Advanced Mathematical Statistics

Bing-Yi JING

HKUST
majing@ust.hk

May 11, 2021

Contents

Chapter 1

Statistical Models

1.1 Data

Given a set of data,

$$X_i = (X_{i1}, X_{i2}, ..., X_{ip}), \qquad i = 1, ..., n,$$

we wish to draw useful information about the underlying population. For instance, we collect information of n people consisting of

(Age, Gender, Occupation, Income, Address, Health, ...)

The data can be written in a $n \times p$ matrix form:

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & ...... & X_{1p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ X_{n1} & X_{n2} & ...... & X_{np} \end{bmatrix}$$

Some of the questions might include

- what is the distribution of each of the p features?
- what are the relationships between the features?
- can we make prediction of one of the component from the rest?
- can we explain the variations?

To answer these questions, we need first to introduce stastistical models.

1.2 Statistical Models

The essence of statistics is distributions, i.e., statistical models, which are used to describe data. Broadly speaking, models can be classified into

- *Parametric models.*
- *Non-parametric models.*
- *Semi-parametric models.*

(a) Parametric models

Parametric models are a family of distribution functions (d.f.s), indexed by a finite-dimensional parameter θ:

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}$$

where $\Theta \subset R^d$ is the *parameter space*.

Examples

1. A random sample
$$X_i \sim_{i.i.d.} N(\mu, \sigma^2), \qquad i = 1, ..., n,$$
where $\theta = (\mu, \sigma^2)$.

2. Linear regression
$$Y_i \sim_{indep} N(\alpha + \beta^T x_i, \sigma_i^2), \qquad i = 1, ..., n.$$
$\Longleftrightarrow$
$$Y_i = \alpha + \beta^T x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i = 1, ..., n,$$
where $\theta = (\alpha, \beta_1, ..., \beta_p, \sigma^2)$.

Pros and Cons

- simple, easy to understand and interpret

  (once a model is chosen, one only needs to estimate the parameter)
- prone to model mis-specification (not robust) parameter

(b) Non-parametric models

Non-parametric models are families of distributions which can not be characterised by a finite-dimensional parameter.

Examples

1. A random sample
$$X_i\text{'s are i.i.d. with } EX_i^2 < \infty.$$

2. Linear regression
$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, ..., n,$$
where ϵ_i are i.i.d. For instance,
   - $E\epsilon_i = 0, \quad V(\epsilon_i) = \sigma^2$, or
   - median$(\epsilon_i) = 0$, and IQR$(\epsilon_i) = \sigma^2$.

3. Non-linear regression
$$Y_i = g_1(x_{i1}) + ... + g_p(x_{ip}) + \epsilon_i, \qquad \epsilon_i \sim (0, \sigma^2).$$

Pros and Cons

- more flexible, harder to interpret
- prone to overfitting

(c) Semi-parametric models

Semi-parametric models are mixtures of parametric and nonparametric models. On one hand, they could be thought to include parametric and nonparametric models as special cases. On the other hand, they can also be regarded as nonparametric models.

Examples

1. Partial linear model:
$$Y_i = \mathbf{x}_i^T \beta + g(t_i) + \epsilon_i.$$

2. Cox's proportional hazard model:
$$h(t|\mathbf{x_i}) = \lambda(t) e^{\mathbf{x_i}^T \beta}.$$
where
$$h(t|x) = \lim_{\Delta t \to 0} P(T \le t + \Delta t | T > t, \mathbf{X} = \mathbf{x}) = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t, \mathbf{X} = \mathbf{x})}{P(T \ge t, \mathbf{X} = \mathbf{x})}$$

1.3 Model selection

These models all have their advantages and disadvantages, and are useful in different situations. **Exactly which model to use purely depends on the problem at hand.**

1.4 Exponential family

Definition

1. A k**-parameter exponential family** has pdf

$$f_\theta(\mathbf{x}) = \exp\left\{\eta(\theta)\cdot T(\mathbf{x}) - B(\theta)\right\} h(\mathbf{x}),$$

   where

$$\eta(\theta) = (\eta_1(\theta), ..., \eta_k(\theta)), \qquad \text{and} \qquad T(\mathbf{x}) = (T_1(\mathbf{x}), ..., T_k(\mathbf{x})).$$

2. From $\int f_\theta(\mathbf{x})d\mathbf{x} = 1$, we get

$$B(\theta) = \ln\left\{\int \exp\left\{\sum_{j=1}^{k}\eta_j(\theta)T_j(\mathbf{x})\right\} h(\mathbf{x})d\mathbf{x}\right\}$$

3. **Canonical family**: Reparametrizing $\eta_j = \eta_j(\theta)$, we get a canonical family

$$f_\eta(\mathbf{x}) = \exp\left\{\eta\cdot T(\mathbf{x}) - A(\eta)\right\} h(\mathbf{x}),$$

   where $\eta = (\eta_1, ..., \eta_k)$ is called the natural parameter.

4. **The natural parameter space**

$$\Xi = \left\{\eta = (\eta_1, ..., \eta_k) : \int \exp\left\{\sum_{j=1}^{k}\eta_j T_j(\mathbf{x})\right\} h(\mathbf{x})d\mathbf{x} < \infty\right\}$$

   is convex.

5. A canonical family is of **full rank** if

   - it is minimal (i.e., neither T's nor η's satisfy a linear constraint), and
   - Ξ contains a k-dimensional rectangle.

Examples

Exponential family includes

- Gaussian
- Gamma
- Binomial
- Poisson
- etc.

Exponential family DOES NOT include

- Uniform
- Student-t
- Cauchy
- Double-exponential.
- etc.

Examples

EXAMPLE **1.1** *Normal family $N(\mu, \sigma^2)$:*

$$f_\theta(\mathbf{x}) = (\sqrt{2\pi}\sigma)^{-n} e^{-\sum_{i=1}^n (x_i-\mu)^2/(2\sigma^2)} = C(\theta) \exp\left\{ \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 \right\}$$

So $\theta = (\mu, \sigma^2)$ and $k = 2$,

$$\eta(\theta) = (\eta_1, \eta_2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right)$$

and

$$T(x) = (T_1(x), T_2(x)) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 \right).$$

The family is of full rank since Ξ is a lower half space in R^2, containing a 2-dim rectangle.
∎

EXAMPLE **1.2** *Let $X_1, \ldots, X_n \sim N(\theta, \theta^2)$. Then,*

$$f_\theta(\mathbf{x}) = (\sqrt{2\pi}|\theta|)^{-n} e^{-\sum_{i=1}^n (x_i-\theta)^2/(2\theta^2)} = C(\theta) \exp\left\{ \frac{1}{\theta} \sum_{i=1}^n x_i - \frac{1}{2\theta^2} \sum_{i=1}^n x_i^2 \right\}.$$

So $k = 2$,

$$\eta(\theta) = (\eta_1, \eta_2) = \left(\frac{1}{\theta}, -\frac{1}{2\theta^2} \right)$$

and

$$T(x) = (T_1(x), T_2(x)) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 \right).$$

The family is not of full rank since Ξ is a curve in R^2, which DOES NOT contain a 2-dim rectangle. ∎

Exponential family is closed under independent, or marginal, or conditional operations

Theorem **1.1** *If $X, Y \in Exp(k)$, and are independent ($X \perp Y$), then $(X, Y) \in Exp$.* ∎

Theorem **1.2** *If $X_1, ..., X_n \sim_{i.i.d.} Exp(k)$, then $(X_1, ..., X_n) \in Exp(k)$ as*

$$f_\theta(\mathbf{x}) = \prod_{i=1}^{n} f_\theta(x_i) = \exp\left\{\eta(\theta) \cdot \left[\sum_{i=1}^{n} T(\mathbf{x}_i)\right] - nB(\theta)\right\} \prod_{i=1}^{n} h(\mathbf{x}_i). \quad ∎$$

$T(\mathbf{X})$ is also in another exponential family

Theorem **1.3** *If $X \in Exp(k)$, so is $T(X) = (T_1(X), ..., T_k(X))$:*

$$f_T(\mathbf{t}) = \exp\left\{\eta(\theta) \cdot \mathbf{t} - B(\theta)\right\} h_0(\mathbf{t}).$$

Proof. For simplicity, we only prove this for the discrete r.v. For $\mathbf{t} = (t_1, ..., t_k)$,

$$\begin{aligned} P(T = \mathbf{t}) &= \sum_{T(\mathbf{x})=\mathbf{t}} P\left(\mathbf{X} = \mathbf{x}\right) \\ &= \sum_{T(\mathbf{x})=\mathbf{t}} \exp\left\{\eta(\theta) \cdot T(\mathbf{x}) - B(\theta)\right\} h(\mathbf{x}) \\ &= \sum_{T(\mathbf{x})=\mathbf{t}} \exp\left\{\eta(\theta) \cdot \mathbf{t} - B(\theta)\right\} h(\mathbf{x}) \\ &= \exp\left\{\eta(\theta)\mathbf{t} - B(\theta)\right\} h_0(\mathbf{t}), \end{aligned}$$

where $h_0(\mathbf{t}) = \sum_{T(\mathbf{x})=\mathbf{t}} h(\mathbf{x})$. ∎

Calculating Moments of Exponential Families

THEOREM **1.4** *Let* $X \sim f_\eta(\mathbf{x}) = \exp\{\eta \cdot T(\mathbf{x}) - A(\eta)\} h(\mathbf{x})$. *The moment-generating function (m.g.f.) of* $T(X)$*:*

$$M(s) := Ee^{s \cdot T(X)} = \exp[A(s+\eta) - A(\eta)]. \quad \blacksquare$$

Proof. Note

$$\begin{aligned} M(s) &= Ee^{s\cdot T(X)} = \int \dots \int e^{s\cdot T(x)} f_\eta(\mathbf{x}) dx = \int \dots \int e^{(s+\eta)\cdot T(x) - A(\eta)} h(x) dx \\ &= e^{A(s+\eta)-A(\eta)} \int \dots \int e^{(s+\eta)^\tau T(x) - A(s+\eta)} h(x) dx \\ &= e^{A(s+\eta)-A(\eta)}. \quad \blacksquare \end{aligned}$$

So its cumulant generating function (c.g.f.) is $C(s) =: \log M(s) = A(s+\eta) - A(\eta)$. Hence

$$E(T(X)) = C'(s)|_{s=0} = A'(\eta), \qquad Var(T(X)) = C''(s)|_{s=0} = A''(\eta).$$

It is clear that exponential family requires the existence of all moments (very thin-tailed distributions). This excludes cases like Student-t, Cauchy, etc.

