $\{X_n\}$ is a sequence of random vectors and $\{Y_n\}$ is a sequence of random variables. Then

- $X_n = O(Y_n)$ almost surely if and only if

$$\mathsf{P}\left(\omega : \|X_n(\omega)\| = O\left(\|Y_n(\omega)\|\right)\right) = 1$$

- $X_n = o(Y_n)$ almost surely if and only if $X_n/Y_n \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$.
- $X_n = O_p(Y_n)$ if and only if for any $\varepsilon > 0$, there are $C_\varepsilon, N_\varepsilon$ such that

$$\sup_{n \geq N_\varepsilon} \mathsf{P}\left(\frac{\|X_n\|}{\|Y_n\|} \geq C_\varepsilon\right) < \varepsilon$$

- $X_n = o_p(Y_n)$ if and only if $X_n/Y_n \xrightarrow{\mathsf{P}} 0$ as $n \to \infty$.

Exercise 1.1

Some implications

- If $X_n = o_p(Y_n)$, then $X_n = O_p(Y_n)$
- If $X_n = O_p(Y_n)$ and $Y_n = O_p(Z_n)$, then $X_n = O_p(Z_n)$.
- Suppose $\{X_n\}$ is a sequence of random variables and $X_n \xrightarrow{d} X$, then $X_n = O_p(1)$. In such case, we say $\{X_n\}$ is bounded in probability.
- Establish whether $o(1)/n \stackrel{?}{=} o(n^{-1})$ and similarly $O(1)/n \stackrel{?}{=} O(n^{-1})$ using the definitions.

1.6 Weak convergence

Tighness

Definition 1.15 (Tighness)

A sequence $\{\mathsf{P}_n\}$ of probability measures on $(\mathbb{R}^k, \mathbb{B}^k)$ is called **tight** if for an $\varepsilon > 0$, there is a compact set $C \subset \mathbb{R}^k$ such that $\inf_n \mathsf{P}_n(C) > 1 - \varepsilon$. If $\{X_n\}$ is the corresponding sequence of random vectors to $\{\mathsf{P}_n\}$, then tightness of $\{\mathsf{P}_n\}$ is equivalent to boundedness of $\{\|X_n\|\}$ in probability.

Theorem 1.16

$\{\mathsf{P}_n\}$ is tight if and only if for any subsequence $\{\mathsf{P}_{n_j}\}$ there is a further subsequence $\{\mathsf{P}_{n_{j_k}}\}$ such that $\mathsf{P}_{n_{j_k}} \xrightarrow{w} \mathsf{P}$ as $k \to \infty$. Furthermore, if all these subsequences converge to the same P, then $\mathsf{P}_n \xrightarrow{w} \mathsf{P}$.

Theorem 1.17

$X_n \xrightarrow{w} X$ if and only if

$$\int_\Omega h(X_n(\omega))\, \mathrm{d}\mathsf{P}_{X_n}(\omega) \Rightarrow \int_\Omega h(X(\omega)) d\mathsf{P}(\omega)$$

for all bounded continous functions. Or equivalently,

$$\limsup_{n\to\infty} \mathsf{P}_{X_n}(C) \leq \mathsf{P}_X(C), \qquad \forall \text{ closed sets } C$$

or equivalently

$$\liminf_{n\to\infty} \mathsf{P}_X(O) \geq \mathsf{P}_X(O), \qquad \forall \text{ open sets } O$$

Theorem 1.18 (Lévy-Cramer continuity theorem)

Let $\Phi_X, \Phi_{X_1}, \ldots$ be the characteristic functions of $X, X_1, \ldots$ respectively. Then $X_n \xrightarrow{d} X$ if and only if $\Phi_{X_n}(t) \to \Phi_X(t), \ \forall\, t \in \mathbb{R}^k$

Proposition 1.19 (Cramer-Wold device)

$X_n \xrightarrow{d} X$ if and only if $\boldsymbol{c}^\top X_n \xrightarrow{d} \boldsymbol{c}^\top X$ for all $\boldsymbol{c} \in \mathbb{R}^k$. [5]

1.7 Convergence of transformations

Theorem 1.20 (Continuous mapping theorem)

Let $\boldsymbol{X}, \{\boldsymbol{X}_n\}_{n=1}^\infty$ be a sequence of random vectors on $(\mathbb{R}^k, \mathbb{B}^k)$. Suppose $g : (\mathbb{R}^k, \mathbb{B}^k) \to (\mathbb{R}^l, \mathbb{B}^l)$ is a continuous map. Then

$$\boldsymbol{X}_n \xrightarrow{a} \boldsymbol{X} \quad \Rightarrow \quad g(\boldsymbol{X}_n) \xrightarrow{a} g(\boldsymbol{X})$$

where a is almost surely or in probability (p) or weakly (w).

[5] In particular, we cannot say that if $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} Y$, then $(X_n\ Y_n)^\top \xrightarrow{d} (X\ Y)^\top$ and we need the device

Theorem 1.21 (Slutsky)
If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{\mathsf{P}} c$, a constant. Then

(a) $X_n + Y_n \xrightarrow{d} X + c$

(b) $X_n Y_n \xrightarrow{d} cX$

(c) $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$

(d) **Slutsky**: if $X_n \xrightarrow{d} X$ and $|X_n - Y_n| \xrightarrow{\mathsf{P}} 0$, then $Y_n \xrightarrow{d} X$

In particular, confidence intervals for proportions can be found using (c) with

$$\frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} \frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \xrightarrow{d} \mathcal{N}(0,1)$$

approximating the denominator as the latter variance doesn't depend on the unknown parameter.

Proposition 1.22 (Delta method)
Let $\{\boldsymbol{X}_n\}$ and $\boldsymbol{Y}$ be random vectors, such that $a_n(\boldsymbol{X}_n - \boldsymbol{c}) \xrightarrow{d} \boldsymbol{Y}$, where $\boldsymbol{c} \in \mathbb{R}^k$ and $a_n > 0$, $\lim_{n\to\infty} a_n = \infty$. Let $g : \mathbb{R}^k \to \mathbb{R}$, $g \in \mathcal{C}_1$ (meaning continuous and differentiable at $\boldsymbol{c}$). Then

$$a_n(g(\boldsymbol{X}_n) - g(\boldsymbol{c})) \xrightarrow{d} [\nabla g(\boldsymbol{c})]^\top \boldsymbol{Y}$$

where $[\nabla g(\boldsymbol{c})]$ is the gradient of g at $\boldsymbol{c}$.

If $\boldsymbol{Y} \sim \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, then

$$a_n\big[g(\boldsymbol{X}_n) - g(\boldsymbol{c})\big] \xrightarrow{w} [\nabla g(\boldsymbol{c})]^\top \boldsymbol{Y} \sim \mathcal{N}\left(0, [\nabla g(\boldsymbol{c})]^\top \boldsymbol{\Sigma} [\nabla g(\boldsymbol{c})]\right)$$

This special case is often called δ-method. [6]

Indeed, we can then use a Taylor's series approximation

$$g(\bar{X}_n) = g(\mu) + g'(\mu)(\bar{X}_n - \mu) + R$$

and we can then say

$$\mathsf{E}\left(g(\bar{X}_n)\right) = g(\mu) + g'(\mu)\mathsf{E}\left(\bar{X}_n - \mu\right) + \mathsf{E}\left(R\right).$$

[6] To establish convergence for random processes, one need to establish convergence of the random variables for finite-sequence, then combine with tightness to get the result for infinite dimensions. Also, we keep the sequence a_n as the rate may not be $\sqrt{n}$

The remainder is of importance, indeed $1/n$ is the maximum rate one can get. We have

$$\mathsf{E}\left(g(\bar{X}_n)\right) \approx g(\mu) + O\left(\frac{1}{n}\right).$$

Theorem 1.23 (Second-order δ-method)
Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. For a given function g and a specific value of θ, suppose $g'(\theta) = 0$ and $g''(\theta)$ exists and is not zero. Then

$$n[g(Y_n) - g(\theta)] \xrightarrow{d} \sigma^2 \frac{g''(\theta)}{2} \chi_1^2$$

Application
Consider $Y_n \overset{\text{iid}}{\sim} \mathcal{B}(p)$ (Bernoulli) for $n = 1, 2, \ldots$, and $g(p) = p(1-p)$ so $g'(p) = 1 - 2p$ and hence $g'(1/2) = 0$ *i.e.* $g' = 0$ if $p = \frac{1}{2}$. If one wanted to find $[\bar{X}_n(1 - \bar{X}_n) - p(1-p)]$, one needs to go to higher δ-methods.

The δ-method is essentially based on Taylor's expansion. The first order expansion, in particular, is often used.

Proposition 1.24 (Taylor's series expansion)
If $g : U \subset \mathbb{R}^k \to \mathbb{R}$ has continuous partial derivatives of third order, then

$$g(\boldsymbol{x}_0 + \boldsymbol{h}) = g(\boldsymbol{x}_0) + [\nabla g(\boldsymbol{x}_0)]^\top \boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^\top \boldsymbol{H} g(\boldsymbol{x}_0)\boldsymbol{h} + R_2(\boldsymbol{h}, \boldsymbol{x}_0)$$

where $R_2(\boldsymbol{h}, \boldsymbol{x}_0)/\|\boldsymbol{h}\|^2 \to 0$ as $\boldsymbol{h} \to \boldsymbol{0}$ (namely $o(\|\boldsymbol{h}\|^2)$).

Note that we denote the gradient

$$[\nabla g(\boldsymbol{x}_0)]^\top = \left(\frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_k}\right)\Bigg|_{\boldsymbol{x}_0}$$

and the Hessian matrix $\boldsymbol{H} g(\boldsymbol{x}_0)$ is defined as

$$\boldsymbol{H} g(\boldsymbol{x}_0) = \left[\frac{\partial^2 g}{\partial x_i \partial x_j}\right]_{i,j=1,\ldots,k}\Bigg|_{\boldsymbol{x}_0}$$

Thus, we have

$$g(\boldsymbol{X}_n) = g(\boldsymbol{c}) + [\nabla g(\boldsymbol{c})]^\top + \frac{1}{2}(\boldsymbol{X}_n - \boldsymbol{c})^\top \boldsymbol{H} g(\boldsymbol{c})(\boldsymbol{X}_n - \boldsymbol{c}) + R_2$$

where

$$\frac{R_2}{\|\boldsymbol{X}_n - \boldsymbol{c}\|^2} \to 0$$

as

$$a_n(\boldsymbol{X}_n - \boldsymbol{c}) \xrightarrow{w} \boldsymbol{Y}$$

as $|\boldsymbol{X}_n - \boldsymbol{c}| = O_p\left(\frac{1}{a_n}\right)$. If $\nabla g(\boldsymbol{c}) = 0$, then

$$g(\boldsymbol{X}_n) - g(\boldsymbol{c}) = \frac{1}{2}(\boldsymbol{X}_n - \boldsymbol{c})^\top \boldsymbol{H} g(\boldsymbol{c})(\boldsymbol{X}_n - \boldsymbol{c}) + R_n$$

then $a_n(\boldsymbol{X}_n - \boldsymbol{c}) \xrightarrow{w} \boldsymbol{Y}$ and $\|\boldsymbol{X}_n - \boldsymbol{c}\| = O_p\left(\frac{1}{a_n^2}\right)$ and $(\boldsymbol{X}_n - \boldsymbol{c})^2 = O_p\left(\frac{1}{a_n^2}\right)$ For example, with hazard function, we usually have remainder of the order $\sqrt[3]{n}$; this causes many problems, for example the usual bootstrap will not work.

Suppose one performs a cross-sectional study (which are cheaper compared to other methods to implement). You will recruit the prevalent cases (not incident), that is they already suffer from some disease, rather than getting them in the early stage of study and assessing whether they have the disease. This is a problem of length bias censored data. One can think for example of studies on dementia or on unemployment in economics (in the latter, one may have biased sampling as well).

Theorem 1.25

Let Y be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

For a given function $g \in \mathcal{C}_2$ and a specific value of θ. Suppose $g'(\theta) = 0$ and $g''(\theta)$ exists and are not zero. Then

$$n[g(Y_n) - g(\theta)] \xrightarrow{d} \sigma^2 \frac{g''(\theta)}{2} \chi_1^2$$

$$g(Y_n) - g(\theta) = \frac{\sigma^2}{2}\left(\frac{Y_n - \theta}{\sigma}\right)^2 g''(\theta) + R_2$$

and so

$$n[g(Y_n) - g(\theta)] = \frac{\sigma^2}{2}\left[\frac{\sqrt{n}(Y_n - \theta)}{\sigma}\right]^2 g''(\theta) + nR_2$$

This was from a distributional perspective, but we can also use the Taylor expansion to get expectation and variance results. The δ-method, being essentially based on Taylor's expansion, can be readily used to approximate $\mathsf{E}\left(g(X)\right)$ and $\mathsf{Var}\left(g(X)\right)$.

Example 1.1

Suppose X is a random variable with $\mathsf{E}(X) = \mu \neq 0$. Then,

$$g(X) \approx g(\mu) + g'(\mu)(x - \mu)$$

if g is $\mathcal{C}_1$. This then leads us to $\mathsf{E}(g(X)) \approx g(\mu)$ and $\mathsf{Var}(g(X)) \approx [g'(\mu)]^2 \mathsf{Var}(X)$.

Exercise 1.2

Consider for example $g(x) = 1/x$. Then

$$\mathsf{E}\left(\frac{1}{X}\right) \approx \frac{1}{\mu}$$

while

$$\mathsf{Var}\left(\frac{1}{X}\right) \approx \left(\frac{1}{\mu}\right)^4 \mathsf{Var}(X).$$

This type approximation can easily be extended to higher dimensions and be used for approximation. Here $\boldsymbol{Z}_n = (X_n \; Y_n)^\top$, with $a_n(\boldsymbol{Z}_n - \boldsymbol{\theta}) \xrightarrow{d} V$ and take $g(\boldsymbol{z}) = x/y, \boldsymbol{z} = (x \; y)^\top$, then $\mathsf{E}(X/Y)$ and $\mathsf{Var}(X/Y)$. For this purpose, we need multivariate version of Taylor's expansion.

If X, Y are two random variables with $\mathsf{E}(X) = \mu_X$ and $\mathsf{E}(Y) = \mu_Y$ and $g(x, y) = \frac{x}{y}$, then

$$\begin{aligned}
\mathsf{E}\left(\frac{X}{Y}\right) &\approx \frac{\mu_X}{\mu_Y} \\
\mathsf{Var}\left(\frac{X}{Y}\right) &\approx \frac{1}{\mu_Y^2}\mathsf{Var}(X) + \frac{\mu_X^2}{\mu_Y^4}\mathsf{Var}(Y) - 2\frac{\mu_X}{\mu_Y^3}\mathsf{Cov}(X, Y)
\end{aligned}$$

But be careful as weak convergence doesn't imply convergence in moments. We need that $\mathsf{E}(Y/X)$ to exist for this estimate to be meaningfull, for example if $X_n, Y_n \sim \mathcal{N}(0, 1)$ and $X_n/Y_n \sim \mathcal{C}(0, 1)$. For more on the rate of error of these approximations, see Rickel and Dokson (2007) on pages 306-309.

1.8 Generation of a random sample

Probability Integral Transform

If Y is a continuous random variable with CDF F_Y, then $U = F_Y(Y) \sim \mathcal{U}(0, 1)$. For example, if $Y \sim \mathcal{E}(\lambda)$, that is exponentially distributed with

$$f_Y(y) = \lambda e^{-\lambda y} \mathbf{1}_{y \geq 0}$$

and

$$F_Y(y) = 1 - e^{-\lambda y} \mathbf{1}_{y \geq 0}$$

then $U = 1 - e^{-\lambda Y}$, $U \sim \mathcal{U}(0,1)$.

To generate random variables from $\mathcal{E}(\lambda)$

- Generate $U_i \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$, for $i = 1, 2, \ldots, n$
- $Y_i = -\frac{1}{\lambda}\log(1 - U_i)$, for $i = 1, \ldots, n$.
- Having done the first step, we can generate random samples from Gamma $\mathcal{G}(k, \lambda)$. If $X_i \overset{\text{iid}}{\sim} \mathcal{E}(\lambda)$, $\sum_{i=1}^{k} X_i \sim \mathcal{G}(k, \lambda)$.
- For each i, we generate k samples from $\mathcal{U}(0,1)$, say $U_{i1}, \ldots, U_{ik}$ and then set $w_i = -\frac{1}{\lambda}\sum_{j=1}^{k}\log(U_{ij})$.
- Having performed the second step, we can generate random sample from Beta $\mathcal{B}(k, l)$ where both k, l are integers by generating $n(k + l)$ iid samples from $\mathcal{U}(0,1)$, then compute

$$V_i = \frac{\sum_{j=1}^{k}\log U_{ij}}{\sum_{j=1}^{k+l}\log U_{ij}}$$

  for $i = 1, 2, \ldots, n$.

Remark

Note that we cannot perform the same procedure for generating random samples from Beta distribution if either k or l is not an integer.

Accept-Reject algorithm

Algorithm 1.1

- Generate U, V independent $\mathcal{U}(0,1)$
- If $U < \frac{1}{c} f_Y(v)$, set $Y = v$, otherwise return to step 1. Here $c = \max_{0 \le y \le 1} f_Y(y)$. Recall

$$f_Y(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}\mathbf{1}_{y \in [0,1]}$$

  if $Y \sim \mathcal{B}(a, b)$. Thus

$$\begin{aligned} c &= \max_{0 \le y \le 1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y_*^{a-1}(1-y_*)^{b-1} \end{aligned}$$

where

$$y_* = y_{\text{mod}} = \frac{a-1}{a+b-2}$$

if both a and $b > 1$.

Theorem 1.26

Let $Y \sim f_Y(y)$ (the target distribution) and $V \sim f_V(v)$ (candidate distribution), where f_Y and f_V have common support with

$$M = \sup_y \frac{f_Y(y)}{f_V(y)} < \infty.$$

To generate a random variable $Y \sim f_Y$

a. Generate $U \sim \mathcal{U}(0,1)$ and $V \sim f_V$ independent

b. If $U < \frac{1}{M}\frac{f_Y(v)}{f_V(v)}$, set $Y = v$, otherwise return to step a.

Proof

$$\begin{aligned}
\mathsf{P}(Y \le y) &= \mathsf{P}(V \le y | \text{stop}) \\
&= \mathsf{P}\left(V \le y | U < \frac{1}{M}\frac{f_Y(v)}{f_V(v)}\right) \\
&= \frac{\mathsf{P}\left(V \le y, U < \frac{1}{M}\frac{f_Y(v)}{f_V(v)}\right)}{\mathsf{P}\left(U < \frac{1}{M}\frac{f_Y(v)}{f_V(v)}\right)} \\
&= \frac{\int_{-\infty}^{y}\int_0^{\frac{1}{M}\frac{f_Y(v)}{f_V(v)}} \mathrm{d}u f_V(v)\,\mathrm{d}v}{\int_{\mathbb{R}}\int_0^{\frac{1}{M}\frac{f_Y(v)}{f_V(v)}} \mathrm{d}u f_V(v)\,\mathrm{d}v} \\
&= \frac{\int_{-\infty}^{y} \frac{1}{M}\frac{f_Y(v)}{f_V(v)} f_V(v)\,\mathrm{d}v}{\int_{-\infty}^{\infty} \frac{1}{M}\frac{f_Y(v)}{f_V(v)} f_V(v)\,\mathrm{d}v} \\
&= \frac{\int_{-\infty}^{y} f_Y(v)\,\mathrm{d}v}{\int_{\mathbb{R}} f_Y(v)\,\mathrm{d}v} \\
&= \int_{-\infty}^{y} f_Y(v)\,\mathrm{d}v \\
&= F_Y(y)
\end{aligned}$$

■

Remark

The probability of stopping

$$\begin{aligned}
\mathsf{P}(\text{stop}) &= \mathsf{P}\left(U < \frac{1}{M}\frac{f_Y(V)}{f_V(V)}\right) \\
&= \mathsf{E}\left(\mathsf{E}\left(\mathbf{1}_{U<\frac{1}{M}\frac{f_Y(V)}{f_V(V)}}\middle|V\right)\right) \\
&= \mathsf{E}\left(\mathsf{E}\left(\mathbf{1}_{U<\frac{1}{M}\frac{f_Y(v)}{f_V(v)}}\middle|V=v\right)\right) \\
&= \int_v \frac{1}{M}\frac{f_Y(v)}{f_V(v)} f_V(v)\,\mathrm{d}v \\
&= \frac{1}{M}\int_v f_Y(v)\,\mathrm{d}v \\
&= \frac{1}{M}
\end{aligned}$$

since $U \perp\!\!\!\perp V$. Recall that $\mathsf{P}(A) = \mathsf{E}(\mathbf{1}_A) = \int_A \mathrm{d}F(x)$ and indeed

$$\begin{aligned}
\mathsf{E}\left(\mathbf{1}_{U<\frac{1}{M}\frac{f_Y(V)}{f_V(V)}}\middle|V=v\right) &= \mathsf{P}\left(U < \frac{1}{M}\frac{f_Y(v)}{f_V(v)}\middle| V = v\right) \\
&= \mathsf{P}\left(U < \frac{1}{M}\frac{f_Y(v)}{f_V(v)}\right) \\
&= \frac{1}{M}\frac{f_Y(v)}{f_V(v)}
\end{aligned}$$

Remark

Let N be the number of trials needed to generate one Y. Then $N \sim \mathcal{G}(p = M^{-1})$. Note that $M = 1/\mathsf{P}(\text{stop})$ and also $M = \mathsf{E}(N)$. Therefore, the smaller the M, the lesser the number of trials needed to generate Y.

Remark

The candidate density f_V should be chosen such that f_V has heavier tail than f_Y, the target density. This esures a good representation of samples from f_Y.

1.9 Markov-Chain Monte Carlo (MCMC)

Definition 1.27 (Stochastic process)

A stochastic process is a family of random variables $\{X_t : t \in T\}$. A Markov process enjoys the Markov property, that is

$$\mathsf{P}\left(X^{(t+1)} = x | X^{(t)} = x_t, X^{(t-1)} = x_{t-1}, \ldots, X^{(1)} = x_1\right) = \mathsf{P}\left(X^{(t+1)} = x | X^{(t)} = x_t\right)$$

so that forecasting the next value depends only of the present. We call this **chain** if the process is has discrete state space, **process** for continuous state space processes and in higher dimensions, **fields** (Gaussian fields (brain mapping – points in manifolds)) .

Metropolis-Hastings algorithm

The MH algorithm constructs a Markov chain (MC) with state space $\mathcal{X}$ and equilibrium distribution $\pi(x)$. The algorithm works as follows; let $q(x, x')$ denote a transition probability function such that if $X^{(t)} = x$, x' is drawn from $q(x, x')$, *i.e.* $\mathsf{P}\left(X^{(t+1)} = x' | X^t = x\right)$ is considered as a proposed possible value. for $X^{(t+1)}$. A further randomization, however, takes place with some probability $\alpha(x, x')$; we accept $X^{(t+1)} = x'$, otherwise we reject the value generated from $q(x, x')$ and set $X^{(t+1)} = x$. This construction defines a Markov Chain with transition probability given by

$$p(x, x') = \begin{cases} q(x, x')\alpha(x, x') & \text{if } x' \neq x \\ 1 - \sum_{x' \neq x} q(x, x')\alpha(x, x') & \text{if } x' = x \end{cases}$$

If we now set

$$\alpha(x, x') = \begin{cases} \min\left\{\frac{\pi(x')q(x', x)}{\pi(x)q(x, x')}, 1\right\} & \text{if } \pi(x)q(x, x') \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

It is not hard (exercise) to check that

$$\frac{\pi(x')}{\pi(x)} = \frac{p(x, x')}{p(x', x)}$$

called Fouler's principle of detailed balance, provided that $q(x, x')$ is irreducible and aperiodic (that is ergodic). The principle of detailed balance is a sufficient condition for π to be the equilibrium distribution of the constructed chain, *i.e.* $\pi P = \pi$.

For a stationary process, we can find the probability to go from y to x from time t to $t + 1$. The process is **stationary** if $\mathsf{P}\left(X^{(t+1)} = x | X^{(t)} = y\right) = p(y, x) \quad \forall\ t \geq 0$. The transition matrix of conditional probabilities is

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \ddots & \ddots & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix}$$

then there a stable distribution for ergodic chains and $p_{ij}^n \to \pi(j)$, the equilibrium probability,

where π is the equilibrium distribution of the process.

Remark

The implementation of the algorithm only requires knowledge of $\pi(x')/\pi(x)$.

When the dimension of $\boldsymbol{\theta}$, the vector of parameters is high (say 100), then the posterior is

$$\pi(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \propto f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

and working with the ratio allows cancellation of the normalizing constant.

Here are special cases of the algorithm.

- Tierney (1991) – Taking $q(\boldsymbol{x}, \boldsymbol{x}') = q(\boldsymbol{x}', \boldsymbol{x})$ entail $\alpha(\boldsymbol{x}, \boldsymbol{x}') = \frac{\pi(\boldsymbol{x}')}{\pi(\boldsymbol{x})} \wedge 1$. This is known as Metropolis algorithm
- Muler (1991) – Take $q(\boldsymbol{x}, \boldsymbol{x}') = q'(\boldsymbol{x}' - \boldsymbol{x})$, this translation-invariance – "random walk process" with multivariate Normal or Student-t, or split Student$-t$ among other choices for $q'(\boldsymbol{x}' - \boldsymbol{x})$.
- Take $q(\boldsymbol{x}, \boldsymbol{x}') = q(\boldsymbol{x}')$, which leads to

  $$\min\left\{\frac{\omega(\boldsymbol{x}')}{\omega(\boldsymbol{x})}, 1\right\}$$

  where $w(\boldsymbol{x}) = \pi(\boldsymbol{x})/q(\boldsymbol{x})$. This is like the importance sampling via the importance weights $\omega(\boldsymbol{x})$. Again, possible choices for q can be Student$-t$, or split Student$-t$ -with small degrees of freedom so that we explore the whole region including the tails more easily.

Convergence is a big issue with these MCMC algorithms; there are valid results about convergence, which are completely useless.[7]

The last case, the importance sampling type, is what is given in Casella and Berger.

Algorithm 1.2 (Importance sampling)

Let $Y \sim f_Y(y)$ and $V \sim f_V(v)$ with common support

0. Generate $V \sim f_V$. Set $Z_0 = V$. For $i = 1, \ldots$
1. Generate $U_i \sim \mathcal{U}(0,1), V_i \sim f_V$ and calculate

   $$\rho_i = \min\left\{\left(\frac{f_Y(V_i)}{f_V(V_i)}\frac{f_V(Z_{i-1})}{f_Y(Z_{i-1})}\right), 1\right\}$$

[7] A variant of simulated annealing (a method arising from metalogy) is the more recent proposal of equi-energy sampler in *Annals of Statistics*

2. Set

$$Z_i = \begin{cases} V_i & \text{if } U_i \leq \rho_i \\ Z_{i-1} & \text{if } U_i > \rho_i \end{cases}$$

Note that the probability of acceptance is ρ_i. $\mathsf{P}\,(\text{move}) = \mathsf{P}\,(Z_{i+1} = V_i) = \mathsf{P}\,(U_i \leq \rho_i) = \rho_i$.

An interesting article and another useful algorithm is the **Gibbs sampler**, when the conditional $\pi(\theta_i|\boldsymbol{\theta}_{-i}, \boldsymbol{x})$ is available $\forall\, \theta_i$ is available. See to that effect the paper by Smith and Roberts, *Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods* (1993).

Section 2
Data reduction principle

2.1 Sufficiency

Consider first the easy (yet insightful) case of a Bernoulli random variable,

$$\mathsf{P}(X = x) = p^x(1-p)^{1-x}\mathbf{1}_{x\in\{0,1\}}$$

This distribution is used for example to analyse referendum and voting for a given party. For $X_i \in \{0,1\}, \mathsf{P}(X_i = 1) = p$, where p is also the average. The average has all the good sample property: unbiased, MLE estimate, $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$ where we do estimate based on an infinite population. From this, we can obtain confidence intervals. The Hypergeometric could be used for sampling-resampling procedure (yet the gain in efficiency for using the population is negligeable).

We also have Polya's urn problem. As long as we do not know the probability of the first draw, the probability of drawing doesn't change, where drawing from an urn blue and red marbles, as

$$\mathsf{P}(X_1 = 1) = \frac{R}{R+B}$$
$$\mathsf{P}(X_2 = 1) = \mathsf{P}(X_1 = 1|X_1 = 1)\,\mathsf{P}(X_1 = 1) + \mathsf{P}(X_2 = 1|X_1 = 0)\,\mathsf{P}(X_1 = 0) = \frac{R}{R+B}$$

which is still the same probability, yet we do not have independence.

Assuming that n is known, the likelihood

$$\begin{aligned}\mathcal{L}(p; x_1, \ldots, x_n) &= \prod_{i=1}^n \mathsf{P}(X_i = x_i)\\ &= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}\\ &= p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}\\ &= \left(\frac{p}{1-p}\right)^{\Sigma x_i}(1-p)^n\end{aligned}$$

and define $T = \sum_{i=1}^n X_i$. Then, to check for sufficient statistic, we need to assert whether conditional on the postulated statistic, the likelihood is independent on the unknown parameter (in this case p).

Indeed,

$$\begin{aligned}
\mathsf{P}_p\left(X_i = x_i, i = 1, \ldots, n | T = t\right) &= \frac{\mathsf{P}_p\left(X_i = x_i; i = 1, \ldots, n, T = t\right)}{\mathsf{P}_p(T = t)} \mathbf{1}_{\Sigma x_i = t} \\
&= \frac{\mathsf{P}_p\left(X_i = x_i; i = 1, \ldots, n\right)}{\mathsf{P}_p(T = t)} \mathbf{1}_{\Sigma x_i = t} \\
&= \frac{p^{\Sigma\, x_i}(1-p)^{n - \Sigma\, x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} \mathbf{1}_{\Sigma x_i = t} \\
&= \frac{1}{\binom{n}{t}} \mathbf{1}_{\Sigma x_i = t}
\end{aligned}$$

In this case, the cardinality of the set

$$\mathcal{A}_t = \left\{ (x_1, \ldots, x_n) : x_i \in \{0, 1\}, \sum_{i=1}^{n} x_i = t \right\}$$

is $\binom{n}{t}$. In this case, the dimension-reduction for the problem was from n to 1. However, even if it always possible to get a sufficient statistic in every case, we may not be able to do data-compression.

