§5 Estimation

§5.1 Estimator

5.1.1 $\boldsymbol{X}$: observed dataset

θ: unknown parameter

$\psi(\theta)$: target for estimation, for some specified function ψ of θ

Point estimation of $\psi(\theta)$: selection of a "good" value to *estimate* $\psi(\theta)$ based on $\boldsymbol{X}$.

5.1.2 **Definition.** An *estimator* of $\psi(\theta)$ is a statistic $T = T(\boldsymbol{X})$ for estimating $\psi(\theta)$.

5.1.3 $T(\boldsymbol{X})$ is random (varies from sample to sample) and has its own sampling distribution, based on which the quality of $T(\boldsymbol{X})$ is assessed in the frequentist sense.

5.1.4 Treating $T(\boldsymbol{X})$ as a "decision rule", we may assess its quality by its risk function

$$R(\theta, T) = \mathbb{E}_\theta\big[L(\theta, T(\boldsymbol{X}))\big],$$

where $L(\theta, a)$ defines the "loss" resulting from estimating $\psi(\theta)$ by the "action" (estimate) a.

§5.2 Bias and mean squared error

5.2.1 **Definition.** Let T be an estimator of $\psi(\theta)$. Then the *bias* of T is:

$$\mathrm{bias}_\theta(T) = \mathbb{E}_\theta[T] - \psi(\theta).$$

If T has zero bias, it is *unbiased*.

5.2.2 Bias of T measures its *accuracy*.

For $T \in \mathbb{R}$, $\mathrm{Var}_\theta(T)$ (or *standard deviation* of T, $\sqrt{\mathrm{Var}_\theta(T)}$) measures its *precision*.

A good estimator should be both *accurate* and *precise*.

5.2.3 If we set the loss function to be $L(\theta, a) = \|\psi(\theta) - a\|_2^2$, the risk function $R(\theta, T)$ reduces to the *mean squared error*.

Definition. An estimator $T = (T_1, \ldots, T_d)^\top$ of $\psi(\theta) \in \mathbb{R}^d$ has the *mean squared error* (MSE)

$$\begin{aligned}
\mathrm{MSE}_\theta(T) &= \mathbb{E}_\theta\big[\|T - \psi(\theta)\|_2^2\big] \\
&= \mathbb{E}_\theta\big[\|T - \mathbb{E}_\theta[T]\|_2^2\big] + 2\,\mathbb{E}_\theta\big[(T - \mathbb{E}_\theta[T])^\top(\mathbb{E}_\theta[T] - \psi(\theta))\big] + \big\|\mathbb{E}_\theta[T] - \psi(\theta)\big\|_2^2 \\
&= \sum_{j=1}^d \mathrm{Var}_\theta(T_j) + \sum_{j=1}^d \{\mathrm{bias}_\theta(T_j)\}^2 = \sum_{j=1}^d \mathrm{MSE}_\theta(T_j).
\end{aligned}$$

MSE provides a measure of the quality of an estimator by taking into account both accuracy (bias) and precision (variance).

Small MSE $\Rightarrow$ sampling distribution of T highly concentrated near $\psi(\theta)$.

§5.3 Rao-Blackwell Theorem

5.3.1 **Lemma.** Let $T = T(\boldsymbol{X})$ be complete sufficient for θ. Then, for any function $\psi(\theta)$ of θ, there exists at most one unique function of T which is unbiased for $\psi(\theta)$.

Proof:

Suppose $g(T)$, $h(T)$ are unbiased estimators of $\psi(\theta)$ such that

$$\mathbb{E}_\theta\big[\,g(T)\big] = \mathbb{E}_\theta\big[\,h(T)\big] = \psi(\theta) \quad \forall\,\theta.$$

Then $\quad \mathbb{E}_\theta\big[\,g(T) - h(T)\big] = 0 \quad \forall\,\theta \quad \Longrightarrow \quad \mathbb{P}_\theta\big(g(T) = h(T)\big) = 1.$ ∎

5.3.2 $\boldsymbol{X} \sim f(\cdot|\theta)$

Target for estimation: $\psi(\theta)$

Loss function: $L(\theta, a)$ [= loss incurred when $\psi(\theta)$ is estimated by a]

Theorem. *(Rao-Blackwell)*

Suppose that $T(\boldsymbol{X})$ is sufficient for θ and $L(\theta, a)$ is convex in a.

Let $\rho(\boldsymbol{X})$ be an estimator of $\psi(\theta)$. Define

$$\rho^*(t) = \mathbb{E}\big[\rho(\boldsymbol{X})\big|T(\boldsymbol{X}) = t\big].$$

Then, if $\rho(\boldsymbol{X})$ has finite expectation and risk, i.e. both $\mathbb{E}_\theta\big[\rho(\boldsymbol{X})\big]$ and $\mathbb{E}_\theta\big[L(\theta, \rho(\boldsymbol{X}))\big]$ are finite, then $\rho^*(T(\boldsymbol{X}))$ is an estimator of $\psi(\theta)$ such that

$$\mathbb{E}_\theta\big[L(\theta, \rho^*(T(\boldsymbol{X})))\big] \le \mathbb{E}_\theta\big[L(\theta, \rho(\boldsymbol{X}))\big].$$

Moreover, if $\rho(\boldsymbol{X})$ is unbiased and $T(\boldsymbol{X})$ is complete, then

(i) $\rho^*(T(\boldsymbol{X}))$ is the **unique** unbiased estimator of $\psi(\theta)$ which is a function of $T(\boldsymbol{X})$;

(ii) $\mathbb{E}_\theta\big[L(\theta, \rho^*(T(\boldsymbol{X})))\big] \le \mathbb{E}_\theta\big[L(\theta, S(\boldsymbol{X}))\big]$ for any unbiased estimator $S(\boldsymbol{X})$ of $\psi(\theta)$.

[i.e. $\rho^*(T(\boldsymbol{X}))$ either strictly dominates, or has at least the same risk function as, any unbiased estimator $S(\boldsymbol{X})$.]

. . . .

Proof:

Sufficiency of $T(\boldsymbol{X})$ for θ implies that the conditional distribution of $\rho(\boldsymbol{X})$ given $T(\boldsymbol{X})$ is free of θ, so $\rho^*(T(\boldsymbol{X}))$ is a **legitimate** *estimator of $\psi(\theta)$.*

Applying Jensen's inequality,

$$L(\theta, \rho^*(T(\boldsymbol{X}))) \le \mathbb{E}\big[L(\theta, \rho(\boldsymbol{X}))\big|T(\boldsymbol{X})\big] \quad \Rightarrow \quad \mathbb{E}_\theta\big[L(\theta, \rho^*(T(\boldsymbol{X})))\big] \le \mathbb{E}_\theta\big[L(\theta, \rho(\boldsymbol{X}))\big].$$

Next suppose $\rho(\boldsymbol{X})$ is unbiased and $T(\boldsymbol{X})$ is complete. Then

$$\mathbb{E}_\theta\big[\rho^*(T(\boldsymbol{X}))\big] = \mathbb{E}_\theta\big[\,\mathbb{E}[\rho(\boldsymbol{X})\big|T(\boldsymbol{X})]\,\big] = \mathbb{E}_\theta\big[\rho(\boldsymbol{X})\big] = \psi(\theta),$$

i.e. $\rho^*(T(\boldsymbol{X}))$ is also unbiased for $\psi(\theta)$.