Location-scale family

Definition **1.1** *Let $X \sim F(x)$ be any c.d.f.. Then we define*

$$\begin{aligned}
&\textit{Location family:} && \{F(x-a) : a \in R\}; \\
& && (i.e., \;\; X + a \sim F(x-a)) \\
&\textit{Scale family:} && \left\{F\left(\frac{x}{b}\right) : b > 0\right\}; \\
& && (i.e., \;\; bX \sim F(x/b)) \\
&\textit{Location-scale family:} && \left\{F\left(\frac{x-a}{b}\right) : \mu \in R, \sigma > 0\right\}; \\
& && (i.e., \;\; a + bX \sim F((x-a)/b)).
\end{aligned}$$

where a is called the location parameter, $b > 0$ the scale parameter.

Let $f(x)$ be the corresponding p.d.f. to $F(x)$. Then we have

$$\begin{aligned}
&\textit{Location family:} && \{f(x-a) : a \in R\}; \\
&\textit{Scale family:} && \left\{\frac{1}{\sigma} f\left(\frac{x}{b}\right) : b > 0\right\}; \\
&\textit{Location-scale family:} && \left\{\frac{1}{b} f\left(\frac{x-a}{b}\right) : a \in R, b > 0\right\}.
\end{aligned}$$

Note a and b are NOT necessarily the mean and standard deviation, which may not even exist, e.g., for $F =$ Cauchy.

Lemma **1.1** *If $U_1, ..., U_n \sim Unif(0,1)$, then*

$$U_{(k)} \sim \;\; Beta(k, n-k+1), \;\; where \; 1 \le k \le n$$

with $EU_{(k)} = \frac{k}{n+1}$. ∎

Example **1.3** *It follows easily from the above that*

1. *If $X_1, ..., X_n \sim U(0,\theta)$ (a scale family), then $X_i/\theta \sim Unif(0,1)$, and hence*

$$EX_{(k)} = \frac{k}{n+1}\theta.$$

2. *If $Y_1, ..., Y_n \sim U(\theta, \theta+1)$ (a location family), then $Y_i - \theta \sim Unif(0,1)$, and hence*

$$EY_{(k)} = \theta + \frac{k}{n+1}.$$

Chapter 2

Principles of Data Reduction

Suppose a random sample

$$\mathbf{X} =: (X_1, ..., X_n) \sim_{iid} F, \qquad X_k \in R^p.$$

We wish to estimate F.

1. As a first step in exploratory data analysis (EDA), we can look at various plots.

   (a) Histograms of k-th component, $k = 1, ..., p$
   How the data is distributed in each dimension.

   (b) Scatter plots of two components
   How each pair is related.

2. We can also use some summary statistics to have a very good description of F (non-parametric approach).

   For example,

   (a) Summary using sample q-quantiles: $Q_p = F_n^{-1}(p/n)$

   - The 0-quantile = the mimimum
   - The only 2-quantile = the median
   - The 3-quantiles = tertiles or terciles
   - The 4-quantiles = quartiles
   - The 5-quantiles = quintiles
   - The 6-quantiles = sextiles
   - The 7-quantiles = septiles
   - The 8-quantiles = octiles
   - The 10-quantiles = deciles
   - The 12-quantiles = duo-deciles or dodeciles
   - The 16-quantiles = hexadeciles
   - The 20-quantiles = ventiles, vigintiles, or demi-deciles
   - The 100-quantiles = percentiles
   - The 1-quantile is called maximum

   Comments:

   - Of course, the bigger the p is, the more information, the more storage required.

- The most commonly used is the stem-leaf plot ($q = 4$):

  $$T = (Q_0, Q_1, Q_2, Q_3, Q_4) = (X_{(1)}, F_n^{-1}(1/4), F_n^{-1}(1/2), F_n^{-1}(3/4), X_{(n)})$$

  (minimum, first quartile, median, third quartile, maximum)

(b) Summary using sample moments

$$T = \{\hat{\mu}_1, \cdots, \hat{\mu}_p\} = \{\hat{\gamma}_k, \cdots, \hat{\gamma}_p\}$$

where

$$\hat{\mu}_k = n^{-1} \sum_{i=1}^{n} X_i^k, \qquad or \qquad \hat{\gamma}_k = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^k$$

are sample k-th moments and central moments.

Comments:

- $F(x) \Longleftrightarrow M(t) = Ee^{tX}$, moment generating function (m.g.f.) if $M(t)$ exists.
  $M(t)$ (hence F) can be approximated by its first p-th sample moments.
  In practice, $p \leq 10$ is enough.
- Fatal problems:
  moments are very sensitive to outliers, and hence not reliable.

3. Suppose that we have some fair ideas about the family of the distribution (parametric approach)

   $$\mathbf{X} =: (X_1, ..., X_n) \sim_{iid} F_\theta,$$

   where F is known, and θ is the unknown signature or the I.D. of F.

   Then our objective is to do inference on θ, or more generally $g(\theta)$ (estimation or hypothesis testing).

   To do this, it turns out that we can reduce the original data substantially to a few summary statistics (sufficient statistics). For example,

   - If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, we only need to keep

     $$T = (\bar{X}, S^2),$$

     and throw away the rest of the data.
     T is called the sufficient statistics.

In this chapter, we will study sufficiency and related topics.

- Sufficiency principle

  One can reduce $\mathbf{X}$ to a "sufficient" one $T(\mathbf{X})$ without losing information on θ.

  That is, $T(\mathbf{X})$ contains all information about θ. (If we cannot reduce it any further, we get the minimal sufficient statistics.)

- Auxiliary statistics

  They contain no information about θ. Properly chosen, this proves to be very useful in (conditional) statistical inference. This is opposite to sufficiency.

- Basu Theorem

  Sufficient and auxiliary statistics are independent for a complete family.

2.1 Sufficiency

Definition **2.1** *$T(\mathbf{X})$ is sufficient for θ if the distribution of $\mathbf{X}$ conditional on $T(\mathbf{X})$ does not depend on θ.*

(That is, once T is known, the data $\mathbf{X}$ contains no more information about θ.)

Example **2.1**

1. *$T(\mathbf{X}) = \mathbf{X}$ is always sufficient.*

   Proof. $P(\mathbf{X}=\mathbf{x}|\mathbf{T}=\mathbf{t}) = P(\mathbf{X}=\mathbf{x}|\mathbf{X}=\mathbf{t}) = I\{\mathbf{x}=\mathbf{t}\}$ *does not depend on θ.*

2. *$\mathbf{T} = (X_{(1)}, ..., X_{(n)})$ is sufficient, if F_θ is continuous (i.e., no ties a.s.).*

   Proof. *Now*

$$\begin{aligned} P(\mathbf{X}=\mathbf{x}|\mathbf{T}=\mathbf{t}) &= P((X_1,...,X_n)=(x_1,...,x_n)|(X_{(1)},...,X_{(n)})=(t_1,...,t_n)) \\ &= (1/n!)I\{\{x_1,...,x_n\}=\{t_1,...,t_n\}\}, \end{aligned}$$

   not depending on θ.

   (Continuity of d.f. ensures no ties. One can remove the assumption by the Factorization Theorem later.)

3. *$T = \sum_{i=1}^n X_i$ if $\mathbf{X} \sim Bernoulli(p)$.*

   Proof. *$T \sim Bin(n,p)$.*

$$P(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=t) = \frac{P(\mathbf{X}=\mathbf{x}, T(\mathbf{X})=t)}{P(T(\mathbf{X})=t)}$$

   - *If $\sum_{i=1}^n x_i \neq t$, $P(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=t) = 0$.*
   - *If $\sum_{i=1}^n x_i = t$, then $P(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=t)$ equals*

$$\begin{aligned} P(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=t) &= \frac{P(\mathbf{X}=\mathbf{x}, T(\mathbf{X})=t)}{P(T(\mathbf{X})=t)} && (1.1) \\ &= \frac{P(\mathbf{X}=\mathbf{x})}{P(T(\mathbf{X})=t)} && (1.2) \\ & \quad \textit{as } (\{\mathbf{X}=\mathbf{x}\} \subset \{T(\mathbf{X})=t\}) && (1.3) \\ &= \frac{\prod_{i=1}^n P(X_i=x_i)}{P(\sum_{i=1}^n X_i = t)} = \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t}p^t(1-p)^{n-t}} \\ &= \frac{1}{\binom{n}{t}}. \end{aligned}$$

   Hence, T is sufficient.

2.2 Factorization Theorem

An easy way to find sufficient statistics is offered by the Factorization Theorem.

THEOREM **2.1 (Factorization Theorem)** $T(\mathbf{X})$ *is sufficient for* θ *iff*

$$f_\theta(\mathbf{x}) = g\left(T(\mathbf{x}), \theta\right)\ h(\mathbf{x})\,. \quad \blacksquare$$

(Therefore, sufficiency makes joint pdf or likelihood as simple as possible.)

Proof. We prove it for the discrete case only. Now $P\left(\mathbf{X} = \mathbf{x} | T = t\right) = \dfrac{P\left(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t\right)}{P\left(T(\mathbf{X}) = t\right)}$.

- Suppose $f_\theta(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = g\left(T(\mathbf{x}), \theta\right) h\left(\mathbf{x}\right)$.
  If $T(\mathbf{x}) \neq t$, then $P\left(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t\right) = 0$.
  If $T(\mathbf{x}) = t$, then similarly to (1.1)-(1.3), $P\left(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t\right)$ becomes

  $$\frac{P\left(\mathbf{X} = \mathbf{x}\right)}{P\left(T(\mathbf{X}) = t\right)} = \frac{P\left(\mathbf{X} = \mathbf{x}\right)}{\sum_{T(\mathbf{x})=t} P\left(\mathbf{X} = \mathbf{x}\right)} = \frac{g(t,\theta) h(\mathbf{x})}{g(t,\theta)\sum_{T(\mathbf{x})=t} h(\mathbf{x})} = \frac{h\left(\mathbf{x}\right)}{\sum_{T(\mathbf{x})=t} h(\mathbf{x})},$$

  Hence T is sufficient.

- Suppose $T(\mathbf{X})$ is sufficient. Then

  $$\begin{aligned} f_\theta\left(\mathbf{x}\right) &= P\left(\mathbf{X} = \mathbf{x}\right) = P\left(\mathbf{X} = \mathbf{x}, T\left(\mathbf{X}\right) = T\left(\mathbf{x}\right)\right) \\ &\quad\ (as\ \{\mathbf{X} = \mathbf{x}\} \subset \{T\left(\mathbf{X}\right) = T\left(\mathbf{x}\right)\}) \\ &= P\left(T(\mathbf{X}) = T(\mathbf{x})\right) P\left(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x})\right) \\ &=: g(T(\mathbf{x}), \theta)\ h(\mathbf{x}), \end{aligned}$$

  where sufficiency implies that $h(\mathbf{x})$ is free of θ. $\blacksquare$

EXAMPLE **2.2** $f_\theta(\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i)$.