Example 2.1

Let $\{X_i\}_{i=1}^n$ be an IID sample $X_i \overset{\text{iid}}{\sim} f$, f continuous.

Looking at the order statistics, let us rederive the joint distribution of two order statistic using the multinomial distribution

$$\begin{aligned}
f_{X_{(r)}, X_{(s)}}(x, y)\, \mathrm{d}x\, \mathrm{d}y &\approx \mathsf{P}\left(x < X_{(r)} < x + \Delta x, y < X_{(s)} + y + \Delta y\right) \\
&= \binom{n}{r-1, s-r-1, n-s} \mathsf{P}\left(X \leq x\right)^{r-1} \mathsf{P}\left(x < X < x + \Delta x\right) \\
&\quad \times \mathsf{P}\left(x + \Delta x < X < y\right)^{s-r-1} \mathsf{P}\left(y < X < y + \Delta y\right) \mathsf{P}\left(Y > y\right)^{n-s}
\end{aligned}$$

so that $f_{X_{(r)}, X_{(s)}}(x, y)$

$$\begin{aligned}
&= \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \mathsf{P}\left(x < X_{(r)} < x + \Delta x, y < X_{(s)} < y + \Delta y\right) \Delta x \Delta y \\
&= \binom{n}{r-1, s-r-1, n-s} F_X^{r-1} f_X(x) [F_X(y) - F_X(x)]^{s-r-1} f_X(y) [1 - F_X(y)]^{n-s}
\end{aligned}$$

and in the case where we have all order statistics, $f_{X_{(1)}, \ldots, X_{(n)}}(y_1, \ldots, y_n) = n! \prod_{i=1}^n f_X(y_i) =$

$\prod_{i=1}^n f_X(x_i)$ and

$$\begin{aligned} f_{\boldsymbol{X}|X_{(1)},\ldots,X_{(n)}}(\boldsymbol{x}|\boldsymbol{y}) &= \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{(\boldsymbol{X})}(\boldsymbol{y})} \\ &= \frac{\prod_{i=1}^n f_X(x_i)}{n! \prod_{i=1}^n f_X(x_i)} \\ &= \frac{1}{n!} \end{aligned}$$

The spacing of the empirical cumulative distribution function, the non-parametric estimate for the CDF, takes into account the space, while $\widehat{f}_n = 1/n$.

Example 2.2 (Sufficient statistics for the Normal distribution)
Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$ and consider the random variable $T = \sum_{i=1}^n X_i$. Then, the likelihood is

$$\begin{aligned} \mathcal{L}(\mu; \boldsymbol{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \mu)^2}{2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2\right) \end{aligned}$$

and we easily see that $T \sim \mathcal{N}(n\mu, n)$ through moment generating function arguments. We also have

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2$$

which entails that the ratio is

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}|t) &= \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_T(t)} \mathbf{1}_{\Sigma x_i = t} \\ &= \frac{\left(\sqrt{2\pi}\right)^{-n} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2\right)}{(2\pi n)^{-\frac{1}{2}} \exp\left(-\frac{(t-n\mu)^2}{2n}\right)} \mathbf{1}_{\Sigma x_i = t} \\ &= C \frac{\exp\left(-\frac{1}{2}\left[\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2\right]\right)}{\exp\left(-\frac{1}{2n}(t^2 - 2nt\mu + n^2\mu^2)\right)} \mathbf{1}_{\Sigma x_i = t} \\ &= C \frac{\exp\left(-\frac{1}{2}\left[\sum_{i=1}^n x_i^2\right]\right)}{\exp\left(-\frac{1}{2}t^2\right)} \mathbf{1}_{\Sigma x_i = t} \end{aligned}$$

We now consider a case for the Poisson, where $X_1, X_2 \overset{\text{iid}}{\sim} \mathcal{P}(\lambda)$, it is easy to show $T = X_1 + X_2$

is sufficient (exercise). For a Poisson, however, taking

$$\mathsf{P}\left(X_1 = x_1, X_2 = x_2 | T = t\right) = \binom{t}{x_1} \frac{1}{2^t} \mathbf{1}_{x_1 + x_2 = t}$$

and zero otherwise and in this example,

$$\mathcal{A}_t = \{(x_1, x_2) : x_1 + x_2 = t\}.$$

However, taking $X_1 + 2X_2$ is not sufficient. Indeed,

$$\begin{aligned}
\mathsf{P}\left(X_1 = 0, X_2 = 1 | X_1 + 2X_2 = 2\right) &= \frac{\mathsf{P}\left(X_1 = 0, X_2 = 1\right)}{\mathsf{P}\left(X_1 + 2X_2 = 2\right)} \\
&= \frac{\mathsf{P}\left(X_1 = 0, X_2 = 1\right)}{\mathsf{P}\left(X_1 = 0, X_2 = 1\right) + \mathsf{P}\left(X_1 = 2, X_2 = 0\right)} \\
&= \frac{e^{-\lambda}\lambda e^{-\lambda}}{e^{-\lambda}\lambda e^{-\lambda} + \frac{e^{-\lambda}\lambda^2}{2!} e^{-\lambda}} \\
&= \frac{1}{1 + \frac{\lambda}{2}}
\end{aligned}$$

which is still a function of λ.

2.2 Fisher-Neyman factorization theorem

Here is a general tool that gives us recipe to find a sufficient statistic.

Theorem 2.1 (Fisher-Neyman Factorization criterion)

Let $X_1, \ldots, X_n$ be distributed according to $\mathsf{P}_\theta(x_1, \ldots, x_n), \theta \in \Theta$. Then $T(X_1, \ldots, X_n)$ is sufficient for θ if and only if we can write

$$\mathsf{P}_\theta(x_1, \ldots, x_n) = h(x_1, \ldots, x_n) g_\theta(T(x_1, \ldots, x_n)) \tag{2.2}$$

where h is a **non-negative** function of $x_1, \ldots, x_n$ only and does not depend on θ and g is a non-negative function of θ and $T(x_1, \ldots, x_n)$ only.

Proof The proof in the general case is quite involved. We have a proof by Halmos and Savage, with $\{\mathsf{P}_\theta : \theta \in \Theta\}$ and $\mathsf{P}_\theta << \nu, \theta \in \Theta$ and is also available in Lehman's book in *Testing of Statistical Hypothesis* and in Shao's book. We present the version for the discrete case.

Let T be sufficient for θ. Then $\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}|T = t)$ is independent of θ and we may write

$$\begin{aligned}\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) &= \mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = t) \\ &= \mathsf{P}\left(T = t\right)\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}|T = t) \\ &= g_\theta(T(\boldsymbol{x}))h(\boldsymbol{x})\end{aligned}$$

Conversely, suppose (2.2) holds. Then

$$\begin{aligned}\mathsf{P}_\theta(T = t) &= \sum_{\boldsymbol{x}:T(\boldsymbol{x})=t} \mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) \\ &= \sum_{\boldsymbol{x}:T(\boldsymbol{x})=t} g_\theta(T(\boldsymbol{x}))h(\boldsymbol{x}) \\ &= g_\theta(t) \sum_{\boldsymbol{x}:T(\boldsymbol{x})=t} h(\boldsymbol{x})\end{aligned}$$

Suppose $\mathsf{P}_\theta(T = t) > 0$, for some θ. Then

$$\begin{aligned}\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}|T = t) &= \frac{\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x}, T = t)}{\mathsf{P}_\theta(T = t)} \\ &= \frac{\mathsf{P}_\theta(\boldsymbol{X} = \boldsymbol{x})}{\mathsf{P}_\theta(T = t)}\mathbf{1}_{T(\boldsymbol{x})=t} \\ &= \frac{g_\theta(T(\boldsymbol{x}))h(\boldsymbol{x})}{g_\theta(T(\boldsymbol{x}))\sum_{\boldsymbol{x}:T(\boldsymbol{x})=t} h(\boldsymbol{x})}\mathbf{1}_{T(\boldsymbol{x})=t} \\ &= \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{x}:T(\boldsymbol{x})=t} h(\boldsymbol{x})}\mathbf{1}_{T(\boldsymbol{x})=t}\end{aligned}$$

■

Fisher defined information such that it increases linearly with the sample size, taking logs yields a sum. The idea of derivative is for sensitivity.

Example 2.3 (Sufficient statistic for Normal)
Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then

$$\begin{aligned}\mathcal{L}(\mu, \sigma^2; \boldsymbol{x}) = f_{\mu,\sigma^2}(\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)\end{aligned}$$

and breaking up the sum, we get

$$\sum_{i=1}^{n}(x_i-\mu)^2=\sum_{i=1}^{n}x_i^2+n\mu^2-2\mu\sum_{i=1}^{n}x_i$$

so that the likelihood[8] is rewritten in the form

$$\begin{aligned}\mathcal{L}(\mu,\sigma^2;\boldsymbol{x})&=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2+n\mu^2-2\mu\sum_{i=1}^{n}x_i\right)\right)\\&=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n g_{\mu,\sigma^2}\left(\sum_{i=1}^{n}x_i,\sum_{i=1}^{n}x_i^2\right).\end{aligned}$$

Using the factorization theorem, the pair $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ is a sufficient statistic for (μ,σ^2)

Remark

It is false to say that $\sum_{i=1}^n x_i$ is sufficient for μ and $\sum_{i=1}^n x_i^2$ for σ^2; we will define later the notion of partial sufficiency; writing out the problem, we keep dependence remaining.

Exercise 2.1

Show that if S is sufficient and h is a one-to-one function, then $h(S)$ is also sufficient.

Using the result of the above exercise, we can show that $(\bar{X}_n, S_n^2)$ is sufficient for (μ,σ^2), namely there is a injective map from

$$\left(\sum_{i=1}^{n}x_i,\sum_{i=1}^{n}x_i^2\right)\longleftrightarrow(\bar{X}_n,S_n^2)$$

where

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

Before proving results for more general members of the exponential family, we review the notion of

Definition 2.2 (Exponential family)

Let $\Theta\subseteq\mathbb{R}^d$ and $\{x\in\mathbb{R}^m : f_{\boldsymbol{\theta}}(\boldsymbol{x})>0\}$ does not depend on $\boldsymbol{\theta}$. If there exists real-valued functions $\omega_1,\ldots,\omega_k$ and c on Θ and real-valued functions $t_1,\ldots,t_k$ and h on $\mathbb{R}^k$, where

[8] Recall that likelihood is a post-experimental approach

$d \leq k$,[9] such that

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = h(\boldsymbol{x})c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} \omega_i(\boldsymbol{\theta}) t_i(\boldsymbol{x})\right) \tag{2.3}$$

we say that the family $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is a d-parameter exponential family.

Remark

We like exponential family since they are the maximum **entropy** distributions, where entropy is of the form $H(p_1, \ldots, p_k) = -\sum_{i=1}^{k} p_i \log(p_i)$, which is maximized for discrete distribution assigning equal weights to all points on the support. We require continuity. This is an extension of Laplace principle of insufficient reason (give equal chance to all cases). In many cases, we have some extra information about the mean of the variance, and in such case H is maximized under some additional criterion.

Imposing second order conditions, and maximizing entropy, we get the Normal distribution and setting conditions on the first order, the maximum entropy yields exponential distribution. We use then the most "honest" distributions.

Remark

All GLMs and quasi-likelihood methods are based on the entropy principle. For more on quasi-likelihood, refer to Wedderburn (1974) article *Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method* in Biometrika

Exercise 2.2 (Sufficient statistics for exponential family)

Suppose that $X_i \overset{\text{iid}}{\sim} f_{\boldsymbol{\theta}}(\boldsymbol{x})$ where $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is given by (2.3). Show that

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

is sufficient.

Example 2.4

Let $f_\theta(x) = q(\theta)h(x), 0 \leq x \leq \theta$ where q and h are both positive functions and $\theta \in \Theta = \mathbb{R}^+$. Suppose $X_i \overset{\text{iid}}{\sim} f_\theta(x)$. Then

$$\mathcal{L}(\theta; \boldsymbol{x}) = q^n(\theta) \prod_{i=1}^{n} h(x_i) \mathbf{1}_{0 \leq x_i \leq \theta \ \forall \ i}$$

[9] This requirement is for identifiability of w_i, then we need from those equations to identify the θ from those w_i

and note that the support of θ depends on the observation. We can write this as

$$\mathcal{L}(\theta;\boldsymbol{x}) = q^n(\theta)\prod_{i=1}^n h(x_i)\prod_{i=1}^n \mathcal{E}(x)\mathcal{E}(\theta - x_i)$$
$$\left[\mathcal{E}\left(\bigwedge_{i=1}^n x_i\right)\prod_{i=1}^n h(x_i)\right]\left[q^n(\theta)\mathcal{E}\left(\theta - \bigvee_{i=1}^n x_i\right)\right]$$

where

$$\mathcal{E}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\bigwedge_{i=1}^n x_i = \min_{1\leq i\leq n} x_i, \qquad \bigvee_{i=1}^n x_i = \max_{1\leq i\leq n} x_i$$

so that the sufficient statistic $T(\boldsymbol{x}) = \bigvee_{i=1}^n x_i$.

Take the mapping $f : X \to Y$ such that the values are partitionned into almost disjoint sets. When conditioning on T, we restrict ourselves to a partition set. When the sets are bigger (meaning there are more values mapped to the same point a_i), we have more compression taking the minimal sufficient statistic. If a statistic is complete, then it is also minimal

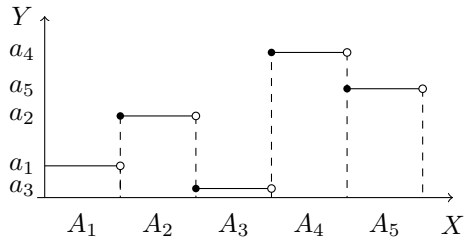


Figure 2: Partition sets for minimal statistics

sufficient. Ancillary statistics are perpendicular to the minimal sufficient statistic.

Taking the joint distribution of all order statistics, $n!\prod_{i=1}^n f(x_i)$ for $X_{(1)},\ldots,X_{(n)}$. Let $R(X_i)$ =rank of X_i, then the joint distribution

$$\mathsf{P}\left(R_i = r_i\ \forall\ i\right) = \frac{1}{n!}$$

The distribution of this does not depend on any unknown parameter (it is ancillary), while the others (the order statistic) are complete sufficient statistics and thus minimal. This means that the joint distribution of the order statistic with the ranks, we get from independence back the data, despite the fact that the ranks have no information. There is orthogonality (independence) between the sufficient and ancillary statistics. Putting them together, we get exactly everything and we can recover the sample.

For minimal sufficient statistic, which we define shortly, the partitions sets are chubbiest.

Definition 2.3 (Minimal sufficient statistic)

A sufficient statistic $T(X)$ is called **minimal sufficient statistic** if for any other sufficient statistic $T'(X)$, then $T(X)$ is a function of $T'(X)$.

Theorem 2.4

Let $f_\theta(\boldsymbol{x})$ be the PMF or PDF of a sample $\boldsymbol{x}$. **Suppose**[10] there exists a function $T(\boldsymbol{X})$ such that for every two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the ratio $f_\theta(\boldsymbol{x})/f_\theta(\boldsymbol{y})$ is constant as a function of θ if and only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then $T(\boldsymbol{X})$ is a minimal sufficient statistic for θ.

Proof For the sake of simplicity, assume that $f_\theta(\boldsymbol{x}) > 0 \; \forall \; \boldsymbol{x} \in \mathcal{X}$ (the sample space) and $\theta \in \Theta$ (the parameter space).

First, we show that T is sufficient. Let

$$\mathcal{T} = \{t : t = T(\boldsymbol{x}), \text{ for some } \boldsymbol{x} \in \mathcal{X}\}$$

Define the partition sets induced by $T(\boldsymbol{x})$ as

$$A_t = \{\boldsymbol{x} \in \mathcal{X} : T(\boldsymbol{x}) = t\} = T^{-1}(\{t\}).$$

For each A_t, choose and fix one element $\boldsymbol{x}_t \in A_t$. For any $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{x}_{T(\boldsymbol{x})}$ is the fixed element that is in the same set, A_t, as $\boldsymbol{x}$. Since $\boldsymbol{x}$ and $\boldsymbol{x}_{T(\boldsymbol{x})}$ are in the same set A_t, $T(\boldsymbol{X}) = T(\boldsymbol{X}_{T(\boldsymbol{x})})$ and hence the ratio $f_\theta(\boldsymbol{x})/f_\theta(\boldsymbol{x}_{T(\boldsymbol{x})})$ is constant as a function of θ. Thus we can define a function on $\mathcal{X}$ by $h(\boldsymbol{x}) = f_\theta(\boldsymbol{x})/f_\theta(\boldsymbol{x}_{T(\boldsymbol{x})})$ and h does not depend on θ. Define a function on $\mathcal{T}$ by $g_\theta(t) = f_\theta(\boldsymbol{x}_{T(\boldsymbol{x})})$. Then

$$f_\theta(\boldsymbol{x}) = \frac{f_\theta(\boldsymbol{x}_{T(\boldsymbol{x})}) f_\theta(\boldsymbol{x})}{f_\theta(\boldsymbol{x}_{T(\boldsymbol{x})})} = g_\theta(T(\boldsymbol{x}))h(\boldsymbol{x})$$

and by the factorization theorem, $T(\boldsymbol{X})$ is a sufficient statistic for θ.

[10] Based on the axiom on choice

Now to show that T is minimal, let T' be another sufficient statistic. By the factorization theorem, there are functions f' and g' such that

$$f_\theta(\boldsymbol{x}) = g'_\theta(T'(\boldsymbol{x}))h'(\boldsymbol{x}).$$

Let $\boldsymbol{x}, \boldsymbol{y}$ be any two sample points with $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$. Then

$$\frac{f_\theta(\boldsymbol{x})}{f_\theta(\boldsymbol{y})} = \frac{g'_\theta(T'(\boldsymbol{x}))h'(\boldsymbol{x})}{g'_\theta(T'(\boldsymbol{y}))h'(\boldsymbol{y})} = \frac{h'(\boldsymbol{x})}{h'(\boldsymbol{y})}.$$

Since the ratio does not depend on θ, the assumption of the theorem implies that $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Thus $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$ (why?) and $T(\boldsymbol{x})$ is minimal. ■

2.3 Completeness

The following notion is due to Lehmann & Scheffé, in an article published in 1950 in *Sankhya*

Definition 2.5 (Completeness)
Let $\{f_\theta : \theta \in \Theta\}$ be a family of PDF's (or PMF's) (or bounded). We say this family is **complete** (or boundedly complete; if for any g measurable) if

$$\mathsf{E}_\theta\left(g(X)\right) = 0, \quad \forall\, \theta \in \Theta \quad \text{implies that } \mathsf{P}_\theta\left(g(X) = 0\right) = 1,\ \forall\, \theta \in \Theta$$

Note that the probability is under the distribution of the sufficient statistic. This notion of completeness is a uniqueness requirement.

This is a bias-variance tradeoff; in the $\mathcal{L}^2$-norm,

$$\mathsf{E}\left((T(X) - \psi(\theta))^2\right) = \mathsf{Var}\left(T\right) + \left[\mathsf{Bias}_{\psi(\theta)}T\right]^2$$

Recall that **unbiasedness** is the condition

$$\mathsf{E}_\theta(T) = \psi(\theta),\ \forall\, \theta.$$

we will try to minimize the mean-square error, and so a restriction to the subclass of unbiased estimators; we will see UMVUE later in the course.. Rao and Blackwell showed that conditioning on a sufficient statistic yields another statistic which has lower variability. The missing piece of the puzzle was the notion of completeness; in such case, we get the maximum reduction and conditioning once will be sufficient to get the minimum variance estimator.

Definition 2.6 (Complete statistic)

A statistic T is said to be complete if the family of distribution of T is complete.

Example 2.5

Let $X_i \overset{\text{iid}}{\sim} \mathcal{B}(p)$ for a sample of size n, and let $T = \sum_{i=1}^n X_i \sim \mathcal{B}(n,p)$. Consider

$$\begin{aligned} 0 = \mathsf{E}_p(g(T)) &= \sum_{t=0}^n g(t)\binom{n}{t}p^t(1-p)^{n-t} \\ &= (1-p)^n \sum_{t=0}^n g(t)\binom{n}{t}\left(\frac{p}{1-p}\right)^t = 0 \end{aligned}$$

Define $\theta = \frac{p}{1-p}$, $a_t = g(t)\binom{n}{t}$ and consider

$$\sum_{t=0}^n a_t \theta^t = 0, \qquad \forall\, \theta \in \mathbb{R}^+$$

by the Fundamental theorem of algebra, has at most n roots, therefore $a_t = 0,\ \forall\, t = 0, 1, \ldots, n$ and as such $g(t) = 0$. We have thus found a minimal complete sufficient statistic.

Example 2.6

Let $X \sim \mathcal{N}(0,\theta)$, $\theta \in \mathbb{R}^+$. Since $\mathsf{E}_\theta(X) = 0$, one can easily see in this case that X is not complete, however $T = X^2$ is complete. Indeed, $X^2/\theta \sim \chi^2(1)$ and $X^2 \sim \theta\chi^2(1)$ which yields PDF

$$f_T(t) = \frac{e^{-\frac{t}{2\theta}}}{\sqrt{2\pi\theta}} t^{-\frac{1}{2}} \mathbf{1}_{t>0}$$

and as such

$$\mathsf{E}_\theta(g(T)) = \frac{1}{\sqrt{2\pi\theta}} \int_0^\infty g(t) t^{-\frac{1}{2}} e^{-\frac{t}{2\theta}} dt = 0, \qquad \forall\, \theta \in \mathbb{R}^+$$

which implies

$$\int_0^\infty g(t) t^{-\frac{1}{2}} e^{-\frac{t}{2\theta}} dt = 0$$

and using the uniqueness property of the Laplace transform, we have $g(t)t^{-\frac{1}{2}} = 0,\ \forall\, t > 0$, which implies $g(t) = 0,\ \forall\, t > 0$.

In many cases, we will restrain ourselves to continuous functions, since we do not gain more insight from looking at measurable function, and it takes more time (and requires measure theory). Using Lusin and Egorov theorem, we can approximate by continuous functions, using Littlewood's principles.

Example 2.7

Let $X \sim \mathcal{U}(0, \theta)$ for $\theta \in \mathbb{R}^+$. Then

$$\mathsf{E}_\theta(g(X)) = \int_0^\theta \frac{1}{\theta} g(x)\,\mathrm{d}x = 0 \Rightarrow \int_0^\theta g(x)\,\mathrm{d}x = 0.$$

Using the fundamental theorem of calculus, then $g(\theta) = 0$ implies $g(x) = 0$, $\forall\, x \in \mathbb{R}^+$. If we only had measurable function, we could break this into $g^+(x)$ and $g^-(x)$ and these functions are equal on any interval. Looking at this using Egorov's theorem on a set where we have almost continuous function, we can separate the interval into two sets U, A, one on which we have continuity and the other of measure zero; looking at $m_*(U \Delta A) < \varepsilon$, the outer measure of the symmetric difference, and we can provided that $g(x)$ is bounded find such ε and get the result for $g(X)$ an arbitrary bounded measurable function.

Example 2.8

Consider again a case where $X_i \overset{\text{iid}}{\sim} \mathcal{U}(0, \theta)$ and the statistic $T = \vee_{i=1}^n X_i$, which we have shown to be sufficient Then

$$f_T(x; \theta) = \frac{n}{\theta^n} x^{n-1} \mathbf{1}_{0<x<\theta}$$

and so

$$\mathsf{E}_\theta(g(T)) = \int_0^\theta g(x) \frac{n}{\theta^n} x^{n-1}\,\mathrm{d}x = 0, \quad \forall\, \theta \in \mathbb{R}^+$$

which is proportional to

$$\int_0^\theta g(x) x^{n-1}\,\mathrm{d}x = 0$$

and using the Fundamental theorem of Calculus

$$g(\theta)\theta^{n-1} = 0$$

and conclude that $g(\theta) = 0\ \forall\, \theta \in \mathbb{R}^+$.

Example 2.9

Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \theta^2)$. One can show (exercise) that $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient. Then

$$g(t) = 2\left(\sum_{i=1}^n x_i\right)^2 - (n+1)\sum_{i=1}^n x_i^2$$

and $\mathsf{E}_\theta(g(T)) = 0$, $\forall\, \theta$

Example 2.10

Consider the random variables $X_i \overset{\text{iid}}{\sim} \mathcal{U}(N)$, discrete uniform random variables. Again, it is

easy to show that $M_n = \bigvee_{i=1}^n X_i$ is sufficient. Then, since the variable is discrete, write

$$\begin{aligned}\mathsf{P}\left(M_n = x\right) &= \mathsf{P}\left(M_n \leq x\right) - \mathsf{P}\left(M_n \leq x-1\right) \\ &= \left(\frac{x}{N}\right)^n - \left(\frac{x-1}{N}\right)^n.\end{aligned}$$

Then, $\forall\, N \in \mathbb{N}$, we require

$$\begin{aligned}0 &= \mathsf{E}_n(g(M_n)) \\ &= \sum_{x=1}^{N} g(x)\left[\frac{x^n}{N^n} - \frac{(x-1)^n}{N^n}\right] \\ &= \frac{1}{N^n}\sum_{x=1}^{N} g(x)[x^n - (x-1)^n]\end{aligned}$$

We thus have for all $N = 1, 2, \ldots$

$$\sum_{x=1}^{N} g(x)[x^n - (x-1)^n] = 0 \tag{2.4}$$

$$\sum_{x=1}^{N-1} g(x)[x^n - (x-1)^n] = 0 \tag{2.5}$$

for $N = 2, 3, \ldots$ and using (2.4) and (2.5), we get $g(N)[N^n - (N-1)^n] = 0$ using the telescoping sum. This imply that $g(N) = 0$ for $N = 2, 3, \ldots$ and it follows from (2.4) that $g(1) = 0$.

2.4 Ancillarity and Basu's theorem

Definition 2.7 (Ancillary statistic)

A statistic $V(\boldsymbol{X})$ is said to be **ancillary** if its distribution doesn't depend on any unknown parameter. $V(\boldsymbol{X})$ is called **first-order ancillary** if $\mathsf{E}_\theta(V)$ does not depend on θ.

An example of such statistic is the rank. We now go for a few more examples.

Example 2.11

If we have a location family, where $X_i \overset{\text{iid}}{\sim} F_\theta(x) = F(x-\theta)$. For example, if $X \sim \mathcal{N}(\mu, 1)$ then we have $Z = X - \mu \sim \mathcal{N}(0, 1)$ does not depend on μ.

We have seen the range $R = X_{(n)} - X_{(1)}$ and

$$\begin{aligned} F_R(r) &= \mathsf{P}\left(R \leq r\right) \\ &= \mathsf{P}\left(X_{(n)} - X_{(1)} \leq r\right) \\ &= \mathsf{P}\left((X_{(n)} - \theta) - (X_{(1)} - \theta) \leq r\right) \\ &= \mathsf{P}\left(Z_{(n)} - Z_{(1)} \leq r\right) \end{aligned}$$

does not depend on θ, where $Z_i = X_i - \theta$.

Theorem 2.8 (Basu)
If T is a complete sufficient statistic for the family $\mathrm{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$, then any ancillary statistic V is independent of T.

Proof If V is ancillary, the probability $p_A = \mathsf{P}\left(V \in A\right)$ is independent of θ for all A. Let $\eta_A(t) = \mathsf{P}\left(V \in A | T = t\right)$. Then

$$\mathsf{E}_\theta\left(\eta_A(T)\right) = p_A$$

hence

$$\mathsf{E}_\theta\left(\eta_A(T) - p_A\right) = 0$$

for all $\theta \in \Theta$. Then $\eta_A(T) = p_A$ almost everywhere P. This means that $\mathsf{P}\left(V \in A | T = t\right) = \mathsf{P}\left(V \in A\right)$. ■

Note
We can replace completeness by boundedly completeness for the step $\mathsf{E}_\theta\left(\eta_A(T) - p_A\right) = 0$.

With the ranks, we can get complete decompression and not only get similar data generation, but the original data using a sufficient or a minimal sufficient statistic. When ancillary statistic are placed next to them, those statistics (like ranks) become quite informative. If we have pairs (X_i, Y_i), knowing how concordant or disconcordant the ranks are, we can assess the dependence structure.

Example 2.12 (Nonparametric completeness)
Consider $\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R}^+; \int_{\mathbb{R}} d\mu = 1, f \text{ is continuous a.e. } \mu\}$. We show this family $\mathcal{F}$ is complete, so for $X_i \overset{\text{iid}}{\sim} f \in \mathcal{F}$, then $T = (X_{(1)}, \ldots, X_{(n)})$ is complete. We want to show that $\mathsf{E}_f(g(T)) = 0$, $\forall\, f \in \mathcal{F}$ implies $g \equiv 0$. The complete proof for general measurable functions is available in Lehmann *Testing Statistical Hypotheses* (second edition), on pages 143-144 or in Lehmann and Romano (third edition) on page 118.

Order statistics are sufficient, but not always complete. This is the case for example with the logistic distribution, given by

$$f(x) = \frac{e^x}{1+e^x}\mathbf{1}_{x\in\mathbb{R}}.$$

The logistic distribution was suggested by Berkson; it ressembles a Normal distribution, yet has heavier tails.

Proof Let $T = (X_{(1)}, \ldots, X_{(n)})$ and h be a continuous function of T. Suppose $\mathsf{E}_f(h(T)) = 0,\ \forall\ f \in \mathcal{F}$. Wlog, suppose $f : [0,1] \to [0,1]$. Then, by definition we have

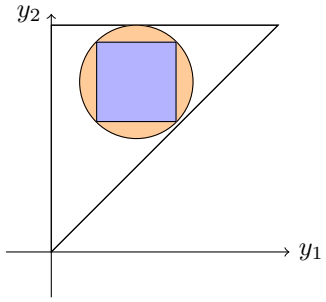
$$\mathsf{E}_f(h(T)) = \int_0^1 \int_{y_1}^1 \cdots \int_{y_{n-1}}^1 h(y_1, \ldots, y_n) n! \prod_{i=1}^n f(y_i)\, \mathrm{d}y_i = 0,\ \forall\ f \in \mathcal{F}$$

implies

$$\int_{[0,1]^n} \mathbf{1}_A(y_1, \ldots, y_n) h(y_1, \ldots, y_n) \prod_{i=1}^n f(y_i)\, \mathrm{d}y_i = 0,\ \forall\ f \in \mathcal{F}$$

where $A = \{(y_1, \ldots, y_n) \in [0,1]^n : 0 \leq y_1 \leq \cdots \leq y_n < 1\}$. Suppose for a contradiction $h(\boldsymbol{y}^*) > 0$ for some $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*) \in A$. Then, there exists $\delta > 0$ such that $h(\boldsymbol{y}) > 0,\ \forall\ \boldsymbol{y} \in B_{\boldsymbol{y}^*}(\delta)$ where $B_{\boldsymbol{y}^*}(\delta) = \{\boldsymbol{y} \in [0,1]^n : \|\boldsymbol{y} - \boldsymbol{y}^*\| < \delta\}$. Suppose $C_{\boldsymbol{y}^*}(\delta)$ is the greatest hypercube inscribed in $B_{\boldsymbol{y}^*}(\delta)$ boundary faces, since data is IID.