Using Lemma §5.3.1, $\rho^*(T(\boldsymbol{X}))$ is the unique function of $T(\boldsymbol{X})$ which is unbiased for $\psi(\theta)$. This proves part (i).

To prove (ii), for any unbiased estimator $S(\boldsymbol{X})$, define $S^*(T(\boldsymbol{X})) = \mathbb{E}\big[S(\boldsymbol{X})\big|T(\boldsymbol{X})\big]$. Then $S^*(T(\boldsymbol{X}))$ is unbiased for $\psi(\theta)$ and is a function of $T(\boldsymbol{X})$. By part (i), we have $S^* = \rho^*$ with probability one for all θ. Thus

$$\mathbb{E}_\theta\big[L(\theta, \rho^*(T(\boldsymbol{X})))\big] = \mathbb{E}_\theta\big[L(\theta, S^*(T(\boldsymbol{X})))\big] \le \mathbb{E}_\theta\big[L(\theta, S(\boldsymbol{X}))\big]. \qquad \blacksquare$$

5.3.3 Important applications of Rao-Blackwell Theorem —

- We may reduce, or at least not increase, the risk of an arbitrary estimator by evaluating its expected value conditional on a sufficient statistic.
- If a complete sufficient statistic T exists, we can obtain an unbiased estimator which has the smallest risk among all unbiased estimators. This can be done by

  either

  finding the expected value of any unbiased estimator conditional on T,

  or

  finding an unbiased estimator which is a function of T.

5.3.4 Consider a scalar target $\psi(\theta) \in \mathbb{R}$. Setting $L(\theta, a) = \{\psi(\theta) - a\}^2$ (squared loss), we obtain immediately the following.

Corollary. Let $\rho(\boldsymbol{X})$ be an estimator of $\psi(\theta)$ with finite second moment (i.e. $\mathbb{E}_\theta\big[\rho(\boldsymbol{X})^2\big] < \infty$). Then

$$\text{MSE of } \rho^*\big(T(\boldsymbol{X})\big) \leq \text{MSE of } \rho(\boldsymbol{X}).$$

If we assume further that $\rho(\boldsymbol{X})$ is unbiased and $T(\boldsymbol{X})$ is complete, then

(i) $\rho^*(T(\boldsymbol{X}))$ is the unique unbiased estimator of $\psi(\theta)$ which is a function of $T(\boldsymbol{X})$;

(ii) $\mathrm{Var}_\theta\big(\rho^*(T(\boldsymbol{X}))\big) \leq \mathrm{Var}_\theta\big(S(\boldsymbol{X})\big)$ for any unbiased estimator $S(\boldsymbol{X})$ of $\psi(\theta)$, i.e.

$\rho^*(T(\boldsymbol{X}))$ is a *uniformly minimum variance unbiased* (UMVU) estimator, and is also the unique UMVU estimator which is a function of $T(\boldsymbol{X})$.

5.3.5 **Examples.**

(i) Example §4.5.2(1) has proved that $\max_i X_i$ is complete sufficient for θ for X_i iid $\sim U[0, \theta]$. Since $\mathbb{E}_\theta[2X_1] = 2(\theta/2) = \theta$, $2X_1$ is unbiased for θ.

Note that

$$\mathbb{P}_\theta(X_1 = \max_i X_i) = 1/n, \quad \mathbb{P}_\theta(X_1 < \max_i X_i) = 1 - 1/n.$$

Given $\max_i X_i = t$ and $X_1 < \max_i X_i$, $X_1 \sim U[0, t]$. Thus

$$\mathbb{E}\big[2X_1 \big| \max_i X_i = t\big] = (2\,t)(1/n) + t\,(1 - 1/n) = (1 + 1/n)\,t,$$

so that

$(1 + 1/n)\max_i X_i$ has the smallest risk among all unbiased estimators of θ, for any convex loss function. In particular, it is a UMVU estimator of θ.

(ii) $\boldsymbol{X} = (X_1, \ldots, X_n)$ iid $\sim N(\mu, \sigma^2)$ $\leftarrow$ here $\theta = (\mu, \sigma)$

By exponential family properties, $T = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ is complete sufficient for $\theta = (\mu, \sigma)$.

- The sample mean $\bar{X}$ is unbiased for $\psi(\theta) = \mu$ and is a function of T. Thus

  $\bar{X}$ has the smallest risk among all unbiased estimators of μ, for any convex loss function. In particular, it is a UMVU estimator of μ.

- The sample variance $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ is unbiased for $\psi(\theta) = \sigma^2$ and is a function of T. Thus

  S^2 has the smallest risk among all unbiased estimators of σ^2, for any convex loss function. In particular, it is a UMVU estimator of σ^2.

§5.4 Information inequality

5.4.1 **Regularity assumptions.**

- $\boldsymbol{X} \sim f(\cdot|\theta)$, $\theta = (\theta_1, \ldots, \theta_k)^\top \in \Theta \subset \mathbb{R}^k$, where Θ contains an **open rectangle**,
- sample space $\{\boldsymbol{x} : f(\boldsymbol{x}|\theta) > 0\} = \mathcal{S}$ is common to all $\theta \in \Theta$,
- for any $\boldsymbol{x} \in \mathcal{S}$, $\theta \in \Theta$ and $i = 1, \ldots, k$, $\dfrac{\partial f(\boldsymbol{x}|\theta)}{\partial \theta_i}$ exists and is finite.

5.4.2 **Definitions.** Under regularity assumptions, the *score function* is defined to be

$$\boldsymbol{U}(\theta) = \big[U_1(\theta), \ldots, U_k(\theta)\big]^\top = \left[\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial \theta_1}, \ \ldots\ , \ \frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial \theta_k}\right]^\top,$$

and the *(Fisher) information matrix* is the $k \times k$ matrix

$$I(\theta) = \mathbb{E}_\theta\big[\boldsymbol{U}(\theta)\boldsymbol{U}(\theta)^\top\big],$$

whose (i, j)th entry is given by

$$I_{ij}(\theta) = \mathbb{E}_\theta\big[U_i(\theta)U_j(\theta)\big] = \mathbb{E}_\theta\left[\left(\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial \theta_i}\right)\left(\frac{\partial \ln f(\boldsymbol{X}|\theta)}{\partial \theta_j}\right)\right].$$

Note: Equivalently, we may define

$$\boldsymbol{U}(\theta) = \left[\frac{\partial \ln \ell_{\boldsymbol{X}}(\theta)}{\partial \theta_1}, \ \ldots\ , \ \frac{\partial \ln \ell_{\boldsymbol{X}}(\theta)}{\partial \theta_k}\right]^\top = \left[\frac{\partial S_{\boldsymbol{X}}(\theta)}{\partial \theta_1}, \ \ldots\ , \ \frac{\partial S_{\boldsymbol{X}}(\theta)}{\partial \theta_k}\right]^\top,$$

where $\ell_{\boldsymbol{X}}(\theta) \propto f(\boldsymbol{X}|\theta)$ denotes the likelihood function and $S_{\boldsymbol{X}}(\theta)$ the loglikelihood function.