1. $\mathbf{X} \sim$ *Bernoulli*(p)*, then*

   $$f_\theta(\mathbf{x}) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

   So $T = \sum_{i=1}^n X_i$ *is sufficient.*

2. $\mathbf{X} \sim U[0, \theta]$*, then*

   $$f_\theta(\mathbf{x}) = \prod_{i=1}^n \theta^{-1} I\{0 \le x_i \le \theta\} = \theta^{-n} I\{0 \le x_{(1)} \le x_{(n)} \le \theta\} = \theta^{-n} I\{0 \le x_{(1)}\} I\{x_{(n)} \le \theta\}.$$

   So $T = X_{(n)}$ *is sufficient.*

2.3 Minimal Sufficiency

Minimal sufficient statistics provide the maximal reduction to the data.

DEFINITION **2.2** *$T(X)$ is* **minimal sufficient** *if, given any other sufficient statistic $S(X)$, $T(X)$ is a function of $S(X)$.*

THEOREM **2.2** *Let $T = l(S)$.*

1. *If T is sufficient, So is S.*
2. *l is 1-1. Then T is (minimal) sufficient iff S is (minimal) sufficient.*

Proof.

1. $f_\theta(x) = g_\theta(T(x))h(x) = g_\theta(l(S(x)))h(x) \Longrightarrow S(X)$ is sufficient (by Factorization Theorem).
2. The proof follows from (i) and the definition. ∎

2.3.1 Examples

EXAMPLE **2.3** *Clearly,* $(X_{(1)}, ..., X_{(n)})$ *is sufficient, so the original data* $(X_1, ..., X_n)$ *is never minimal sufficient.*

EXAMPLE **2.4** $\mathbf{X} \sim$ *$U(\theta, \theta+1)$, i.e., $f_\theta(x) = I\{\theta < x < \theta+1\}$. Then*

$$\begin{aligned} f_\theta(\mathbf{x}) &= \prod_{i=1}^{n} I\{\theta < x_i < \theta+1\} = I\{\theta < x_{(1)} < x_{(n)} < \theta+1\} \\ &= I\{x_{(n)} - 1 < \theta < x_{(1)}\}. \end{aligned} \tag{3.4}$$

By the factorization theorem, $T = \left(X_{(1)}, X_{(n)}\right)$ is sufficient.

To show minimal sufficiency, suppose that $S(\mathbf{X})$ is also sufficient. By the factorization theorem,

$$f_\theta\left(\mathbf{x}\right) = g\left(S(\mathbf{x}), \theta\right) h\left(\mathbf{x}\right).$$

We only consider those x s.t. $h(x) > 0$, so all conclusions hold a.s. From (3.4),

$$\begin{aligned} x_{(1)} &= \sup\{\theta : f_\theta\left(\mathbf{x}\right) > 0\} = \sup\{\theta : g\left(S(\mathbf{x}), \theta\right) > 0\}, \\ x_{(n)} &= \inf\{\theta : f_\theta\left(\mathbf{x}\right) > 0\} + 1 = \inf\{\theta : g\left(S(\mathbf{x}), \theta\right) > 0\} + 1. \end{aligned}$$

So $T\left(\mathbf{x}\right) = \left(x_{(1)}, x_{(n)}\right)$ is a function of $S(\mathbf{x})$. Hence, $T\left(\mathbf{X}\right)$ is minimal sufficient. ∎

An equivalent minimal sufficient statistic is $T' = (R, M)$, where

$$R = X_{(n)} - X_{(1)}, \qquad \text{and} \qquad M = (X_{(n)} + X_{(1)})/2$$

are the range and mid-range point, respectively.

2.3.2 A simple rule to find minimal sufficient statistics

Sufficient statistics can be obtained easily by Factorization theorem. Then one can apply the following to verify minimal sufficiency.

Theorem **2.3** *Suppose*

1. $T(\mathbf{X})$ *is sufficient.*
2. $\dfrac{f_\theta(\mathbf{x})}{f_\theta(\mathbf{y})}$ *is free of* θ, $\forall \mathbf{x}, \mathbf{y} \qquad \Longrightarrow \qquad T(\mathbf{x}) = T(\mathbf{y})$.

Then $T(\mathbf{X})$ *is minimal sufficient.*

Proof. $\widetilde{T}(X)$ is any sufficient statistics. For each $x \in \mathcal{X}$, we have a pair $(\widetilde{T}(\mathbf{x}), T(\mathbf{x}))$. To show that $T(\mathbf{x})$ is a function of $\widetilde{T}(\mathbf{x})$, it suffices to prove that

$$\widetilde{T}(\mathbf{x}) = \widetilde{T}(\mathbf{y}) \qquad \Longrightarrow \qquad T(\mathbf{x}) = T(\mathbf{y}). \tag{3.5}$$

By the Factorization Theorem, $f_\theta(\mathbf{x}) = \widetilde{g}_\theta(\widetilde{T}(\mathbf{x}))\widetilde{h}(\mathbf{x})$. Now if $\widetilde{T}(\mathbf{x}) = \widetilde{T}(\mathbf{y})$, then

$$\frac{f_\theta(\mathbf{x})}{f_\theta(\mathbf{y})} = \frac{\widetilde{g}_\theta(\widetilde{T}(\mathbf{x}))\widetilde{h}(\mathbf{x})}{\widetilde{g}_\theta(\widetilde{T}(\mathbf{y}))\widetilde{h}(\mathbf{y})} = \frac{\widetilde{h}(\mathbf{x})}{\widetilde{h}(\mathbf{y})}, \quad \text{free of } \theta.$$

By assumption, we get $T(\mathbf{x}) = T(\mathbf{y})$, which proves (3.5), So $T(\mathbf{x})$ is minimal. ∎

2.3.3 Examples

EXAMPLE **2.5** *Find minimal sufficient statistics for the following cases:*

1. $\mathbf{X} \sim U(\theta, \theta+1)$
2. $\mathbf{X} \sim U(0, \theta)$
3. $\mathbf{X} \sim U(0, \theta)$ *with* $\theta \geq 1$.

Solution.

1. *Since* $f_\theta(\mathbf{x}) = I\{x_{(n)} - 1 < \theta < x_{(1)}\}$, *so* $T(\mathbf{x}) = \left(x_{(1)}, x_{(n)}\right)$ *is sufficient (Factorisation Theorem).*

   Next, if

$$r(\theta) \equiv \frac{f_\theta(\mathbf{x})}{f_\theta(\mathbf{y})} = \frac{I\{x_{(n)} - 1 < \theta < x_{(1)}\}}{I\{y_{(n)} - 1 < \theta < y_{(1)}\}}.$$

   is free of θ, *we must have* $T(\mathbf{x}) = T(\mathbf{y})$.

   Thus, $T = \left(X_{(1)}, X_{(n)}\right)$ *is minimal sufficient.* ∎

   (For example, if $x_{(1)} > y_{(1)}$ *and* $x_{(n)} < y_{(n)}$, *then* $r(\theta)$ *takes values* 0, 1 *and* ∞, *depending where* θ *is, so* $r(\theta)$ *is not free of* θ. *)*

2. *First* $f_\theta(x) = \theta^{-1} I\{0 < x < \theta\}$, *then* $f_\theta(\mathbf{x}) = \theta^{-n} I(0 < x_{(1)} \leq x_{(n)} < \theta) = \theta^{-n} I(0 < x_{(1)}) I(x_{(n)} < \theta)$. *Hence* $T(\mathbf{x}) = x_{(n)}$ *is sufficient.*

   Now if

$$r(\theta) \equiv \frac{f_\theta(\mathbf{x})}{f_\theta(\mathbf{y})} = \frac{I(0 < x_{(1)}) I(x_{(n)} < \theta)}{I(0 < y_{(1)}) I(y_{(n)} < \theta)}.$$

   is free of θ, *we must have* $T(\mathbf{x}) = T(\mathbf{y})$.

   Thus $T = X_{(n)}$ *is minimal sufficient.*

3. *First the pdf of* X *is* $f_\theta(x) = \theta^{-1} I\{0 < x < \theta\} I\{\theta > 1\}$. *Hence,*

$$\begin{aligned} f_\theta(\mathbf{x}) &= \theta^{-n} I(0 < x_{(1)} \leq x_{(n)} < \theta) I\{\theta > 1\} \\ &= \theta^{-n} I(0 < x_{(1)}) I(x_{(n)} < \theta) I\{\theta > 1\} \\ &= \theta^{-n} I(x_{(1)} > 0) I(\widetilde{T}(\mathbf{x}) < \theta). \end{aligned}$$

   Clearly, $\widetilde{T}(\mathbf{x}) = \widetilde{\mathbf{T}}(\mathbf{y})$ *is sufficient.*

   Now if

$$r(\theta) \equiv \frac{f_\theta(\mathbf{x})}{f_\theta(\mathbf{y})} = \frac{I(x_{(1)} > 0)}{I(y_{(1)} > 0)} \frac{I(\widetilde{T}(\mathbf{x}) < \theta)}{I(\widetilde{T}(\mathbf{y}) < \theta))}$$

   is free of θ, *we must have* $\widetilde{T}(\mathbf{x}) = \widetilde{\mathbf{T}}(\mathbf{y})$.

   Thus, $\widetilde{T}(\mathbf{X}) = \max\{X_{(n)}, 1\}$ *is minimal sufficient.* ∎

It is intuitively clear why $\widetilde{T}(\mathbf{X}) = \max\{X_{(n)}, 1\}$, rather than $X_{(n)}$, is a minimal sufficient statistic: if $X_{(n)} < 1$, say $X_{(n)} = 1/2$, then $1/2$ would not be a good estimator of θ since we know $\theta > 1$. A better choice would be 1. Of course, if $X_{(n)} \geq 1$, then $X_{(n)}$ would be a very good estimator of θ.

2.4 Minimal Sufficiency for Exponential Family

For full rank exponential families, finding minimal sufficient statistics is very easy.

THEOREM **2.4** *If P is in an exponential family of full rank with p.d.f.'s given by*

$$f_\theta(\mathbf{x}) = \exp\left\{\sum_{i=1}^{k} \eta_j(\theta)T_j(\mathbf{x}) - A(\eta)\right\} h(\mathbf{x}).$$

Then $T(\mathbf{X}) = (T_i(\mathbf{X}), ..., T_k(\mathbf{X}))$ is minimal sufficient for θ.