Figure 3: Picture of simplex: illustration for sample of size two



Then, there exists $0 < \alpha < \beta < 1$ with $C_{\boldsymbol{y}^*} = \{(y_1, \ldots, y_n) \in [0,1]^n, \alpha < y_i < \beta\ \forall\ i\}$. Now suppose $f(y) = (\beta - \alpha)^{-1}\mathbf{1}_{\alpha<y<\beta}$. Then

$$\int_{C_{\boldsymbol{y}^*}(\delta)} \mathbf{1}_{A(\boldsymbol{y})} h(\boldsymbol{y}) \prod_{i=1}^n \mathrm{d}y_i = 0.$$

This is a contradiction since $h(\boldsymbol{y}) > 0$ on $C_{\boldsymbol{y}^*}$. ■

To summarize, the idea of the proof is the following: look at greatest cube that fit in ball,

and choose f that only puts mass on the sides, since it is uniform, since inside and uniformly positive; since we integrate a positive function and the result is zero, the function must be null.

$T(X_{(1)}, \ldots, X_{(n)})$ is complete sufficient for $\mathcal{F}$ and $(R(X_1), \ldots, R(X_n))$ is ancillary. Then, using Basu's theorem

$$\boldsymbol{T} = (X_{(1)}, \ldots, X_{(n)}) \perp\!\!\!\perp (R(X_1), \ldots, R(X_n)) = \boldsymbol{R}.$$

Thus $f_{\boldsymbol{T},\boldsymbol{R}} = f_{\boldsymbol{X}} = \prod_{i=1}^{n} f(y_i)$. [11]

2.5 Complete-sufficient statistic in an exponential family

Recall that a parametric family $\mathrm{P} = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ dominated by σ-finite measure ν on $(\Omega, \mathcal{F})$ is called an exponential family if and only if

$$\frac{\mathrm{d}\mathsf{P}_\theta}{\mathrm{d}\nu}(\boldsymbol{\omega}) = h(\boldsymbol{\omega}) \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^\top T(\boldsymbol{\omega}) - \xi(\boldsymbol{\theta})\right)$$

where $T(\boldsymbol{\omega})$ is a random p-vector, $\boldsymbol{\eta}$ is a function from Θ to $\mathbb{R}^p$, h is a non-negative Borel function on $(\Omega, \mathcal{F})$ and

$$\xi(\boldsymbol{\theta}) = \log\left(\int_\Omega \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^\top T(\boldsymbol{\omega})\right) h(\boldsymbol{\omega})\, \mathrm{d}\nu(\boldsymbol{\omega})\right)$$

Note that any transformation $\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}) = D\boldsymbol{\eta}(\boldsymbol{\theta})$ with $p \times p$ nonsingular matrix D gives another valid representation with T replaced by $\tilde{T} = (D^\top)^{-1} T$ (why?). A change of measure that dominates the family also changes the representation. For example, if

$$\lambda(A) = \int_A h\, \mathrm{d}\nu, \ \forall\ A \in \mathcal{F},$$

we obtain

$$\frac{\mathrm{d}\mathsf{P}_\theta}{\mathrm{d}\nu}(\boldsymbol{\omega}) = \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^\top T(\boldsymbol{\omega}) - \xi(\boldsymbol{\theta})\right)$$

(why?). Consider the reparametrization $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\theta})$ and

$$f_{\boldsymbol{\eta}}(\boldsymbol{\omega}) = \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^\top T(\boldsymbol{\omega}) - \zeta(\boldsymbol{\theta})\right) h(\boldsymbol{\omega}), \tag{2.6}$$

$\boldsymbol{\omega} \in \Omega$ where

$$\zeta(\boldsymbol{\eta}) = \log\left(\int_\Omega \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{\omega})\right) h(\boldsymbol{\omega})\, \mathrm{d}\nu(\boldsymbol{\omega})\right)$$

[11] To show independence, one must be careful without Basu's theorem, since $\boldsymbol{R}$ is discrete, while $\boldsymbol{T}$ is continuous.

This is called the **canonical form** for the family. The new parameter $\boldsymbol{\eta}$ is called the **natural parameter**. The new parameter space $\mathcal{Z} = \{\boldsymbol{\eta}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, subset of $\mathbb{R}^p$ is called the **natural parameter space.** An exponential family in canonical form is called a **natural exponential family** If there is an open set contained in the natural parameter space of an exponential family, the family is said to be of **full rank.**

Example 2.13

Consider $\mathrm{P} = \{\mathsf{P}_\theta : \mathcal{B}(n, \theta), \theta \in (0, 1)\}$ and

$$f_\theta(x) = \frac{\mathrm{d}\mathsf{P}_\theta}{\mathrm{d}N}(x) = \exp\left(x \log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right)\binom{n}{x}\mathbf{1}_{x\in\{0,1,\ldots,n\}}$$

where N is the counting measure on $\mathbb{N}$.

Note that $T(x) = x, \eta(\theta) = \text{logit}(\theta)$, $\xi(\theta) = -n\log(1-\theta)$ and $h(x) = \binom{n}{x}\mathbf{1}_{x\in\{0,1,\ldots,n\}}$. If we let $\eta = \text{logit}(\theta)$, then $\mathcal{Z} = \mathbb{R}$ and the family P with PMF's

$$f_\eta(x) = \exp\left(x\eta - n\log(1+e^\eta)\right)\binom{n}{x}\mathbf{1}_{0,\ldots,n}(x)$$

is a natural exponential family of full rank. Then $\boldsymbol{\eta}(\theta) = \text{logit}(\theta)$ is called the natural link in GLMs.

Theorem 2.9

If P is in an exponential family with PDF's given by (2.6) and it is of full rank, then $T(X)$ is a complete sufficient statistic for $\boldsymbol{\eta} \in \mathcal{Z}$.

Proof We have already shown that T is sufficient (using the factorization theorem, left as an exercise). Suppose that there is a function f such that $\mathsf{E}_{\boldsymbol{\eta}}(f(T)) = 0,\ \forall\, \boldsymbol{\eta} \in \mathcal{Z}$. Then

$$\int_t f(t)\exp\left(\boldsymbol{\eta}^\top \boldsymbol{t} - \zeta(\boldsymbol{\eta})\right)\,\mathrm{d}\lambda = 0,\ \forall\, \boldsymbol{\eta} \in \mathcal{Z}$$

where $\lambda(A) = \int_A h\,\mathrm{d}\nu$ is a measure on $(\mathbb{R}^p, \mathbb{B}^p)$. In fact,

$$\begin{aligned}\mathsf{P}\left(T \in A\right) &= \int_{\boldsymbol{x}:T(\boldsymbol{x})\in A} \exp\left(\boldsymbol{\eta}^\top T(x) - \zeta(\boldsymbol{\eta})\right) h(\boldsymbol{x})\,\mathrm{d}\nu(\boldsymbol{x}) \\ &= \exp\left(\boldsymbol{\eta}^\top \boldsymbol{t} - \zeta(\boldsymbol{\eta})\right)\int_A h(\boldsymbol{x})\,\mathrm{d}\nu(\boldsymbol{x})\end{aligned}$$

Let $\boldsymbol{\eta}_0$ be an interior point of $\mathcal{Z}$. Then

$$\int f_+(t)e^{\boldsymbol{\eta}^\top \boldsymbol{t}}\,\mathrm{d}\lambda = \int f_-(t)e^{\boldsymbol{\eta}^\top \boldsymbol{t}}\,\mathrm{d}\lambda, \qquad \forall\, \eta \in N(\boldsymbol{\eta}_0) \tag{2.7}$$

where $N(\boldsymbol{\eta}_0) = \{\boldsymbol{\eta} \in \mathbb{R}^p : \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\| < \varepsilon\}$ for some $\varepsilon > 0$. In particular,

$$\int f_+(t)e^{\boldsymbol{\eta}^\top \boldsymbol{t}}\,\mathrm{d}\lambda = \int f_-(t)e^{\boldsymbol{\eta}^\top \boldsymbol{t}}\,\mathrm{d}\lambda = c$$

If $c = 0$, then $f = 0$ almost everywhere λ (why?). If $c > 0$, then $c^{-1}f_+(t)e^{\boldsymbol{\eta}_0^\top \boldsymbol{t}}$ and $c^{-1}f_-(t)e^{\boldsymbol{\eta}_0^\top \boldsymbol{t}}$ and both PDF's with respect to λ and (2.7) imply that their MGF's are the same in a neighborhood of $\boldsymbol{\eta}_0$. Using uniqueness of Laplace transforms,

$$c^{-1}f_+(t)e^{\boldsymbol{\eta}^\top \boldsymbol{t}} = c^{-1}f_-(t)e^{\boldsymbol{\eta}^\top \boldsymbol{t}} \Rightarrow f = f_+ - f_- = 0$$

almost everywhere λ. Thus T is complete. ■

Lemma 2.10

Suppose

$$\mathrm{d}\mathsf{P}_{\boldsymbol{X},\boldsymbol{\eta}}(\boldsymbol{x}) = \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x}) - \zeta(\boldsymbol{\eta})\right) h(\boldsymbol{x})\,\mathrm{d}\nu(\boldsymbol{x}).$$

Then $\mathrm{d}\mathsf{P}_{T,\boldsymbol{\eta}}(t) = \exp\left(\boldsymbol{\eta}^\top t - \zeta(\boldsymbol{\eta})\right)\,\mathrm{d}\tilde{\lambda}(t)$ where $\mathrm{d}\tilde{\lambda}(t) = \mathrm{d}\lambda T^{-1}(t)$ and $\mathrm{d}\lambda(\boldsymbol{x}) = h(\boldsymbol{x})\,\mathrm{d}\nu(\boldsymbol{x})$

Theorem 2.11 (Change of variable)

Let f be a measurable function from $(X, \mathcal{F}, \mu)$ to $\mathbb{R}$ and g be a Borel measurable function on $\mathbb{R}$. Then

$$\int g\,\mathrm{d}\mu f^{-1} = \int g \circ f\,\mathrm{d}\mu$$

in the sense that if either exists, so does the other and the two are equal. The measure μf^{-1}, defined by $\mu f^{-1}(B) = \mu(f^{-1}(B))$ for all Borel sets B; $B \in \mathbb{B}$, is the measure induced by f on $\mathbb{R}$.

Corollary 2.12

If B is a Borel set, then $\int_B g\,\mathrm{d}\mu f^{-1} = \int_{f(\boldsymbol{\theta})} g \circ f\,\mathrm{d}\mu$. If μf^{-1} is dominated by the Lebesgue measure, $\mu f^{-1} << m$, then

$$\int g \circ f\,\mathrm{d}\mu = \int g(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}(\boldsymbol{x})$$

where $p(\boldsymbol{x}) = \mathrm{d}\mu f^{-1}/\,\mathrm{d}m(\boldsymbol{x})$.

Proof We prove the lemma. Let P_T be the probability measure induced by T.

$$\mathsf{P}_T(A) = \mathsf{P}_X(T^{-1}(A)) = \int_{\boldsymbol{x} \in T^{-1}(A)} \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x}) - \zeta(\boldsymbol{\eta})\right) h(\boldsymbol{x})\,\mathrm{d}\nu(\boldsymbol{x})$$

Define $\mathrm{d}\lambda(\boldsymbol{x}) = h(\boldsymbol{x})\,\mathrm{d}\nu(\boldsymbol{x})$. Then

$$\mathsf{P}_T(A) = \int_{T^{-1}(A)} \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x}) - \zeta(\boldsymbol{\eta})\right)\,\mathrm{d}\lambda(\boldsymbol{x})$$

Define $\boldsymbol{y} = T(\boldsymbol{x})$, then using the theorem of change of variable, we have

$$\begin{aligned}\mathsf{P}_T(A) &= \int_{T^{-1}(A)} \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x}) - \zeta(\boldsymbol{\eta})\right)\,\mathrm{d}\lambda(\boldsymbol{x}) \\ &= \int_A \exp\left(\boldsymbol{\eta}^\top \boldsymbol{y} - \zeta(\boldsymbol{\eta})\right)\,\mathrm{d}\lambda T^{-1}(\boldsymbol{y})\end{aligned}$$

Let $\mathrm{d}\tilde{\lambda}(\boldsymbol{y}) = \mathrm{d}\lambda T^{-1}(\boldsymbol{y})$. Then

$$\begin{aligned}\mathsf{P}_T(A) &= \int_A \exp\left(\boldsymbol{\eta}^\top \boldsymbol{y} - \zeta(\boldsymbol{\eta})\right)\,\mathrm{d}\tilde{\lambda}(\boldsymbol{y}) \\ \frac{\mathrm{d}\mathsf{P}_T}{\mathrm{d}\tilde{\lambda}}(t) &= \exp\left(\boldsymbol{\eta}^\top \boldsymbol{t} - \zeta\eta\right)\end{aligned}$$

■

If δ is an estimate of $\boldsymbol{\theta}$, one way to look is to minimize MSE

$$\mathsf{mse}_\theta(\delta) = \mathsf{E}_\theta[(\delta(\boldsymbol{x}) - \theta)^2]$$

but one could have $\mathsf{E}_\theta(|\delta(\boldsymbol{x}) - \theta|)$, which is another loss function; if we look at the deviation from target, we are thus picking the ones which have minimum variability, subject to extreme weights. Hubert thus proposes to look at

$$\begin{cases}(\delta(\boldsymbol{x}) - \theta)^2 & \text{if } |\delta(\boldsymbol{x}) - \theta| \leq d \\ d^2 & \text{if } |\delta(\boldsymbol{x}) - \theta| > d\end{cases}$$

Follow Lehmann and Casella to get a more general account of this.

Definition 2.13 (Loss function)

Let $L : \Theta \times \Theta \to \mathbb{R}^+$. Thus if δ is an estimator of $\boldsymbol{\theta}$ based on $\boldsymbol{X} = X_1, \ldots, X_n$, then $\delta : \delta^n \to \Theta$, where δ is the sample space. $L(\delta(\boldsymbol{X}), \boldsymbol{\theta})$ is the **loss function** and define the risk of an estimator as

$$R(\delta, \boldsymbol{\theta}) = \mathsf{E}_\theta\left(L(\delta(\boldsymbol{X}), \boldsymbol{\theta})\right)$$

If $R(\delta_1, \boldsymbol{\theta}) \leq R(\delta_2, \boldsymbol{\theta}),\ \forall\ \boldsymbol{\theta} \in \Theta$ and the inequality is strict for some $\boldsymbol{\theta}_0 \in \Theta$, we say δ_2 is **inadmissible**.

For example, take $L(a, \theta) = (a - \theta)^2$ and $R(\delta, \theta) = \mathsf{mse}_\theta(\delta)$.

This entire field is called **decision theory**. If we have a unique Bayes,

$$R_T(\delta) = \int_\theta R(\delta, \theta)\pi(\,\mathrm{d}\theta)$$

then it is admissible. However, some estimators are not unique Bayes and still admissible. The fundamental theorem is complete class theorem.

Proposition 2.14 (Stein's paradox)

If we consider $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$, then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ (which we will show is UMVUE using Rao-Blackwell and Lehmann-Scheffé) is admissible for dimension 1 and 2, but not for higher dimensions. One can show that the so-called James-Stein estimator perform better than $\bar{X}_n$ in terms of MSE, since we are not restricting ourselves to unbiased estimates. This shows that naive statistic is not possible.

Definition 2.15 (Convex loss functions)

A real valued-function ϕ defined on an open interval $I = (a, b)$ with $-\infty \leq a < b \leq \infty$ is convex if for any $a < x < y < b$ and any $0 < \gamma < 1$,

$$\varphi[\lambda x + (1-\gamma)y] \leq \gamma\varphi(x) + (1-\gamma)\varphi(y) \tag{2.8}$$

An example of convex function is $\varphi(x) = x^2$.

Theorem 2.16 (Convexity)

1. If φ is defined and differentiable on (a, b), then a necessary and sufficient condition for φ to be convex is that $\varphi'(x) \leq \varphi'(y)$ for all $a < x < y < b$. The function is **stricly convex** if and only if (2.8) is strict for all $x < y$.

2. If in addition φ is twice differentiable, then the necessary and sufficient condition, (2.8) is equivalent to $\varphi''(x) \geq 0$ for all $a < x < b$.

Theorem 2.17

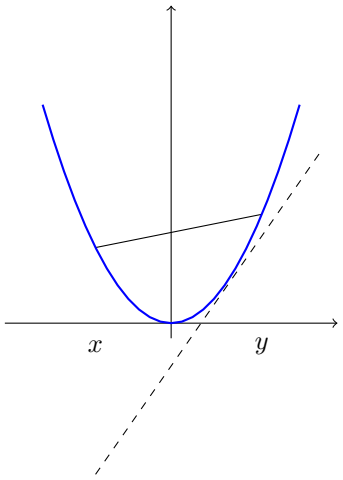
Let φ be a convex function defined on $I = (a, b)$ and let t be any fixed point in I. Then, there exists a straight line $y = L(x) = c(x - t)s + \varphi(t)$ through the point $[t, \varphi(t)]$ such that $L(x) \leq \varphi(x)$, $\forall\, x \in I$.

Theorem 2.18 (Jensen's inequality)

If φ is a convex function defined over an open interval I, and X is a random variable with $\mathsf{P}(X \in I) = 1$ and finite expectation, then

$$\varphi(\mathsf{E}(X)) \leq \mathsf{E}(\varphi(X)) \tag{2.9}$$

If φ is strictly convex, the inequality is strict unless X is a constant with probability one.



Proof (from MATH 556) Suppose g is convex, then

$$g'_+(x) = \lim_{\substack{h \downarrow 0 \\ h>0}} \frac{g(x+h) - g(x)}{h}$$

exists for all $x \in \mathbb{R}$. If we consider the graph of a continuous convex function and a tangent, then the graph will lie above the tangent. Tangent in x is given by the expression: $g(x) + g'_+(x)(z-x)$ for $z \in \mathbb{R}$. Thus,

$$g(x) + g'_+(x)(z-x) \leq g(z), \qquad \forall\, z \in \mathbb{R},\ \forall\, x \in \mathbb{R}$$

For all $\omega \in \Omega$, $g(x) + g'_+(x)(X(\omega) - x) \leq g(X(\omega))$, therefore cleverly chose $x = \mathsf{E}(X)$, then

$$g(\mathsf{E}(X)) + g'_+(\mathsf{E}(X))(X(\omega) - \mathsf{E}(X)) \leq g(X(\omega))$$

and taking expectation on both sides, on the LHS we get

$$g(\mathsf{E}(X)) + g'_+(\mathsf{E}(X))\mathsf{E}(X - \mathsf{E}(X)) \leq \mathsf{E}(g(X))$$

and since $\mathsf{E}(X - \mathsf{E}(X)) = 0$, we recover the Jensen's inequality. To remember it, notice that $\mathsf{E}\left(X^2\right) \geq (\mathsf{E}(X))^2$, since $\mathsf{Var}(X) \geq 0$ and it can be composed into $\mathsf{E}\left(X^2\right) - (\mathsf{E}(X))^2 \geq 0$. ■

Corollary 2.19

We have

$$\frac{1}{\mathsf{E}(X)} < \mathsf{E}\left(\frac{1}{X}\right) \text{ and } \mathsf{E}(\log(X)) < \log(\mathsf{E}(X)) \tag{2.10}$$

Definition 2.20 (Entropy distance)

This notion is also know as Kullback-Leibler information of g at f (or KL distance between g and f)

$$\mathsf{E}_f\left(\log\left(\frac{f(X)}{g(X)}\right)\right) = \int \log\left(\frac{f(x)}{g(x)}\right) f(x)\,\mathrm{d}x$$

and using (2.10),

$$\begin{aligned}
\mathsf{E}_f\left(\log\left(\frac{f(X)}{g(X)}\right)\right) &= -\mathsf{E}_f\left(\log\left(\frac{g(X)}{f(X)}\right)\right) \\
&\geq -\log \mathsf{E}_f\left(\frac{g(X)}{f(X)}\right) \\
&= -\log \int_{\mathcal{X}} \frac{g(x)}{f(x)} f(x)\,\mathrm{d}x \\
&= -\log(1) = 0
\end{aligned}$$

which shows that the Kullback-Leibler divergence between two probability measures is positive.

Theorem 2.21 (Rao-Blackwell)

Let X be a random variable with distribution $\mathsf{P}_\theta \in \mathrm{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ and let T be sufficient for P. Let δ be an estimator of an estimand $g(\theta)$, and let the loss function $L(a, \theta)$ be a strictly convex function of a. Then, if δ has finite expectation and risk

$$R_g(\delta, \theta) = \mathsf{E}(L_g(\delta(X), \theta)) = \mathsf{E}(L(\delta(X), g(\theta))) < \infty$$

and if $\eta(t) = \mathsf{E}(\delta(X)|t)$, the risk of the estimator $\eta(T)$ satisfies $R_g(\eta, \theta) < R_g(\delta, \theta)$ unless $\delta(X) = \eta(T)$ with probability one.

Proof In Jensen's inequality, let $\varphi(A)$ be $L(a, \theta)$, let $\delta = \delta(X)$ and let X given $T = t$ have the conditional distribution $p(X|T)$. Then

$$L(\eta(t), \theta) < \mathsf{E}(L(\delta(x), \theta)|t)$$

unless $\delta(x) = \eta(T)$ with probability one. Taking the expectation from both sides of the above inequality, we find the desired result. ■

Note

- Sufficiency of T is used in the proof only to ensure $\eta(T)$ does not depend on θ and hence is an estimator.
- If the loss function is convex rather than strictly convex, the above holds if we replace strict inequality by

$$L(\eta(t), \theta) \leq \mathsf{E}\left(L(\delta(x), \theta)|t\right)$$

Why is this result useful? By conditioning, we reduce the risk and the variance. Conditioning on a sufficient statistic, we guarantee that the resulting is a statistic and an estimator, and not a variable.

Theorem 2.22 (Lehmann-Scheffé)

If T is a complete sufficient statistic and there exists an unbiased estimate h of θ, there exists a unique UMVUE of θ, which is given by $\mathsf{E}\left(h|T\right)$.

Definition 2.23 (Uniformly minimum variance unbiased estimate)

An unbiased estimator $\delta(\boldsymbol{x})$ of $g(\boldsymbol{\theta})$ is the uniform minimum variance unbiased (UMVU) estimator of $g(\boldsymbol{\theta})$ if

$$\mathsf{Var}_\theta\left(\delta(\boldsymbol{x})\right) \leq \mathsf{Var}_\theta\left(\delta'(\boldsymbol{x})\right), \quad \forall\, \boldsymbol{\theta} \in \Theta$$

where $\delta' \in \mathcal{U}_g$ and

$$\mathcal{U}_g = \{\delta' : \mathcal{S}^n \to \Theta : \mathsf{E}_\theta(\delta'(\boldsymbol{x})) = g(\boldsymbol{\theta})\}$$

where $\mathcal{S}$ is the sample space.

The requirement that there exists unbiased estimate is necessary ($\mathcal{U}$ estimable). One can cook up examples where this does not hold. For example

Example 2.14

Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a sample where $X \sim \mathcal{B}(n, p)$ for $p \in (0, 1)$. Looking at $g(p) = 1/p$, then

$$\mathsf{E}_p(\delta(\boldsymbol{X})) = \frac{1}{p}, \ \forall\, p \in (0, 1)$$

Then

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

Now as $p \to 0$, the left hand side goes to $\delta(0) \in \mathbb{R}$, while the right hand side goes to ∞.

We now proceed with the proof of Lehmann-Scheffé.

Proof If $h_1, h_2 \in \mathcal{U}$, where $\mathcal{U}$ is the class of all unbiased estimators of θ, then $\mathsf{E}(h_1|T)$ and $\mathsf{E}(h_2|T)$ are both unbiased estimators of $\boldsymbol{\theta}$. Thus

$$0 = \mathsf{E}_\theta\left(\mathsf{E}(h_1|T) - \mathsf{E}(h_2|T)\right), \qquad \forall\, \theta \in \Theta$$

Since T is a complete sufficient, it follows that

$$\mathsf{E}(h_1|T) = \mathsf{E}(h_2|T)$$

by Rao-Blackwell theorem. Thus $\mathsf{E}(h|T)$ is the UMVUE. ■

Example 2.15

If $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is complete sufficient. Thus consider the function

$$g_1\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n$$
$$g_2\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = S_n^2$$

then we see that $\mathsf{E}_\theta(\bar{X}_n) = \mu, \qquad \forall\, \theta \in \Theta = \mathbb{R} \times \mathbb{R}^+$ and $\mathsf{E}_\theta(S_n^2) = \sigma^2$ are UMVUE.

Example 2.16

Again consider the Normal case $\psi(\mu, \sigma^2) = \sigma$. Unfortunately, we cannot look at $\sqrt{S_n^2}$ since UMVUE is not invariant; we need to start afresh every time. If $\psi(\boldsymbol{\theta})$ is a **linear function** of $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$ is the UMVUE of $\boldsymbol{\theta}$, then $\psi(\widehat{\boldsymbol{\theta}})$ is UMVUE of $\psi(\boldsymbol{\theta})$. It boils down to $\mathsf{E}(\psi(\boldsymbol{x})) = \psi(\mathsf{E}(X))$. If we want to find the UMVUE for σ, starting with S_n, we try to find a linear function fo σ, then we may invert it; in other cases, its not good.

We have $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $V \equiv (n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$. Looking at the inverse map for S

$$y = \frac{(n-1)x^2}{\sigma^2} \Rightarrow x = \sqrt{\frac{\sigma^2 y}{n-1}}$$

and so as a result

$$\begin{aligned}\mathsf{E}\left(S_n\right) &= \int_0^\infty \sqrt{\frac{\sigma^2 v}{n-1}} \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} v^{\frac{n-1}{2}-1} e^{-\frac{v}{2}} dv \\ &= \frac{\sigma}{\sqrt{n-1}} \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} \int_0^\infty v^{\frac{n-1}{2}-\frac{1}{2}} e^{-\frac{v}{2}} dv \\ &= C(n) \int_0^\infty v^{\frac{n}{2}-1} e^{-\frac{v}{2}} \frac{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} dv\end{aligned}$$

which we recognize to be the kernel of a chi-square distribution which has density given by

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \mathbf{1}_{x\geq 0}$$

and so $\mathsf{E}\left(S_n\right) = d(n)\sigma$ where

$$\begin{aligned}d(n) &= \frac{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)}2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)} \\ &= \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\end{aligned}$$

and $S_n/d(n)$ is the UMVUE of σ.

Example 2.17

Suppose we look again as the Normal case $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and we consider $\mathfrak{z}_p$, the quantile

$$p = \mathsf{P}\left(X \leq \mathfrak{z}_p\right) = \mathsf{P}\left(Z \leq \frac{\mathfrak{z}_p - \mu}{\sigma}\right)$$

for $Z \sim \mathcal{N}(0,1)$. Thus $\mathfrak{z}_p = \sigma Z_{(1-p)} + \mu$. The UMVUE of $\mathfrak{z}_p$ is

$$\frac{S_n}{d(n)} Z_{(1-p)} + \bar{X}_n$$

and

$$\mathsf{E}\left(\frac{S_n}{d(n)}\mathfrak{z}_p + \bar{X}_n\right) = \sigma Z_{(1-p)} + \mu = \mathfrak{z}_p$$

Theorem 2.24

Let $\mathcal{U}$ be the class of all unbiased estimators T of a parameter $\boldsymbol{\theta} \in \Theta$ with $\mathsf{E}_\theta(T^2),\ \forall\, \theta \in \Theta$ and suppose that $\mathcal{U}$ is non-empty. Let $\mathcal{U}_0$ be the set of all unbiased estimators ν of 0, that is

$$\mathcal{U}_0 = \{\nu : \mathsf{E}_\theta(\nu) = 0, \mathsf{E}_\theta(\nu^2) < \infty,\ \forall\, \theta \in \Theta\}$$

Then $T_0 \in \mathcal{U}$ is a UMVUE if and only if $\mathsf{E}_\theta(\nu T_0) = 0,\ \forall\, \theta \in \Theta,\ \forall\, \nu \in \mathcal{U}_0$.

Proof The conditions of the theorem guarantee the existence of $\mathsf{E}_\theta(\nu T_0),\ \forall\, \theta \in \Theta, \nu \in \mathcal{U}_0$ (by Cauchy-Schwartz inequality). By contradiction: suppose that $T_0 \in U$ is a UMVUE and $\mathsf{E}_\theta(\nu_0 T_0) \neq 0$ for some $\theta_0 \in \Theta$ and some $\nu_0 \in \mathcal{U}_0$. Then $T_0 + \lambda\nu_0 \in \mathcal{U}\ \forall\ \lambda \in \mathbb{R}$. If $\mathsf{E}_{\theta_0}(\nu_0^2) = 0$, then $\mathsf{E}_{\theta_0}(\nu_0 T_0)$ must hold since $\mathsf{P}\,(\theta_0)\,(\nu_0 = 0) = 1$.[12]

Let $\mathsf{E}_{\theta_0}(\nu_0^2) > 0$; choose $\lambda_0 = -\mathsf{E}_{\theta_0}(\nu_0 T_0)/\mathsf{E}_{\theta_0}\left(\nu_0^2\right)$. Then

$$\mathsf{E}_{\theta_0}(T_0 + \lambda_0\nu_0)^2 = \mathsf{E}_{\theta_0} T_0^2 - \frac{\left(\mathsf{E}_{\theta_0}(\nu_0 T_0)\right)^2}{\mathsf{E}_{\theta_0}(\nu_0^2)} < \mathsf{E}_{\theta_0}(T_0^2) \tag{2.11}$$

Since $T_0 + \lambda_0\nu_0 \in \mathcal{U}$ and $T_0 \in \mathcal{U}$; it follows from (2.11) that $\mathsf{Var}_{\theta_0}(T_0 + \lambda_0\nu_0) < \mathsf{Var}_{\theta_0}(T_0)$ which is a contradiction.