5.4.3 **Lemma.** Under regularity assumptions and assuming that $\dfrac{\partial}{\partial \theta_i}\displaystyle\int f(\boldsymbol{x}|\theta)\,d\boldsymbol{x} = \int \frac{\partial f(\boldsymbol{x}|\theta)}{\partial \theta_i}\,d\boldsymbol{x}$, we have

$$\mathbb{E}_\theta\big[\boldsymbol{U}(\theta)\big] = \mathbf{0} \quad \text{and} \quad I(\theta) = \mathrm{Var}_\theta\big(\boldsymbol{U}(\theta)\big).$$

If, in addition, $f(\boldsymbol{x}|\theta)$ has second derivatives $\dfrac{\partial^2}{\partial \theta_i \partial \theta_j} f(\boldsymbol{x}|\theta)$ for all i, j, then

$$I(\theta) = -\,\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta \partial \theta^\top} \ln f(\boldsymbol{X}|\theta)\right].$$

Note: the lemma says that componentwise,

$$I_{ij}(\theta) = \mathrm{Cov}_\theta\big(U_i(\theta), U_j(\theta)\big) = -\,\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f(\boldsymbol{X}|\theta)\right].$$

Proof:

$$\mathbb{E}_\theta\big[U_i(\theta)\big] = \mathbb{E}_\theta\left[\frac{1}{f(\boldsymbol{X}|\theta)}\frac{\partial f(\boldsymbol{X}|\theta)}{\partial\theta_i}\right] = \int \frac{\partial f(\boldsymbol{x}|\theta)}{\partial\theta_i}\,d\boldsymbol{x} = \frac{\partial}{\partial\theta_i}\int f(\boldsymbol{x}|\theta)\,d\boldsymbol{x} = \frac{\partial}{\partial\theta_i}(1) = 0,$$

which implies immediately $I(\theta) = \mathbb{E}_\theta\big[\boldsymbol{U}(\theta)\boldsymbol{U}(\theta)^\top\big] = \mathrm{Var}_\theta\big(\boldsymbol{U}(\theta)\big)$.

To prove the last assertion, consider

$$\begin{aligned}
\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f(\boldsymbol{X}|\theta)\right] &= -\,\mathbb{E}_\theta\left[\frac{1}{f(\boldsymbol{X}|\theta)^2}\frac{\partial f(\boldsymbol{X}|\theta)}{\partial\theta_i}\frac{\partial f(\boldsymbol{X}|\theta)}{\partial\theta_j}\right] + \mathbb{E}_\theta\left[\frac{1}{f(\boldsymbol{X}|\theta)}\frac{\partial^2 f(\boldsymbol{X}|\theta)}{\partial\theta_i\partial\theta_j}\right]\\
&= -\,\mathbb{E}_\theta\big[U_i(\theta)U_j(\theta)\big] + \int \frac{\partial^2 f(\boldsymbol{x}|\theta)}{\partial\theta_i\partial\theta_j}\,d\boldsymbol{x}\\
&= -\,I_{ij}(\theta) + \frac{\partial^2}{\partial\theta_i\partial\theta_j}\int f(\boldsymbol{x}|\theta)\,d\boldsymbol{x} = -\,I_{ij}(\theta).
\end{aligned}$$

∎

5.4.4 **Theorem.** *(Information Inequality)*

Let $T = T(\boldsymbol{X})$ be a statistic with $\mathbb{E}_\theta[T^2] < \infty$. Assume that $\boldsymbol{\alpha}(\theta) = \left[\frac{\partial}{\partial\theta_1}\mathbb{E}_\theta[T], \ldots, \frac{\partial}{\partial\theta_k}\mathbb{E}_\theta[T]\right]^\top$ exists, and can be obtained by differentiating under the integral sign.

Then, under regularity assumptions and assuming that $I(\theta)$ is positive definite,

$$\mathrm{Var}_\theta(T) \geq \boldsymbol{\alpha}(\theta)^\top I(\theta)^{-1}\boldsymbol{\alpha}(\theta).$$

Proof:

Note that

$$\mathbb{E}_\theta\big[TU_i(\theta)\big] = \int T(\boldsymbol{x})\left\{\frac{1}{f(\boldsymbol{x}|\theta)}\frac{\partial f(\boldsymbol{x}|\theta)}{\partial\theta_i}\right\} f(\boldsymbol{x}|\theta)\,d\boldsymbol{x} = \frac{\partial}{\partial\theta_i}\int T(\boldsymbol{x})f(\boldsymbol{x}|\theta)\,d\boldsymbol{x} = \frac{\partial}{\partial\theta_i}\mathbb{E}_\theta[T],$$

so that $\boldsymbol{\alpha}(\theta) = \mathbb{E}_\theta\big[\boldsymbol{U}(\theta)T\big] = \mathbb{E}_\theta\big[\boldsymbol{U}(\theta)(T - \mathbb{E}_\theta[T])\big]$.

Consider

$$\begin{aligned}
0 &\leq \mathbb{E}_\theta\Big[\big\{T - \mathbb{E}_\theta[T] - \boldsymbol{\alpha}(\theta)^\top I(\theta)^{-1}\boldsymbol{U}(\theta)\big\}^2\Big]\\
&= \mathrm{Var}_\theta(T) - 2\,\boldsymbol{\alpha}(\theta)^\top I(\theta)^{-1}\mathbb{E}_\theta\big[\boldsymbol{U}(\theta)(T - \mathbb{E}_\theta[T])\big] + \boldsymbol{\alpha}(\theta)^\top I(\theta)^{-1}\mathbb{E}_\theta\big[\boldsymbol{U}(\theta)\,\boldsymbol{U}(\theta)^\top\big]I(\theta)^{-1}\boldsymbol{\alpha}(\theta)\\
&= \mathrm{Var}_\theta(T) - \boldsymbol{\alpha}(\theta)^\top I(\theta)^{-1}\boldsymbol{\alpha}(\theta),
\end{aligned}$$

which implies the Information Inequality. ∎

5.4.5 **Corollary.** Let $\psi(\theta)$ be a differentiable function of θ, and T be an unbiased estimator of $\psi(\theta)$. Then

$$\mathrm{Var}_\theta(T) \geq \left[\frac{\partial\psi(\theta)}{\partial\theta_1}, \ldots, \frac{\partial\psi(\theta)}{\partial\theta_k}\right] I(\theta)^{-1} \left[\frac{\partial\psi(\theta)}{\partial\theta_1}, \ldots, \frac{\partial\psi(\theta)}{\partial\theta_k}\right]^\top.$$

Proof: Immediate from Theorem §5.4.4, noting that $\mathbb{E}_\theta[T] = \psi(\theta)$. ∎

Practical implication:

The above corollary provides a lower bound for the variance of **any** *unbiased estimator of* $\psi(\theta)$.

If an unbiased estimator has variance **equal** *to this lower bound, it must be a UMVU estimator of* $\psi(\theta)$.

5.4.6 For the special case where $k = 1$ and $\psi(\theta) = \theta$, the lower bound given in Corollary §5.4.5, i.e. $I(\theta)^{-1} = 1/I(\theta)$, is known as the *Cramér-Rao Lower Bound* (CRLB).