Proof. Sufficiency follows from the Factorization Theorem. To prove minimality, we denote $a_j \equiv a_j(\mathbf{x}, \mathbf{y}) = T_j(\mathbf{x}) - T_j(\mathbf{y})$, $1 \le j \le n$. Then

$$\frac{f_\theta(\mathbf{x})}{f_\theta(\mathbf{y})} = \frac{\exp\left\{\sum_{j=1}^{k} \eta_j(\theta)T_j(\mathbf{x})\right\}}{\exp\left\{\sum_{j=1}^{k} \eta_j(\theta)T_j(\mathbf{y})\right\}} \frac{h(\mathbf{x})}{h(\mathbf{y})} = \exp\left\{\sum_{j=1}^{k} a_j\eta_j(\theta)\right\} \frac{h(\mathbf{x})}{h(\mathbf{y})}.$$

is free of θ iff $\sum_{i=1}^{k} a_j\eta_j(\theta)$ is free of θ, i.e.,

$$\sum_{j=1}^{k} a_j\eta_j(\theta) = C(\mathbf{x}, \mathbf{y}),$$

Fix $\theta_0 \in \Theta$, then $\sum_{j=1}^{k} a_j\eta_j(\theta_0) = C(\mathbf{x}, \mathbf{y})$. Then

$$0 = \sum_{j=1}^{k} a_j [\eta_j(\theta) - \eta_j(\theta_0)] = \mathbf{a} \cdot \widetilde{\eta}(\theta) \tag{4.6}$$

where $\mathbf{a} = (a_1, ..., a_k)'$ and $\widetilde{\eta}(\theta) = (\eta_1(\theta) - \eta_1(\theta_0), ..., \eta_k(\theta) - \eta_k(\theta_0))'$.

Since $f_\theta(\mathbf{x})$ is of full rank k, there exist $\theta_1, ..., \theta_k$ such that

$$\{\widetilde{\eta}(\theta_1), ..., \widetilde{\eta}(\theta_k)\}$$

are linearly independent. Then it follows from (4.6) that $\mathbf{a} = 0$. That is, $T_i(\mathbf{x}) = T_i(\mathbf{y})$ for all $1 \le i \le k$. ∎

Examples

Example **2.6** *Let* $\mathbf{X} \sim N(\mu, \sigma^2)$*, find a minimal sufficient statistics for* $\theta = (\mu, \sigma^2)$*.*

Solution. $f_\theta(x) = (\sqrt{2\pi}\sigma)^{-1} e^{-(x-\mu)^2/(2\sigma^2)}$. Therefore,

$$f_\theta(\mathbf{x}) = (\sqrt{2\pi}\sigma)^{-n} e^{-\sum_{i=1}^n (x_i-\mu)^2/(2\sigma^2)} = C(\theta) \exp\left\{ \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 \right\}$$

So $k = 2$,

$$\eta(\theta) = (\eta_1, \eta_2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right)$$

and

$$T(x) = (T_1(x), T_2(x)) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 \right).$$

Clearly, the natural parameter space Ξ is a half space in R^2, which contains a 2-dim rectangle. By Theorem 2.4, $T(X)$ is a minimal sufficient statistic. ∎

Example **2.7** *If* $X_1, \ldots, X_n \sim N(\theta, \theta^2)$*, find a minimal sufficient statistics for* θ*.*

Solution. Recall that

$$f_\theta(\mathbf{x}) = (\sqrt{2\pi}|\theta|)^{-n} e^{-\sum_{i=1}^n (x_i-\theta)^2/(2\theta^2)} = C(\theta) \exp\left\{ \frac{1}{\theta} \sum_{i=1}^n x_i - \frac{1}{2\theta^2} \sum_{i=1}^n x_i^2 \right\}$$

with sufficient statistics $T(x) = (T_1(x), T_2(x)) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)$. Since it is a curved exponential family (not a full rank), one can not use the very last theorem to find the minimal sufficient statistic. However, we can use the simple rule given earlier. Now

$$r(\theta) \equiv \frac{f_\theta(\mathbf{x})}{f_\theta(\mathbf{y})} = \exp\left\{ \frac{1}{\theta} \left(T_1(\mathbf{x}) - T_1(\mathbf{y})\right) - \frac{1}{2\theta^2} \left(T_2(\mathbf{x}) - T_2(\mathbf{y})\right) \right\}$$

is free of θ iff

$$\frac{1}{\theta} \left(T_1(\mathbf{x}) - T_1(\mathbf{y})\right) - \frac{1}{2\theta^2} \left(T_2(\mathbf{x}) - T_2(\mathbf{y})\right) = C\left(\mathbf{x}, \mathbf{y}\right) = 0,$$

where the last equality is obtained by letting $\theta \to \infty$ on both sides. Again, since θ is arbitrary, we get (by taking $\theta = 1$ and $\theta = 2$)

$$T_1(\mathbf{x}) - T_1(\mathbf{y}) = 0, \qquad T_2(\mathbf{x}) - T_2(\mathbf{y}) = 0.$$

Therefore, $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is a minimal sufficient statistic. ∎

2.5 Minimal Sufficiency for Location Family

If $\mathbf{X} = (X_1, \ldots, X_n) \sim f(x-\theta)$, then

$$f_\theta(\mathbf{x}) = \Pi_{i=1}^n f(x_i - \theta) = \Pi_{i=1}^n f(x_{(i)} - \theta).$$

Then $T(\mathbf{X}) = \left(X_{(1)}, ..., X_{(n)}\right)$ is sufficient. Further reduction may or may not be possible.

Further reduction is possible

1. If $\mathbf{X} \sim N(\mu, 1)$ with pdf $\phi(x-\mu)$, $\bar{X}$ is minimal sufficient (also in exponential family).
2. If $\mathbf{X} \sim \exp(\theta, 1)$ with pdf $f(x-\theta) = e^{-(x-\theta)} I(x > \theta)$, then $X_{(1)}$ is minimal sufficient. (Similar proof to $U(0, \theta)$ example.)
3. If $\mathbf{X} \sim U(\theta, \theta+1)$, $\left(X_{(1)}, X_{(n)}\right)$ is minimal sufficient (shown earlier).

Further reduction is impossible

$T(\mathbf{X}) = \left(X_{(1)}, ..., X_{(n)}\right)$ is minimal sufficient if $\mathbf{X} \sim f_\theta(x) = f(x-\theta)$, where

1. **(Double exponential)** $f(x) = \frac{1}{2} e^{-|x|}$ (all moments exist, and $f_\theta(x)$ is not differentiable at $\theta = x$).
2. **(Cauchy)** $f(x) = \frac{1}{\pi} \frac{1}{(1+x^2)}$ (the mean does not exist).
3. **(Logistic)** $f(x) = \frac{e^{-x}}{(1+e^{-x})^2}$ ($\mu = EX_1$ exists).

Note that $f(-x) = f(x)$ in all cases.

Proof. We only prove the double exponential case here. Now

$$r(\theta) = \frac{f_\theta(y)}{f_\theta(x)} = \frac{\exp\{\sum_{i=1}^n |x_i - \theta|\}}{\exp\{\sum_{i=1}^n |y_i - \theta|\}}.$$

If $r(\theta)$ is free of θ, then

$$\lim_{\theta \to \infty} r(\theta) = \frac{\exp\{\sum_{i=1}^n (\theta - x_i)\}}{\exp\{\sum_{i=1}^n (\theta - y_i)\}} = \exp\{\sum_{i=1}^n (y_i - x_i)\} =: A,$$
$$\lim_{\theta \to -\infty} r(\theta) = \frac{\exp\{\sum_{i=1}^n (x_i - \theta)\}}{\exp\{\sum_{i=1}^n (y_i - \theta)\}} = \exp\{\sum_{i=1}^n (x_i - y_i)\} = A^{-1}.$$

So $A = 1/A$, hence, $A = 1$ (as $A > 0$). That is,

$$r(\theta) = \frac{\exp\{\sum_{i=1}^n |x_i - \theta|\}}{\exp\{\sum_{i=1}^n |y_i - \theta|\}} = r(\infty) = r(-\infty) = 1.$$

Hence,

$$g(\theta) =: \sum_{i=1}^n |x_i - \theta| = \sum_{i=1}^n |y_i - \theta|$$

The value minimizing the above is the sample median. Thus, the two data sets X's and Y's have the same medians. Consider two cases:

- $n = 2m + 1$ is odd

  By minimizing w.r.t. θ, we see that x's and y's have the same sample medians, i.e.,

  $$x_{(m+1)} = y_{(m+1)}.$$

  Therefore, we can remove this common value by considering $g(\theta) - |x_{(m+1)} - \theta| = g(\theta) - |y_{(m+1)} - \theta|$. Once this is done, we have an even number of observations left. Thus, it suffices to consider the case where n is even, as we shall do below.

- $n = 2m$ is even

  Any value in $[x_{(m)}, x_{(m+1)}]$ is a median of x's, which minimizes $g(\theta)$ w.r.t. θ.

  Any value in $[y_{(m)}, y_{(m+1)}]$ is a median of y's, which minimizes $g(\theta)$ w.r.t. θ.

  These two sets must be the same, thus $x_{(m)} = y_{(m)}$ and $x_{(m+1)} = y_{(m+1)}$.

  Removing these two points, we still have even number of points left.

  Continuing doing this, we get $x_{(k)} = y_{(k)}$ for all $k = 1, ..., n$. ∎

Remarks.

- For double exponential d.f., the sample median is better than the sample mean.

  The sample median is in fact the **MLE**, and hence asymptotically efficient (optimal).

- For Cauchy d.f., the sample mean is a very bad choice. The sample median is better, but it is not **MLE**.

- For logistic d.f., the sample median is better than the sample mean, although neither is **MLE** (to be discussed later).

2.6 Ancillary Statistics

Definition **2.3** *$S(X)$ is ancillary if its distribution does not depend on θ.*

Example **2.8**

1. *If $X_1, \dots, X_n \sim F(x-\theta)$ (a location family), the range, $R = X_{(n)} - X_{(1)}$, is ancillary.*

   Proof. *$Z_i = X_i - \theta \sim F(x)$ free of θ. Then $R = Z_{(n)} - Z_{(1)}$ is ancillary.*

2. *If $X_1, \dots, X_n \sim F(x/\sigma)$ (a scale family), $(X_1/X_n, \cdots, X_{n-1}/X_n)$ is ancillary.*

   Proof. *$Z_i = X_i/\sigma \sim F(x)$ free of θ. Then $(X_1/X_n, \cdots, X_{n-1}/X_n) = (Z_1/Z_n, \cdots, Z_{n-1}/Z_n)$ is ancillary.*

Ancillary statistics are useful in conditional inference and also in hypotheses testing.