Conversely, let T_0 be orthogonal to $\mathcal{U}_0$ and let $T \in \mathcal{U}$. Then $T_0 - T \in \mathcal{U}_0$ and for every $\theta \in \Theta$, we have

$$\mathsf{E}_\theta(T_0(T_0 - T)) = 0.$$

We have

$$\mathsf{E}_\theta(T_0^2) = \mathsf{E}_\theta(TT_0) \leq \left(\mathsf{E}_\theta(T_0^2)\right)^{\frac{1}{2}} \left(\mathsf{E}_\theta(T^2)\right)^{\frac{1}{2}}$$

by Cauchy-Schwartz inequality. If $\mathsf{E}_{\theta_0}\left(T_0^2\right) = 0$, then $\mathsf{P}\,(T_0 = 0) = 1$ and there is nothing to prove. Otherwise,

$$\left(\mathsf{E}_\theta(T_0^2)\right)^{\frac{1}{2}} \leq \left(\mathsf{E}_\theta(T^2)\right)^{\frac{1}{2}}$$

or equivalently

$$\mathsf{Var}_\theta(T_0) \leq \mathsf{Var}_\theta(T).$$

■

Before proceeding with a geometric proof of the previous theorem, we need some preliminary results.

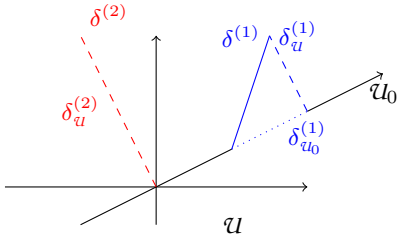
Theorem 2.25

Let M be a non-empty closed convex subset of the Hilbert space $\mathcal{H}$. If $\boldsymbol{x} \in \mathcal{H}$, there is a unique element $\boldsymbol{y}_0 \in M$ such that

$$\|\boldsymbol{x} - \boldsymbol{y}_0\| = \inf\{\|\boldsymbol{x} - \boldsymbol{y}\| : \boldsymbol{y} \in M\}$$

[12] Note that E is a projection operator.

Figure 4: Geometric interpretation of UMVUE, $\delta^{(2)} \perp \mathcal{U}_0$



Theorem 2.26

Let M be a closed subspace of the Hilbert space $\mathcal{H}$ and $\boldsymbol{y}_0$ an element of M. Then

$$\|\boldsymbol{x} - \boldsymbol{y}_0\| = \inf\{\|\boldsymbol{x} - \boldsymbol{y}\| : \boldsymbol{y} \in M\} \quad \Leftrightarrow \quad \boldsymbol{x} - \boldsymbol{y}_0 \perp M$$

that is $\langle \boldsymbol{x} - \boldsymbol{y}_0 \rangle = 0$ for all $\boldsymbol{y} \in M$.

Remark

The element $\boldsymbol{y}_0$ is called **projection** of $\boldsymbol{x}$ on M.

Theorem 2.27 (Projection theorem)

Let M be a closed subspace of the Hilbert space $\mathcal{H}$. If $\boldsymbol{x} \in \mathcal{H}$, then $\boldsymbol{x}$ has a unique representation $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$ where $\boldsymbol{y} \in M$ and $\boldsymbol{z} \perp M$. Furthermore, $\boldsymbol{y}$ is the projection of $\boldsymbol{x}$ on M

We restrict ourselves to a subspace since for δ_i unbiased estimator, then any linear combination $\sum_{i=1}^{K} a_i \delta_i$ is unbiased if $\sum a_i = 1$. Recall that a Hilbert space is heuristically the generalization of the Euclidian space, and we have the norm defined based on the inner product. Since we have angles, this gives a geometric interpretation. Hilbert space also enjoys nice properties, such as completeness.

Let

$$\mathcal{H}_{\theta_0} = \left\{\delta : \mathsf{E}_{\theta_0}\left((\delta - \theta_0)^2\right) < \infty\right\}$$

and

$$M_{\theta_0} = \left\{\nu : \mathsf{E}_{\theta_0}\left(\nu^2\right) < \infty\right\}$$

Then, $\forall\ \delta \in \mathcal{H}_{\theta_0}, \exists \nu_\delta \in M_{\theta_0}$ such that $\delta = \nu_\delta + (\delta - \nu_\delta)$ (using the projection theorem) where $\delta - \nu_\delta \perp M_{\theta_0}$. Note

$$\mathsf{mse}_{\theta_0}(\delta) = \mathsf{E}_{\theta_0}\left((\delta - \theta_0)^2\right) = \|\delta\|^2$$

and

$$\begin{aligned}\|\delta\|^2 &= \|\nu_\delta\|^2 + \|\delta - \nu_\delta\|^2 \\ &\geq \|\delta - \nu_\delta\|^2\end{aligned}$$

which implies $\|\delta\| \geq \|\delta - \nu_\delta\|$. Note that $\mathsf{E}_{\theta_0}(\delta) = \mathsf{E}_{\delta_0}(\delta - \nu_\delta)$. Then

$$\|\delta_{\min}\| = \inf_{\delta \in \mathcal{U}} \|\delta\| \quad \Leftrightarrow \quad \delta_{\min} \perp M_{\theta_0}$$

where $\mathcal{U} = \{\delta \in \mathcal{H}_{\theta_0} : \mathsf{E}_\theta(\delta) = \theta_0\}$.

Example 2.18

Let X be a random variable with PMF

$$\mathsf{P}_\theta(X = x) = \begin{cases} \theta & \text{if } x = -1 \\ (1-\theta)^2\theta^x & \text{if } x = 0, 1, 2, \ldots \end{cases}$$

The family $\mathrm{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ is not complete.

Proof Let $\mathsf{E}_\theta(g(X)) = 0, \ \forall\, \theta \in (0,1)$. Thus

$$\begin{aligned}0 &= \sum_{x=-1}^{\infty} g(x)\mathsf{P}_\theta(X = x) \\ &= \theta g(-1) + \sum_{x=0}^{\infty} g(x)(1-\theta)^2\theta^x\end{aligned}$$

Thus

$$\sum_{x=0}^{\infty} g(x)\theta^x = \frac{\theta}{(1-\theta)^2}(-g(-1))$$

and define $c = -g(-1)$ and use

$$\frac{1}{(1-\theta)^2} = \sum_{x=0}^{\infty}(x+1)\theta^x. \tag{2.12}$$

Then

$$\begin{aligned}\sum_{x=0}^{\infty} g(x)\theta^x &= \sum_{x=0}^{\infty} c(x+1)\theta^{x+1} \\ &= \sum_{x=1}^{\infty} cx\theta^x\end{aligned}$$

Therefore

$$g(0) + \sum_{x=1}^{\infty} (g(x) - cx)\,\theta^x = 0$$

for all $\theta \in (0,1)$. This then implies that $g(x) = cx, x = 0, 1, \ldots$. In other words,

$$\mathcal{C}_0 = \{T(x) : T(-1) = -c \in \mathbb{R} \text{ and } T(x) = cx, x = 0, 1, \ldots\}$$

is the class of all unbiased estimators of zero. ■

Now suppose δ is an estimator. For δ to be the UMVUE of its expectation, $\mathsf{E}_\theta(\delta(X))$, a necessary and sufficient condition is $\delta \perp \mathcal{C}_0$, *i.e.*

$$\begin{aligned}
0 &= \mathsf{E}_\theta(\delta(X)T(X)) \\
&= \sum_{x=-1}^{\infty} \delta(x)T(x)\mathsf{P}_\theta(X = x) \\
&= \delta(-1)T(-1)\theta + \sum_{x=0}^{\infty} \delta(x)cx(1-\theta)^2\theta^x \\
&= -c\delta(-1)\theta + c(1-\theta)^2 \sum_{x=1}^{\infty} \delta(x)x\theta^x
\end{aligned}$$

and $c = 0$ corresponds to $T \equiv 0$. If $c \neq 0$, then

$$\delta(-1)\theta = (1-\theta)^2 \sum_{x=1}^{\infty} \delta(x)x\theta^x,$$

for all $\theta \in (0,1)$

$$\frac{\delta(-1)\theta}{(1-\theta)^2} = \sum_{x=1}^{\infty} \delta(x)x\theta^x$$

Applying (2.12),

$$\delta(-1) \sum_{x=1}^{\infty} x\theta^x = \sum_{x=1}^{\infty} \delta(x)x\theta^x$$

This then implies $\delta(x) = \delta(-1), x = 1, 2, \ldots$. This $\delta(0)$ and $\delta(-1)$ are arbitrary and $\delta(x) = \delta(-1)\ \forall\ x = 1, 2, \ldots$. The class of all estimators orthogonal to $\mathcal{C}_0$ is therefore

$$\mathcal{O} = \{\delta : \delta(0) = a, \delta(x) = b, x = -1, 1, 2, \ldots \text{ and } a, b \in \mathbb{R}\}$$

If $\delta \in \mathcal{O}$, then

$$\begin{aligned}\mathsf{E}_\theta\left(\delta(X)\right) &= \delta(-1)\theta + \delta(0)(1-\theta)^2 + \delta(-1)(1-\theta)^2 \sum_{x=1}^{\infty} \theta^x \\ &= a(1-\theta)^2 + b\theta + b(1-\theta)\theta \\ &= (a-b)\theta^2 - 2(a-b)\theta + a \\ &= g(\theta)\end{aligned}$$

So the only UMVUE are quadratic forms of θ. This fact illustrate that UMVUE may not exist for many estimand of interest in given problems.

Example 2.19

Let $X_i \overset{\text{iid}}{\sim} \mathcal{B}(p)$. Then $T = \sum_{i=1}^n X_i \sim \mathcal{B}(n,p)$ and $g(p) = p^r$. By Rao-Blackwell, if we have a function of T sufficient such that for $r \in \mathbb{N}$

$$\mathsf{E}_p\left(\delta(T)\right) = p^r$$

$$\sum_{t=0}^{n} \binom{n}{t} \delta(t) p^t (1-p)^{n-t} = p^r,$$

for all $p \in (0,1)$

$$\sum_{t=0}^{n} \delta(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \frac{p^r}{(1-p)^n}$$

and let now $\theta = p/(1-p)$, so that $p = \theta/(1+\theta)$ and $1/(1-p) = 1+\theta$. Then,

$$\begin{aligned}\sum_{t=0}^{n} \delta(t) \binom{n}{t} \theta^t &= \theta^r (1+\theta)^{n-r} \\ &= \theta^r \sum_{s=0}^{n-r} \binom{n-r}{s} \theta^s \\ &= \sum_{s=0}^{n-r} \binom{n-r}{s} \theta^{r+s} \\ &= \sum_{t=r}^{n} \binom{n-r}{t-r} \theta^t\end{aligned}$$

for all $\theta \in (0,1)$ using the binomial expansion.[13] Then

$$\delta(t)\binom{n}{t} = \binom{n-r}{t-r}$$

for $t = r, \ldots n$. We thus have

$$\delta(t) = \begin{cases} 0 & \text{if } t = 0, 1, \ldots, r-1 \\ \frac{\binom{n-r}{t-r}}{\binom{n}{t}} & \text{if } t = r, \ldots, n \end{cases}$$

Note that we cannot estimate p^r if $n < r$.

Before moving to the next topic, that of M-estimators and maximum likelihood, we look at a result for Generalized Estimating Equations (GEE). Coming back to loss function and convex functions, we let a be an action for $\delta(\boldsymbol{x})$ with $L(\theta, a) = \rho(\theta - a)$ and ρ a convex function; we look at the estimating equation, a function of the data and the unknown parameter.

Theorem 2.28

Let ρ be a convex function defined on $(-\infty, \infty)$ and X a random variable such that $\varphi(a) = \mathsf{E}(\rho(x - a))$ is finite for some a. If ρ is not monotone, $\varphi(a)$ takes on its minimum value and the set on which this values is taken in a closed interval. If ρ is stricly convex, the minimizing value is unique.

Lemma 2.29

Let φ be a convex function on $(-\infty, \infty)$ which is bounded below, $\varphi(x) \geq k$ for some $k \in \mathbb{R}$, and suppose that φ is not monotone. Then φ takes on its minimum value, the set $\mathcal{S}$ on which this value is taken is a closed interval and is a single point if φ is strictly convex.

Proof Since φ is convex and not monotone, it tends to ∞ as $x \to \pm\infty$ (why?) [14]. Since φ is also continuous, it takes on its minimizing value.

Note

This result does not follow from max/min value theorem (any continuous functions on a compact set achieves its minimum and maximum) since we need compactness.

This result follows from continuity and being bounded below and not monotone as follows. Since φ is bounded below,

$$\varphi(x) \geq K, \ \forall\, x \in D(\varphi)$$

[13] Recall that the binomial expansion is given by $(x+y)^n = \Sigma_{k=0}^n \binom{n}{k} x^{n-k} y^k = \Sigma_{k=0}^n \binom{n}{k} x^k y^{n-k}$. Taking $y = 1$, this translates into $(1+x)^n = \sum_{k=0}^n \binom{n}{k} x^k$

[14] Otherwise, this violates condition from Theorem (2.17)

Thus we have $\inf \varphi(x) = \alpha$. Then $\forall\ n \in \mathbb{N}, \exists x_n \in D(\varphi)$ such that $\alpha < \varphi(x_n) < \alpha + \frac{1}{n}$. Consider $\{x_n\}_{n=1}^{\infty}$. This sequence is bounded, otherwise $x_n \to \infty$ and since $\varphi(x) \to \infty$, we have a contradiction. Since $\{x_n\}$ is bounded, by Bolzano-Weirstrass theorem, we have a convergent subsequence $\{x_{n'}\}$. Let $x^* = \lim_{n\to\infty} x_{n'}$ and

$$\alpha < \varphi(x_{n'}) < \alpha + \frac{1}{n}$$

and $\lim_{n\to\infty} \varphi(x_{n'}) = \alpha$. On the other hand, using continuity,

$$\varphi(x^*) = \lim_{n\to\infty} \varphi(x_{n'})$$

■

2.6 MLE, M-estimators and Generalized Estimating Equations

Suppose that $\boldsymbol{X} \overset{\text{iid}}{\sim} f_{\boldsymbol{\theta}}(\boldsymbol{x})$ We can link the observed quantities, $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) = \mathsf{P}\left(X_i = x_i; \boldsymbol{\theta}\right) = \mathsf{P}_{\boldsymbol{\theta}}\left(X_i = x_i\right)$ and we treat what we have observed as being typical of what one would expect to observe. We thus wish to maximize the likeliness of observations. The **maximum likelihood estimate**

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x})$$

is recognized as a post-experimental approach. Unbiasedness or UMVUE are properties of procedures, rather than of the point estimates. The post-experimental approaches are good for large size of the populations, and need to guard again the cases not observed. Pre-experimental approach are better for small sample size, as opposed to post-experimental approach. We do a simple example.

Example 2.20
If $X_i \overset{\text{iid}}{\sim} \mathcal{B}(p)$, then

$$\begin{aligned}\mathcal{L}(p, \boldsymbol{x}) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{x_i} \\ &= p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}\end{aligned}$$

where we denote $\sum x_i = t$ since applying monotone transformations does not affect the maximization, therefore

$$\begin{aligned}
\log(\mathcal{L}(p;\boldsymbol{x})) \equiv \ell_n(p;\boldsymbol{x}) = t\log(p) - (n-t)\log(1-p) = 0 \\
\Rightarrow \qquad \frac{t}{p} - \frac{n-t}{1-p} = 0 \\
\Rightarrow \qquad t(1-p) - (n-t)p = 0 \\
\Rightarrow \qquad \widehat{p} = \frac{t}{n}
\end{aligned}$$

since we have a smooth function, if the function has an extrema, then we can solve analytically. Note that for the MLE, we could also have different distribution, for consistency, we need smoothness and identifiability. For asymptotic normality, we need existence of the first derivative.

Let $\rho(x,t)$ be a function $\mathbb{R}^d \times \mathbb{R} \to \Theta \subseteq \mathbb{R}$. An ***M*-functional** is defined to be a solution of [15]

$$\int_{\mathcal{X}} \rho(x, T(G))\,\mathrm{d}G(x) = \min_{t\in\Theta} \int_{\mathcal{X}} \rho(x,y)\,\mathrm{d}G(x)$$

for $G \in \mathcal{F}_0$, $\mathcal{F}_0$ contains all CDFs on $\mathbb{R}^d$ for which the above integrals are well-defined. If $X_1, \ldots, X_n \sim F \in \mathcal{F}_0$, then $T(F_n)$ is called an M-estimator of $T(F)$. For $F_n = \frac{1}{n}\sum_{i=1}^n \mathcal{E}(x - X_i)$, the empirical CDF. Then

$$\int_{\mathcal{X}} \rho(x,y)\,\mathrm{d}F_n(x) = \frac{1}{n}\sum_{i=1}^n \rho(x_i, t)$$

and $\rho(x,t) = -\log f_t(x)$ which corresponds to maximum likelihood.

We can get the counterpart of the score equations; if ρ is differentiable a.e.,

$$\psi(x,y) = \frac{\partial \rho(x,y)}{\partial t}$$

and

$$\begin{aligned}
\lambda_G(t) = \int_{\mathcal{X}} \psi(x,y)\,\mathrm{d}G(x) \\
\frac{\partial}{\partial t}\int_{\mathcal{X}} \rho(x,y)\,\mathrm{d}G(x).
\end{aligned}$$

Then $\lambda(T(G)) = 0$.

[15] M stands for minimization, functional is for function input. Here $\mathrm{d}G(x)$ is the Lebesgue Stieltjes integral.

Example 2.21

(i) Consider $\rho(x,y) = \frac{1}{2}(x-t)^2$ with $\psi(x,y) = t - x$. Then

$$T(F) = \int_{\mathcal{X}} x \, \mathrm{d}G(x)$$
$$T(F_n) = \bar{X}_n$$

we have adding and subtracting x

$$\begin{aligned}\Phi(a) &= \mathsf{E}\left((X-a)^2\right) \\ &= \mathsf{Var}\,(X) + (\mu - a)^2 \\ &\geq \mathsf{Var}\,(X)\end{aligned}$$

where $\mu = \mathsf{E}\,(X)$ and the minimum is obtained at $a = \mu$.

(ii) For the median, $\rho(x,y) = |x-t|$, $T(G)$ is the media, $T(F_n)$ is the sample median. In this example, $\Phi(a) = \mathsf{E}\,(|X-a|)$,

$$\Phi(a) = \begin{cases} \phi(m) + (a-m)[\mathsf{P}\,(X \geq m) - \mathsf{P}\,(X > m)] + 2\int_m^a (a-x)\,\mathrm{d}p & \text{if } a \geq m \\ \phi(m) + (m-a)[\mathsf{P}\,(X \geq m) - \mathsf{P}\,(X > m)] + 2\int_a^m (x-a)\,\mathrm{d}p & \text{if } a < m \end{cases}$$

and $\Phi(a) \geq \phi(m)$ where m is a median. This absolute value is less sensitive to tails and extreme value, known as L_1 regression. Another loss function could be

$$\rho_1(x,y) = \begin{cases} (x-t)^2 & \text{if } |x-t| \leq K \\ 2K|x-t| - K^2 & \text{if } |x-t| > K \end{cases}$$
$$\rho_2(x,t) = |x-t|^p, \qquad \text{if } 1 < p < 2$$

or ρ_2

We can show that $\sqrt{n}[T(F_n) - T(F)] \xrightarrow{d} \mathcal{N}(0, \sigma_F^2$ where $\sigma_F^2 = \mathsf{E}\left(\Phi_F(x)\right)^2$ is the variance of the influence function, where [16]

$$\Phi_F(x) = \frac{-\psi(x, T(F))}{\lambda_F'(T(F))}$$

[16] We did not define since it requires Gâteau derivative, since T is a functional

We can view the least-squares in a more general form, known as **generalized estimating equations**

$$\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\gamma}^\top \boldsymbol{Z}_i)\boldsymbol{Z}_i = 0$$

which leads to

$$\sum_{i=1}^{n}\psi(\boldsymbol{X}_i, \boldsymbol{\gamma}) = 0$$

This gives approximation to the margin of error. A further generalization the covers many more interesting cases is as follows:

Definition 2.30 (Generalized estimating equations)
Let $\Theta \subseteq \mathbb{R}^k, \psi_i$ be a function from $\mathbb{R}^{d_i} \times \Theta \to \mathbb{R}^k$ for $i = 1, \ldots, n$ and

$$S_n(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\psi_i(\boldsymbol{X}_i, \boldsymbol{\gamma}) \tag{2.13}$$

for $\boldsymbol{\gamma} \in \Theta$. If $\boldsymbol{\theta}$ is estimated by $\widehat{\boldsymbol{\theta}} \in \Theta$ satisfying $S_n(\widehat{\boldsymbol{\theta}}) = 0$, then $\widehat{\boldsymbol{\theta}}$ is called a **generalized estimating equations** (GEE) estimate. The equation $S_n(\boldsymbol{\gamma}) = 0$ is called a GEE.

Note
When choosing a GEE, we need $\mathsf{E}\left(S_n(\boldsymbol{\theta})\right) = 0$ for convergence.

Analogous to the earlier result on UMVUE, we have

Theorem 2.31 (Blyth (1974))
A necessary and sufficient condition for $\mathsf{Cov}\left(\delta, \psi\right)$ to depend on δ only through $g(\theta) = \mathsf{E}_\theta\left(\delta\right)$ is that for all θ, $\psi \perp \mathcal{U}_0$ where $\mathcal{U}_0$ is the space of unbiased estimators of 0.

Indeed, using Cauchy-Schwartz

$$\mathsf{Var}\left(\delta\right) \geq \frac{\left[\mathsf{Cov}\left(\delta, \psi\right)\right]^2}{\mathsf{Var}\left(\psi\right)}$$

and we get a lower bound. This result does not require likelihood, as opposed to Fréchet-Rao-Carmer inequality.

Theorem 2.32
Let ρ be a non-negative convex function defined on $\mathbb{R}$ and X a random variable such that $\Phi(a) = \mathsf{E}\left(\rho(X - a)\right)$ is finite. If ρ is not monotone, $\Phi(a)$ takes its minimum value and the set of minimum values is a closed set. If ρ is strictly convex, the minimizer is unique.

We need first to prove the following lemma

Lemma 2.33

Let Φ be a convex function on $\mathbb{R}$ which is bounded below. Suppose Φ is not monotone. Then Φ takes its minimum value. The set S of minimum values is closed and is a singleton if Φ is strictly convex.

Consistency of GEE

We need first recall equicontinuity of a class of functions.

Definition 2.34 (Equicontinuity)

A sequence of functions $\{g_i\}$ from $\mathbb{R}^k \to \mathbb{R}^k$ is called **equicontinuous** on an open set $O \subseteq \mathbb{R}^k$ if and only if, for any $\varepsilon > 0$, there exists a $\delta_\varepsilon > 0$ such that $\sup_i \|g_i(t) - g_i(s)\| < \varepsilon$ for all $t, s \in O$ such that $\|t - s\| < \delta_\varepsilon$.

Since a continuous function on a compact set is uniformly continuous, functions such as $g_i(\gamma) = g(t_i, \gamma)$ form an equicontinuous sequence on O if t_i's vary in a compact set containing O and $g(t_i, \gamma)$ is a continuous function in (t, γ).

We recall a version of the WLLN: if there is a sequence $\{X_i\}$ of independent random variables with $p \in (1, 2]$ such that

$$\lim_{n\to\infty} \frac{1}{n^p} \sum_{i=1}^{n} \mathsf{E}\left(|X_i|^p\right) = 0,$$

then $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathsf{E}\left(X_i\right)) \xrightarrow{\mathsf{P}} 0$.

Proof Indeed, recall from Markov inequality.

$$\begin{aligned}\mathsf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mathsf{E}\left(X_i\right)\right)\right| > \varepsilon\right) &\le \mathsf{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \mathsf{E}\left(X_i\right)\right| > \varepsilon\right) \\ &\le \frac{1}{n\varepsilon}\sum_{i=1}^{n}\mathsf{E}\left(\left|X_i - \mathsf{E}\left(X_i\right)\right|\right)\end{aligned}$$

and we have

$$\mathsf{E}\left(\left|X_i - \mathsf{E}\left(X_i\right)\right|\right) \le 2\mathsf{E}\left(|X_i|\right) \le \frac{2}{n\varepsilon}\sum_{i=1}^{n}\mathsf{E}\left(|X_i|\right)$$

■

Lemma 2.35 (Consistency of generalized estimating equations)

Suppose $\Theta \subseteq \mathbb{R}^k$ is compact. Let $h_i(x_i) = \sup_{\gamma\in\Theta}\|\psi_i(x_i, \gamma)\|$ for $i = 1, 2, \ldots$

Let $\sup_i \mathsf{E}\left(|h_i(X_i)|^{1+\delta}\right) < \infty$ and $\sup_i \mathsf{E}\left(\|X_i\|^{\delta}\right) < \infty$ and $\psi_i \equiv \psi$). Suppose further that for **any** $c > 0$ and **sequence** $\{x_i\}$ satisfying $\|x_i\| \leq c$, the sequence of functions $\{g_i(\gamma) = \psi_i(x_i, \gamma)\}$ is equicontinuous on any open subset of Θ. Then

$$\sup_{\gamma\in\Theta} \left\| \frac{1}{n} \sum_{i=1}^{n} \left(\psi_i(X_i, \gamma) - \mathsf{E}\left(\psi_i(X_i, \gamma)\right) \right) \right\| \xrightarrow{\mathsf{P}} 0$$

We want that the sequence of $\psi_i(X_i, \gamma)$ converge to the target $\mathsf{E}\left(\psi_i(X_i, \gamma)\right)$ and as such the roots also converge for the GEE equations.

Proof Since we need only consider components of ψ_i's, without loss of generality we can assume that ψ_i's are functions from $\mathbb{R}^{d_1} \times \Theta \to \mathbb{R}$. For any $c > 0$,

$$\sup_n \mathsf{E}\left(\frac{1}{n} \sum_{i=1}^{n} h_i(X_i) \mathbf{1}_{\|X_i\|\in(c,\infty)} \right) \leq \sup_i \mathsf{E}\left(h_i(X_i) \mathbf{1}_{\|X_i\|\in(c,\infty)} \right)$$

Let $c_0 = \sup_i \mathsf{E}\left(h_i(X_i)1 + \delta\right)$ and $c_1 = \sup_i \mathsf{E}\left(\|X_i\|^{\delta}\right)$. Recall Hölder's inequality, namely

$$\mathsf{E}\left(|XY|\right) \leq \left(\mathsf{E}\left(|X|^p\right)\right)^{\frac{1}{p}} \left(\mathsf{E}\left(|Y|^q\right)\right)^{\frac{1}{q}}$$

where p, q are conjugate exponents, that is $p^{-1} + q^{-1} = 1$, $p \geq 1$. The proof of this inequality follows from the concavity of $f(x) = \log(x)$. Using Hölder's inequality with $p = 1 + \delta$

$$\begin{aligned}
\mathsf{E}\left(h_i(X_i) \mathbf{1}_{\|X_i\|\in(c,\infty)} \right) &\leq \left(\mathsf{E}\left(|h_i(X_i)|^{1+\delta}\right)\right)^{\frac{1}{1+\delta}} \left(\left[\mathsf{P}\left(\|X_i\| > c\right)\right]^{\frac{1+\delta}{\delta}} \right)^{\frac{\delta}{1+\delta}} \\
&\leq \left(\mathsf{E}\left(|h_i(X_i)|^{1+\delta}\right)\right)^{\frac{1}{1+\delta}} \left(\mathsf{P}\left(\|X_i\| > c\right)\right)^{\frac{\delta}{1+\delta}} \\
&\leq c_0^{\frac{1}{1+\delta}} c_1^{\frac{1}{1+\delta}} c^{-\frac{\delta}{1+\delta}}
\end{aligned}$$

using Markov's inequality as $\mathsf{P}\left(\|X_i\| > c\right) \leq \frac{1}{c}\mathsf{E}\left(\|X_i\|\right)$.

Let $\varepsilon, > 0, \tilde{\varepsilon} > 0$ be given. Choose a c such that

$$c_0^{\frac{1}{1+\delta}} c_1^{\frac{1}{1+\delta}} c^{-\frac{\delta}{1+\delta}} < \varepsilon \frac{\tilde{\varepsilon}}{4}$$

Then, for any $O \subset \Theta$,

$$\mathsf{P}\left(\frac{1}{n} \sum_{i=1}^{n} \left(\sup_{\gamma\in O} \psi_i(X_i, \gamma) - \inf_{\gamma\in O} \psi_i(X_i, \gamma) \right) \mathbf{1}_{\|X_i\|\in(c,\infty)} > \frac{\varepsilon}{2} \right) < \tilde{\varepsilon} \tag{2.14}$$

using Markov inequality and the fact that

$$\sup_{\gamma\in O}\psi_i(X_i,\gamma)-\inf_{\gamma\in O}\psi_i(X_i,\gamma)\leq 2h_i(X_i).$$

Using the equicontinuity of $\{\psi_i(x_i,\gamma)\}$ we have a δ_ε such that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sup_{\gamma\in O_\varepsilon}\psi_i(X_i,\gamma)-\inf_{\gamma\in O_\varepsilon}\psi_i(X_i,\gamma)\right)\mathbf{1}_{\|X_i\|\in(0,c)}<\frac{\varepsilon}{2} \tag{2.15}$$

for sufficiently large n, where O_ε denotes any open ball in $\mathbb{R}^k$ with radius less than δ_ε. These results, together with the WLLN and $\|\psi_i(X_i,\gamma)\|\leq h_i(X_i)$ imply that

$$\mathsf{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\sup_{\gamma\in O_\varepsilon}\psi_i(X_i,\gamma)-\mathsf{E}\left(\inf_{\gamma\in O_\varepsilon}\psi_i(X_i,\gamma)\right)\right)>\varepsilon\right)\to 0 \tag{2.16}$$

For this, add and substract $\inf_{\gamma\in O_\varepsilon}\psi_i(X_i,\gamma)$ and then use (2.14) and (2.15).