5.4.7 **Example §5.4.1** (*Normal mean regression*)

$(u_1, \ldots, u_n)$:	u_i a non-random covariate vector associated with ith response
$\mu_\beta(u)$:	(possibly non-linear) regression function parameterised by $\beta \in \mathbb{R}^q$
$\boldsymbol{X} = (X_1, \ldots, X_n)$:	independent responses, with $X_i \sim N\big(\mu_\beta(u_i), \nu\big)$

Joint pdf: $f(\boldsymbol{x}|\beta,\nu) = (2\pi\nu)^{-n/2} \exp\left[-\frac{\sum_{i=1}^n \{x_i - \mu_\beta(u_i)\}^2}{2\nu}\right]$. Then

$$\frac{\partial}{\partial\beta} \ln f(\boldsymbol{X}|\beta,\nu) = \nu^{-1} \sum_{i=1}^n \big\{X_i - \mu_\beta(u_i)\big\} \frac{\partial\mu_\beta(u_i)}{\partial\beta},$$

$$\frac{\partial}{\partial\nu} \ln f(\boldsymbol{X}|\beta,\nu) = -\frac{n}{2\nu} + \frac{1}{2\nu^2} \sum_{i=1}^n \big\{X_i - \mu_\beta(u_i)\big\}^2,$$

$$\frac{\partial^2}{\partial\beta\partial\beta^\top} \ln f(\boldsymbol{X}|\beta,\nu) = \nu^{-1} \sum_{i=1}^n \left[-\frac{\partial\mu_\beta(u_i)}{\partial\beta}\frac{\partial\mu_\beta(u_i)}{\partial\beta^\top} + \big\{X_i - \mu_\beta(u_i)\big\} \frac{\partial^2\mu_\beta(u_i)}{\partial\beta\partial\beta^\top}\right],$$

$$\frac{\partial^2}{\partial\beta\partial\nu} \ln f(\boldsymbol{X}|\beta,\nu) = -\frac{1}{\nu^2} \sum_{i=1}^n \big\{X_i - \mu_\beta(u_i)\big\} \frac{\partial\mu_\beta(u_i)}{\partial\beta},$$

$$\frac{\partial^2}{\partial\nu^2} \ln f(\boldsymbol{X}|\beta,\nu) = \frac{n}{2\nu^2} - \frac{1}{\nu^3} \sum_{i=1}^n \big\{X_i - \mu_\beta(u_i)\big\}^2,$$

so that

$$I(\beta,\nu) = \begin{bmatrix} \nu^{-1}\displaystyle\sum_{i=1}^{n} \frac{\partial \mu_\beta(u_i)}{\partial \beta}\frac{\partial \mu_\beta(u_i)}{\partial \beta^\top} & \mathbf{0} \\ \mathbf{0}^\top & n/(2\nu^2) \end{bmatrix}$$

and

$$I(\beta,\nu)^{-1} = \begin{bmatrix} \nu\left\{\displaystyle\sum_{i=1}^{n} \frac{\partial \mu_\beta(u_i)}{\partial \beta}\frac{\partial \mu_\beta(u_i)}{\partial \beta^\top}\right\}^{-1} & \mathbf{0} \\ \mathbf{0}^\top & 2\nu^2/n \end{bmatrix}.$$

Let T be any unbiased estimator of $\boldsymbol{c}^\top\beta$ for a given vector $\boldsymbol{c} \in \mathbb{R}^q$, i.e. $\mathbb{E}_{\beta,\nu}[T] = \boldsymbol{c}^\top\beta$ for all β, ν. Note

$$\boldsymbol{\alpha}(\beta,\nu) = \begin{bmatrix} \dfrac{\partial \boldsymbol{c}^\top\beta}{\partial\beta} \\ \dfrac{\partial \boldsymbol{c}^\top\beta}{\partial\nu} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ 0 \end{bmatrix}, \quad \boldsymbol{\alpha}(\beta,\nu)^\top I(\beta,\nu)^{-1}\boldsymbol{\alpha}(\beta,\nu) = \nu\,\boldsymbol{c}^\top\left\{\sum_{i=1}^{n} \frac{\partial \mu_\beta(u_i)}{\partial \beta}\frac{\partial \mu_\beta(u_i)}{\partial \beta^\top}\right\}^{-1}\boldsymbol{c}.$$

Thus

$$\mathrm{Var}_{\beta,\nu}(T) \geq \nu\,\boldsymbol{c}^\top\left\{\sum_{i=1}^{n} \frac{\partial \mu_\beta(u_i)}{\partial \beta}\frac{\partial \mu_\beta(u_i)}{\partial \beta^\top}\right\}^{-1}\boldsymbol{c}.$$

Special cases:

(i) (*Multiple linear regression*) $\mu_\beta(u) = u^\top\beta$

$$\rightarrow \quad \frac{\partial \mu_\beta(u_i)}{\partial\beta} = u_i \quad \text{and} \quad \mathrm{Var}_{\beta,\nu}(T) \geq \nu\,\boldsymbol{c}^\top\Big(\sum_{i=1}^{n} u_i u_i^\top\Big)^{-1}\boldsymbol{c}.$$

Take $T = \boldsymbol{c}^\top\Big(\sum_{i=1}^{n} u_i u_i^\top\Big)^{-1}\sum_{i=1}^{n} X_i u_i$ (*least squares estimator*). Then

$$\mathbb{E}_{\beta,\nu}[T] = \boldsymbol{c}^\top\beta \quad \text{and} \quad \mathrm{Var}_{\beta,\nu}(T) = \nu\,\boldsymbol{c}^\top\Big(\sum_{i=1}^{n} u_i u_i^\top\Big)^{-1}\boldsymbol{c} \longleftarrow \text{lower bound},$$

i.e. T gives the **minimum** variance among all unbiased estimators of $\boldsymbol{c}^\top\beta$,
i.e. T is a UMVU estimator of $\boldsymbol{c}^\top\beta$.

(ii) (*Two-sample comparison*) — a special case of (i)

$$\begin{cases} X_1,\ldots,X_m \text{ iid from } N(\beta_1,\nu) & \rightarrow \quad u_1 = \cdots = u_m = [1,0]^\top, \\ X_{m+1},\ldots,X_n \text{ iid from } N(\beta_2,\nu) & \rightarrow \quad u_{m+1} = \cdots = u_n = [0,1]^\top. \end{cases}$$

Target for estimation: $\boldsymbol{c}^\top\beta = [1,-1]\beta = \beta_1 - \beta_2$.
Take $T = \boldsymbol{c}^\top\Big(\sum_{i=1}^n u_i u_i^\top\Big)^{-1}\sum_{i=1}^n X_i u_i = m^{-1}\sum_{i=1}^m X_i - (n-m)^{-1}\sum_{i=m+1}^n X_i$. Then

$$\mathbb{E}_{\beta,\nu}[T] = \beta_1 - \beta_2 \quad \text{and} \quad \mathrm{Var}_{\beta,\nu}(T) = \nu\,\boldsymbol{c}^\top\Big(\sum_{i=1}^n u_i u_i^\top\Big)^{-1}\boldsymbol{c} = \nu\left(\frac{1}{m} + \frac{1}{n-m}\right) \leftarrow \text{lower bound,}$$

i.e. T gives the **minimum** variance among all unbiased estimators of $\beta_1 - \beta_2$,
i.e. T is a UMVU estimator of $\beta_1 - \beta_2$.

§5.5 Maximum likelihood estimator

5.5.1 **Definition.** Suppose $\hat\theta$ maximises $\ell_{\boldsymbol{X}}(\theta)$, or equivalently, $S_{\boldsymbol{X}}(\theta)$. We say $\hat\theta$ is the *maximum likelihood estimator* (mle) of θ.