2.7 Complete Statistics

Definition **2.4** *Let $f_\theta(t)$ be a family of pdf's for a statistic T.*

1. T is *complete* for θ if

$$E_\theta g(T) = 0 \qquad \Longrightarrow \qquad g(T) = 0 \;\; a.s. \qquad \forall\theta.$$

2. T is *boundedly complete* if the previous statement holds for all bounded g. ∎

Clearly, completeness implies bounded completeness, but not vice versa.

2.7.1 Examples

EXAMPLE **2.9** *If* $\mathbf{X} \sim$ *Bernoulli*(p)*,* $T = \sum_{i=1}^n X_i$ *is complete.*

Proof. $T \sim Bin(n, p)$. If $E_p g(T) = 0$, i.e.,

$$E_p g(T) = \sum_{t=0}^n \binom{n}{t} p^t q^{n-t} g(t) = q^n \sum_{t=0}^n \binom{n}{t} r^t g(t) = 0, \qquad \text{where } r = p/q > 0,$$

we have $g(t) = 0$ for $t = 0, 1, \cdots, n$. That is, $g(T) = 0$ a.s. ∎

EXAMPLE **2.10** *If* $\mathbf{X} \sim U(0, \theta)$*,* $T = X_{(n)}$ *is complete.*

Proof. $F(x) = (x/\theta)I(0 < x < \theta) + I(x \geq \theta)$, $P(T \leq t) = F(t)^n$. If $E_\theta g(T) = 0$ for $\theta > 0$, then

$$\int_0^\theta g(t) n t^{n-1} \theta^{-n} dt = 0, \qquad \Longrightarrow \qquad \int_0^\theta g(t) t^{n-1} dt = 0.$$

Differentiating w.r.t. θ, one gets $g(\theta)\theta^{n-1} = 0$, i.e., $g(\theta) = 0$ for all $\theta > 0$. ∎

2.7.2 Exponential family of full rank is complete

THEOREM **2.5** *If $\mathbf{X} \sim Exp(k)$ of full rank:*

$$f_\eta(\mathbf{x}) = \exp\left\{\sum_{i=1}^{k} \eta_j T_j(\mathbf{x}) - A(\eta)\right\} h(\mathbf{x}).$$

Then $T(\mathbf{X}) = (T_1(X), ..., T_k(X))$ is complete (and minimal sufficient).

Proof. If $E_\eta[g(T)] = 0$, recalling $f_T(\mathbf{t}) = \exp\{\mathbf{t}\eta' - A(\eta)\} h_0(\mathbf{t})$ for $\mathbf{t} = (t_1, ..., t_k)$ and $\eta = (\eta_1, ..., \eta_k)$, we have

$$E_\eta[g(T)] = \int g(\mathbf{t}) f_T(\mathbf{t}) d\mathbf{t} = \int g(\mathbf{t}) \exp\{\mathbf{t}\eta' - A(\eta)\} h_0(\mathbf{t}) d\mathbf{t} = \int g(\mathbf{t}) \exp\{\mathbf{t}\eta' - A(\eta)\}\, d\lambda(\mathbf{t}) = 0,$$

where $\lambda(A) = \int_A h_0(\mathbf{t}) d\mathbf{t}$ is a measure on $(R^k, \mathcal{B}^k)$. Let η_0 be an interior point in Ξ. Then

$$\int g^+(\mathbf{t}) \exp\{\mathbf{t}\eta' - A(\eta)\}\, d\lambda = \int g^-(\mathbf{t}) \exp\{\mathbf{t}\eta' - A(\eta)\}\, d\lambda \tag{7.7}$$

for all $\eta \in N(\eta_0)$, where $N(\eta_0) = \{\eta : ||\eta - \eta_0|| < \epsilon\}$ for some $\epsilon > 0$. In particular,

$$\int g^+(\mathbf{t}) \exp\{\mathbf{t}\eta_0' - A(\eta_0)\}\, d\lambda\mathbf{t} = \int g^-(\mathbf{t}) \exp\{\mathbf{t}\eta_0' - A(\eta_0)\}\, d\lambda\mathbf{t} = C \geq 0.$$

If $C = 0$, then $g^+(\mathbf{t}) = g^-(\mathbf{t}) = 0$ a.e. λ, i.e., $g(\mathbf{t}) = 0$ a.e. λ. If $C > 0$, then $C^{-1} g^+(\mathbf{t}) \exp\{\mathbf{t}\eta_0' - A(\eta_0)\}$ and $C^{-1} g^-(\mathbf{t}) \exp\{\mathbf{t}\eta_0' - A(\eta_0)\}$ are p.d.f. with respect to λ. However, (7.7) implies that

$$\int \exp\{\mathbf{t}(\eta - \eta_0)'\} \frac{g^+(\mathbf{t}) \exp\{\mathbf{t}\eta_0'\} d\lambda(\mathbf{t})}{\int g^+(\mathbf{t}) \exp\{\mathbf{t}\eta_0'\} d\lambda(\mathbf{t})} = \int \exp\{\mathbf{t}(\eta - \eta_0)'\} \frac{g^-(\mathbf{t}) \exp\{\mathbf{t}\eta_0'\} d\lambda(\mathbf{t})}{\int g^-(\mathbf{t}) \exp\{\mathbf{t}\eta_0'\} d\lambda(\mathbf{t})}$$

for all $\eta \in N(\eta_0)$. That is, the m.g.f.'s of these two p.d.f.'s, $\frac{g^+(\mathbf{t}) \exp\{\mathbf{t}\eta_0'\} d\lambda(\mathbf{t})}{\int g^+(\mathbf{t}) \exp\{\mathbf{t}\eta_0'\} d\lambda(\mathbf{t})}$ and $\frac{g^-(\mathbf{t}) \exp\{\mathbf{t}\eta_0'\} d\lambda(\mathbf{t})}{\int g^-(\mathbf{t}) \exp\{\mathbf{t}\eta_0'\} d\lambda(\mathbf{t})}$, are the same in a neighborhood of 0. This implies that $g^+(\mathbf{t}) = g^-(\mathbf{t})$ a.e. λ. That is, $g(\mathbf{t}) = 0$ a.e. λ. That is, T is complete. ∎

Examples of Non-Complete Families

Sufficiency does not imply completeness.

EXAMPLE **2.11** *Let $X_1, ..., X_n \sim Bin(1, p)$. Then $(\sum_{i=1}^{n-1} X_i, X_n)$ is not complete.*

Proof. For simplicity, we take $n = 2$. Clearly, $E(X_1 - X_2) = 0$. However,

$$P(X_1 - X_2 = 0) = P(X_1 = 0, X_2 = 0) + P(X_1 = 1, X_2 = 1) < 1.$$

Hence, $X_1 - X_2$ is NOT complete. The general case $n \geq 2$ can be shown similarly. ∎

Even minimal sufficiency may not imply completeness.

EXAMPLE **2.12** *Given $X_1, ..., X_n \sim U(\theta, \theta + 1)$, $T = \left(X_{(1)}, X_{(n)}\right)$ is minimal sufficient, but not complete.*

Proof. If $U_1, ..., U_n \sim Unif(0, 1)$, then $U_{(k)} \sim \text{Beta}(k, n - k + 1)$ where $1 \leq k \leq n$. Therefore, $EU_{(k)} = k/(n + 1)$.

Now $X_k - \theta = U_k \sim U(0, 1)$, we have $EX_{(k)} = \theta + EU_{(k)} = k/(n + 1)$. Similarly, we can get $E(X_{(1)}) = \theta + EU_{(1)} = \theta + \frac{1}{n+1}$. Therefore, $E\left[X_{(n)} - X_{(1)} - \frac{n-1}{n+1}\right] = 0$. But $X_{(n)} - X_{(1)} - (n - 1)(n + 1) \neq 0$. So T is not complete. ∎

However, **a complete and sufficient statistic is minimal sufficient**.

2.8 Basu Theorem

There is an interesting relationship between minimal sufficient and ancillary statistics.

Sufficient statistics contain all the information about θ while ancillary statistics contain no information about θ. It is natural to ask whether (minimal) sufficient statistics and ancillary statistics are independent. This does not hold true in general unless bounded completeness is imposed.

EXAMPLE **2.13** *For $X_1, \ldots, X_n \sim U(\theta, \theta+1)$, a minimal sufficient is $(X_{(1)}, X_{(n)})$, or equivalently $(X_{(n)} - X_{(1)},\ X_{(n)} + X_{(1)})$, while an ancillary statisitc is $\left(X_{(n)} - X_{(1)}\right)$ (i.e. range of a location family). Clearly, the minimal sufficient and ancillary statistics are not independent.* ∎

THEOREM **2.6 (Basu's Theorem)** *If $T(X)$ is bounded complete and sufficient and V is ancillary, then $T(X) \perp V$.*

Proof. It suffices to show that

$$g(T) =: P(V \in B|T) - P(V \in B) = 0 \qquad a.s., \quad \text{for any Borel set } B. \tag{8.8}$$

Note that $g(T)$ is free of θ as T is sufficient and V is ancillary. It is also bounded. Clearly,

$$E_\theta g(T) = EE(I\{V \in B\}|T) - P(V \in B) = P(V \in B) - P(V \in B) = 0.$$

Since T is bounded complete, we have $g(T) = 0$ a.s., i.e., hence (8.8) holds. ∎

Examples

Example **2.14** *Given $\mathbf{X} \sim N(\mu, \sigma^2)$, $\bar{X}$ and $S^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are independent.*

Proof. Fix σ^2, and treat μ as the only parameter. Then, $\bar{X}$ is complete and sufficient for μ. On other other hand, S^2 is ancillary for μ. The result follows from Basu's theorem.

Alternative proof. Let $Y_i = X_i/\sigma$, and define $\bar{Y}$ and S_Y^2 correspondingly. Note that $Y_i \sim N(\mu/\sigma, 1) =: N(\widetilde{\mu}, 1)$. Treating $\widetilde{\mu}$ as our new parameter, we can use Basu's Theorem as above. ∎

Example **2.15** *Given $X_1, \ldots, X_n \sim f_\theta(x) = e^{-(x-\theta)} I\{x > \theta\}$, it is easy to show that $X_{(1)}$ is complete and sufficient. Hence, $X_{(1)} \perp g(\mathbf{X})$, where $g(\mathbf{X})$ is any ancillary statistics, e.g.,*

$$g(\mathbf{X}) = S^2, \quad \textit{or} \quad (X_{(n)} - X_{(1)}), \quad \textit{or} \quad X_2 - X_1, \quad \textit{etc.}$$

By Basu's Theorem,

$$X_{(1)} \perp g(\mathbf{X}).$$

For example,

$$e^{X_{(1)}} - \cos\left(X_{(1)}\right), \qquad \textit{and} \qquad e^{S^2}\sin^2(X_{(n)} - X_{(1)}) + \log\left(1 + |X_2 - X_1|\right).$$

In this case, Basu's Theorem allows us to deduce independence of two statistics without finding their joint distributions, which is typically very messy to do.