Let

$$H_n(\gamma)=\frac{1}{n}\sum_{i=1}^{n}\left(\psi_i(X_i,\gamma)-\mathsf{E}\left(\psi_i(X_i,\gamma)\right)\right)$$

Then

$$\sup_{\gamma\in O_\varepsilon}H_n(\gamma)\leq\sum_{i=1}^{n}\left(\sup_{\gamma\in O_\varepsilon}\psi_i(X_i,\gamma)-\mathsf{E}\left(\inf_{\gamma\in O_\varepsilon}\psi_i(X_i,\gamma)\right)\right)$$

and using (2.16) we have

$$\mathsf{P}\left(\sup_{\gamma\in O_\varepsilon}H_n(\gamma)>\varepsilon\right)\xrightarrow{\mathsf{P}}0$$

as $n\to\infty$. Likewise,

$$\mathsf{P}\left(\sup H_n(\gamma)<-\varepsilon\right)\xrightarrow{\mathsf{P}}0$$

Since Θ is compact, there exists m_ε, open balls $O_{\varepsilon,j}$ such that $\Theta\subset\bigcup_j O_{\varepsilon,j}$. Then the result follows from

$$\mathsf{P}\left(\sup_{\gamma\in\Theta}|H_n(\gamma)>\varepsilon\right)\leq\sum_{j=1}^{m_\varepsilon}\mathsf{P}\left(\sup_{\gamma\in O_{\varepsilon,j}}|H_n(\gamma)|>\varepsilon\right)\to 0$$

■

We shall not prove the asymptotic normality of generalized estimating equations. Let $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \dots \boldsymbol{X}_{id})$ for $i = 1, \dots, n$ where $\mathsf{E}(\boldsymbol{X}_{it}) = \mu(\eta_{it}) = g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{z}_{it})$ and $\mathsf{Var}(\boldsymbol{X}_{it}) = \boldsymbol{\Phi}_i \mu'(\eta_{it})$ and $\boldsymbol{z}_{it}$ are k-vector value of covariates.

Suppose one looks at repeated measurement for calcium suplement on blood pressure. We have many measurements on the longitudinal setting, with few covariates that are not time dependent, $t = 1, \dots d$ time points. Balanced data when measurements are not equally spaced.

Which race show earlier reaction to this treatment. If we knew whe the drug kicks in, knowing the change the mean blood pressure, we have a mixture distribution, with a changepoint problem. If we have unbalance, this problem becomes quite difficult. If we have the mixture $\sum_{j=1}^{d_i} p_j f_j(x)$. and d_i are not the same, the dimensin of the parameter space is not fixed. We lose the asymptotics since the point may not be inside the space.

In biostatistics and lifetime testing problems, components of $\boldsymbol{X}_i$ are repeated measurements at different time points, from subject i and are called **longitudinal data**.

Let $\boldsymbol{R}_i$ be the $d_i \times d_i$ correlation matrix whose $(t,l)^{\text{th}}$ element is the correlation coefficient between X_{it} and X_{il}. Then, the covariance matrix

$$\mathsf{Var}(\boldsymbol{X}_i) = \boldsymbol{\Phi}_i [\boldsymbol{D}_i(\boldsymbol{\beta})]^{\frac{1}{2}} \boldsymbol{R}_i [\boldsymbol{D}_i(\boldsymbol{\beta})]^{\frac{1}{2}}$$

where $\boldsymbol{D}_i$ is the $d_i \times d_i$ diagonal matrix with the t^{th} diagonal denotes $(g^{-1})^\top(\boldsymbol{\gamma}^\top \boldsymbol{Z}_{it})$. If $\boldsymbol{R}_i$ are known,

$$\sum_{i=1}^{n} \boldsymbol{G}_i(\boldsymbol{\gamma}) \left([\boldsymbol{D}_i(\boldsymbol{\gamma})]^{\frac{1}{2}} \boldsymbol{R}_i [\boldsymbol{D}_i(\boldsymbol{\gamma})]^{\frac{1}{2}} \right)^{-1} [\boldsymbol{x}_i - \mu_i(\boldsymbol{\gamma})] = 0. \tag{2.17}$$

where $\boldsymbol{\mu}_i(\boldsymbol{\gamma}) = (\mu(\psi(\boldsymbol{\gamma}^\top \boldsymbol{z}_{i1})), \dots, \mu(\psi(\boldsymbol{\gamma}^\top \boldsymbol{z}_{id_i})))$ so that $\psi(g \circ \mu)^{-1}$ (which becomes the identity link if one choses the canonical link function), and $\boldsymbol{G}_i = \partial \boldsymbol{\mu}_i(\boldsymbol{\gamma}) / \partial \boldsymbol{\gamma}$.

Consider (2.17) in which the true $\boldsymbol{R}$ is replaced by a known $\widetilde{\boldsymbol{R}}_i$ and define

$$\psi_i(\boldsymbol{x}_i, \boldsymbol{\gamma}) = \boldsymbol{G}_i(\boldsymbol{\gamma}) \left([\boldsymbol{D}_i(\boldsymbol{\gamma})]^{\frac{1}{2}} \widetilde{\boldsymbol{R}}_i [\boldsymbol{D}_i(\boldsymbol{\gamma})]^{\frac{1}{2}} \right)^{-1} [\boldsymbol{x}_i - \mu_i(\boldsymbol{\gamma})]$$

and suppose the parameter space is compact and $\sup_i \|z_i\|$. Then we can show that all the conditions of Lemma 2.35 are fulfilled.

2.7 Maximum likelihood

We have four types of results for maximum likelihood estimates, under continuity assumption, namely the following, which we will show also extend to the non-parametric case.

1. Consistency
2. Asymptotic normality.
3. Invariance, if $\widehat{\theta}_{\text{ML}}$ is the MLE for θ, then $\psi(\widehat{\theta}_{\text{ML}})$ is the MLE for $\psi(\theta)$
4. Convolution theorem: confining ourself to a class of estimators, if $\widetilde{\theta}$ is regular and converges at the same rate as ML estimate, given

$$\sqrt{n}(\widehat{\theta}_{\text{ML}} - \theta) = O_p(1)$$

and we can show that $\widetilde{\theta}_n = \widehat{\theta}_{\text{ML}} + X$, where $X \perp\!\!\!\perp \widehat{\theta}_{\text{ML}}$ and as such always have larger variance.

The minimal conditions that we will need are **identifiability**, **smoothness**, **positive definitiveness** of the Information matrix.

Definition 2.36 (Identifiability)

For $\theta \in \Theta \to f_\theta(x)$, $\theta_1 = \theta_2 \Rightarrow f_{\theta_1}(x) = f_{\theta_2}(x)$ a.s. ν, and $f_{\theta_1}(x) = f_{\theta_2}(x)$ a.e. ν implies $\theta_1 = \theta_2$ given a one-to-one map $\theta \overset{\pi}{\leftrightarrow} f_\theta$, where $\pi : \Theta \to \mathcal{L}^1$.

This is crucial condition for consistency, as if it was to fail, then we could have the entire population and yet not have enough to know the parameter. This guarantees some form of uniqueness to the target.

Definition 2.37 (Information matrix)

We define Fisher information as

$$\mathcal{I}(\theta) = \mathsf{E}_\theta \left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2$$

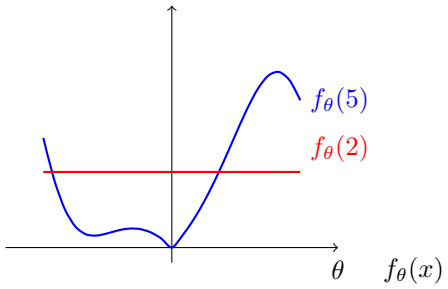
and under regularity conditions, we have also

$$\mathcal{I}(\theta) = -\mathsf{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right).$$

Thus, the information is

$$\mathcal{I}(\boldsymbol{\theta}) = \mathsf{Var} \left(\frac{\partial}{\partial \theta_1} \log f_{\boldsymbol{\theta}}(\boldsymbol{X}), \ldots, \frac{\partial}{\partial \theta_k} \log f_{\boldsymbol{\theta}}(\boldsymbol{X}) \right)$$

We have sensitivity with changes in θ. Given for example f_θ at $x = 2$ and $x = 5$ with the corresponding plot for values of θ displayed below, one would choose 2 on the basis of sensitivity for estimation purpose. One may also wonder the observed information, or the observed information, a paper by Efron and Hinckley in Biometrika discuss about the



efficiency of the estimators. Compute the **observed information**, the empirical estimate of the Information matrix,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\log f_\theta(x_i)\right)^2$$

versus the **estimated information** that uses the MLE estimator for θ.

$$-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2}{\partial\theta^2}\log f_\theta(x_i)\right)$$

and for more, refer to Rothenberg (1971) in a paper in Econometrica.

2.8 Consistency of the MLE

The proof is due to Ibragimov and Hasminski, 1981

Lemma 2.38
Suppose $X_i \overset{\text{iid}}{\sim} f(\boldsymbol{x};\theta)$ for $i = 1, 2, \ldots, n$ and $\theta \in \Theta \subseteq \mathbb{R}^p$ is a bounded open set. Define

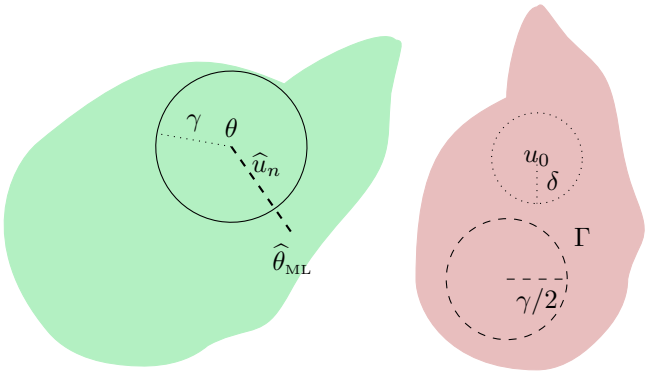
$$R_{n,\theta}^{\frac{1}{2}}(u) = \left[\frac{\mathcal{L}(\theta+u; x_1, \ldots, x_n)}{\mathcal{L}(\theta; x_1, \ldots, x_n)}\right]^{\frac{1}{2}} = \prod_{i=1}^{n}\frac{f^{\frac{1}{2}}(\boldsymbol{x}_i;\theta+u)}{f^{\frac{1}{2}}(\boldsymbol{x}_i;\theta)} \tag{2.18}$$

where $u \in \mathcal{U} = \Theta - \theta$ and we understand the latter as being the set of all points of Θ shifted by θ . Then

$$\widehat{\theta}_{\text{ML}} \xrightarrow{\mathsf{P}} \theta \quad \text{if} \quad \mathsf{P}\left(\sup_{\|u\|>\gamma} R_{n,\theta}^{\frac{1}{2}} \geq 1\right) \xrightarrow{n\to\infty} 0 \ \ \forall\, \gamma > 0.$$

If the last bit holds true uniformly in $\theta \in K \subseteq \Theta$, then $\widehat{\theta}_{\text{ML}}$ is uniformly consistent in K.

Proof The idea of the proof is simple. We want to show that $\widehat{\theta}_{\text{ML}}$ should be in the vicinity of θ. In other words, the chance that $\widehat{\theta}_{\text{ML}}$ falls outside of a neighbourhood around



θ tends to zero, no matter how small is the neighbourhood. Set $\widehat{u}_n = \widehat{\theta}_{\mathrm{ML}} - \theta$ so that $R_{n,\theta}(\widehat{u}_n) = \sup_u R_{n,\theta}(u)$. Then

$$\begin{aligned}\mathsf{P}_\theta^*\left(\|\widehat{\theta}_{\mathrm{ML}} - \theta\| > \gamma\right) &= \mathsf{P}_\theta\left(\|\widehat{u}_n\| > \gamma\right) \\ &\leq \mathsf{P}\left(\sup_{\|u\|>\gamma} R_{n,\theta}(u) \geq 1\right) \xrightarrow{n\to\infty} 0\end{aligned}$$

■

Theorem 2.39 (Consistency of the Maximum Likelihood Estimators)

Suppose Θ is a bounded open subset of $\mathbb{R}^p$ and $f(\boldsymbol{x};\theta)$ is the PDF of P_θ with respect to the σ-finite measure ν. Let $f(\boldsymbol{x};\theta)$ be a continuous function of θ on $\overline{\Theta}$; [17] for almost all $\boldsymbol{x} \in \mathcal{X}$ and let the following conditions be satisfied

1. For all $\theta \in \Theta$ and all $\gamma > 0$,

$$0 < K_\theta(\gamma) = \inf_{\|\theta-\theta'\|>\gamma} r_2^2(\theta,\theta') = \inf_{\|\theta-\theta'\|>\gamma} \int_{\boldsymbol{x}} \left[f^{\frac{1}{2}}(\boldsymbol{x};\theta) - f^{\frac{1}{2}}(\boldsymbol{x};\theta')\right] \nu(\,\mathrm{d}\boldsymbol{x})$$

is positive and thus distinguishable. This is an identifiability condition. [18]

[17] $\overline{\Theta}$ denotes the closure of the set Θ

[18] Here r_2^2 is the Hellinger distance between two distributions given as above. Since densities are in $\mathcal{L}^1$, we could try to embed into the unit ball of a Hilbert space; we can then invoke the notion of orthogonality, which is the $\mathcal{L}^2$ norm in that space.

2. For all $\theta \in \overline{\Theta}$,[19]

$$\left\{ \int_{\mathcal{X}} \sup_{\|t\| \le \delta} \left[f^{\frac{1}{2}}(\boldsymbol{x}; \theta) - f^{\frac{1}{2}}(\boldsymbol{x}; \theta + t) \right]^2 \right\}^{\frac{1}{2}} = \omega_\theta(\delta) \xrightarrow{\delta \to 0} 0$$

Then $\widehat{\theta}_{\mathrm{ML}} \xrightarrow{\mathsf{P}} \theta$ (or in fact almost surely) and as such this proves consistency of the maximum likelihood estimates.

Proof In view of the above lemma and Markov's inequality, it suffices to find an upper bound for the expectation $\mathsf{E}_\theta \left(\sup_\Gamma R_n(u) \right)$ which tends to zero as $n \to \infty$. This is the core of the proof which is given below.

Suppose θ is fixed. Consider $R_{n,\theta}(u)$ as a function of u. Let Γ be a sphere of small radius δ situated in its entirety in the region $\|u\| > \frac{1}{2}\gamma$. We shall bound the expectation $\mathsf{E}_\theta \left(\sup_\Gamma R_n(u) \right)$. If u_0 is the center of Γ, then

$$\begin{aligned} \sup_\Gamma R_n^{\frac{1}{2}}(u) &= \sup_\Gamma \prod_{i=1}^n \left[\frac{f(\boldsymbol{x}_i; \theta + u)}{f(\boldsymbol{x}_i; \theta)} \right]^{\frac{1}{2}} \\ &\le \prod_{i=1}^n f^{-\frac{1}{2}}(\boldsymbol{x}_i; \theta) \left[f^{\frac{1}{2}}(\boldsymbol{x}_i; \theta + u_0) \right. \\ &\quad \left. + \sup_{\|t\| \le \delta} \left| f^{\frac{1}{2}}(\boldsymbol{x}_i; \theta + u_0 + t) - f^{\frac{1}{2}}(\boldsymbol{x}_i; \theta + u_0) \right| \right] \end{aligned}$$

for $t = u - u_0$. Note that $f(\boldsymbol{x}; \theta + u) = f(\boldsymbol{x}; \theta + u_0) + [f(\boldsymbol{x}; \theta + u_0 + t) - f(\boldsymbol{x}; \theta + u_0)]$. Therefore, we obtain by Markov inequality

$$\begin{aligned} \mathsf{E}_\theta \left(\sup_\Gamma R_n(u) \right) \le \prod_{i=1}^n \left[\int_{\mathcal{X}} f^{\frac{1}{2}}(\boldsymbol{x}_i; \theta) f^{\frac{1}{2}}(\boldsymbol{x}_i; \theta + u_0) \nu(\,\mathrm{d}\boldsymbol{x}) \right. \\ \left. + \int_{\mathcal{X}} \sup_{\|t\| \le \delta} \left| f^{\frac{1}{2}}(\boldsymbol{x}_i; \theta + u_0 + t) - f^{\frac{1}{2}}(\boldsymbol{x}_i; \theta + u_0) \right| f^{\frac{1}{2}}(\boldsymbol{x}_i; \theta) \nu(\,\mathrm{d}\boldsymbol{x}) \right] \end{aligned}$$

[19] This condition is some type of modulus of continuity condition. Indeed, if we could invoke the mean value theorem, we would have the distance between the two giving boundedness of the derivative, and so when $\delta \to 0$, we would get the result.

Now, it is easy to see that

$$\begin{aligned}
&\int_{\mathcal{X}} f^{\frac{1}{2}}(\boldsymbol{x}_i;\theta) f^{\frac{1}{2}}(\boldsymbol{x}_i;\theta+u_0)\nu(\,\mathrm{d}\boldsymbol{x}) \\
&\quad = \frac{1}{2}\left\{\int_{\mathcal{X}} f(\boldsymbol{x};\theta)d\nu + \int_{\mathcal{X}} f(\boldsymbol{x};\theta+u_0)d\nu \right. \\
&\qquad \left. - \int_{\mathcal{X}} \left[f^{\frac{1}{2}}(\boldsymbol{x};\theta+u_0) - f^{\frac{1}{2}}(\boldsymbol{x};\theta)\right]^2 d\nu\right\} \\
&\quad = \frac{1}{2}[2 - r_2^2(\theta,\theta+u_0)] \\
&\quad \leq 1 - \frac{\kappa_\theta\left(\frac{\gamma}{2}\right)}{2}
\end{aligned}$$

On the other hand, using the Cauchy-Schwartz inequality and condition 2,

$$\int_{\mathcal{X}} \sup_{\|t\|\leq\delta} \left| f^{\frac{1}{2}}(\boldsymbol{x}_i;\theta+u_0+t) - f^{\frac{1}{2}}(\boldsymbol{x}_i;\theta+u_0)\right| f^{\frac{1}{2}}(\boldsymbol{x}_i;\theta)d\nu \leq \omega_{\theta+u_0}(\delta).$$

Taking this into account together with the elementary inequality $1+a \leq e^a$ for $a \in \mathbb{R}$, we obtain

$$\mathsf{E}_\theta\left[\sup_{\Gamma} R_n^{\frac{1}{2}}(u)\right] \leq \exp\left\{-n\left[\frac{\kappa_\theta\left(\frac{\gamma}{2}\right)}{2} - \omega_{\theta+u_0}(\delta)\right]\right\}. \tag{2.19}$$

It follows from (2.19) that to each point ξ of the set $\overline{\mathcal{U}} \setminus \{\|u\| \leq \gamma\}$ there corresponds a sphere $\Gamma(\xi)$ with center ξ such that $\mathsf{E}_\theta\left[\sup_{\Gamma(\xi)} R_n^{\frac{1}{2}}(u)\right] \to 0$ in $\mathsf{P}_\theta-$probability as $n \to \infty$. Using compactness of $\overline{\Theta}$, select a finite cover $\Gamma(\xi_q)$ for $q = 1, 2, \ldots, N$ of the set $\overline{\mathcal{U}} \setminus \{\|u\| \leq \gamma\}$ from the collection $\{\Gamma(\xi)\}$. Then

$$\sup_{\|u\|\geq\gamma} R_n^{\frac{1}{2}}(u) \leq \sum_{q=1}^{N} \sup_{\Gamma(\xi_q)} R_n^{\frac{1}{2}}(u) \xrightarrow{n\to\infty} 0 \text{ in } \mathsf{P}_\theta.$$

Since each of the expectation of each term on the right hand side tends to zero, the proof is complete.[20] ■

[20]Note that κ is positive and ω goes to zero, so $-n\kappa$ is negative and $e^{-n\kappa} \to 0$ as $n \to \infty$.

Consistency for quasi-identifiable models

In a regression context, if we have $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$ for $i = 1, \ldots, n$, then

$$\prod_{i=1}^{n} f(\boldsymbol{x}_i, y_i) = \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i) f(\boldsymbol{x}_i) = \left[\prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i)\right] \left[\prod_{i=1}^{n} f(\boldsymbol{x}_i)\right] \propto \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i)$$

which can be the case for people with different covariates, this leads to a conditional analysis. In this context, we have independent, but not identically distributed. The proof given above can be arranged to work, with some additional conditions. See paper for the statement and proof. If possible values for covariates can be discretized and is finite, then as long as the number of observations in the categories increase faster than $\log(n)$, we still have consistency.

Consistency of the MLE

We have under the regularity conditions established below that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\boldsymbol{\theta}))$$

where the information is

$$\mathcal{I}(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\boldsymbol{x}) \right)^2 = [\mathcal{I}_{ij}(\boldsymbol{\theta})]$$

where

$$\mathcal{I}_{ij} = \mathsf{Cov} \left(\frac{\partial}{\partial \theta_i} \log f_{\boldsymbol{\theta}}(\boldsymbol{x}), \frac{\partial}{\partial \theta_j} \log f_{\boldsymbol{\theta}}(\boldsymbol{x}) \right).$$

2.9 Cramer-Fréchet-Rao lower bound

If we have the following conditions, namely $\mathsf{E}_{\boldsymbol{\theta}}(T) = g(\boldsymbol{\theta})$,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathcal{X}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_{\mathcal{X}} \frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$
$$\frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathcal{X}} T(\boldsymbol{x}) f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_{\mathcal{X}} T(\boldsymbol{x}) \frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

We then have the following uncertainty bound, the so-called Fréchet-Cramer-Rao bound, which we give the proof of in the one dimensional case: if $\mathsf{E}_{\boldsymbol{\theta}}(T^2) < \infty$, then

$$\mathsf{Var}(T) \geq \frac{(g'(\boldsymbol{\theta}))^2}{\mathcal{I}(\boldsymbol{\theta})}$$

Proof Consider the case where

$$\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{x})\right)=0$$
$$\int_{\mathcal{X}}\frac{f'_\theta(\boldsymbol{x})}{f_\theta(\boldsymbol{x})}fs_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=0$$

Then

$$\mathsf{E}_\theta\left(T(\boldsymbol{x})\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{x})\right)=g'(\theta)$$

and thus

$$\begin{aligned}\int_{\mathcal{X}}T(\boldsymbol{x})\frac{f'_\theta(\boldsymbol{x})}{f_\theta(\boldsymbol{x})}f_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}&=\int_{\mathcal{X}}T(\boldsymbol{x})\frac{\partial}{\partial\theta}f_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\&=\frac{\partial}{\partial\theta}\int_{\mathcal{X}}T(\boldsymbol{x})f_\theta(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\&=g'(\theta)\end{aligned}$$

thus

$$\mathsf{E}_\theta\left([T(\boldsymbol{x})-g(\theta)]\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{x})\right)=g'(\theta)$$

and then applying Cauchy-Schwartz inequality,

$$(g'(\theta))^2\leq\mathsf{Var}\left(T\right)\mathsf{Var}\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{x})\right)$$

and as such

$$\mathsf{Var}\left(T\right)\geq\frac{(g'(\theta))^2}{\mathcal{I}(\theta)}$$

and if $g(\theta)=\theta$, we have $\mathsf{Var}\left(T\right)\geq\mathcal{I}(\theta)^{-1}$. ■

Suppose that $\mathsf{E}_\theta\left(\delta\right)=g(\theta)$, then the Fréchet-Cramer-Rao lower bound is given by

$$\mathsf{Var}\left(\delta\right)\geq\frac{[g'(\theta)]^2}{\mathcal{I}(\theta)}$$

In the proof, what we used was really the fact that $|\mathsf{Cov}\left(X,Y\right)|^2\leq\mathsf{Var}\left(X\right)\mathsf{Var}\left(Y\right)$ using Y as the score function, provided that $\mathsf{Var}\left(Y\right)\neq 0$. Suppose we want to generalize the above

and replace the score function

$$\frac{\partial}{\partial\theta}\log f_\theta(x)$$

Take $\psi(x,\theta)$ and δ an estimator of $g(\theta)$, then

$$\mathsf{Var}\left(\delta\right) \geq \frac{\mathsf{Cov}\left(\delta,\psi\right)}{\mathsf{Var}\left(\psi\right)}$$

if $\mathsf{Var}\left(\psi\right) \neq 0$. However, this is not useful since the covariance depends on δ, so we do not have a gold standard.

Blyth's theorem states the necessary and sufficient conditions for the $\mathsf{Cov}\left(\delta,\psi\right)$ to be a function solely of ψ. Suppose $\mathsf{E}_\theta\left(T\right) = 0$, then if

$$\mathcal{U}_0 = \left\{T : \mathsf{E}_\theta\left(T\right) = 0, \mathsf{E}_\theta\left(T^2\right) < \infty\right\}$$

and thus we have

$$\begin{aligned}
\mathsf{E}_\theta\left(T(X)\frac{\partial}{\partial\theta}\log f_\theta(X)\right) &= \int_{\mathcal{X}} T(x)\frac{\partial}{\partial\theta}\log f_\theta(x) f_\theta(x)\,\mathrm{d}x \\
&= \int_{\mathcal{X}} T(x)\frac{f'_\theta(x)}{f_\theta(x)} f_\theta(x)\,\mathrm{d}x \\
&= \int_{\mathcal{X}} T(x) f'_\theta(x)\,\mathrm{d}x
\end{aligned}$$

and under regularity conditions

$$\begin{aligned}
&= \frac{\partial}{\partial\theta}\int_{\mathcal{X}} T(x) f_\theta(x)\,\mathrm{d}x \\
&= \frac{\partial}{\partial\theta}\mathsf{E}_\theta\left(T\right) = 0
\end{aligned}$$

and so $\psi \perp \mathcal{U}_0$, $\widehat{\theta}$ is the solution $\psi(x;\widehat{\theta}) = 0$, which implies

$$\mathsf{Var}\left(\widehat{\theta}\right) = \frac{\mathsf{Cov}\left(\delta,\psi\right)}{\mathsf{Var}\left(\psi\right)}$$

and $\mathsf{Cov}\left(\delta,\psi\right)$ only depends on $\mathsf{E}\left(\widehat{\theta}\right)$.

The result presented next is due to Blyth (1974)

Theorem 2.40

A necessary and sufficient condition for $\mathsf{Cov}\left(\delta,\psi\right)$ to depend on δ only through $g(\theta)\left(=\mathsf{E}_\theta\left(\delta\right)\right)$

is that for all θ,

$$\mathsf{Cov}\left(U, \psi\right) = 0, \ \forall \, U \in \mathcal{U}_0$$

where

$$\mathcal{U}_0 = \left\{U : \mathsf{E}_\theta\left(U\right) = 0, \mathsf{E}_\theta\left(U^2\right) < \infty, \ \forall \, \theta \in \Theta\right\}$$

Therefore, if $\psi \perp \mathcal{U}_0$,

$$\mathsf{Var}\left(\delta\right) \geq \frac{h(\mathsf{E}_\theta\left(\delta\right))}{\mathsf{Var}\left(\psi\right)}$$

where $g(\theta) \equiv h(\mathsf{E}_\theta\left(\delta\right))$.

Proof See assignment. ■

We can now extend the FCR result to higher dimension.

Theorem 2.41

For any unbiased estimator δ of $g(\boldsymbol{\theta})$ and any function $\psi_i(\boldsymbol{x}; \boldsymbol{\theta})$ with finite second moments, we have

$$\mathsf{Var}\left(\delta\right) \geq \boldsymbol{\gamma}^\top \mathbf{C}^{-1} \boldsymbol{\gamma}$$

where $\boldsymbol{\gamma}^\top = (\gamma_1, \ldots, \gamma_r)$ and

$$\mathbf{C} = \Big[\mathbf{C}_{ij}\Big]_{i,j=1,\ldots,r}$$

with

$$\gamma_i = \mathsf{Cov}\left(\delta, \psi_i\right) \qquad \mathbf{C}_{ij} = \mathsf{Cov}\left(\psi_i, \psi_j\right). \tag{2.20}$$

The right hand side of (2.20) depend on δ only through $g(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}\left(\delta\right)$ provided each of the functions $\psi_i \perp \mathcal{U}_0$ for $i = 1, \ldots, r$. Now, if $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ and $\psi_i = \frac{\partial}{\partial \theta_i} \log f_{\boldsymbol{\theta}}(x)$ and

$$\mathbf{C}_{ij} = \mathsf{Cov}\left(\psi_i, \psi_j\right) = \mathsf{Cov}\left(\frac{\partial}{\partial \theta_i} \log f_{\boldsymbol{\theta}}(\boldsymbol{x}), \frac{\partial}{\partial \theta_j} \log f_{\boldsymbol{\theta}}(\boldsymbol{x})\right)$$

and $\boldsymbol{\psi}(\boldsymbol{x};\boldsymbol{\theta}) = \nabla \log f_{\boldsymbol{\theta}}(\boldsymbol{x})$ where we recover the information matrix, as

$$\mathcal{I}(\boldsymbol{\theta}) = \left[\mathcal{I}_{ij}(\boldsymbol{\theta})\right]_{i,j=1,\ldots,r}$$
$$\mathcal{I}_{ij}(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(\boldsymbol{x})\frac{\partial}{\partial\theta_j}\log f_{\boldsymbol{\theta}}\right)$$

Proof For any constants $a_1,\ldots,a_j$, it follows that

$$\mathsf{Var}\,(\delta) \geq \frac{\left[\mathsf{Cov}\,(\delta,\sum_{i=1}^r a_i\psi_i)\right]^2}{\mathsf{Var}\,(\sum_{i=1}^r a_i\psi_i)}. \tag{2.21}$$

We thus have

$$\mathsf{Cov}\left(\delta,\sum_{i=1}^r a_i\psi_i\right) = \sum_{i=1}^r a_i\gamma_i = \boldsymbol{a}^\top\boldsymbol{\gamma}$$
$$\mathsf{Var}\left(\sum_{i=1}^r a_i\psi_i\right) = \boldsymbol{a}^\top\mathbf{C}\boldsymbol{a}$$

Since (2.21) is true for any vector $\boldsymbol{a}$, we have

$$\mathsf{Var}\,(\delta) \geq \max_{\boldsymbol{a}}\frac{[\boldsymbol{a}^\top\boldsymbol{\gamma}]^2}{\boldsymbol{a}^\top\mathbf{C}\boldsymbol{a}} = \boldsymbol{\gamma}^\top\mathbf{C}^{-1}\boldsymbol{\gamma}$$

The above result is based on spectral decomposition. ■

Lemma 2.42
Suppose $\mathbf{P} = \boldsymbol{p}\boldsymbol{p}^\top$ where $\boldsymbol{p}$ is a $(r\times 1)$ vector and $\mathbf{Q}$ is a positive definite matrix, then

$$\max_{\boldsymbol{a}}\frac{\boldsymbol{a}^\top\mathbf{P}\boldsymbol{a}}{\boldsymbol{a}^\top\mathbf{Q}\boldsymbol{a}}$$

equal to the largest eigenvalue of $\mathbf{Q}^{-1}\mathbf{P} = \boldsymbol{p}^\top\mathbf{Q}^{-1}\boldsymbol{p}$. For more, see C.R. Rao *Linear Statistical Inference and its applications*(1965), p.48.