5.5.2 Suppose $\theta \in \mathbb{R}^k$. Usually $\hat\theta$ can be obtained by solving

$$(*)\ \textit{likelihood equations}: \boldsymbol{U}(\theta) = \boldsymbol{0}, \quad \text{i.e.}\ \frac{\partial}{\partial\theta_j}S_{\boldsymbol{X}}(\theta) = 0, \quad j = 1,\ldots,k.$$

- $(*)$ may have > 1 solutions for $\theta \in \Theta$, so we must check for maximality.
- $(*)$ may be nonsense, e.g. $\boldsymbol{X} = (X_1,\ldots,X_n)$ iid from Binomial $(\theta, 1/2)$ and θ is an **integer**.
- for some Θ, e.g. $\Theta = [0,1]$, $S_{\boldsymbol{X}}(\theta)$ may be maximised at the boundary of Θ rather than at a solution to $(*)$.

5.5.3 Suppose T is minimal sufficient for θ. Then $\ell_{\boldsymbol{X}}(\theta) \propto g(T,\theta)$. Thus $\hat\theta$ maximises $g(T,\theta)$ with respect to θ and is a function of T only.

[Note: $\hat\theta$ must be a function of minimal sufficient statistic but itself needs **not** be sufficient for θ.]

5.5.4 Maximum likelihood estimators can be biased, e.g. $X_1,\ldots,X_n$ iid from $N(\mu,\sigma^2)$

$$\to \text{ mle of } \sigma^2 \text{ is } \hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar X)^2 \quad\Rightarrow\quad \mathbb{E}\big[\hat\sigma^2\big] = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

5.5.5 If $\hat\theta$ is mle of θ, then $\psi(\hat\theta)$ is mle of $\psi(\theta)$ for any transformation $\psi(\cdot)$.

5.5.6 $\boldsymbol{X} = (X_1,\ldots,X_n) \sim$ joint probability function $p_1(x_1|\theta)\times\cdots\times p_n(x_n|\theta)$
Likelihood function: $\ell_n(\theta) \propto \prod_{i=1}^n p_i(X_i|\theta)$, loglikelihood function: $S_n(\theta) = \ln \ell_n(\theta)$

Note: $X_1, \ldots, X_n$ are independent but may not be identically distributed.

Let θ_0 be true value of $\theta \in \Theta \subset \mathbb{R}^k$. Assume that $S_n(\theta)$ is differentiable in an open neighbourhood of θ_0 and that the likelihood equations have a root at $\theta = \hat{\theta}_n$ (mle).

Theorem. Subject to regularity conditions[1] on $p_1(\cdot|\theta), \ldots, p_n(\cdot|\theta)$, we have

(i) $\hat{\theta}_n$ *converges in probability* to θ_0,

(ii) $n^{1/2}(\hat{\theta}_n - \theta_0)$ *converges in distribution* to $N\big(\mathbf{0}, \mathscr{I}(\theta_0)^{-1}\big)$,

(iii) $n^{-1/2}\boldsymbol{U}(\theta_0)$ converges in distribution to $N\big(\mathbf{0}, \mathscr{I}(\theta_0)\big)$,

where $\mathscr{I}(\theta) = \lim\limits_{n\to\infty} n^{-1} I(\theta) = -\lim\limits_{n\to\infty} n^{-1} \sum\limits_{\ell=1}^{n} \mathbb{E}_\theta \left[\dfrac{\partial^2 \ln p_\ell(X_\ell|\theta)}{\partial\theta\partial\theta^\top}\right]$ has (i,j)th entry

$$\mathscr{I}_{ij}(\theta) = -\lim_{n\to\infty} n^{-1} \sum_{\ell=1}^{n} \mathbb{E}_\theta \left[\frac{\partial^2 \ln p_\ell(X_\ell|\theta)}{\partial\theta_i\partial\theta_j}\right].$$

Note: If $X_1, \ldots, X_n$ are iid from $p(\cdot|\theta)$, then $\mathscr{I}(\theta) = -\mathbb{E}_\theta \left[\dfrac{\partial^2 \ln p(X_1|\theta)}{\partial\theta\partial\theta^\top}\right]$.

. . . .

Proof: (outline)

(i) Without loss of generality write $S_n(\theta) = \sum_{i=1}^n \ln p_i(X_i|\theta) = \sum_{i=1}^n s_i(\theta)$. *For any* $\theta \in \Theta$, *Jensen's inequality and concavity of the logarithm imply that, for* $i = 1, \ldots, n$,

$$\mathbb{E}_{\theta_0}\big[s_i(\theta) - s_i(\theta_0)\big] \le \ln \mathbb{E}_{\theta_0}\left[\frac{p_i(X_i|\theta)}{p_i(X_i|\theta_0)}\right] = 0.$$

Thus the function $\theta \mapsto \lim_{n\to\infty} n^{-1}\mathbb{E}_{\theta_0}\big[S_n(\theta)\big]$ *is maximised at* $\theta = \theta_0$. *We may assume under regularity conditions that* $\lim_{n\to\infty} n^{-1}\mathbb{E}_{\theta_0}\big[S_n(\theta)\big] < \lim_{n\to\infty} n^{-1}\mathbb{E}_{\theta_0}\big[S_n(\theta_0)\big]$ *for* $0 < \|\theta - \theta_0\| < \Delta$, *some* $\Delta > 0$.

Fix an arbitrary $\epsilon \in (0, \Delta)$. *Then, by the (uniform) Weak Law of Large Numbers,*

$$n^{-1}\sum_{i=1}^{n}\big\{s_i(\theta) - s_i(\theta_0)\big\} \xrightarrow{\text{p}} \lim_{n\to\infty} n^{-1}\mathbb{E}_{\theta_0}\big[S_n(\theta) - S_n(\theta_0)\big] < 0 \quad \textit{uniformly over } \|\theta - \theta_0\| = \epsilon.$$

[1]For details see Bradley, R.A. and Gart, J.J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika*, **49**, 205–214.

Put $\delta_\epsilon = -\sup\Big\{ \lim_{n\to\infty} n^{-1}\mathbb{E}_{\theta_0}\big[S_n(\theta) - S_n(\theta_0)\big] : \|\theta-\theta_0\| = \epsilon \Big\} > 0$. *Assume that* $\hat\theta_n$ *solves the likelihood equations uniquely. Then*

$$\begin{aligned}
&\mathbb{P}_{\theta_0}\big(\|\hat\theta_n - \theta_0\| \le \epsilon\big)\\
&\quad\ge\ \mathbb{P}_{\theta_0}\big(S_n(\theta) < S_n(\theta_0)\ \forall\, \|\theta-\theta_0\| = \epsilon\big)\\
&\quad\ge\ \mathbb{P}_{\theta_0}\Big(n^{-1}\sum_{i=1}^n \big\{s_i(\theta) - s_i(\theta_0)\big\} \le -\delta_\epsilon/2\ \ \forall\, \|\theta-\theta_0\| = \epsilon\Big)\\
&\quad\ge\ \mathbb{P}_{\theta_0}\Big(n^{-1}\sum_{i=1}^n \big\{s_i(\theta) - s_i(\theta_0)\big\} - \lim_{n\to\infty} n^{-1}\mathbb{E}_{\theta_0}\big[S_n(\theta) - S_n(\theta_0)\big] \le \delta_\epsilon/2\ \ \forall\, \|\theta-\theta_0\| = \epsilon\Big) \to 1.
\end{aligned}$$