2.9 Exercises

1. Let $X_1, ..., X_n \sim_{iid} F_\theta$. Find a non-trivial sufficient statistic for θ, where

   (a) $f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \ x > 0, \ \theta = (\alpha, \beta)$.

   (b) $f(x) = \frac{1}{\sigma} e^{-(x-\mu)/\sigma}$ for $x > \mu$, $\theta = (\mu, \sigma)$.

   (c) $F_\theta = \text{Uniform}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$.

2. Let $X_1, \ldots, X_n$ be independent r.v.'s with densities $f_{X_i}(x) = e^{i\theta - x}$, for $x \geq i\theta$. Prove that $T = \min_i\{X_i/i\}$ is a sufficient statistic for θ.

3. For each of the following distributions let $X_1, \ldots, X_n$ be a random sample. Find a minimal sufficient statistic for the following problems.

   (a) $f(x) = (2\pi)^{-1/2} e^{-(x-\theta)^2/2}$ (normal)

   (b) $f(x) = P(X = x) = e^{-\theta}\theta^x/x!$, for $x = 0, 1, ...$ (Poisson)

   (c) $f(x) = e^{-(x-\theta)}$, for $x > \theta$ (location exponential)

   (d) $f(x) = c(\mu, \sigma) \exp\{-(x-\mu)^4/\sigma^4\}$, where $\sigma > 0$.

   (e) $f(x) = (\pi)^{-1}\left[1 + (x-\theta)^2\right]^{-1}$ (Cauchy)

   (f) $f_\theta(x) = C(\alpha)x^\alpha(1-x)^{1-\alpha}$, $0 < x < 1$, where $\alpha > 0$ and $C(\alpha)$ is a constant depending only on α (Beta(α, α) distribution.)

4. Let $X_1, \ldots, X_n \overset{iid}{\sim} F_\theta$. Then $T = (X_1, \cdots, X_{n-1})$ is not sufficient for θ. (T_n only uses part of the data, and is wasting information).

5. Let $T = (X_1, \ldots, X_n) \overset{iid}{\sim} F_\theta$. Then, T is sufficient, but not complete.

6. Let $X_1, \cdots, X_n$ be i.i.d. random variables from $U(a, b)$, where $a < b$. Show that $Z_i =: \dfrac{X_{(i)} - X_{(1)}}{X_{(n)} - X_{(1)}}$, $i = 2, ..., n-1$, are independent of $(X_{(1)}, X_{(n)})$ for any a and b.

7. $X_1, \cdots, X_n$ are i.i.d. r.v.s with pdf $f(x) = \theta^{-1}e^{-(x-a)/\theta} I_{(a,\infty)(x)}$. Show that

   (a) $\sum_{i=1}^n (X_i - X_{(1)})$ are $X_{(1)}$ are independent

   (b) $Z_i =: (X_{(n)} - X_{(i)})/(X_{(n)} - X_{(n-1)})$ $i = 1, 2, ..., n-2$, are independent of $((X_{(1)}, \sum_{i=1}^n (X_i - X_{(1)}))$

8. Let X be a discrete r.v.'s with p.d.f.

$$\begin{aligned} f_\theta(x) &= \theta, & x = 0 \\ &= (1-\theta)^2\theta^{x-1}, & x = 1, 2, ... \\ &= 0 & \textit{otherwise} \end{aligned}$$

   where $\theta \in (0, 1)$. Show that X is boundedly complete, but not complete.

Chapter 3

Unbiased Estimation

3.1 Optimality Criterion

Given $X_1, \cdots, X_n \sim_{iid} F_\theta$, we wish to find an "optimal" estimator $T = T(X_1, \cdots, X_n)$ for θ or more generally $g(\theta)$. There are various criterions for optimality, e.g.

- Least mean square error (MSE) estimator [=minimal L^2 loss]
- Least absolute deviation estimator (LADE) [=minimal L^1 loss]
- Maximal likelihood estimator (MLE) [=minimal cross entropy]
- Bayesian estimator [=minimal Bayesian risk]
- ......

One could unify these under the term

”**Loss function**”.

A few comments are in order.

- Optimal estimators under different criterions may not agree with each other.
- Optimal estimators may not exist, and/or may not be unique (when they exist).
- Which criterion to choose largely depends on the particular problem at hand.
- MSE and LAD estimators are model-free (nonparametric).
- MLE and Bayesian estimators are model dependent (parametric).
- Different estimators have different properties, e.g.
  - They may have good finite-sample or asymptotic properties.
  - Some are more robust, while others are more efficient.

3.2 Mean square error (MSE)

Let T be an estimate of $g(\theta)$. Its **Mean Squared Errors (MSE)** is

$$MSE[T] \ := \ E[T-g(\theta)]^2 = Var(T) + bias^2[T]. \tag{2.1}$$

where $bias[T] = ET - g(\theta)$ is the bias.

T is MSE-optimal if it has the smallest MSE for all $\theta \in \Theta$.

A globally optimal MSE estimator does not exist

EXAMPLE **3.1** *Let $X_1, \cdots, X_n$ be iid with $EX_i = \mu$ and $Var(X_i) = 1$. Take three estimators of μ:*

$$T_1 = X_1, \qquad T_2 = 3, \qquad T_3 = \bar{X}.$$

Clearly,

$$MSE(T_1) = 1, \qquad MSE(T_2) = (3-\mu)^2, \qquad MSE(T_3) = \frac{1}{n}.$$

We can plot their MSEs against μ.

- *T_3 is always better than T_1 (not sufficient) as*

  $$MSE(T_3) \le MSE(T_1).$$

- *Neither T_2 nor T_3 dominates the other for all θ.*
  - *If we have prior knowledge that $\mu \approx 3$, then $MSE(T_2) \le MSE(T_3)$, we choose T_2 (Bayesian view).*
  - *Otherwise, T_3 is preferred as T_2 is just a blind guess, particularly for large n (Frequentist view.)*

- *We might combine prior knowledge (if available) and data to produce*

  $$T_4 = \lambda T_2 + (1-\lambda) T_3$$

  where λ typically depends on n. This is Bayesian estimator.

3.3 UMVUE

MSE-optimal estimators do not exist in general since the class of all estimators under consideration (with second moments) is too big. One possible solution is to focus on a smaller class of estimators.

One such class is the U-class

$$\mathcal{U}\{g(\theta)\} = \text{the class of all unbiased estimators of } g(\theta).$$

- $T(\mathbf{X})$ is **unbiased** for $g(\theta)$ if $ET(\mathbf{X}) = g(\theta)$.
  Then $MSE(T) = Var(T)$.
- $T(\mathbf{X})$ is **uniformly minimum variance unbiased estimator** (UMVUE) for $g(\theta)$ if it is unbiased and has the smallest variance for all θ.

Remarks.

- $\mathcal{U}\{g(\theta)\}$ might be empty, i.e., unbiased estimates may not exist. See the example below.
  Hence, UMVUEs may not exist.

- Unbiasedness is not transformation-invariant, i.e., if $\hat{\theta}$ is unbiased for θ, but $g(\hat{\theta})$ may be biased for $g(\theta)$.

- UMVUEs apply to all fixed sample size n. No asymptotics is involved.

- UMVUE may not be MSE-optimal (called inadmissible).
  e.g., James-Stein shrinkage estimator.

- An MSE-optimal estimator strikes a right balance between bias and variance in (2.1).
  Thus, 0 bias may not be MSE-optimal.

- There are many ways to introduce bias. Most common is by L_1- or L_2-shrinkage.

Example: $\mathcal{U}\{g(\theta)\}$ might be empty

EXAMPLE **3.2** *$X \sim$ Bin(n,p). Then, $\mathcal{U}\{g(\theta)\}$ is nonempty (i.e., $g(p)$ is U-estimable) iff it is a polynomial in p of degree m (an integer), where $0 \le m \le n$.*

Proof. $\eta(X)$ is an unbiased estimator iff

$$g(p) = E\eta(X) = \sum_{k=0}^{n} \binom{n}{k} \eta(k) p^k (1-p)^{n-k},$$

i.e. $g(p)$ is a polynomial in p of degree m, where $0 \le m \le n$.

Thus, $g(p) = p^{-1}$, $p/(1-p)$ and $\sqrt{p}$ are not U-estimable. ∎

3.4 Characterization of a UMVUE.

Theorem **3.1**

1. *$\delta(X)$ is a UMVUE iff*
   - *(a) $E\delta(X) = g(\theta)$*
   - *(b) $Cov(\delta, U) = E[\delta(X)U(X)] = 0$, $\forall U \in \mathcal{U}\{0\}$ and $\forall \theta \in \Theta$.*
2. *T is sufficient. $\eta(T)$ is a UMVUE iff*
   - *(a) $E\eta(T) = g(\theta)$;*
   - *(b) $Cov[\eta(T), U(T)] = E[\eta(T)U(T)] = 0$, $\forall \theta \in \Theta$ and $\forall U \in \widetilde{\mathcal{U}}\{0\} = \{h(T) : Eh(T) = 0\}$.*

Proof. We shall only prove the first part. The second part can be shown similarly.

Necessity. Suppose that δ is a UMVUE for $g(\theta)$. Fix $U \in \mathcal{U}\{0\}$ and $\theta \in \Theta$. Let $\delta_1 = \delta + \lambda U$. Clearly $E\delta_1 = g(\theta)$, and so

$$Var(\delta_1) = Var(\delta) + 2\lambda Cov(U, \delta) + \lambda^2 Var(U) \geq Var(\delta),$$

that is,

$$2\lambda Cov(U, \delta) + \lambda^2 Var(U) \geq 0,$$

for all λ. The quadratic in λ has two roots $\lambda_1 = 0$ and $\lambda_2 = -2Cov(U, \delta)/Var(U)$, and hence takes on negative values unless $\lambda_1 = \lambda_2$, that is, $Cov(U, \delta) = 0$.