If $\boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{\theta}) = \nabla \log f_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $\mathbf{C} = \mathcal{I}(\boldsymbol{\theta})$, then

$$\mathsf{Var}\,(\delta) \geq \boldsymbol{\gamma}^\top\mathcal{I}(\boldsymbol{\theta})^{-1}\boldsymbol{\gamma}$$

where

$$\gamma_i = \mathsf{Cov}\,(\delta,\psi_i) = \mathsf{Cov}\left(\delta,\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(\boldsymbol{x})\right)$$

Under the FCR regularity conditions,

$$\mathsf{E}_\theta\left(\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(\boldsymbol{X})\right)=0$$

and

$$\begin{aligned}\mathsf{Cov}\left(\delta,\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(\boldsymbol{X})\right)&=\mathsf{E}_{\boldsymbol{\theta}}\left(\delta(\boldsymbol{X})\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(\boldsymbol{x})\right)\\&=\frac{\partial}{\partial\theta_i}g(\boldsymbol{\theta})\end{aligned}$$

as it is easily seen that

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\theta}}\left(\delta(\boldsymbol{X})\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(\boldsymbol{x})\right)&=\int_{\mathcal{X}}\delta(\boldsymbol{x})\frac{f'_{\theta_i}(\boldsymbol{x})}{f_{\boldsymbol{\theta}}(\boldsymbol{x})}f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\&=\frac{\partial}{\partial\theta_i}\int_{\mathcal{X}}\delta(x)f_{\boldsymbol{\theta}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\&=\frac{\partial}{\partial\theta_i}g(\boldsymbol{\theta})\end{aligned}$$

Corollary 2.43
If $\boldsymbol{\psi}=\nabla\log f_{\boldsymbol{\theta}}(\boldsymbol{x})$, then

$$\mathsf{Var}\,(\delta)\geq[\nabla g(\boldsymbol{\theta})]^\top\mathcal{I}(\boldsymbol{\theta})^{-1}[\nabla g(\boldsymbol{\theta})]$$

We need positive definitiveness of the matrix. For exponential family, unless one has linear dependence. For mixtures, this is a difficult problem. Recall that an $(r\times r)$ matrix $\mathbf{C}$ is positive-semi definite if for any vector $\boldsymbol{a}^\top\mathbf{C}\boldsymbol{a}\geq 0$ and the inequality is strict for positive definite. Other condition is that $\det\mathbf{C}>0$, so we require in our case that $\det\mathcal{I}(\boldsymbol{\theta})>0$. If we cannot establish positive definitiveness

$$\begin{aligned}&\Gamma\{\boldsymbol{\theta}\in\Theta:\det(\mathcal{I}(\boldsymbol{\theta}))=0\}\\&\mathrm{F}=\{f_{\boldsymbol{\theta}}(\boldsymbol{x}):\boldsymbol{\theta}\in\Theta\}\end{aligned}$$

then Λ is a nowhere dense set, $\Gamma^\circ=\emptyset$.

We now aim at showing that the MLE has asymptotic distribution is given by

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\mathcal{I}^{-1}(\boldsymbol{\theta})\right)$$

Weak consistency does not imply $\mathsf{E}\left(|X_n-X|\right)$ to be finiete, however if $X_n\xrightarrow[\mathcal{L}^2]{}X$, then $\mathsf{E}\left(|X_n-X|^2\right)\to 0$ and $X_n\xrightarrow{\mathsf{P}}X$.

Hodges gave an example of super efficient estimators that goes as follows.[21] Suppose that $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$, then $\mathcal{I}(\theta) = 1$, since information in the 1d case for Normal is reciprocal to the variance as

$$\mathcal{I}(\theta) = \mathsf{E}_\theta \left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2$$

where

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2} \right)$$

If σ is known,

$$\log f_\theta(x) = -\log(\sqrt{2\pi}\sigma) - \frac{(x-\theta)^2}{2\sigma^2}$$

and thus

$$\begin{aligned} \frac{\partial}{\partial \theta} \log f_\theta(x) &= \frac{2(x-\theta)}{2\sigma^2} \\ &= \frac{1}{\sigma^2}(x-\theta) \end{aligned}$$

This means that

$$\begin{aligned} \mathcal{I}(\theta) &= \mathsf{E}_\theta \left(\frac{1}{\sigma^4}(X-\theta)^2 \right) \\ &= \frac{1}{\sigma^4} \mathsf{E}\left((X-\theta)^2 \right) = \frac{1}{\sigma^2} \end{aligned}$$

Back now to Hodges' example. The FCR lower bound for unbiased estimators of θ is 1, $\mathsf{E}_\theta(\delta) = \theta$, $\mathsf{Var}(\delta) \geq \frac{1}{\mathcal{I}(\theta)} = 1$ and so

$$\delta_n = \begin{cases} \bar{X}_n & \text{if } |\bar{X}_n| \geq \frac{1}{n^{\frac{1}{4}}} \\ a\bar{X}_n & \text{if } |\bar{X}_n| < \frac{1}{n^{\frac{1}{4}}} \end{cases}$$

and $\sqrt{n}(\delta_n - \theta) \xrightarrow{d} \mathcal{N}(0, \nu(\theta))$ where

$$\nu(\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \\ a^2 & \text{if } \theta = 0. \end{cases}$$

If $a^2 < 1$, then the FCR inequality is violated. These estimators are called **shrinkage estimator.**

[21] There is an article connecting this to penalization and variable selection by R. Beran

Theorem 2.44 (Le Cam theorem)

Let $X_1, \ldots, X_n$ be IID with density $f_\theta(x)$ where θ is real-valued and suppose the following conditions hold:

1. The parameter space Θ is an open interval
2. The distribution $f_\theta(x)$ have common support $\mathrm{A} = \{x : f_\theta(x) > 0\}$ does not depend on θ.
3. For every $x \in \mathrm{A}$, the density $f_\theta(x)$ is twice continuously differentiable with respect to θ
4. The integral $\int f_\theta(x)\,\mathrm{d}x$ can be twice differentiated under the integral sign.
5. $0 < I(\theta) < \infty$
6. For any $\theta_0 \in \Theta$, there exists a positive number c and a function $M(x)$ such that

$$\left|\frac{\partial^2}{\partial\theta^2}\log f_\theta(x)\right| \leq M(x) \qquad \forall\, x \in \mathrm{A}, \theta \in (\theta_0 - c, \theta_0 + c)$$

   and $\mathsf{E}_{\theta_0}\left(M(X)\right) < \infty$.

Under these assumptions if $\delta_n = \delta_n(x_1, \ldots, x_n)$ is any estimator such that

$$\sqrt{n}(\delta_n - \theta) \xrightarrow{d} \mathcal{N}(0, \nu(\theta))$$

then $\nu(\theta) \geq \mathcal{I}^{-1}(\theta)$ except on a set of Lebesgue measure zero.

We now tackle the problem of showing that $\sqrt{n}(\widehat{\theta}_{\mathrm{ML}} - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}\theta)$, which gives asymptotic confidence intervals, $P|Z| < 1.96 = 0.95$, then

$$\mathsf{P}\left(\sqrt{\mathcal{I}(\theta)}|\widehat{\theta}_{\mathrm{ML}} - \theta| < 1.96\right) = 0.95$$

$$\mathsf{P}\left(\widehat{\theta}_{\mathrm{ML}} - \frac{1.96}{\sqrt{\mathcal{I}(\theta)}} < \theta < \widehat{\theta}_{\mathrm{ML}} + \frac{1.96}{\sqrt{\mathcal{I}(\theta)}}\right) = 0.95$$

and replacing $\mathcal{I}(\theta)$ by the consistent estimate $\mathcal{I}(\widehat{\theta}) = \widehat{\mathcal{I}(\theta)}$. If $\widehat{\theta}$ is the maximum likelihood of θ, $h(\widehat{\theta})$ is the maximum likelihood of $h(\theta)$, under regularity conditions

$$\mathsf{E}\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2 = -\mathsf{E}\left(\frac{\partial^2}{\partial\theta^2}\log f_\theta * X)\right)$$

if we do not have regularity, we can use the Chapman-Kiefer-Robins hold, which may not be as sharp. This was not satisfactory for Hajek and Le Cam; they introduced regular estimators, not far from MLE (in the sense they converge at the same rate as MLE, $|\tilde{\theta} - \theta| = O_p(n^{-\frac{1}{2}})$). If we denote $f_{\tilde{\theta}}$ the asymptotic distribution of $\tilde{\theta}$, we then can use the convolution theorem to show that $f_{\tilde{\theta}} = f_{\hat{\theta}} * h$, then $\tilde{\theta} = \hat{\theta} + X$ and so since the two variables are independent, the variance of $f_{\tilde{\theta}}$ estimator must be bigger.

Confidence interval is also discussed, since they are approximate. It either covers the true value or not, so it should be zero or 1. The interpretation that one should have in mind is that of Buehler of **betting**, since the confidence interval is a pre-experimental approach; it holds before collecting the data, being a procedure of the estimators rather than the estimate. It is also approximate in the sense that the procedure depends on the sample size. θ is also in a Bayesian perspective not a random variable, assumed to be drawn from a prior $\pi(\theta)$; the posterior is

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{\int_\theta f(\boldsymbol{x}|\theta)\pi(\theta)\,\mathrm{d}\theta}$$

In machine learning and computer science, since the flow of data is continuous, this is the approach. In survival analysis, this is not the right approach. From a scientist perspective, since the choice of a prior that match the data can always be made, and always against the data. The choice of the prior is thus a form of scientific pollution which should be avoided.

Credible intervals work with the inference being made on the posterior distribution, so we can find an interval that covers 95% of the mass. We then have a probability statement.

For Bayesian statistics, under minimal conditions, we have $\pi(\theta|\boldsymbol{x}) \xrightarrow{\mathsf{P}} \delta_{\theta_0}$ and convergence to the mass when $\theta \in \Theta \subseteq \mathbb{R}^k$. However, when Θ is infinite, like often the case in non-parametric (like the Kaplan-Meier estimator) and survival analysis (we have finite data, but infinite dimension parameter space), we get this very often. $F(x)$ is estimated by $\widehat{F}_n(x)$, and the survival function, our parameter $S(x) = 1 - F(x)$ is infinite dimensional since we are making no assumption; we have robustness.

The problem with non-parametric is $\|\widehat{F}_n - F\| = O\left(\sqrt{\frac{\log\log(n)}{n}}\right)$ has a constant in front which grows rapidly when the dimension go up. Most of the time, the amount of data required is huge. See "Consistency of Bayes estimates for nonparametric regression: normal theory" by Diaconis and Freedman in *Bernoulli*(1998). We now sketch the proof for one dimension, can be extended easily for finite dimension.

Proposition 2.45 (Asymptotic Normality of the MLE)
Let $X_i \overset{\text{iid}}{\sim} f_\theta$ where $f \in \mathcal{C}^3$ (*i.e.* $f_\theta(x)$ is three times continuously differentiable w.r.t. θ for

almost all x). Then, using a Taylor expansion around the MLE,

$$\frac{\partial}{\partial\theta}\log f_\theta(x_i) = \frac{\partial}{\partial\theta}\log f_{\hat\theta}(x_i) + (\theta-\hat\theta)\frac{\partial^2}{\partial\theta^2} f_{\hat\theta}(x_i) + \epsilon_i$$

and since $\hat\theta$ is the solution of the first derivative of the likelihood (MLE is solution to the score function), we therefore have summing both sides

$$\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f_\theta(x_i) = 0 + (\hat\theta-\theta)\left[-\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f_{\hat\theta}(x_i)\right] + \varepsilon_n$$

where $\varepsilon_n = \sum_{i=1}^n \epsilon_i$, using the fact that $\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f_{\hat\theta}(x_i) = 0$ and the left hand side is the so-called score function and thus

$$\frac{\displaystyle\sum_{i=1}^n \frac{\partial}{\partial t}\log f_\theta(x_i)}{\displaystyle\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f_{\hat\theta}(x_i)} = (\hat\theta-\theta) + \frac{\varepsilon_n}{\displaystyle\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f_{\hat\theta}(x_i)}.$$

This then implies that if we can get the rightmost term to go to zero and apply CLT to show that the LHS converges to Normal, then by Slutski we are done. However, we need to scale properly. We want that for $\bar Y$, where

$$\bar Y = \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f_\theta(X_i)$$

and $\mathsf{Var}\left(\bar Y\right) = \frac{\mathsf{Var}(Y)}{n}$. Using the properties of information, we have

$$\mathsf{Var}\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right) = \mathsf{E}\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2 = -\mathsf{E}\left(\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right)$$

Then, back to the problem, we can rearrange as to get

$$\underbrace{\frac{1}{\sqrt{-\displaystyle\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f_{\hat\theta}(x_i) n}}}_{A} \cdot \underbrace{\frac{\frac{1}{n}\displaystyle\sum_{i=1}^n \frac{\partial}{\partial t}\log f_\theta(x_i)}{\sqrt{\frac{1}{n}\displaystyle\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f_{\hat\theta}(x_i)\Big/\sqrt{n}}}}_{B} = \sqrt{n}(\hat\theta-\theta) + \frac{\sqrt{n}\varepsilon_n}{-\displaystyle\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f_{\hat\theta}(x_i)}.$$

We now note,[22] using WLLN, that $A \xrightarrow{\mathsf{P}} \left(-\mathsf{E}\left(\frac{\partial^2}{\partial\theta^2}\log f_\theta(x)\right)\right)^{-\frac{1}{2}}$ and using the CLT that $B \xrightarrow{d} \mathcal{N}(0,1)$. The second term on the right hand side is equal to

$$\frac{\varepsilon_n}{\sqrt{n}} \cdot \left(-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f_{\hat{\theta}}(x_i)\right)^{-1} \xrightarrow{\mathsf{P}} \frac{\varepsilon_n}{\sqrt{n}}\mathcal{I}(\theta)^{-1}.$$

It then suffices to show that $\varepsilon_n/\sqrt{n} \xrightarrow{\mathsf{P}} 0$, or equivalently $\varepsilon_n = O_p(\sqrt{n})$, provided that $\mathcal{I}(\theta)$ is non-zero, since otherwise the variance is infinite.

Theorem 2.46 (Asymptotic normality of MLE)

Under the assumptions of the previous proposition, we have

$$\frac{1}{\sqrt{\mathcal{I}(\theta)}}Z \overset{d}{=} \sqrt{n}(\widehat{\theta}-\theta)$$

for Z standard Gaussian as $n \to \infty$, or equivalently that

$$\sqrt{n\mathcal{I}(\theta)}(\widehat{\theta}_n-\theta) \xrightarrow{d} \mathcal{N}(0,1).$$

We must also be wary that the above proof sketch does not work if $\mathcal{I}(\theta)=0$. Indeed, recall that consistency requires **identifiability** and some **smoothness**. For asymptotic normality, we need **consistency**, and further smoothness and **positive definitiveness** of the information matrix (linked to identifiability, see Rothenberg [23]) that $\mathcal{I}(\theta) > 0$. If the information is zero, then the rate goes down and we need to go further in the Taylor series expansion and for $(\widehat{\theta}-\theta)^2$, we need to multiply by n to get stabilization, but the rates will change. We have if $\mathcal{I}(\theta)$ is smooth map (analytic function here), as we have $\mathcal{I}(\boldsymbol{\theta}) = \mathsf{Cov}\left(\nabla\frac{\partial}{\partial\boldsymbol{\theta}}\log f_\theta(\boldsymbol{X})\right)$, we have positive semi-definitiveness. Then $\Gamma = \{\theta : \det\mathcal{I}(\theta)=0\}$ and Λ is nowhere dense sets. However, the proof requires differential topology, so the result will be skipped.

Imposing some prior, uniform on λ, and reparametrize $\theta = e^{-\lambda}$, then you impose exponential distribution. The information vehiculed by the prior depends quite heavily on the parametrization. A result from analysis shows that mapping from nowhere dense set to nowhere dense set and measure zero are meager.

Consider also if $\mathsf{E}(X_i)=\mu, \mathsf{Var}(X_i)=\sigma^2 \ \ \forall\, i$ where X_i are orthogonal.

[22] Warning! some things swept under the carpet, since we have $\widehat{\theta}$ and not θ_0

[23] Conditions include checking the rank of $\mathcal{I}(\theta)$ is locally constant

Then

$$\begin{aligned}\mathsf{P}\left(|\bar{X}_n - \mu| > \varepsilon_n\right) &\leq \frac{\mathsf{Var}\left(\bar{X}_n\right)}{\varepsilon_n^2} \\ &= \frac{\sigma^2}{n\varepsilon_n^2}\end{aligned}$$

where $\varepsilon_n = \delta_n/\sqrt{n}$ where δ_n is *e.g.* $\log n$ such that

$$\varepsilon_n = \frac{\delta_n}{\sqrt{n}} \to 0, \quad \delta_n \to \infty.$$

Then we have

$$\mathsf{P}\left(|\bar{X}_n - \mu| > \frac{\delta_n}{\sqrt{n}}\right) \leq \frac{\sigma^2}{n\left(\frac{\delta_n}{\sqrt{n}}\right)^2} = \frac{\sigma^2}{\delta_n^2} \to 0 \text{ as } n \to \infty$$

and thus $\mathsf{P}\left(\sqrt{n}|\bar{X}_n - \mu| > \delta_n\right) \to 0$. This implies that $\bar{X}_n - \mu = O_p(n^{-\frac{1}{2}})$. Indeed, by definition, $\forall\, \varepsilon > 0, \exists C_\varepsilon$ and N_ε mean

$$\sup_{n \geq N_\varepsilon} \mathsf{P}\left(\sqrt{n}|\bar{X}_n - \mu| > C_\varepsilon\right) < \varepsilon$$

Then $\varepsilon = \frac{\sigma^2}{\delta_n^2} = \frac{\sigma^2}{(\log(n)^2}$ where if $\delta_n = \log(n)$, then $\log(n) = \frac{\sigma}{\sqrt{\varepsilon}}$ and $n = e^{\frac{\sigma}{\sqrt{\varepsilon}}}$ then $N_\varepsilon \geq e^{\frac{\sigma}{\sqrt{\varepsilon}}}$ with $C_\varepsilon = \log(n) = \frac{\sigma}{\sqrt{\varepsilon}}$.

Back to the case where information is zero. We have

$$\varepsilon_n = (\widehat{\theta}_n - \theta)^2 \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \log f_\theta(X_i)$$

If we have

$$\left|\frac{\partial^3}{\partial \theta^3} \log f_\theta(X)\right| \leq M(X)$$

where $\mathsf{E}\left(M(X)\right) < \infty$. Then

$$\begin{aligned}\varepsilon_n &= n(\widehat{\theta} - \theta)^2 \mathsf{E}\left(M(X)\right) \\ &= n\left(\frac{1}{\sqrt{n}}\right)^2 \mathsf{E}\left(M(X)\right) \\ &= \mathsf{E}\left(M(X)\right).\end{aligned}$$

Another way (simpler) to prove it would be to expand the score function around θ_0, and

then to evaluate the equation θ at $\widehat{\theta}$; this allows to use WLLN on the latter term, without requiring smoothness.

To relax the assumptions of the above theorem, one can show only consistency, $f \in \mathcal{C}_1$ and $\mathcal{I}(\boldsymbol{\theta}) > 0$. Using Morse's lemma would allow to approximate the likelihood by quadratic functions. Relaxing the continuity assumption, we cannot perform the Taylor series expansion anymore. Since $\mathcal{I}(\boldsymbol{\theta}) = \mathsf{Cov}\left(\nabla \log f_\theta\right)$ requires only the first derivative. Using Fréchet derivative and embeding onto a unit ball of a Hilbert space, if the map is Fréchet differentiable (guaranteed by consistency), then establishing that the score function is Normal, then show that appropriately normalized and standardized, $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta})$ minus the score goes to zero in probability. Using Slutski theorem, we can then show that the two have the same asymptotic distribution. Why are relaxing these conditions useful? If we have for example Laplace distribution, we do not have differentiable everywhere, but almost everywhere, so this allows generalizations. For more, see Ibragimov and Has'minskii (1981) or Bickel, Klaassen, Ritov, and Wellner (1993).

For more on the FCR bound when the dimension space is not finite, see U. Grenander (1981) in *Abstract Inference*, defined for normed linear space, using the Bochner integral.

Theorem 2.47

Suppose $X_1, \ldots, X_n$ are iid and satisfy the assumptions of Le Cam's theorem 2.44 and (c) and (d) are replaced by the corresponding assumptions on the 3^{rd} (rather than the 2^{nd}) derivative, thais is by the existence of a 3^{rd} derivative satisfying

$$\left| \frac{\partial^3}{\partial \theta^3} \log f_\theta x) \right| \leq M(x)$$

for all $x \in A = \{x : f_\theta(x) > 0\}$ for $\theta \in (\theta_0 - c, \theta_0 + c)$ with $\mathsf{E}\left(M(X)\right) < \infty$. Then, any consistent sequence $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$ of roots of the likelihood equation satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\theta))$$

2.10 Invariance property of MLE

We will show that if $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then for any estimand h, $h(\widehat{\boldsymbol{\theta}})$ is the MLE of $h(\boldsymbol{\theta})$. The property of $\widehat{\theta}$ is that of the estimate and not the estimator, since maximum likelihood is a post-experimental approach. If the sample is small in comparison with the dimension of $\boldsymbol{\theta}$, then UMVUE or other pre-experimental approaches in frequentist statistics will be proposed.

Before we present Zehna's theorem, here is the foreword of the original article that illustrate the need to develop results for maps that are not injective.

> "One of the distinguishing features of the method of maximum likelihood in

statistical estimation is the fact that it enjoys a certain invariance property. likelihood estimator for $u(\widehat{\theta})$ where u is some function of $u(\theta)$. Some textbooks on the subject avoid any explicit mention of properties that u must possess in order for invariance to hold. When a proof of the property is given, it is at least assumed, either explicitly or implicitly that u is 1-1 thereby defining a unique inverse.

Now if the assumption that u be 1-1 is really necessary, then the invariance principle could not be invoked to find the maximum likelihood estimator for even as common a case as the variance, $p(1-p)$, of a Bernoulli random variable.

Indeed, there may be some doubt as to the meaning of maximum likelihood in such a case. The purpose of this note is to point out that the notion of a maximum likelihood estimator for $u(\theta)$ when u is not 1-1 can and should be made explicit.

The method used for accomplishing this task has the desirable feature that it coincides with the usual method employed when u is 1-1.

Theorem 2.48 (Zehna's theorem (1969))
Let $\{f_\theta : \theta \in \Theta\}$ be a family of PDF's (PMF's) and let $\mathcal{L}(\theta)$ be the likelihood function. Suppose that $\Theta \subset \mathbb{R}^k$, for $k \geq 1$. [24] Let $h : \Theta \to \Lambda$ be an (arbitrary) mapping of Θ onto Λ. If $\hat{\theta}$ is an MLE of θ, then $h(\hat{\theta})$ is an MLE of $h(\theta)$.

This result does not hold in infinite dimensions. See Søren Johansen on this.

Proof For each $\lambda \in \Lambda$, let us define

$$\Theta_\lambda = \{\theta : \theta \in \Theta, h(\theta) = \lambda\} \text{ and } M(\lambda; \boldsymbol{x}) = \sup_{\theta \in \Theta_\lambda} \mathcal{L}(\theta, \boldsymbol{x})$$

Then M defined on Λ is called the likelihood function induced by h.

If $\hat{\theta}$ is any MLE of θ, then $\hat{\theta}$ belongs to one and only one set $\Theta_{\hat{\lambda}}$ say. [25] Since $\hat{\theta} \in \Theta_{\hat{\lambda}}, \hat{\lambda} = h(\hat{\theta})$. Now

$$M(\hat{\lambda}, \boldsymbol{x}) = \sup_{\theta \in \Theta_{\hat{\lambda}}} \mathcal{L}(\theta, \boldsymbol{x}) \geq \mathcal{L}(\hat{\theta}, \boldsymbol{x}) \tag{2.22}$$

On the other hand

$$M(\hat{\lambda}, \boldsymbol{x}) \leq \sup_{\lambda \in \Lambda} M(\lambda, \boldsymbol{x}) = \sup_{\theta \in \Theta} \mathcal{L}(\theta, \boldsymbol{x}) = \mathcal{L}(\hat{\theta}, \boldsymbol{x}) \tag{2.23}$$

[24] This precision is important as we usually don't have this property in ∞-dimension
[25] Indeed, if $\theta \in \Theta_{\lambda_1} \cap \Theta_{\lambda_2}$ and so $h(\theta) = \lambda_1, h(\theta) = \lambda_2$, which contradicts the partition for our function h.

Thus, using (2.22) and (2.23),

$$M(\hat{\lambda}, \boldsymbol{x}) = \sup_{\lambda \in \Lambda} M(\lambda, \boldsymbol{x})$$

It then follows that $\hat{\lambda} \leq h(\hat{\theta})$ is the MLE of $h(\theta)$. ■

Section 3
Computational statistics

3.1 Newton-Raphson algorithm

We present two methods here, the Newton-Raphson iterative process and Expectation-Maximization (EM) algorithm. We want to find the root of the function

$$l(\theta) \coloneqq \frac{\partial}{\partial\theta}\log\mathcal{L}(\theta;\boldsymbol{x}) = 0$$

using a Taylor series expansion

$$l'(\widehat{\theta}) \approx l'(\tilde{\theta}) + (\widehat{\theta}-\tilde{\theta})l''(\tilde{\theta})$$

which yields

$$\widehat{\theta} = \tilde{\theta} - \frac{l'(\tilde{\theta})}{l''(\tilde{\theta})}$$

or in the multivariate setting

$$\widehat{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}} - \nabla l(\tilde{\theta})\boldsymbol{H}^{-1}(\tilde{\boldsymbol{\theta}})$$

and applying this procedure iteratively yields a fixed point $x_{n+1} = f(x_n)$.

Theorem 3.1 (Newton-Raphson algorithm)

Suppose that the assumptions of the theorem of asymptotic normality (Theorem 2.46) hold and $\tilde{\theta}_n$ is $\sqrt{n}$-consistent estimator of θ, *i.e.* $\sqrt{n}(\tilde{\theta}_n - \theta) = O_p(1)$. Then

$$\delta_n = \tilde{\theta}_n - \frac{l'(\tilde{\theta}_n)}{l''(\tilde{\theta}_n)}$$

is asymptotically efficient, that is

$$\sqrt{n}(\delta_n - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\theta)).$$

Remark

If the function is concave of convex, this is good. If we have mixtures, the Newton-Raphson or EM algorithms will give local maxima. Other approaches, like simulated annealing (see Ingher (1993)), work well for global maximization.

If we have missing data (such as censoring where you observe $X_i = T_i \wedge C_i$ where T_i is follow-up and C_i a competing variable. Instead, we observe (X_i, δ_i) where $\delta_i = \mathbf{1}_{T_i \leq C_i}$. We lose longer survivors, so we are underestimate values; throwing the data would induce bias. We can use EM to augment the incomplete data with the complete data we which to have, for which the likelihood is easier to maximize. We augment the data, get a simple likelihood to maximize. One then expectate over the unknown values. Then maximize and we repeat

the procedure. One can show that it converges to the MLE. This is also useful for mixtures; if we have $f_\theta(x) = \sum_{j=1}^k p_j g_j(x; \boldsymbol{\theta}_j)$ from a mixture of Normal populations, where p_j are the proportion. If we knew for observations x_1, x_2, x_3 which population they arise from (say respectively $x_1 \mapsto k^{\text{th}}$ population, $x_2 \mapsto 3^{\text{rd}}$ and $x_3 \mapsto 4^{\text{rd}}$; then the likelihood would be much easier to work with (being some $g_k(x_1, \boldsymbol{\theta}_k) g_3(x_2, \boldsymbol{\theta}_3) g_4(x_3, \boldsymbol{\theta}_4)$, rather than the product of three mixtures $f_{\boldsymbol{\theta}}(x_1) f_{\boldsymbol{\theta}}(x_2) f_{\boldsymbol{\theta}}(x_3)$.

3.2 Expectation-Maximization (EM) algorithm

The EM algorithm was designed for missing data cases. Consider the case of a simple mixture with two distribution, with $X \sim f(x) = pg(x) + (1-p)h(x)$. In this case, identifiability, maximization and invariance is difficult. If we observed from which distribution the sample observations were drawn, it would be then easy to augment the data as (X_i, δ_i) and write down the likelihood. The EM algorithm we will see is guaranteed to converge to a solution of the MLE equations, but it may not be the global maximum.

Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be bivariate data with (X_i, Y_i) for $i = 1, \dots, n$ with $X_i \sim \mathcal{P}(\tau_i)$ and $Y_i \sim \mathcal{P}(\beta\tau_i)$ independent between and within pairs (X_i, Y_i). Then

$$\mathcal{L}_c(\beta, \boldsymbol{\tau}; (\boldsymbol{x}, \boldsymbol{y})) = \prod_{i=1}^n f_{\beta,\boldsymbol{\tau}}(x_i, y_i) = \prod_{i=1}^n \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!} \frac{e^{-\tau_i}\tau_i^{x_i}}{x_i!}$$

and maximization of the log-likelihood yields

$$\widehat{\beta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}, \qquad \tau_i = \frac{x_i + y_i}{\widehat{\beta} + 1}$$

Suppose X_1 is missing. Then we have

$$\sum_{x_i=0}^{\infty} \prod_{i=1}^n f_{\beta,\boldsymbol{\tau}}(x_i, y_i)$$

The incomplete data consists of $\{y_1, (x_2, y_2), \dots, (x_n, y_n)\}$ while the complete data comprises of all pairs $\{(x_i, y_i), i = 1, \dots, n\}$. The likelihood for the incomplete data is

$$\mathcal{L}(\beta, \boldsymbol{\tau}; y_1, (x_2, y_2), \dots, (x_n, y_n)) = \left[\prod_{i=1}^n \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!}\right] \left[\prod_{i=2}^n \frac{e^{-\tau_i}\tau_i^{x_i}}{x_i!}\right]$$

and the corresponding MLE equations are

$$\widehat{\beta} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n \widehat{\tau}_i}, \qquad y_1 = \widehat{\beta}\widehat{\tau}_1, \qquad x_i + y_i = \widehat{\tau}_i(\widehat{\beta} + 1) \text{ for } i = 2, 3, \dots$$

We have for $\boldsymbol{\theta} = (\beta, \boldsymbol{\tau})$ the likelihoods $\mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}) = f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$ and $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = g(\boldsymbol{y}|\boldsymbol{\theta})$ where here $\boldsymbol{y} = \{y_1, (x_j, y_j) \text{ for } j = 2, \ldots, n\}$ and $\boldsymbol{x} = x_1$. Let

$$k(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}) = \frac{f(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta})}{g(\boldsymbol{y}; \boldsymbol{\theta})}$$

and consider the log-likelihood

$$\log(g(\boldsymbol{y}; \boldsymbol{\theta})) = \log(f(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta})) - \log(k(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}))$$

and so

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}) - \log(k(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}))$$

and we have

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) \approx \mathsf{E}_{X|Y=y,\boldsymbol{\theta}'}\left(\ell(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X})\right) - \mathsf{E}_{X|Y=y,\boldsymbol{\theta}'}\left(\log(k(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}))\right)$$

and θ^{r+1} is the value that maximizes

$$\mathsf{E}_{X|Y=y,\boldsymbol{\theta}^r}\left(\ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})\right) = \int_{\boldsymbol{x}} \ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) k(\boldsymbol{x}|\boldsymbol{\theta}', \boldsymbol{y})\, \mathrm{d}x$$
$$\mathsf{E}_{X|Y=y,\boldsymbol{\theta}'}\left(\log(k(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}))\right) = \int_{\boldsymbol{x}} \log(k(\boldsymbol{x}|\boldsymbol{\theta}', \boldsymbol{y})) k(\boldsymbol{x}|\boldsymbol{\theta}', \boldsymbol{y})\, \mathrm{d}x$$

taking $\boldsymbol{\theta} = \theta^{(r)}$ and $\boldsymbol{\theta}' = \boldsymbol{\theta}^{(r+1)}$. The choice of a $\boldsymbol{\theta}^{(0)}$ is a difficult one if we do not have unimodal distribution and concavity.

At each step, $\boldsymbol{\theta}^{(r+1)}$ is the value that maximizes

$$\mathsf{E}_{X|Y=y,\boldsymbol{\theta}^{(r)}}\left(\ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})\right) = \arg\max_{\boldsymbol{\theta}} \int_{\mathcal{X}} \ell_c(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}) k(\boldsymbol{x}|\boldsymbol{\theta}^{(r)}, \boldsymbol{y})\, \mathrm{d}x$$

Back to the toy example, we have for the E step

$$\mathsf{E}_{X_1|\boldsymbol{y},\beta^{(r)},\boldsymbol{\tau}^{(r)}}\left(\log \mathcal{L}(\beta, \boldsymbol{\tau}; \boldsymbol{x}, \boldsymbol{y})\right) = \sum_{x_1} \log\left(\prod_{i=1}^{n} \frac{e^{-\beta\tau_i}(\beta\tau_i)^{y_i}}{y_i!} \frac{e^{-\tau_i}\tau_i^{x_i}}{x_i!}\right) \frac{e^{-\tau_1^{(r)}} {\tau_i^{(r)}}^{x_1}}{x_1!}$$

Simplification yields

$$
\begin{aligned}
&= \sum_{i=1}^{n}[-\beta\tau_i + y_i(\log(\beta) + \log(\tau_i) - \log(y_i!)] + \sum_{i=2}^{n}[-\tau_i + x_i\log(\tau_i) - \log(x_i!)] \\
&\quad + \sum_{x_1=0}^{\infty}[-\tau_1 + x_1\log(\tau_1) - \log(x_1!)]\frac{e^{\tau_1^{(r)}}(\tau_1^{(r)})^{x_1}}{x_1!} \\
&\quad = \left(\sum_{i=1}^{n}[-\beta\tau_i + y_i(\log(\beta) + \log(\tau_i)] + \sum_{i=2}^{n}[-\tau_i + x_i\log(\tau_i)] + \underline{\sum_{x_1=0}^{\infty}[-\tau_1 + x_1\log(\tau_1)]\frac{e^{\tau_1^{(r)}}(\tau_1^{(r)})^{x_1}}{x_1!}}\right) \\
&\quad - \left(\sum_{i=1}^{n}\log(y_i!) + \sum_{i=2}^{n}\log(x_i!) + \sum_{x_1=0}^{\infty}\log(x_1!)\frac{e^{\tau_1^{(r)}}(\tau_1^{(r)})^{x_1}}{x_1!}\right)
\end{aligned}
$$

where we can further simplify the underlined term as

$$
-\tau_1 + \log(\tau_1)\sum_{x_1=0}^{\infty} x_1\frac{e^{\tau_1^{(r)}}(\tau_1^{(r)})^{x_1}}{x_1!} = -\tau_1 + \log(\tau_1)\tau_1^{(r)}
$$

Substituting back in the equation and this completes the E step. For the M step, we can maximize to get the sequence

$$
\begin{aligned}
\widehat{\beta}^{(r+1)} &= \frac{\sum_{i=1}^{n} y_i}{\tau_1^{(r)} + \sum_{i=2}^{n} x_i}, \\
\widehat{\tau}_1^{(r+1)} &= \frac{\widehat{\tau}_1^{(r)} + y_1}{\widehat{\beta}^{(r+1)} + 1} \\
\widehat{\tau}_i^{(r+1)} &= \frac{x_i + y_i}{\widehat{\beta}^{(r+1)} + 1}, \qquad \text{for } i = 2, \ldots, n
\end{aligned}
$$

The heart of the algorithm is to build a sequence of θ such that the likelihood increases. If we have the incomplete data log-likelihood $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ and the complete data one $\ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x})$, then using the simple conditional probability distribution.

$$
\log \ell(\boldsymbol{\theta}|\boldsymbol{y}) = \log \ell(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) - \log(K(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y}))
$$

and set

$$
\log \ell(\boldsymbol{\theta}|\boldsymbol{y}) = \mathsf{E}_{\boldsymbol{X}}\left(\log \ell(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})\right) - \mathsf{E}_{X}\left(\log K(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{y})\right)
$$

and now

$$
\mathsf{E}_{\boldsymbol{X}}\left(\log \ell(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y})\right) = \int_{\mathcal{X}} \log \ell\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})K(x|\boldsymbol{\theta}^{(r)}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{x}
$$

We then have

$$\begin{aligned}
&\log(\ell(\boldsymbol{\theta}^{(r+1)}|\boldsymbol{y})) \\
&\quad = \int \log(\ell(\boldsymbol{\theta}^{(r+1)}|\boldsymbol{x},\boldsymbol{y}))K(\boldsymbol{x}|\boldsymbol{\theta}^{(r+1)},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x} - \int \log(K(\boldsymbol{x}|\boldsymbol{\theta}^{(r+1)},\boldsymbol{y}))K(\boldsymbol{x}|\boldsymbol{\theta}^{(r+1)},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x} \\
&\quad > \ell(\boldsymbol{\theta}^{(r)}|\boldsymbol{y}) \\
&\quad = \int_{\mathcal{X}} \log(\ell(\boldsymbol{\theta}^{(r)}|\boldsymbol{x},\boldsymbol{y}))K(\boldsymbol{x}|\boldsymbol{\theta}^{(r)},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x} - \int_{\mathcal{X}} \log(K(\boldsymbol{x}|\boldsymbol{\theta}^{(r)},\boldsymbol{y}))K(\boldsymbol{x}|\boldsymbol{\theta}^{(r)},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}
\end{aligned}$$

and since the first term is the minimizer, and we have using Jensen's inequality the second part is $\mathsf{E}_g\left(\log\frac{f}{g}\right) < 0$. Thus, the EM algorithm converges to a local maximum.

3.3 Presentation on non-parametric MLE

We can use kernel density estimation of the form

$$\widehat{f}_n(t) = \lim_{h\to 0}\int \frac{1}{n}K\left(\frac{x-t}{n}\right)\,\mathrm{d}\widehat{F}_n(x)$$

with *e.g.* $K(U) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$. This can be used to assess multimodality. The Komolos-Major-Tusnady (KMT) result gives tools to determine the best choice of K and the optimal bandwith and optimal choice of rate $h_n = \log(n)/n$.

3.4 Jackknife and bootstrap

Jackknife started with Quenouille in mid 50's for bias reduction purposes. Suppose calculate a statistic from an n-sample, $T_n = T(X_1,\ldots,X_n)$ and $\mathsf{Bias}\,(T_n) = \mathsf{E}\,(T_n) - \theta$. The original suggestion was to look at $T_{(-i)} = T(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)$, which consists in removing the i^{th} observation. One may then look at the mean obtained from the subsample

$$\bar{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{(-i)} \quad \text{ and } b_{\text{jack}} = (n-1)(\bar{T}_n - T_n)$$

and the corresponding bias-corrected estimator $T_{\text{jack}} = T_n - b_{\text{jack}}$.

Efron started from $\bar{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{(-i)}$, which allows for more copies of the sample . Extending this idea, we coud remove multiple copies. Having $\theta = T(F)$, if $\mu = \mathsf{E}\,(X) = \int_{\mathcal{X}} x\,\mathrm{d}F = T(F)$, and $\sigma^2 = \mathsf{Var}\,(X) = \mathsf{E}\left(X^2\right) - \left(\mathsf{E}\,(X)\right)^2$. We however have dependence in the data, so the draws are all not interesting. Since most parameters are of the form of integrals, we can use $\widehat{\theta} = T(\widehat{F})$ and resampling using the ECDF allows for independent draws conditional on the observed covariates. For variance computation, this is important, but not for bias which relies only on the first moment. We can resample $X_1,\ldots,X_n$ from the discrete distribution,

with $x_1, \ldots, x_n$.

For a large class of estimators,

$$\mathsf{Bias}\left(T_n\right) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right)$$

where for example for the variance plug-in estimator

$$\begin{aligned}
\widehat{\sigma^2}_n &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n-1}{n}S^2 \\
\mathsf{E}\left(\widehat{\sigma^2}_n\right) &= \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{\sigma^2}{n} \\
\mathsf{Bias}\left(\widehat{\sigma^2}_n\right) &= -\frac{\sigma^2}{n}
\end{aligned}$$

so here $a = -\sigma^2, b = 0$.

We have

$$\mathsf{Bias}\left(T_{(-i)}\right) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{n^3}\right) \tag{3.24}$$

Now

$$\mathsf{Bias}\left(\bar{T}_n\right) = \mathsf{Bias}\left(\frac{1}{n}\sum_{i=1}^{n} T_{(-i)}\right)$$

also fulfills (3.24) and $b_{\text{jack}} = (n-1)(\bar{T}_n - T_n) = (\bar{T}_n - \theta) - (T_n - \theta)$ adding and substracting θ to have an expression for the bias. Now

$$\begin{aligned}
\mathsf{E}\left(b_{\text{jack}}\right) &= (n-1)\left[\mathsf{E}\left(\mathsf{Bias}\left(\bar{T}_n\right)\right) - \mathsf{E}\left(\mathsf{Bias}\left(T_n\right)\right)\right] \\
&= (n-1)\left[\left(\frac{1}{n-1} - \frac{1}{n}\right)a + \left(\frac{1}{(n-1)^2} - \frac{1}{n^2}\right)b + O\left(\frac{1}{n^3}\right)\right] \\
&= \frac{a}{n} + \frac{2n-1}{n^2(n-1)} + O\left(\frac{1}{n^2}\right) \\
&= \mathsf{Bias}\left(T_n\right) + O\left(\frac{1}{n^2}\right)
\end{aligned}$$

and

$$\mathsf{Bias}\left(T_n\right) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right)$$

and b_{jack} estimates the bias of T_n up to order $O\left(\frac{1}{n^2}\right)$. Similarly, we can show that

$\mathsf{Bias}\left(T_{\text{jack}}\right) = T_n - b_{\text{jack}}$ and get

$$\mathsf{Bias}\left(T_{\text{jack}}\right) = -\frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n^2}\right)$$

Definition 3.2 (Pseudo-values)
Let $\tilde{T}_i = nT_n - (n-1)T_{(-i)}$ are pseudo-observations. Then, we define

$$T_{\text{jack}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{T}_i$$

The jackknife estimate of $\mathsf{Var}\left(F_n\right)$ is

$$v_{\text{jack}} = \frac{\tilde{s}^2}{n} \quad \text{where} \quad \tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{T}_i - T_{\text{jack}})^2$$

Theorem 3.3
Let $\mu = \mathsf{E}\left(X_1\right), \sigma^2 = \mathsf{Var}\left(X_1\right)$ and suppose $T_n g(\bar{X}_n)$ where g has a continuous, non zero derivative at μ. Then $v_{\text{jack}}/\sigma_n^2 \xrightarrow{\text{a.s.}} 1$ where

$$\sigma_n^2 = \frac{1}{n}\left(g'(\mu)\right)^2\sigma^2$$

and

$$\frac{T_n - g(\mu)}{\sigma_n} \xrightarrow{w} \mathcal{N}(0,1)$$

Theorem 3.4 (Inconsistency of the jackknife for quantiles (Efron, 1982))
If $T(F) = F^{-1}(p)$, the p^{th} quantile, then the jacknife estimate is inconsistent. For the median $(p = 0.5)$, we have

$$\frac{v_{\text{jack}}}{\sigma_n^2} \xrightarrow[w]{} \left(\frac{\chi_2^2}{2}\right)^2$$

where σ_n^2 is the asymptotic variance of the sample median.

Recall the following result for the distribution of the difference of quantiles and order statistic

Theorem 3.5
If $X_{(r)}$ denotes the r^{th} order statistic of a sample of size n $X_1, \ldots, X_n$ with PDF F. For $0 < p < 1$, let F be absolutely continuous with PDF f, which is positive at $F^{-1}(p)$ and

is continuous at that point (that is $F(z_p) = p$ exists and is unique at p). For $r = np$, as $n \to \infty$,

$$\sqrt{n} f(z_p) \frac{(X_{(r)} - z_p)}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

If we have a statistical function $\mathcal{T} : \{F : F \in \mathcal{F}\} \to \mathbb{R}^k$. We could take length bias sampling and find a one-to-one map from the sample population estimate distribution function to the target population. We have

$$f_B(x) = \frac{x f_U(x)}{\mu_U}, \qquad \mu_U = \int_x x f(x) \, \mathrm{d}x$$

then

$$\frac{f_U(x)}{\mu_U} = \frac{f_B(x)}{x}$$

and integrating both sides,

$$\frac{1}{\mu_U} = \int_x \frac{f_B(x)}{x}$$

and as such

$$\begin{aligned} \mu_U &= \left(\int_0^\infty \frac{dF_B(x)}{x} \right)^{-1} \\ \Rightarrow f_U(x) &= \frac{\frac{f_B(x)}{x}}{\int_0^\infty \frac{dF_B(x)}{x}} \\ \Rightarrow F_U(x) &= \frac{\frac{\int_0^x \mathrm{d}f_B(t)}{t}}{\int_0^\infty \frac{dF_B(x)}{x}} \end{aligned}$$

If this map is not linear, we need something akin to Taylor series expansion; this requires different types of derivatives.

Definition 3.6 (Fréchet derivative (F-differentiable))

If we have $\mathbb{B}$ Banach spaces (normed complete linear spaces), of infinite dimensions, and $f : \mathbb{B}_1 \to \mathbb{B}_2$, we can define the Fréchet derivative as

$$\| f(x + h) - f(x) - \mathcal{L}_{f,x}(h) \| = o(\|h\|)$$

as $n \to \infty$ where $\mathcal{L}$ is a linear map (like the tangent to the curve f at point x). Recall the

"usual " derivative

$$\frac{f(x+h)-f(x)}{x} \approx f'(x) \qquad \Rightarrow \qquad |f(x+h)-f(x)-f'(x)h| \to 0$$

and in infinite dimension, we have this as the inner product and the above is a hyperplane tangent to the curve; this is simply the gradient. This should be linear as we move to higher dimensions.

This condition is too strong. We could have the directional derivative instead,

Definition 3.7 (Gâteau derivative (G-differentiability))

$$\left\| \frac{f(x+th)-f(x)}{t} - \mathcal{L}^G_{f,x}(h) \right\| = o(1) \text{ as } t \to \infty$$

Definition 3.8 (Hadamard derivative (H-differentiability))
We have

$$\sup_{h \in \mathcal{K}} \left\| \frac{f(x+th)-f(x)}{t} - \mathcal{L}^H_{f,x} \right\| = o(1)$$

for any compact set $\mathcal{K}$. [26]

This then allows to define the notion of an **influence function**, as

Definition 3.9 (Influence function)

$$\Phi_F(x) = \mathcal{L}^G_F(\delta_x - F)$$

where δ_x is the point map at x, degenerate CDF at point x and $\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathcal{E}(t - X_i)$.

If we consider a Taylor expansion using the Gâteau derivative, we have

$$\sqrt{n}\left((\widehat{F}_n) - T(F)\right) = \mathcal{L}^G_F\left(\sqrt{n}\left(\widehat{F}_n - F\right)\right) + o(1)$$

and since $\mathcal{L}$ is a linear map,

$$T(\widehat{F}_n) = T(F) + \mathcal{L}^G(\widehat{F}_n - F) + o\left(n^{-\frac{1}{2}}\right)$$

[26] The superscript denotes the type of derivative, the first subscript the function at which we approximate, the second subscript the point and the argument inside parenthesis the direction.

or

$$\sqrt{n}\left[T(F+\frac{1}{\sqrt{n}}(\sqrt{n}(\widehat{F}_n-F)))-T(F)\right]-\mathcal{L}_F^G(\sqrt{n}(\widehat{F}_n-F))\to 0$$

using the Gâteau derivative with $t=\sqrt{n}^{-1}$. Then

$$\widehat{F}_n(t)=\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i\in(-\infty,t)}=\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$$

and then since we have a linear map, we interchange summations and $\mathcal{L}^G$

$$\mathcal{L}_F^G(\sqrt{n}(\widehat{F}_n-F))=\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathcal{L}_F^G(\delta_{X_i}-F)=\frac{1}{\sqrt{n}}\sum_{i=1}^n \Phi_F(X_i);$$

the resulting for any functional gives a Normal approximation.

We can show under mild conditions if $\mathsf{E}\left(\Phi_F(X_i)\right)=0$ and $\sigma_F^2=\mathsf{E}\left(\Phi_F(X_i)\right)^2<\infty$ then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \Phi_F(X_i)\xrightarrow{w}\mathcal{N}(0,\sigma_F^2)$$

If we have a parameter of interest, $\theta=T(F)$, then with $T_n=T(\widehat{F}_n)$ where $\tilde{T}_i=T(\delta_{X_i})$; then

$$\frac{T(F+\varepsilon(\delta_{X_i}-F)-T(F)}{\varepsilon}-\mathcal{L}_F^G(\delta_X-F)\to 0.$$

Choosing $\varepsilon=-(n-1)^{-1}$ and replace F by $\widehat{F}_n$, we get

$$-(n-1)\left[T\left(\widehat{F}_n-(n-1)^{-1}(\delta_{X_i}-\widehat{F}_n)\right)-T(\widehat{F}_n)\right]-\mathcal{L}_F^G(\delta_{X_i}-\widehat{F}_n)$$

or

$$\frac{T\left(\frac{n}{n-1}\widehat{F}_n-\frac{1}{n-1}\delta_{X_i}\right)-T(\widehat{F}_n)}{-1/(n-1)}-\Phi_F(X_i) \tag{3.25}$$

and if we open up, we get

$$\frac{1}{n-1}\sum_{j=1}^n \delta_{X_j}-\frac{1}{n-1}\delta_{X_i}=\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^n \delta_{X_j}$$

and substituting this back up in (3.25), we getting

$$-(n-1)\left[T\left(\frac{1}{n-1}\sum_{\substack{j=1\\j\neq i}}^{n}\delta_{X_j}\right)-T(\widehat{F}_n)\right]=(n-1)[T_n-T_{(-i)}]$$

and then

$$\tilde{T}_i=nT_n-(n-1)T_{(-i)}=T_n+(n-1)(T_n-T_{(-i)})=T_n+\Phi_{\widehat{F}_n}(X_i);$$

we have the approximation

$$T(\delta_{X_i})\approx T(\widehat{F}_n)+\Phi_{\widehat{F}_n}(X_i)$$

3.5 Bootstrap

We can replace F by $\widehat{F}_n$; to see why it works. If we work with MLE or NPMLE, we have $n^\alpha(\widehat{\theta}-\theta)$ converges weakly to a Gaussian process. However, we need $\mathsf{Var}_F(T_n)$, we can replace with $\mathsf{Var}_{\widehat{F}_n}(T_n)$. The general procedure for bootstrap is as follows:

Algorithm 3.1 (Nonparametric Bootstrap)
Assume we have IID samples,

1. Draw $X_1^*,\ldots,X_n^*\sim\widehat{F}_n$ with replacement.
2. Compute $T_n^*=g(X_1^*,\ldots,X_n^*)$
3. Repeat step 1 and 2 B times and set $T_{n,1}^*,\ldots,T_{n,B}^*$
4. Let

$$v_{\text{boot}}=\frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^*-\frac{1}{B}\sum_{r=1}^{B}T_{n,r}^*\right)^2$$

We have $\|\widehat{F}_n-F\|_\infty=O\left(\frac{\log(\log(n))}{n}\right)$ and so this means that as the empirical distribution function, ECDF, converges to the CDF.

The order is respectively $O(n^{-\frac{1}{2}})$ and $O(B^{-\frac{1}{2}})$. We also have

$$\mathsf{Var}_F(T_n)\approx\mathsf{Var}_{\widehat{F}_n}(T_n)\approx v_{\text{boot}}.$$

We want $F^*-F=(F^*-\widehat{F}_n)+(\widehat{F}_n-F)$. If both rates are $\sqrt[3]{n}$, we have balance for the first term, but the second will blow up and bootstrap blows up. If we take faster convergence,

Figure 5: Bootstrap scheme

Real world: From the unknown probability model

$$\mathsf{P}(\boldsymbol{x}) \xrightarrow{\text{random sampling}} \boldsymbol{x} = (x_1, \ldots, x_n)^\top \xrightarrow{\text{original estimator}} \widehat{T_n} = g(\boldsymbol{x})$$

and $\sqrt{n}(T_n - \widehat{T}_n) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(\mu_{\mathsf{P}}, \sigma^2_{\mathsf{P}}\right)$

Bootstrap world: From the estimated probability model

$$\widehat{\mathsf{P}}(\boldsymbol{x}) \xrightarrow{\text{random sampling}} \boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)^\top \xrightarrow{\text{bootstrap estimator}} \widehat{T}_n^* = g(\boldsymbol{x}^*)$$

and $\sqrt{n}(\widehat{T}_n^* - \widehat{T}_n) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(\mu_B(\mathsf{P}), \sigma^2_B(\mathsf{P})\right)$;

we have convergence to zero. This is the case with Gernander estimators (estimation under shape constraint).

The boostrap can be used to approximate the CDF of a statisitc T_n. Let $G_n(t) = \mathsf{P}\left(T_n \leq t\right)$, the CDF of T_n. The bootstrap approximation to G_n is

$$\widehat{G}_n(t) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{T^*_{n,b} \leq t}$$

Note

We are using plugin estimators, namely if $\theta = T(F)$, then $\widehat{\theta} = T(\widehat{F}_n)$. When B is large, using $B - 1$ (adding the original observation to get precisely B samples), or this does not make much difference.

Parametric bootstrap

Instead of replacing the CDF by the nonparametric MLE, we replace for $X_1, \ldots, X_n \sim f_\theta$ and use the parametric model and resample from $f_{\hat{\theta}}$. For example, if $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $\widehat{F}_n$ is $\mathcal{N}(\bar{X}_n, \frac{n-1}{n}S^2)$ where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.

Confidence interval for bootstrap

Pivotal quantities. A function of unknown parameters and observables such that the distribution of the resulting is completely known.

Example 3.1

For example, if $X_i \overset{\text{iid}}{\sim} F_{\boldsymbol{\theta}}$, then for $\boldsymbol{x}$, $T(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^n F_{\boldsymbol{\theta}}(X_i)$. To see this, take logarithm,

we have

$$-\log T(\boldsymbol{X},\boldsymbol{\theta}) = -\sum_{i=1}^{n}\log F_{\boldsymbol{\theta}}(X_i)$$

Since $X_i \sim F_\theta$ for continuous random variables X's, then by the probability integral transform, we have $F_{\boldsymbol{\theta}}(X_i) \sim \mathcal{U}(0,1)$. Then, if $w_i = -\log F_{\boldsymbol{\theta}}(x_i) \sim \mathcal{E}(1)$ and since any measurable function of independent random variables are also independent, $\sum_{i=1}^{n} W_i \sim \Gamma(n,1)$.

Sometimes, the distribution of the pivotal quantity is known, but of no use for confidence interval because the form is complicated.

Example 3.2

If $X_i \sim \mathcal{E}(\theta)$, with $f_X(x;\theta) = \theta e^{-}\mathbf{1}_{x\geq 0}$ with $F_\theta(x) = 1 - e^{-\theta x}$ and $-\log(T(\boldsymbol{X},\theta) = -\sum_{i=1}^{n}\log(1-e^{-\theta x_i})$. If we work instead with $V(\boldsymbol{X},\theta) = \prod_{i=1}^{n} s_\theta(x_i)$, the survival function, we have $s_\theta(x) = 1 - F_\theta(x)$ and so we have $-\log(V(\boldsymbol{X},\theta)) \sim \Gamma(n,1)$ and since we have $Z \sim \mathcal{U}(0,1)$ implies $1 - Z \sim \mathcal{U}(0,1)$, then

$$-\log(V(\boldsymbol{X},\theta)) = -\sum_{i=1}^{n}\log(e^{-\theta x_i}) = \theta\sum_{i=1}^{n} X_i$$

For the confidence interval, $\mathsf{P}\left(a < \theta\sum_{i=1}^{n} x_i < b\right) = 1-\alpha$ and then confidence interval is

$$\left(\frac{a}{\sum_{i=1}^{n} x_i}, \frac{b}{\sum_{i=1}^{n} x_i}\right)$$

We could also resort to asymptotic, where

$$\sqrt{n\mathcal{I}(\theta)}(\widehat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(0,1)$$

is not useful in applications since $\mathcal{I}(\theta)$ is dependent on the unknown parameter. However, if $\mathcal{I}$ is continuous, then we have using WLLN and Slutski, we can use

$$\sqrt{n\mathcal{I}(\widehat{\theta})}(\widehat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(0,1)$$

If we do not have a parametric model, we need to go to GEE and find the limiting variance. In nonparametric, we can resort for reasonable sample sizes to bootstrap confidence interval. If $X_i \overset{\text{iid}}{\sim} F$, then $T_n \pm \mathfrak{z}_{\alpha/2}\widehat{\text{se}}_{\text{boot}}$ where $\widehat{\text{se}}_{\text{boot}} = \sqrt{v_{\text{boot}}}$ is the simplest bootstrap confidence interval, based on the normal approximation. The approximation follows from

$$\sqrt{n}\left(T(\widehat{F}_n) - T(F)\right) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Phi_{T_F}(X_i)$$

IF $\theta = T(F)$, and we use the plugin estimator $T(\widehat{F}_n)$, if we know $\widehat{\theta}_n - \theta$ then we can easily construct confidence intervals. Let $H(r) = \mathsf{P}_F(R_n \leq r)$ and $C_n^* = (a, b)$ where $a = \widehat{\theta}_n - H^{-1}(1 - \frac{\alpha}{2})$ and $b = \widehat{\theta}_n - H^{-1}(\frac{\alpha}{2})$, then

$$\begin{aligned}
\mathsf{P}(a \leq \theta \leq b) &= \mathsf{P}\left(\widehat{\theta}_n - b \leq R_n \leq \widehat{\theta}_n - a\right) \\
&= H(\widehat{\theta}_n - a) - H(\widehat{\theta}_n - b) \\
&= H\left(H^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha
\end{aligned}$$

If we do not have H, we can use $\widehat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{R_{n,b}^* \leq r}$ and $R_{n,b}^* = \widehat{\theta}_{n,}^* - \widehat{\theta}_n$ where $\widehat{\theta}_n$ is calculated from the original data. This yields

$$\begin{aligned}
\widehat{a} &= \widehat{\theta}_n - \widehat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) \\
\widehat{b} &= \widehat{\theta}_n - \widehat{H}^{-1}\left(\frac{\alpha}{2}\right)
\end{aligned}$$

and $C_n = (\widehat{a}, \widehat{b})$. We have the following theorem

Theorem 3.10

If $T(F)$ is Hadamard differentiable, then

$$\mathsf{P}_F(T(F) \in C_n) \to 1 - \alpha.$$

A third approach to confidence interval is to consider $z_n = \frac{T_n - \theta}{\widehat{\text{se}}_{\text{boot}}}$ and $z_{n,b}^* = \frac{T_{n,b}^* - T_n}{\widehat{\text{se}}_b^*}$ where $\widehat{\text{se}}_b^*$ is an estimate of the standard errors of $T_{n,b}^*$ (**not** T_n).

Using Edgeworth expansions, we can show that this estimate, which used double bootstrap, is more effective, but more costly. We end with two theorems, without proof, regarding the use of the bootstrap.

Theorem 3.11

Suppose that $\mathsf{E}(X_i^2) < \infty$. Let $T_n = g(\bar{X}_n)$ where g is continuously differentiable at $\mu = \mathsf{E}(X_i)$ and $g'(\mu) \neq 0$. Then

$$\sup_u \left| \mathsf{P}_{\widehat{F}_n}\left(\sqrt{n}\left[T(\widehat{F}_n^*) - T(\widehat{F}_n)\right] \leq u\right) - \mathsf{P}_F\left(\sqrt{n}\left[T(\widehat{F}_n) - T(F)\right] \leq u\right) \right| \xrightarrow{\text{a.s.}} 0 \text{ as } n \to \infty \tag{3.26}$$

The proof rely on Mallow's metric or Berry-Esséen theorem. See Wasserman for the proof.

Theorem 3.12

Suppsoe that $T(F)$ is Hadamard differentiable with respect to $d(F, G) = \sup_x |F(x) - G(x)|$

and $0 < \int \mathcal{L}_F^2(x)\,\mathrm{d}F(x) < \infty$. Then (3.26) holds in probability.