(ii, iii) Write $W(\theta)$ *for the* $k\times k$ *matrix* $\dfrac{\partial^2}{\partial\theta\partial\theta^\top} S_n(\theta)$, *with* (i,j)*th entry* $W_{ij}(\theta) = \dfrac{\partial^2}{\partial\theta_i\partial\theta_j} S_n(\theta)$.
By the Strong Law of Large Numbers,

$$-\,n^{-1}W_{ij}(\theta_0) \longrightarrow -\lim_{n\to\infty} n^{-1}\sum_{\ell=1}^n \mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} s_\ell(\theta)\bigg|_{\theta_0}\right] = \mathscr{I}_{ij}(\theta_0)\ \textit{almost surely.}$$

By Lemma §5.4.3, we have, for $\ell, \ell' = 1,\ldots,n$,

$$\begin{aligned}
\mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial\theta_i} s_\ell(\theta)\bigg|_{\theta_0}\right] &= 0\\
\mathrm{Cov}_{\theta_0}\left(\frac{\partial}{\partial\theta_i} s_\ell(\theta)\bigg|_{\theta_0},\ \frac{\partial}{\partial\theta_j} s_{\ell'}(\theta)\bigg|_{\theta_0}\right) &= -\,\mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} s_\ell(\theta)\bigg|_{\theta_0}\right]\mathbf{1}\{\ell = \ell'\}.
\end{aligned}$$

Then, by Central Limit Theorem,

$$n^{-1/2}\boldsymbol{U}(\theta_0) = n^{-1/2}\sum_{\ell=1}^n\left[\frac{\partial}{\partial\theta_1} s_\ell(\theta)\bigg|_{\theta_0}, \ldots, \frac{\partial}{\partial\theta_k} s_\ell(\theta)\bigg|_{\theta_0}\right]^\top \longrightarrow N\big(\mathbf{0}, \mathscr{I}(\theta_0)\big)\ \textit{in distribution,}$$

which proves (iii). Taylor expansion gives

$$\mathbf{0} = n^{-1/2}\boldsymbol{U}(\hat\theta_n) = n^{-1/2}\boldsymbol{U}(\theta_0) + n^{-1/2}W(\theta_0)(\hat\theta_n - \theta_0) + \boldsymbol{\epsilon}_n,$$

where $\boldsymbol{\epsilon}_n \to \mathbf{0}$ *in probability, so that (ii) follows by noting that*

$$\begin{aligned}
&n^{1/2}(\hat\theta_n - \theta_0) = \big\{-n^{-1}W(\theta_0)\big\}^{-1}\Big(n^{-1/2}\boldsymbol{U}(\theta_0) + \boldsymbol{\epsilon}_n\Big)\\
&\quad\longrightarrow\ N\Big(\mathbf{0}, \mathscr{I}(\theta_0)^{-1}\mathscr{I}(\theta_0)\big[\mathscr{I}(\theta_0)^{-1}\big]^\top\Big) \equiv N\big(\mathbf{0}, \mathscr{I}(\theta_0)^{-1}\big)\ \textit{in distribution.}
\end{aligned}$$

∎

5.5.7 <u>Practical implication:</u>

For large n,

- $\hat{\theta}_n$ is distributed **approximately** as $N\big(\theta_0, n^{-1}\mathscr{I}(\theta_0)^{-1}\big) \approx N\big(\theta_0, I(\theta_0)^{-1}\big)$;
- if $\psi(\theta)$ is differentiable near θ_0, then $\hat{\psi}_n = \psi(\hat{\theta}_n)$ is distributed **approximately** as $N\left(\psi(\theta_0), \boldsymbol{\alpha}(\theta_0)^\top I(\theta_0)^{-1}\boldsymbol{\alpha}(\theta_0)\right)$, where $\boldsymbol{\alpha}(\theta) = \left[\dfrac{\partial\psi(\theta)}{\partial\theta_1}, \ldots, \dfrac{\partial\psi(\theta)}{\partial\theta_k}\right]^\top$.

5.5.8 **Example §5.4.1** (cont'd)

True values of (β, ν): (β_0, ν_0)

Likelihood function: $\ell_n(\beta,\nu) \propto (2\pi\nu)^{-n/2} \exp\left[-\dfrac{1}{2\nu}\sum_{i=1}^n \{X_i - \mu_\beta(u_i)\}^2\right]$

Likelihood equations:

$$\begin{cases} \dfrac{\partial}{\partial\beta} S_n(\beta,\nu) = \nu^{-1}\sum_{i=1}^n \left\{X_i - \mu_\beta(u_i)\right\} \dfrac{\partial\mu_\beta(u_i)}{\partial\beta} & = \mathbf{0} \\ \dfrac{\partial}{\partial\nu} S_n(\beta,\nu) = -\dfrac{n}{2\nu} + \dfrac{1}{2\nu^2}\sum_{i=1}^n \left\{X_i - \mu_\beta(u_i)\right\}^2 & = 0 \end{cases}$$

$\Rightarrow$ mle of (β_0, ν_0) is $(\hat{\beta}_n, \hat{\nu}_n)$, where

$$\sum_{i=1}^n \left\{X_i - \mu_{\hat{\beta}_n}(u_i)\right\} \frac{\partial\mu_\beta(u_i)}{\partial\beta}\bigg|_{\beta=\hat{\beta}_n} = \mathbf{0} \quad \text{and} \quad \hat{\nu}_n = n^{-1}\sum_{i=1}^n \left\{X_i - \mu_{\hat{\beta}_n}(u_i)\right\}^2.$$

Results of §5.4.7 imply that

$$\mathscr{I}(\beta,\nu) = \lim_{n\to\infty} n^{-1} I(\beta,\nu) = \begin{bmatrix} \nu^{-1} \lim\limits_{n\to\infty} n^{-1} \sum\limits_{i=1}^n \dfrac{\partial\mu_\beta(u_i)}{\partial\beta}\dfrac{\partial\mu_\beta(u_i)}{\partial\beta^\top} & \mathbf{0} \\ \mathbf{0}^\top & 1/(2\nu^2) \end{bmatrix}$$

and

$$\mathscr{I}(\beta,\nu)^{-1} = \begin{bmatrix} \nu \left\{ \lim\limits_{n\to\infty} n^{-1} \sum\limits_{i=1}^n \dfrac{\partial\mu_\beta(u_i)}{\partial\beta}\dfrac{\partial\mu_\beta(u_i)}{\partial\beta^\top} \right\}^{-1} & \mathbf{0} \\ \mathbf{0}^\top & 2\nu^2 \end{bmatrix}.$$

Large-sample properties of mle (Theorem §5.5.6(ii)) $\Rightarrow$

$$n^{1/2}\left(\begin{bmatrix} \hat{\beta}_n \\ \hat{\nu}_n \end{bmatrix} - \begin{bmatrix} \beta_0 \\ \nu_0 \end{bmatrix} \right) \longrightarrow N\big(\mathbf{0}, \mathscr{I}(\beta_0,\nu_0)^{-1}\big) \text{ in distribution.}$$

Large-sample properties of score function (Theorem §5.5.6(iii)) $\Rightarrow$

$$n^{-1/2}\begin{bmatrix} \dfrac{1}{\nu_0}\sum_{i=1}^n \{X_i-\mu_{\beta_0}(u_i)\}\left.\dfrac{\partial\mu_\beta(u_i)}{\partial\beta}\right|_{\beta=\beta_0} \\ \dfrac{1}{2\nu_0^2}\displaystyle\sum_{i=1}^n\Big(\{X_i-\mu_{\beta_0}(u_i)\}^2-\nu_0\Big)\end{bmatrix} \longrightarrow N\big(\mathbf{0},\mathscr{I}(\beta_0,\nu_0)\big) \text{ in distribution.}$$

§5.6 Exercise: comparison of UMVUE and MLE

5.6.1 Let $\boldsymbol{X}\subset\{1,\ldots,\theta\}$ be a random subset of positive integers, for some unknown $\theta\in\{1,2,\ldots\}$. Let $p=1-q\in(0,1)$ be a known probability such that the θ events $\{1\in\boldsymbol{X}\},\ldots,\{\theta\in\boldsymbol{X}\}$ are independent and each occurs with probability p.