Sufficiency. Suppose that $E_\theta(\delta U) = 0$ for all $U \in \mathcal{U}\{0\}$. For any δ_1 such that $E(\delta_1) = g(\theta)$, we have $\delta_1 - \delta \in \mathcal{U}\{0\}$. Hence,

$$Cov(\delta, \delta_1 - \delta) = Cov(\delta, \delta_1) - Cov(\delta, \delta) = Cov(\delta, \delta_1) - Var(\delta) = 0$$

Hence, by Cauchy-Schwarz inequality,

$$Var(\delta) = Cov(\delta, \delta_1) \leq Var^{1/2}(\delta) Var^{1/2}(\delta_1)$$

That is, $Var(\delta) \leq Var(\delta_1)$. ∎

Three aspects to find a UMVUE for $g(\theta)$

- $E[U(X)] = 0$.
- $E[\delta(X)U(X)] = 0$.
- $E[\delta(X)] = g(\theta)$.

Or

- $E[U(T)] = 0$.
- $E[\eta(T)U(T)] = 0$.
- $E[\eta(T)] = g(\theta)$.

3.5 UMVUE is unique if it exists

THEOREM **3.2** *If δ is a UMVUE for $g(\theta)$, it is unique.*

Proof. If δ_1 and δ_2 are two UMVUEs, then $\delta_1 - \delta_2 \in \mathcal{U}\{0\}$. By Characterization Theorem, we have $E[\delta_1(\delta_1 - \delta_2)] = 0$ and $E[\delta_2(\delta_1 - \delta_2)] = 0$. So

$$E[(\delta_1 - \delta_2)^2] = E[\delta_1(\delta_1 - \delta_2)] - E[\delta_2(\delta_1 - \delta_2)] = 0,$$

which implies $\delta_1 = \delta_2$ a.s. ∎

3.6 How to find UMVUE when it exists?

- If T is complete and sufficient (C-S),
  - Lehmann-Scheffe Theorem, i.e., solving equations
  - Rao-Blackwell Theorem, i.e., calculating conditional expectation.
- Otherwise, use Characterisation Theorem

3.6.1 Lehmann-Scheffe Theorem

THEOREM **3.3** *If T is C-S, and $E_\theta[\eta(T)] = g(\theta)$, then $\eta(T)$ is the UMVUE for $g(\theta)$.*

Proof. Since T is complete, then $E_\theta(U(T)) = 0$ implies $U(T) = 0$. Then $\widetilde{\mathcal{U}}\{0\} = \{0\}$. The proof follows from the UMVUE characterization theorem. The proof of uniqueness is as in the first proof. ∎

Examples

EXAMPLE **3.3** *$X_1, \dots, X_n \sim U(0, \theta)$. Find a UMVUE for $g(\theta)$ where $g(x)$ is differentiable.*

Solution. It is known that $T = X_{(n)}$ is C-S, and

$$f_T(t) = n\theta^{-n}t^{n-1}I(0 < t < \theta).$$

If $E\eta(T) = \int_0^\theta \eta(t)n\theta^{-n}t^{n-1}dt = g(\theta)$, then

$$n\int_0^\theta \eta(t)t^{n-1}dt = \theta^n g(\theta)$$

Differentiating w.r.t. θ, we get

$$n\eta(\theta)\theta^{n-1} = n\theta^{n-1}g(\theta) + \theta^n g'(\theta)$$

Then $\eta(\theta) = g(\theta) + \frac{\theta}{n}g'(\theta)$.

$$\eta(T) = g(X_{(n)}) + \frac{X_{(n)}}{n}g'(X_{(n)})$$

(a). If $g(\theta) = \theta$, then $\eta(T) = \frac{n+1}{n}X_{(n)}$.

(b). If $g(\theta) = \theta^2$, then $\eta(T) = X_{(n)}^2 + \frac{2}{n}X_{(n)}^2 = \frac{n+2}{n}X_{(n)}^2$. ∎

REMARK **3.1** *An MLE of $g(\theta)$ is $g(X_{(n)})$ with bias of order $O(n^{-1})$.*

EXAMPLE **3.4** *Let $X_1, \ldots, X_n \sim$ Poisson(λ), find a UMVUE for $g(\lambda)$, where $g(\lambda)$ is infinitely differentiable.*

Solution. $T = \sum X_i$ is C-S, and $T \sim Poisson(n\lambda)$, that is, $f_T(t) = e^{-n\lambda}(n\lambda)^t/t!$ for $t = 0, 1, 2, \ldots$ Then

$$E\eta(T) = \sum_{k=0}^{\infty} \eta(k) e^{-n\lambda}(n\lambda)^k/k! = g(\lambda).$$

Then

$$\sum_{k=0}^{\infty} \frac{\eta(k)n^k\lambda^k}{k!} = g(\lambda)e^{n\lambda} = \left(\sum_{i=0}^{\infty} \frac{g^{(i)}(0)\lambda^i}{i!}\right)\left(\sum_{j=0}^{\infty} \frac{n^j\lambda^j}{j!}\right) = \sum_{k=0}^{\infty} a_k\lambda^k,$$

where $a_k = \sum_{i+j=k} \frac{g^{(i)}(0)n^j}{i!j!}$. Therefore, $\frac{\eta(k)n^k}{k!} = a_k$. So $\eta(k) = \frac{a_k k!}{n^k}$. That is,

$$\eta(T) = \frac{a_T T!}{n^T} = \frac{T!}{n^T}\left(\sum_{i+j=T} \frac{g^{(i)}(0)n^j}{i!j!}\right) = \sum_{i=0}^{T} \binom{T}{i} \frac{g^{(i)}(0)}{n^i}.$$

We now give two special examples.

1. If $g(\lambda) = \lambda^r$, where $r > 0$ is an integer, then

$$\begin{aligned} g^{(k)}(\lambda) &= r(r-1)\ldots(r-k+1)\lambda^{r-k} \quad \text{if } 0 \le k \le r, \\ &= 0 \quad \text{if } k > r, \end{aligned}$$

   Then $g^{(r)}(0) = r!$ and 0 otherwise. Therefore,

$$\begin{aligned} \eta(T) &= \frac{T!}{n^T}\frac{n^{T-r}r!}{(T-r)!r!} = \frac{T(T-1)\ldots(T-r+1)}{n^r} \qquad \text{if } T \ge r, \\ &= 0 \qquad \text{if } T < r. \end{aligned}$$

   Remark: Note that $\eta(T) \approx (\bar{X})^r$, the MLE.

2. If $g(\lambda) = P(X_1 = k) = \frac{e^{-\lambda}\lambda^k}{k!}$, for any $k = 0, 1, \ldots$, then we can use the above result. For instance, if $k = 0$, $g(\lambda) = e^{-\lambda}$. So $g^{(i)}(\lambda) = (-1)^i e^{-\lambda}$ and $g^{(i)}(0) = (-1)^i$. Therefore, a UMVUE of $e^{-\lambda}$ is

$$\eta(T) = \sum_{i=0}^{T} \binom{T}{i} \frac{(-1)^i}{n^i} = \left(1 - \frac{1}{n}\right)^T.$$

   Similarly, if $g(\lambda) = e^{-\lambda}\lambda^k/k!$, then its UMVUE is

$$\eta(T) = \binom{T}{k}\left(\frac{1}{n}\right)^k\left(1 - \frac{1}{n}\right)^{T-k} = P\left(Bin(T, p = \frac{1}{n}) = k\right). \quad \blacksquare$$

Example **3.5** *$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Find a UMVUE for μ/σ^2.*

Solution. The following fact will be useful in this problem.

- $\bar{X}$ and S^2 are independent, where $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$
- $\bar{X} \sim N(\mu, \sigma^2/n)$, and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, where the pdf of χ^2_r is $\frac{x^{r/2-1}e^{-x/2}}{2^{r/2}\Gamma(r/2)}$.
- $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$

It is known that $T = (\bar{X}, S^2)$ is complete and sufficient. Let

$$\eta(T) = C\frac{\bar{X}}{S^2}.$$

Then $E\eta(T) = C\mu E\left(\frac{1}{S^2}\right)$. Now taking $r = n-1$, we have

$$\begin{aligned} E\left(\frac{\sigma^2}{(n-1)S^2}\right) &= E\left(\frac{1}{\chi^2_{n-1}}\right) \\ &= \int_0^\infty \frac{1}{x}\frac{x^{r/2-1}e^{-x/2}}{2^{r/2}\Gamma(r/2)}\,dx \\ &= \int_0^\infty \frac{x^{(r-2)/2-1}e^{-x/2}}{2^{(r-2)/2}\Gamma((r-2)/2)}\,dx\frac{2^{(r-2)/2}}{2^{r/2}}\frac{\Gamma((r-2)/2)}{\Gamma((r-2)/2+1)} \\ &= \frac{1}{2}\frac{\Gamma((n-3)/2)}{\Gamma((n-1)/2)} \\ &= \frac{1}{n-3}. \end{aligned}$$

Then

$$E\eta(T) = C\mu E\left(\frac{1}{S^2}\right) = C\frac{n-1}{n-3}\frac{\mu}{\sigma^2} = \frac{\mu}{\sigma^2},$$

if $C = \frac{n-3}{n-1}$. That is, $\frac{n-3}{n-1}\frac{\bar{X}}{S^2}$ is a UMVUE of $\frac{\mu}{\sigma^2}$.

Remark: Note the MLE $= \bar{X}/S^2$.

3.6.2 Rao-Blackwell Theorem

Given a sufficient statistic, we can always improve it by conditioning.

THEOREM **3.4 (Rao-Blackwell)** *: Suppose T is sufficient, and $\delta(\mathbf{X})$ is an estimator of $g(\theta)$. Define $\eta(T) = E(\delta(\mathbf{X})|T)$. Then*

1. *$E\eta(T) = E\delta$ (hence they have the same bias).*
2. *$Var(\eta(T)) \leq Var(\delta)$.*
3. *$MSE(\eta(T)) \leq MSE(\delta)$, where equality holds iff $P(\eta(T) = \delta) = 1$.*

(Hence, by conditioning on T, we don't alter bias, but may reduce variance.)

Proof. By sufficiency, $\eta(T) = E(\delta(\mathbf{X})|T)$ if free of θ, and hence is an estimator of $g(\theta)$.