Section 4
Hypothesis test

4.1 Generalized Likelihood Ratio tests

In Neyman-Pearson, we need a device to look at how far the observations are from the hypothesis. Let $X_1, \ldots, X_n \sim f_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \in \Theta$. Fisher, on the other hand, would not specify an "alternative" hypothesis; the null should be the *status quo* (we thus want to control the Type 1 error.[27] The general idea of testing is to look at the distance between the hypothesised and the observed. We have

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \qquad H_A : \boldsymbol{\theta} \in \Theta \setminus \Theta_0$$

We thus want that $\mathsf{P}_{H_0}\left(() \text{ Rejecting } H_0\right) = \mathsf{P}\left(\text{Type I error}\right) < \alpha$.

		True	
		H_0	H_1
Accepted	H_0	Correct	Type II error
	H_1	Type I error	Correct

Table 1: Decision table for hypothesis test errors

Let $\varphi : \mathcal{X}^n \mapsto [0,1]$. Then, for each $\boldsymbol{x} = (X_1, \ldots, X_n)$, we have $\varphi(\boldsymbol{x})$ a value in $[0,1]$ with which we reject H_0. Then

$$\mathsf{E}_{\boldsymbol{\theta}}\left(\varphi(\boldsymbol{X})\right) = \beta_\varphi(\boldsymbol{\theta})$$

If $\boldsymbol{\theta} \in \Theta_0$, then $\beta_\varphi(\boldsymbol{\theta}) = \mathsf{P}_{\boldsymbol{\theta}}\left(\text{rejecting } H_0\right)$. and

$$\sup_{\boldsymbol{\theta} \in \Theta_0} \beta_{[\varphi]}(\boldsymbol{\theta}) < \alpha$$

where $\beta_\varphi(\boldsymbol{\theta})$ is called **power function**

- If $\boldsymbol{\theta} \in \Theta$, then $\beta_\varphi(\boldsymbol{\theta})$ represents rejecting H_0 when H_0 is correct.
- If $\boldsymbol{\theta} \in \Theta_A$, then $\beta_\varphi(\boldsymbol{\theta})$ represents rejecting H_0 when H_0 is false.

[27] Send no innocent to gaz chamber

Definition 4.1 (Most powerful test)

A test φ is called **most powerful at level** α for testing $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_A : \boldsymbol{\theta}_1 \in \Theta_A$ if

$$\sup_{\boldsymbol{\theta} \in \Theta_0} \beta_\varphi(\boldsymbol{\theta}) \leq \alpha$$

and

$$\beta_\varphi(\boldsymbol{\theta}_1) \geq \beta_{\varphi^*}(\boldsymbol{\theta}_1)$$

for any size α, φ^* that is

$$\sup_{\beta \in \Theta_0} \beta_{\varphi^*}(\beta) \leq \alpha$$

This $_1$ is one point in the alternative. Now, if

$$\beta_\varphi(\boldsymbol{\theta}) \geq \beta_{\varphi^*}(\boldsymbol{\theta}), \qquad \forall\, \boldsymbol{\theta} \in \Theta_A,\ \forall\, \varphi^*$$

such that $\sup_{\beta \in \Theta_0} \beta_{\varphi^*}(\beta) \leq \alpha$, we say that φ is UMP test for testing $H_0 : \boldsymbol{\theta} \in \Theta_0$, versus $H_A : \boldsymbol{\theta} \in \Theta_A$. If $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $H_A : \boldsymbol{\theta} = \boldsymbol{\theta}_1$, then

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{f_{\boldsymbol{\theta}_1}(\boldsymbol{x})}{f_{\boldsymbol{\theta}_0}(\boldsymbol{x})} > k \\ \gamma & \text{if } \frac{f_{\boldsymbol{\theta}_1}(\boldsymbol{x})}{f_{\boldsymbol{\theta}_0}(\boldsymbol{x})} = k \\ 0 & \text{if } \frac{f_{\boldsymbol{\theta}_1}(\boldsymbol{x})}{f_{\boldsymbol{\theta}_0}(\boldsymbol{x})} < k \end{cases}$$

is MP for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the alternative $H_A : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ at its level, *i.e.* at $\mathsf{E}_{\boldsymbol{\theta}_0}(\varphi(\boldsymbol{X}))$

We need this k as we do not treat the null and the alternative equally. The bigger the k, the smaller the rejection set (region); it indicates that we assign more weight to the null and not give up. This is for classification.

In most applications, α, the probability of Type 1 error, is given. Now

$$\alpha = E[\boldsymbol{\theta}_0]\varphi(\boldsymbol{X}) = \mathsf{P}\left(\left\{\boldsymbol{X} : \frac{f_{\boldsymbol{\theta}_1}(\boldsymbol{X})}{f_{\boldsymbol{\theta}_0}(\boldsymbol{X})} > k\right\} + \gamma\left\{\boldsymbol{X} : \frac{f_{\boldsymbol{\theta}_1}(\boldsymbol{X})}{f_{\boldsymbol{\theta}_0}(\boldsymbol{X})} = k\right\}\right)$$

This is for one parameter; we need a particular form for the function, an alternative. Otherwise, its a headaque.

Fisher had a different philosophy; He did not have an alternative, and proposed instead to look at the ratio, with this time the ratio with the null on top. Choosing a criterion that measures the "distance" between the facts and the hypothesis; using the MLE, he introduced

the so-called **generalized likelihood ratio test** where

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\sup_{\boldsymbol{\theta}\in\Theta} f_{\boldsymbol{\theta}}(\boldsymbol{x})}$$

Now if $H : \boldsymbol{\theta} \in \Theta_0$, large values are good for the hypothesis. We expect that as $n \to \infty$, we get the ratio to be 1. Now we have

$$\{\boldsymbol{x} : \lambda(\boldsymbol{x}) < c\}$$

The small values of c are not good for the null hypothesis. Fisher used this successfully in few cases, as the following:

Example 4.1
Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and suppose that we wish to test

$$\begin{cases} H_0 & \mu = \mu_0 \\ H_1 & \mu \neq \mu_0 \end{cases};$$

where

$$\Theta_0 = \{\boldsymbol{\theta} = (\mu_0, \sigma^2) : \sigma^2 \in \mathbb{R}^+\}$$
$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$$

and notice that both are composite hypothesis (since σ^2 is unspecified). Here, $\Theta_0 = \{\boldsymbol{\theta} = (\mu_0, \sigma^2) : \sigma^2 \in \mathbb{R}^+$

We present the calculations step by step for finding the ratio of MLE.

Step 1: Calculate the likelihood function:

$$\begin{aligned}\mathcal{L}(\theta = (\mu, \sigma^2), \boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}\end{aligned}$$

Step 2: Find the supremum on Θ:

$$\sup_{\theta\in\Theta} \mathcal{L}(\theta, \boldsymbol{x}) = \mathcal{L}((\bar{x}, \widehat{\sigma^2}_{\text{ML}}), \boldsymbol{x})$$

where

$$\widehat{\sigma^2}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = (\sqrt{2\pi}\widehat{\sigma})^{-n} e^{-\frac{n}{2}}.$$

and $\widehat{\mu} = \bar{X}_n$

Step 3: Find the supremum on Θ_0:

$$\sup_{\theta \in \Theta_0} \mathcal{L}(\theta, \boldsymbol{x}) = \mathcal{L}((\mu, \widehat{\sigma^2}_*), \boldsymbol{x})$$

where

$$\widehat{\sigma^2}_* = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Step 4: We are now set to calculate $\lambda(\boldsymbol{x})$:

$$\begin{aligned}
\lambda(\boldsymbol{x}) &= \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta, \boldsymbol{x})}{\sup_{\theta \in \Theta} \mathcal{L}(\theta, \boldsymbol{x})} \\
&= \frac{\mathcal{L}((\mu, \widehat{\sigma}^2_*), \boldsymbol{x})}{\mathcal{L}((\mu, \widehat{\sigma}^2_{\mathrm{ML}}, \boldsymbol{x})} \\
&= \frac{\left(\sqrt{2\pi}\widehat{\sigma}_*\right)^{-n} \exp\left\{-\frac{1}{2\widehat{\sigma}^2_*} \sum_1^n (x_i - \mu_0)^2\right\}}{\left(\sqrt{2\pi}\widehat{\sigma}_{\mathrm{ML}}\right)^{-n} \exp\left\{-\frac{1}{2\widehat{\sigma}^2_{\mathrm{ML}}} \sum_1^n (x_i - \bar{x}_0)^2\right\}}
\end{aligned}$$

and noticing that $\widehat{\sigma^2}_* \sum_{i=1}^n (x_i - \mu_0)^2 = n$ and similarly for the denominator, we have

$$\begin{aligned}
\lambda(\boldsymbol{x}) &= \left(\frac{\widehat{\sigma}_{\mathrm{ML}}}{\widehat{\sigma}_*}\right)^n \frac{\exp\left(-\frac{n}{2}\right)}{\exp\left(-\frac{n}{2}\right)} \\
&= \left(\frac{\widehat{\sigma}_{\mathrm{ML}}}{\widehat{\sigma}_*}\right)^n \\
&= \left(\frac{\widehat{\sigma}^2_{\mathrm{ML}}}{\widehat{\sigma}^2_*}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \mu_0)^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}\right)^{\frac{n}{2}} \\
&= \left(1 + \left(\frac{1}{n-1}\right) \frac{n(\bar{x} - \mu_0)^2}{\left(\frac{1}{n-1}\right) \sum_{i=1}^n (x_i - \bar{x})^2}\right)^{-\frac{n}{2}}
\end{aligned}$$

which is like a t distribution. Let

$$F = \frac{n(\bar{x} - \mu_0)^2}{\left(\frac{1}{n-1}\right) \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

is a ratio of $\chi^2(1)/\chi^2(n-1)$. The above then reduces to

$$\lambda(\boldsymbol{x}) = \left(\frac{1}{1 + \frac{F}{n-1}}\right)^{-\frac{n}{2}}$$

Note that $\lambda(\boldsymbol{x})$ is a decreasing function of F. Moreover,

$$\{\lambda(\boldsymbol{x}) \leq c\} \Leftrightarrow \{F(1, n-1) \geq k\}$$

and so this is similar to the Fisher-Neyman UMP test formulation.

For a given α, we find k such that $\alpha = \mathsf{P}\left(\mathcal{F}(1, n-1) \geq k\right)$. You mind want to refresh your memory here about the derivation of the F statistic. If $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \sigma^2)$ under H_0, then $\bar{X}_n \sim \mathcal{N}(\mu_0, \sigma^2/n)$

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \Rightarrow \qquad \frac{n(\overline{X} - \mu_0)^2}{\sigma^2} \sim \chi^2(1).$$

and also

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{X}_n)^2/\sigma^2 = \frac{(n-1)S^2}{(n-1)\sigma^2} = \frac{\chi^2(n-1)}{n-1};$$

since $\bar{X}_n \perp\!\!\!\perp \{X_i - \bar{X}_n\}$ is partially ancillary for μ; this can also be derived through MGF argument.

Heuristic from this come from the part that $X_i = \mu + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, we can express $X_i \equiv \bar{X}_n + (X_i - \bar{X}_n)$ and as $n \to \infty$, the second term goes to μ.

Thus the ratio is $F(m,n) = \frac{V/m}{W/n}$ provided $V \perp\!\!\!\perp W$, and that $V \sim \chi^2(m), W \sim \chi^2(n)$. In our specific case, the distribution is $1/(n-1)\mathcal{F}(1, n-1)$ and we can choose k such that $\mathsf{P}_{H_0}\left(V(\boldsymbol{X}) > k\right) = \alpha$.

Indeed, we have We begin with

Theorem 4.2

Let $X_1, \ldots, X_n$ be iid rv $\mathcal{N}(\mu, \sigma^2)$ rv's, then $\overline{X}$ and $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$ are independent.

Proof We use the joint MGF $M_\zeta(t, t_1, \ldots, t_n)$ for $\zeta \equiv \{\overline{X}, X_1 - \overline{X}, \ldots, X_n - \overline{X}\}$.

$$\begin{aligned} M_\zeta(t, t_1, \ldots, t_n) &= \mathsf{E}\left[\exp\left\{t\overline{X} + \sum_{i=1}^n t_i(X_i - \overline{X})\right\}\right] \\ &= \mathsf{E}\left[\exp\left\{\sum_{i=1}^n t_i X_i - \left(\sum_{i=1}^n t_i - t\right)\overline{X}\right\}\right] \\ &= \mathsf{E}\left[\exp\left\{\sum_{i=1}^n X_i\left(t_i - \frac{t_1 + \ldots + t_n - t}{n}\right)\right\}\right] \end{aligned}$$

Let

$$\bar{t} = \frac{1}{n}\sum_{i=1}^n t_i$$

Continuing from above, we have

$$\begin{aligned} M_\zeta(t, t_1, \ldots, t_n) &= \mathsf{E}\left[\prod_{i=1}^n \exp\left\{\frac{X_i(nt_i - n\bar{t} + t)}{n}\right\}\right] \\ &= \prod_{i=1}^n \mathsf{E}\left[\exp\left\{\frac{X_i\left[t + n(t_i - \bar{t})\right]}{n}\right\}\right] \end{aligned}$$

where the last step follows from independence. Using the fact that $X_i \sim \mathcal{N}(\mu, \sigma^2)$, and adopting the convention

$$t^* = \frac{t + n(t_i - \bar{t})}{n},$$

we obtain

$$\begin{aligned} &\prod_{i=1}^n \exp\left\{\mu t^* + \frac{\sigma^2}{2}t^{*^2}\right\} \\ &= \exp\left\{\frac{\mu}{n}\left[nt + n\sum_{i=1}^n (t_i - \bar{t})\right] + \frac{\sigma^2}{2n^2}\sum_{i=1}^n [t + n(t_i - \bar{t})]^2\right\} \\ &= \exp\left\{\mu t + \frac{\sigma^2}{2n}t^2\right\} \cdot \exp\left\{\frac{\sigma^2}{2}\sum_{i=1}^n (t_i - \bar{t})^2\right\} \\ &= M_{\overline{X}}(t) \cdot M_{X_1 - \overline{X}, \ldots, X_n - \overline{X}}(t_1, \ldots, t_n) \end{aligned}$$

which in turn implies that $\overline{X}$ and $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$ are independent. ■

It follows, since for two vectors of independent variables, any measurable functions of those vectors are independent. Thus, $\overline{X}$ and S^2 are independent.

Corollary 4.3

The ratio of the sample variance statistic with the true variance given by

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

follows a Chi-square distribution with $n-1$ degrees of freedom.

Proof $X_i \sim \mathcal{N}(\mu, \sigma^2)$ thus $\frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1)$ hence

$$\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1).$$

The X_i are independent so

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$$

Now, adding and subtracting $\overline{X}$:

$$\begin{aligned}\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 &= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma}\right)^2 + n\frac{(\overline{X} - \mu)^2}{\sigma^2} \\ &= \underbrace{\frac{(n-1)S^2}{\sigma^2}}_{V} + \underbrace{\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2}_{W}\end{aligned}$$

since the cross terms vanishes (that is $\sum(X_i - \overline{X}) = 0$).

$$\begin{aligned} M_U(t) &= M_V(t) \cdot M_W(t) \\ (1-2t)^{-\frac{n}{2}} &= M_V(t) \cdot (1-2t)^{-\frac{1}{2}} \\ \Rightarrow \qquad M_V(t) &= (1-2t)^{-\frac{n-1}{2}}, \quad t < \frac{1}{2}\end{aligned}$$

Therefore $V \sim \chi^2(n-1)$. Why does this hold? Note here that if

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

as

$$\begin{aligned}\mathsf{E}[e^{t_1 v + t_2 w}] &= \mathsf{E}[e^{t_1 v} e^{t_2 w}] \\ &= \mathsf{E}[e^{t_1 v}]\mathsf{E}[e^{t_2 w}]\end{aligned}$$

since $V \perp\!\!\!\perp W$. ■

Corollary 4.4

The scaled difference between the arithmetic mean and the unknown μ given by

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

follows a Student's t distribution.

Proof Given

$$\sqrt{n}\frac{\overline{X}-\mu}{S} \sim \mathcal{N}(0,1)$$

and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

recall that $\overline{X}$ is independent from S^2, thus

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{(\overline{X}-\mu)/\frac{\sigma}{\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} = \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \sim t(n-1)$$

■

Corollary 4.5

If X_i are iid and $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and Y_j are iid with $Y_j \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and $X_i \perp\!\!\!\perp Y_j$, then

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \mathcal{F}(m-1, n-1).$$

Proof

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\left[(m-1)S_1^2/\sigma_1^2\right]/(m-1)}{\left[(n-1)S_2^2/\sigma_2^2\right]/(n-1)} \sim \frac{\frac{\chi^2(m-1)}{m-1}}{\frac{\chi^2(n-1)}{n-1}} \sim \mathcal{F}(m-1, n-1).$$

■

This was the good side of the story, we may not always be that lucky. Under regularity conditions, we can show

$$-2\log(\lambda(\boldsymbol{X})) \xrightarrow{d} \chi^2_{\dim(\boldsymbol{\theta})-\dim(\beta_0)}$$

as $n \to \infty$. If a UMP test exists, in many cases, this will boil to this test. To show this

result, see Vaan der Vaart.

Consider the case with $H_0 : \mu \in (a, b)$ where $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2)$. To define the linear dimensions appropriately for this type of problem. One can resort to simulation if the problem is hard to compute.

Next time, we discuss Neyman-Pearson and goodness-of-fit test.

4.2 Neyman-Pearson lemma

Theorem 4.6 (Neyman-Pearson fundamental lemma)

Let $(\Theta_0, \Theta_1, \alpha)$ denote the class of all size α tests for testing $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_A : \boldsymbol{\theta} \in \Theta$, where $\sup_{\boldsymbol{\theta} \in \Theta_0} \mathsf{E}_{\boldsymbol{\theta}} (\varphi(\boldsymbol{X})) \leq \alpha$. The simplest possible scenario, $\Theta_0 = \{\theta_0\}$ versus $\Theta_1 = \{\theta_1\}$. This comes up often in classification and clustering (in learning).

1. Any test φ of the form

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} > k \\ \gamma(\boldsymbol{x}) & \text{if } \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} = k \\ 0 & \text{if } \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} < k \end{cases} \tag{4.27}$$

   for some $k \geq 0$ and $0 \leq \gamma(\boldsymbol{x}) \leq 1$ is the most powerful of its size for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$.

2. For any given $\alpha \in [0, 1]$, there exists a test of form (4.27) which is MP size α.

Proof We first begin with part a. Let $\varphi : \mathcal{X} \to [0, 1]$ be a Neyman-Pearson test. Let φ^* be any test in $(\Theta_0, \Theta_A, \alpha)$ where $\alpha = \mathsf{E}_{\boldsymbol{\theta}_0} (\varphi^*(\boldsymbol{X}))$. Thus, $\mathsf{E}_{\theta_0} (\varphi^*(\boldsymbol{X})) \leq \mathsf{E}_{\theta_0} (\varphi(X))$. [28] Recall that $\varphi(x) = 1$ if $f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) > k$ is the probability of rejection.

We have

$$\begin{aligned} I &= \int_{\boldsymbol{x}} [\varphi(\boldsymbol{x}) - \varphi^*(\boldsymbol{x})][f_1(\boldsymbol{x}) - k f_0(\boldsymbol{x})] \, \mathrm{d}\boldsymbol{x} \\ &= \left(\int_{\boldsymbol{x}: f_1(\boldsymbol{x}) > k f_0(\boldsymbol{x})} + \int_{\boldsymbol{x}: f_1(\boldsymbol{x}) < k f_0(\boldsymbol{x})} \right) [\varphi(\boldsymbol{x}) - \varphi^*(\boldsymbol{x})][f_1(\boldsymbol{x}) - k f_0(\boldsymbol{x})] \, \mathrm{d}\boldsymbol{x} \end{aligned}$$

For any $x \in \{\boldsymbol{x} : f_1(\boldsymbol{x}) > k f_0(\boldsymbol{x})\}$, $\varphi(\boldsymbol{x}) - \varphi^*(\boldsymbol{x}) = 1 - \varphi(\boldsymbol{x}) \geq 0$ so that the integrand I is non-negative. For $\boldsymbol{x} \in \{\boldsymbol{x} : f_1(\boldsymbol{x}) < k f_0(\boldsymbol{x})\}$, $\varphi(\boldsymbol{x}) - \varphi^*(\boldsymbol{x}) = -\varphi(\boldsymbol{x}) \leq 0$ so that the

[28] If the distributions are absolutely continuous, the set where equality holds has Lebesgue measure zero, so we can ignore the middle term of (4.27)

integrand is again ≥ 0. It follows that

$$I = \mathsf{E}_{\theta_1}\left(\varphi(\boldsymbol{X})\right) - \mathsf{E}_{\theta_1}\left(\varphi^*(\boldsymbol{X})\right) - k\left(\mathsf{E}_{\theta_0}\left(\varphi(\boldsymbol{X})\right) - \mathsf{E}_{\theta_0}\left(\varphi^*(\boldsymbol{X})\right)\right) \geq 0$$

which in turns implies that $\mathsf{E}_{\theta_1}\left(\varphi(\boldsymbol{X})\right) - \mathsf{E}_{\theta_1}\left(\varphi^*(\boldsymbol{X})\right) \geq k\mathsf{E}_{\theta_0}\left(\varphi(\boldsymbol{X})\right) - \mathsf{E}_{\theta_0}\left(\varphi^*(\boldsymbol{X})\right) \geq 0$ and therefore $\varphi(\boldsymbol{x})$ is more powerful as $\mathsf{E}_{\theta_1}\left(\varphi(\boldsymbol{X})\right) \geq \mathsf{E}_{\theta_1}\left(\varphi^*(\boldsymbol{X})\right)$, *i.e.* $\beta_\varphi(\boldsymbol{\theta}_1) \geq \beta_{\varphi^*}(\boldsymbol{\theta}_1)\ \ \forall$

If $k = \infty$, any test φ^* of size 0 must vanish on the set $\{\boldsymbol{x} : f_0(\boldsymbol{x}) > 0\}$. We have

$$\mathsf{E}_{\theta_1}\left(\varphi(\boldsymbol{X})\right) - \mathsf{E}_{\theta_1}\left(\varphi^*(\boldsymbol{X})\right) = \int_{\boldsymbol{x}: f_0(\boldsymbol{x})=0} [1 - \varphi^*(\boldsymbol{x})] f_1(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \geq 0$$

■

Remark

We look at tests of the same size (that is, they are optimal over tests of same size) as φ. If $k = \infty$, the size of $\mathsf{E}_{\theta_0}(\varphi) = 0$ so the term vanishes. Think of it and write the definition of $\mathsf{E}_{\theta_0}\left(\varphi^*(\boldsymbol{X})\right) = \int_{\boldsymbol{x}} \varphi^*(\boldsymbol{x}) f_0(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 0$ since $\varphi^*(\boldsymbol{x}) \geq 0, f_0(\boldsymbol{x}) \geq 0$ almost surely under P_{θ_0}. We can easily extend this result to more than two categories (finite number). See Thomas Ferguson (1967), *Mathematical Statistics: a decision theoretic approach* in the exercises.

We now prove the second part of the Neyman-Pearson lemma.

Proof We confine ourselves to the case $0 < \alpha \leq 1$. Let $\gamma(\boldsymbol{x}) = \gamma$. The size of a test of form (1) is

$$\begin{aligned}\mathsf{E}_{\theta_0}[\varphi(\boldsymbol{X})] &= \mathsf{P}_{\theta_0}\left(f_1(\boldsymbol{X}) > k f_0(\boldsymbol{X})\right) + \gamma \mathsf{P}_{\theta_0}\left(f_1(\boldsymbol{X}) = k f_0(\boldsymbol{X})\right) \\ &= 1 - \mathsf{P}_{\theta_0}\left(f_1(\boldsymbol{X}) \leq k f_0(\boldsymbol{X})\right) + \gamma \mathsf{P}_{\theta_0}\left(f_1(\boldsymbol{X}) = k f_0(\boldsymbol{X})\right)\end{aligned}$$

In the computation of size, this distinction (ratio as opposed to $f_1(\boldsymbol{X}) = k f_0(\boldsymbol{X})$; indeed, [29]

$$\mathsf{P}_{\theta_0}\left(\boldsymbol{X} : f_0(\boldsymbol{X}) = 0\right) = 0 = \int_{\boldsymbol{x}: f_0(\boldsymbol{x})=0} f_0(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 0$$

[29] This is true since for $A = \{\boldsymbol{x} : f_0(\boldsymbol{x}) = 0\}$, then $\mathsf{P}_{\theta_0}(A) = \int_A f_0(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 0$.

We need to find k and γ such that

$$\begin{aligned} & \mathsf{P}_{\theta_0}\left(\boldsymbol{X} : f_1(\boldsymbol{X}) > k f_0(\boldsymbol{X})\right) + \gamma \mathsf{P}_{\theta_0}\left(\boldsymbol{X} : f_1(\boldsymbol{X}) = k f_0(\boldsymbol{X})\right) = \alpha \\ \Leftrightarrow \quad & \mathsf{P}_{\theta_0}\left(\boldsymbol{X} : \frac{f_1(\boldsymbol{X})}{f_0(\boldsymbol{X})} > k\right) + \gamma \mathsf{P}_{\theta_0}\left(\boldsymbol{X} : f_1(\boldsymbol{X}) = k f_0(\boldsymbol{X})\right) = \alpha \\ \Leftrightarrow \quad & \mathsf{P}_{\theta_0}\left(\boldsymbol{X} : \frac{f_1(\boldsymbol{X})}{f_0(\boldsymbol{X})} \leq k\right) - \gamma \mathsf{P}_{\theta_0}\left(\boldsymbol{X} : f_1(\boldsymbol{X}) = k f_0(\boldsymbol{X})\right) = 1 - \alpha \end{aligned} \tag{4.28}$$

If there exists a k_α such that

$$\mathsf{P}_{\theta_0}\left(\boldsymbol{X} : \frac{f_1(\boldsymbol{X})}{f_0(\boldsymbol{X})} \leq k_\alpha\right) = 1 - \alpha$$

Then, we choose $k = k_\alpha$ and $\gamma = 0$. Otherwise, there exists a k_α (there is a jump at k_α)

$$\mathsf{P}_{\theta_0}\left(\boldsymbol{X} : \frac{f_1(\boldsymbol{X})}{f_0(\boldsymbol{X})} < k_\alpha\right) < 1 - \alpha < \mathsf{P}_{\theta_0}\left(\boldsymbol{X} : \frac{f_1(\boldsymbol{X})}{f_0(\boldsymbol{X})} \leq k_\alpha\right)$$

Then choose $k = k_\alpha$ and choose γ solving (4.28) for γ, and

$$\gamma = \frac{\mathsf{P}_{\theta_0}\left(\boldsymbol{X} : \frac{f_1(\boldsymbol{X})}{f_0(\boldsymbol{X})} \leq k\right) - (1 - \alpha)}{\mathsf{P}_{\theta_0}\left(\boldsymbol{X} : \frac{f_1(\boldsymbol{X})}{f_0(\boldsymbol{X})} = k_\alpha\right)}.$$

■

4.3 Goodness of fit tests

In the context of regression, we have the hypothesis $H_0 : X_1, \ldots, X_n \sim \mathcal{N}(0, 1)$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. We may rely on the nonparametric MLE, looking at the Kolmogorov distance given by

$$\sup_{\boldsymbol{x}} \left| \widehat{F}_n(\boldsymbol{x}) - \Phi(\boldsymbol{x}) \right|$$

where

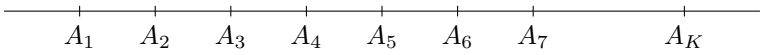
$$\lambda = \|\widehat{F}_n - F_0\|_\infty$$

and we have $H_0 : X_i \overset{\text{iid}}{\sim} F_0$ completely known. We can show the so-called **Kolmogorov-Smirnov test** with

$$\sqrt{n}(\widehat{F}_n(\boldsymbol{x}) - F(\boldsymbol{x})) \xrightarrow{d} \mathfrak{B}(F)$$

where $\mathfrak{B}(F)$ is a Brownian bridge. One could use other metric here; using the $\mathcal{L}^2$ for example leads to the Cramer Von-Mise test. For censored data, we need a large number of observations.

We could adopt another approach for a sample $X_1, \dots, X_n \sim F$ completely known by categorizing the observations.



We then look at $O_j \coloneqq \#$ of X_i's in A_j and ϵ_j the expected number of observations in A_j, equal to $n\mathsf{P}_{H_0}(A_j)$ and

$$\epsilon_j = n\mathsf{P}_{H_0}(A_j) = n\int_{A_j} \mathrm{d}F(\boldsymbol{x})$$

which is a Binomial test, which can also be extended to multinomial case. Looking at the quantity

$$\sum_{j=1}^{K} \frac{(O_j - \epsilon_j)^2}{\epsilon_j} \xrightarrow{d} \chi^2_{K-1}$$

as $n \to \infty$. This result is based on the Central Limit theorem and the Poisson approximation, using independence.

The number of bins and the size of the bins was discussed by Mann and Wald (1948), in *On the choice of the number of class intervals in the application of the chi-square* in AMS.

If $H_0 : X_i \overset{\text{iid}}{\sim} F_{\boldsymbol{\theta}}, \boldsymbol{\theta} = (\theta_1, \dots, \theta_r)$, and the size is large (with $n > r$), we can approximate the expected value by MLE. Lehmann and Chernoff in 195, in the Annals of Mathematical Statistics, *The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit* remarked that the likelihood

$$\mathcal{L} = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(x_i)$$

where $f_{\boldsymbol{\theta}}$ is the pdf of $F_{\boldsymbol{\theta}}$ is **not** the correct likelihood. The correct likelihood is the

multinomial likelihood, that is

$$\mathcal{L}_2 = \prod_{j=1}^{K} p_j(\boldsymbol{\theta})^{O_j}$$

where $p_j(\boldsymbol{\theta}) = \int_{A_j} \mathrm{d}F_{\boldsymbol{\theta}}(\boldsymbol{x})$ and

$$\sum_{j=1}^{k} \left(\frac{O_j - \widehat{\epsilon}_j}{\sqrt{\widehat{\epsilon}_j}} \right)^2 \xrightarrow{d} \chi^2_{K-r-1}$$

as $n \to \infty$ where $\widehat{\varepsilon}_j = n\mathsf{P}_{\widehat{\boldsymbol{\theta}}}(A_j)$.

Index

License