We wish to estimate θ based on the data $\boldsymbol{X}$.

Question 1

Show that $\boldsymbol{X}$ has the mass function

$$f(\boldsymbol{X}|\theta)=p^{\#\boldsymbol{X}}q^{\theta-\#\boldsymbol{X}}\,\mathbf{1}\big\{\boldsymbol{X}\subset\{1,\ldots,\theta\}\big\},$$

where $\#\boldsymbol{X}$ denotes the number of elements in $\boldsymbol{X}$.

Question 2

Find a complete sufficient statistic for θ.

Question 3

Find the maximum likelihood estimator (mle) of θ.

Question 4

Find a uniformly minimum variance unbiased (UMVU) estimator of θ.

Question 5

Compare the MSE's of the maximum likelihood and UMVU estimators.

§5.7 Nonparametric estimation

5.7.1 Classical parametric (or *model-based*) inference requires specification of a parametric model $\big\{f(\cdot|\theta):\theta\in\Theta\big\}$, with parameter space $\Theta\subset\mathbb{R}^k$.

Nonparametric (or *model-free*) inference attempts to yield more *robust* results by dropping such stringent model assumptions.

5.7.2 Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be n independent replicates of $X \sim$ distribution function $F \in \mathscr{F}$, where $\mathscr{F}$ denotes the class of **all** possible distribution functions of X. Based on the observed sample $\boldsymbol{x} = (x_1, \ldots, x_n)$, the *nonparametric likelihood function* is

$$\ell_{\boldsymbol{x}}(F) = \prod_{i=1}^{n} \mathbb{P}_F(X = x_i), \qquad F \in \mathscr{F}.$$

Here $\mathbb{P}_F(X = x_i)$ should be interpreted as

the probability of observing x_i if F were the true underlying distribution function.

Note: If F is continuous, then $\mathbb{P}_F(X = x_i) = 0$ and so $\ell_{\boldsymbol{x}}(F) = 0$.

5.7.3 **Definition.** The *empirical distribution* of $\boldsymbol{X} = (X_1, \ldots, X_n)$ is the distribution of a discrete random variable X^* with the mass function

$$\mathbb{P}\big(X^* = X_i \big| \boldsymbol{X}\big) = \frac{1}{n}, \quad \text{for } i = 1, 2, \ldots, n, \quad \text{conditional on } \boldsymbol{X}.$$

Thus, the empirical distribution places on each observation X_i a probability mass of $1/n$, or equivalently, $X^* = X_J$, where J is randomly selected (with equal probabilities) from the set of integers $\{1, 2, \ldots, n\}$.

5.7.4 The cdf of the empirical distribution is given by

$$\hat{F}_n(t) = \mathbb{P}\big(X^* \le t \big| \boldsymbol{X}\big) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \le t\}.$$

Note: "$X_i \le t$" should be interpreted *componentwise* if X_i and t are vectors.

5.7.5 The nonparametric likelihood function $\ell_{\boldsymbol{X}}(F)$ is maximised by the empirical distribution function $\hat{F}_n$ of the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$.

Thus we may regard $\hat{F}_n$ ($\in \mathscr{F}$ by assumption) as a *nonparametric* mle of F.

Consequently, any given functional $\theta(F)$ of F can be estimated by the *nonparametric* mle $\theta(\hat{F}_n)$.

5.7.6 **Examples**.

(i) If $\theta(F) = \int x \, dF(x)$ (i.e. mean of F), then $\theta(\hat{F}_n) = \int x \, d\hat{F}_n(x) = \bar{X}$ (i.e. sample mean of $X_1, \ldots, X_n$).

(ii) If $\theta(F) = \int x^2\,dF(x) - \big(\int x\,dF(x)\big)^2$ (i.e. variance of F), then $\theta(\hat F_n) = n^{-1}\sum_{i=1}^n X_i^2 - \bar X^2 = n^{-1}\sum_{i=1}^n (X_i - \bar X)^2$ (i.e. sample variance of $X_1,\dots,X_n$).

(iii) If $\theta(F) = F^{-1}(\xi)$ (i.e. ξ^{th} population quantile of F), then $\theta(\hat F_n) = \hat F_n^{-1}(\xi)$ (i.e. ξ^{th} sample quantile of $X_1,\dots,X_n$).

(iv) *(linear regression)*

Let $(\boldsymbol{X}_1, Y_1),\dots,(\boldsymbol{X}_n, Y_n)$ be n i.i.d. replicates of $(\boldsymbol{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$, which follows an unknown $(p+1)$-variate distribution F. Define

$$\theta(F) = \psi(\beta_0, \boldsymbol{\beta}) = \psi\Big(\operatorname*{argmin}_{b_0, \boldsymbol{b}} \mathbb{E}_F\big[L(Y - b_0 - \boldsymbol{b}^\top \boldsymbol{X})\big]\Big),$$

for some loss function $L(\cdot)$ and some given real-valued function $\psi(\cdot)$. Then

$$\theta(\hat F_n) = \psi(\hat\beta_0, \hat{\boldsymbol{\beta}}) = \psi\Big(n^{-1}\operatorname*{argmin}_{b_0, \boldsymbol{b}} \sum_{i=1}^n L(Y_i - b_0 - \boldsymbol{b}^\top \boldsymbol{X}_i)\Big).$$

Exercise: Write down explicit expressions for $(\beta_0, \boldsymbol{\beta})$ and $(\hat\beta_0, \hat{\boldsymbol{\beta}})$ in the special case where $L(u) = u^2$.

(v) Consider two independent random samples, $\boldsymbol{X} = (X_1,\dots,X_n)$ and $\boldsymbol{Y} = (Y_1,\dots,Y_m)$, drawn respectively from the distribution functions F and G. Define a functional

$$\theta(F, G) = \mathbb{P}_{F,G}(X_1 > cY_1),$$

for a given constant c. Denote by $\hat F_n$ and $\hat G_m$ the empirical cdf's of the samples $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Then

$$\theta(\hat F_n, \hat G_m) = m^{-1}\sum_{j=1}^m \big\{1 - \hat F_n(cY_j)\big\} = (mn)^{-1}\sum_{j=1}^m \sum_{i=1}^n \mathbf{1}\{X_i > cY_j\}.$$

If it is known that $F = G$, then we should estimate F and G by the empirical cdf $\hat H_{n+m}$ of the pooled sample $(\boldsymbol{X}, \boldsymbol{Y}) = (W_1,\dots,W_{n+m})$. In this case, we have

$$\theta(\hat H_{n+m}, \hat H_{n+m}) = (n+m)^{-1}\sum_{i=1}^{n+m} \big\{1 - \hat H_{n+m}(cW_i)\big\} = (n+m)^{-2}\sum_{i=1}^{n+m}\sum_{j=1}^{n+m} \mathbf{1}\{W_j > cW_i\}.$$

§5.8 Bootstrap estimation

5.8.1 Suppose $\boldsymbol{X} = (X_1,\dots,X_n)$ are i.i.d. $\sim F \in \mathscr{F}$. Very often inference about F calls for knowledge of the sampling distribution function of some real-valued statistic $T(\boldsymbol{X}, F)$:

$$F_T(y|F) \triangleq \mathbb{P}_F\big(T(\boldsymbol{X}, F) \le y\big), \qquad y \in \mathbb{R}.$$

Example: For a problem of estimating $\theta(F)$, $T(\boldsymbol{X}, F)$ may be defined as the L_p norm $\big\|\hat\theta(\boldsymbol{X}) - \theta(F)\big\|_p$, which measures the estimation error of the estimator $\hat\theta(\boldsymbol{X})$. Then knowledge of $F_T(\cdot|F)$ enables us to investigate, for example, mean squared error or mean absolute deviation of $\hat\theta(\boldsymbol{X})$, among other applications.

5.8.2 If F were known, exact answers would have been available, derived by either *analytical calculations* or *Monte Carlo simulation*.

If F is unknown (as is often the case), one may sometimes resort to asymptotic approximations, usually derived from a certain kind of *central limit theorem* (for some properly chosen normalising constants a_n) :

$$a_n T(\boldsymbol{X}, F) \to N(0, \sigma_T^2) \textit{ in distribution} \quad \Rightarrow \quad F_T(y|F) \approx \Phi\big(a_n y/\sigma_T\big) \textit{ for large } n.$$

What if σ_T is **NOT** available and is very hard to estimate?

What if sample size n is **NOT** large?

5.8.3 The *bootstrap* method provides an approximate solution to the problem by applying the "*plug-in*" idea:

- **plug in** *the empirical distribution* $\hat F_n$ *as a substitute for* F *and*

$$\textit{estimate} \quad F_T(y|F) \quad \textit{by} \quad F_T\big(y\big|\hat F_n\big) = \mathbb{P}_{\hat F_n}\big(T(\boldsymbol{X}^*, \hat F_n) \le y\big), \qquad y \in \mathbb{R},$$

where $\boldsymbol{X}^* = \big(X_1^*, \ldots, X_n^*\big)$ *denotes an i.i.d. sample drawn from* $\hat F_n$*.*

The estimate $F_T\big(\,\cdot\,\big|\hat F_n\big)$ is known as the bootstrap distribution function, and $\boldsymbol{X}^*$ is known as a *bootstrap sample* of size n.

Note: In some problem settings, it may be more appropriate to substitute F by an alternative estimate different from $\hat F_n$ and generate bootstrap samples from that alternative estimate accordingly.

5.8.4 In practice, the *bootstrap distribution* $F_T\big(\,\cdot\,\big|\hat F_n\big)$ can be approximated by *Monte Carlo simulation*:

$$F_T\big(y\big|\hat F_n\big) \approx B^{-1}\sum_{b=1}^{B} \mathbf{1}\Big\{T\big(\boldsymbol{X}^{*b}, \hat F_n\big) \le y\Big\}, \qquad y \in \mathbb{R},$$

where $\boldsymbol{X}^{*1}, \ldots, \boldsymbol{X}^{*B}$ are B independent bootstrap samples, each of size n, drawn from $\hat F_n$.

Noting that $\hat F_n$ is the empirical cdf of $\boldsymbol{X} = (X_1, \ldots, X_n)$, in practice we may generate each $\boldsymbol{X}^{*b}$ by sampling n observations with replacement from $\boldsymbol{X}$. Thus the bootstrap is often referred to as a *resampling method*.

5.8.5 **Example.**

Let $X_1, \ldots, X_n \in \mathbb{R}$ be ordered as $X_{(1)} \le \cdots \le X_{(n)}$.

Consider: $\theta = \theta(F) = \mathbb{E}_F X_1, \quad \hat\theta = \hat\theta(\boldsymbol{X}) = \dfrac{1}{n-2m} \displaystyle\sum_{i=m+1}^{n-m} X_{(i)}, \quad T(\boldsymbol{X}, F) = \big|\hat\theta - \theta\big|.$

Then $\theta(\hat F_n) = \mathbb{E}_{\hat F_n} X_1^* = \bar X$ and

$$\hat\theta^* = \hat\theta(\boldsymbol{X}^*) = \frac{1}{n-2m} \sum_{i=m+1}^{n-m} X^*_{(i)},$$

where $X^*_{(1)} \le \cdots \le X^*_{(n)}$ denotes the ordered sequence of the n observations in the bootstrap sample $\boldsymbol{X}^* = \big(X_1^*, \ldots, X_n^*\big)$.

The bootstrap method estimates $F_T(y|F)$ by:

$$F_T\big(y\big|\hat F_n\big) = \mathbb{P}_{\hat F_n}\Big(\big|\hat\theta^* - \theta(\hat F_n)\big| \le y\Big) = \mathbb{P}_{\hat F_n}\left(\left|\frac{1}{n-2m}\sum_{i=m+1}^{n-m} X^*_{(i)} - \bar X\right| \le y\right).$$

Monte Carlo simulation procedure for approximating $F_T\big(y\big|\hat F_n\big)$:

1. simulate from $\hat F_n$ a large number, B say, of bootstrap samples

$$\boldsymbol{X}^{*1} = \big(X_1^{*1}, \ldots, X_n^{*1}\big), \quad \ldots \quad , \quad \boldsymbol{X}^{*B} = \big(X_1^{*B}, \ldots, X_n^{*B}\big),$$

2. for each $b = 1, \ldots, B$, calculate $\hat\theta^{*b} = \hat\theta\big(\boldsymbol{X}^{*b}\big) = \dfrac{1}{n-2m}\displaystyle\sum_{i=m+1}^{n-m} X^{*b}_{(i)}$ based on the ordered sequence $X^{*b}_{(1)} \le \cdots \le X^{*b}_{(n)}$,
3. approximate $F_T\big(y\big|\hat F_n\big)$ by

$$B^{-1}\sum_{b=1}^{B} \mathbf{1}\Big\{\big|\hat\theta^{*b} - \theta(\hat F_n)\big| \le y\Big\} = B^{-1}\sum_{b=1}^{B} \mathbf{1}\left\{\left|\frac{1}{n-2m}\sum_{i=m+1}^{n-m} X^{*b}_{(i)} - \bar X\right| \le y\right\}.$$

5.8.6 Exercise: (Example §5.7.6(v))

Use the bootstrap method to estimate the distribution of the estimation error $T = \theta(\hat F_n, \hat G_m) - \theta(F, G)$.

Derive from the bootstrap distribution estimates of the following performance indicators of $\theta(\hat F_n, \hat G_m)$:

bias, mean squared error, mean absolute deviation.