1. $E\eta(T) = EE(\delta|T) = E\delta$. Therefore, $bias(\delta) = bias(\eta(T))$.
2. This follows from the last and the next parts, and $MSE(Y) = bias(Y)^2 + Var(Y)$.
3. Now

$$\begin{aligned} MSE(\delta) &= E(\delta - g(\theta))^2 = E\left([\delta - \eta] + [\eta - g(\theta)]\right)^2 \\ &= E[\eta - g(\theta)]^2 + E[\delta - \eta]^2 + 2C \\ &= MSE(\eta) + E[\delta - \eta]^2 + 2C, \end{aligned}$$

where

$$C = E\left([\delta - \eta][\eta - g(\theta)]\right) = EE\left([\delta - \eta][\eta - g(\theta)]|T\right) = E\left\{[\eta - g(\theta)]\left(E[\delta|T] - \eta\right)\right\} = 0.$$

Therefore, $MSE(\delta) \geq MSE(\eta)$ and the equality holds iff $\delta = \eta(T)$ a.s. ∎

THEOREM **3.5** *If T is C-S and $E\delta = g(\theta)$, then $\eta(T) = E(\delta|T)$ is the UMVUE for $g(\theta)$.*

Proof. This follows from Rao-Blackwell Theorem and Lehmann-Scheffe Theorem. ∎

Examples

EXAMPLE **3.6** *Let $X_1, \ldots, X_n \sim Poisson(\lambda)$, find a UMVUE for*

$$g(\lambda) = P(X_1 = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad \textit{for any} \quad k = 0, 1, 2, ...$$

Solution. Let $\delta = I\{X_1 = k\}$. Clearly, $E\delta = g(\lambda)$. Now $T = \sum X_i$ is C-S, and $T \sim Poisson(n\lambda)$, then

$$\begin{aligned} \eta(t) &= E[\delta|T = t] = P(X_1 = k|T = t) \\ &= P(X_1 = k, T = t)/P(T = t) \\ &= P\left(X_1 = k, \sum_{i=2}^{n} X_i = t - k\right) \Big/ P\left(\sum_{i=1}^{n} X_i = t\right) \\ &= \frac{e^{-\lambda}\lambda^k/k!\ e^{-(n-1)\lambda}[(n-1)\lambda]^{t-k}/(t-k)!}{e^{-n\lambda}[n\lambda]^t/t!} \\ &= \binom{t}{k}\left(\frac{1}{n}\right)^k\left(1 - \frac{1}{n}\right)^{t-k}. \end{aligned}$$

So a UMVUE for $g(\lambda)$ is

$$\eta(T) = \binom{T}{k}\left(\frac{1}{n}\right)^k\left(1 - \frac{1}{n}\right)^{T-k}. \quad ∎$$

REMARK **3.2**

- *The above derivation is straightaway by using the following fact: If $X \sim Poisson(\lambda_1)$ and $Y \sim Poisson(\lambda_2)$ and independent, then $X|X + Y = t \sim Bin\{t, \lambda_1/(\lambda_1 + \lambda_2)\}$.*
- *By comparison, the MLE for the above problem is $\frac{e^{-\bar{X}}(\bar{X})^k}{k!}$. .....*

3.6.3 Using Characterisation Theorem

When a C-S statistic is not available or not easily available, we can not use Lehmann-Scheffe Theorem to find UMVUEs. Then we can try the Characterization Theorem.

Example **3.7** *Let $X_1, \dots, X_n \sim U(0, \theta)$, where $\theta > 1$. Find a UMVUE of θ.*

Solution. It is known that $T = X_{(n)}$ is sufficient. But it is not complete. To show this, note that $F_T(t) = F^n(t) = t^n/\theta^n$, and so

$$f_T(t) = \frac{nt^{n-1}}{\theta^n}\, I(0 < t < \theta).$$

If $E_\theta g(T) = 0$, then

$$0 = \int_0^\theta t^{n-1} g(t)dt = \int_0^1 t^{n-1} g(t)dt + \int_1^\theta t^{n-1} g(t)dt.$$

Differentiating w.r.t. θ, we get $g(\theta) = 0$ for any $\theta > 1$, and

$$\int_0^1 t^{n-1} g(t)dt = 0.$$

This certainly has a nonzero solution, $g(t)$. Hence, T is not complete.

It is known that $\tilde{T} = X_{(n)} \vee 1$ is minimal sufficient. However, it may be troublesome to check completeness. Perhaps it might be easier to use the characterization theorem to find a UMVUE in this case.

(1). $E[U(T)] = 0$ implies that (simply take $g(t) = U(t)$ above)

$$U(\theta) = 0 \qquad \text{for any } \theta > 1 \text{ and} \qquad \int_0^1 t^{n-1} U(t)dt = 0.$$

(2). If $E[\eta(T)U(T)] = 0$, then from (1), we get

$$\int_0^\theta t^{n-1} U(t)\eta(t)dt = \int_0^1 t^{n-1} U(t)\eta(t)dt = 0.$$

(3). Next we need to find $\eta(T)$ such that $E\eta(T) = \theta$. We use two approaches.

Approach 1. Clearly, (1) and (2) are both satisfied by

$$\begin{aligned} \eta(t) &= c && \text{if } 0 < t \le 1 \\ &= bt && \text{if } t > 1. \end{aligned}$$

Then from $E\eta(T) = \theta$, we get

$$\begin{aligned} \theta &= E\eta(T) = E\left[\eta(T) I(0 < T \le 1)\right] + E\left[\eta(T) I(T > 1)\right] \\ &= \int_0^1 c f_T(t)dt + \int_1^\theta bt f_T(t)dt \\ &= cP(T < 1) + \int_1^\theta bt \frac{nt^{n-1}}{\theta^n} dt \\ &= \frac{c}{\theta^n} + \frac{bn}{n+1} \frac{\theta^{n+1} - 1}{\theta^n} \\ &= \frac{bn}{n+1}\theta + \left(c - \frac{bn}{n+1}\right)\frac{1}{\theta^n}, \end{aligned}$$

where we used the fact $P(T<1) = F^n(1) = \theta^{-n}$. Therefore, we have

$$\frac{bn}{n+1} = 1, \qquad c - \frac{bn}{n+1} = 0.$$

So we get $b = (n+1)/n$ and $c = 1$. That is,

$$\begin{aligned} \eta(T) &= 1 && \text{if } 0 < X_{(n)} < 1 \\ & \quad (1+n^{-1})X_{(n)} && \text{if } X_{(n)} > 1. \end{aligned}$$

Approach 2. Clearly, we can choose

$$\begin{aligned} \eta(t) &= c && \text{if } 0 < t \le 1 \\ &= \eta(t) && \text{if } t > 1\ . \end{aligned}$$

Here, we choose $\eta(t) = c$ if $0 < t \le 1$ since it satisfies (1). From $E\eta(T) = \theta$, we get

$$\begin{aligned} \theta &= E\eta(T) = E\left[cI(0<T<1)\right] + E\left[\eta(T)I(T>1)\right] \\ &= \int_0^1 cf_T(t)dt + \int_1^\theta \eta(t)f_T(t)dt \\ &= cP(T<1) + \int_1^\theta \eta(t)f_T(t)dt \\ &= \frac{c}{\theta^n} + \int_1^\theta \eta(t)\frac{nt^{n-1}}{\theta^n}dt. \end{aligned}$$

as $P(T<1) = F^n(1) = \theta^{-n}$. That is,

$$\theta^{n+1} = c + \int_1^\theta \eta(t)nt^{n-1}dt. \tag{6.2}$$

Differentiating w.r.t. θ, we get

$$n\eta(\theta)\theta^{n-1} = (n+1)\theta^n.$$

Therefore, $\eta(\theta) = \frac{n+1}{n}\theta$ for $\theta > 1$. Putting it back into (6.2), we get $c = 1$. Hence,

$$\begin{aligned} \eta(T) &= 1 && \text{if } 0 < T \le 1 \\ & \quad (n+1)T/n && \text{if } T > 1. \end{aligned}$$

Remark **3.3** *If $x_{(n)} < 1$, we estimate θ by 1. If $x_{(n)} \ge 1$, we just carry on as usual.*

Remark **3.4** *$T_0 = \max\{X_{(n)}, 1\}$ is minimal sufficient. Clearly, $0 < X_{(n)} \le 1 \Longleftrightarrow T_0 = 1$; and $X_{(n)} > 1 \Longleftrightarrow T_0 > 1$. Then,*

$$\begin{aligned} \eta(T) &= 1 && \textit{if } T_0 = 1 \\ & \quad (1+n^{-1})T_0 && \textit{if } T_0 > 1. \end{aligned}$$

Therefore, $\eta(T)$ is a function of T_0, as can be expected. ∎

3.7 UMVUE in Normal Linear Models

Example **3.8 Simple regression model**. *Suppose*

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i = 1, ..., n.$$

Equivalently,

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), i = 1, ..., n.$$

Find UMVUEs for $\theta = (\alpha, \beta, \sigma^2)$.

Solution. The joint pdf is

$$\begin{aligned} f(\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(\frac{-(Y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right) \\ &= C(\theta) \exp\left(-\frac{\sum_{i=1}^{n} Y_i^2}{2\sigma^2} - \frac{\alpha \sum_{i=1}^{n} Y_i}{2\sigma^2} - \frac{\beta \sum_{i=1}^{n} x_i Y_i}{2\sigma^2} \right) \end{aligned}$$

Hence,

$$T = (\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} x_i Y_i, \sum_{i=1}^{n} Y_i^2)$$

is C-S for $\tilde{\theta} = (1/\sigma^2, \alpha/\sigma^2, \beta/\sigma^2)$ or equivalently for $\theta = (\alpha, \beta, \sigma^2)$ (1 to 1 correspondence).

Since the LSEs of α and β are unbiased and are functions of T, hence they must be UMVUE. Similarly, the UMVUE of σ^2 is given by $\hat{\sigma}^2 = SSE/(n-2)$.

EXAMPLE **3.9 Simple regression model**. *Suppose*

$$X_i = \xi_i + \epsilon_i = \alpha + \beta t_i + \epsilon_i, \qquad i = 1, ..., n,$$

where $\epsilon_i \sim N(0, \sigma^2)$. *Equivalently,* $X_i \sim N(\xi_i, \sigma^2), i = 1, ..., n,$ *where* $\xi_i = \alpha + \beta t_i$. *Find UMVUEs for* $\theta = (\alpha, \beta, \sigma^2)$.

Solution. The joint pdf is

$$\begin{aligned} f(\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(\frac{-(x_i - \alpha - \beta t_i)^2}{2\sigma^2} \right) \\ &= C(\theta) \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2} - \frac{\alpha \sum_{i=1}^n x_i}{2\sigma^2} - \frac{\beta \sum_{i=1}^n t_i x_i}{2\sigma^2} \right) \end{aligned}$$

Hence, $T = (\sum_{i=1}^n x_i, \sum_{i=1}^n t_i x_i, \sum_{i=1}^n x_i^2)$ is C-S for $\tilde{\theta} = (1/\sigma^2, \alpha/\sigma^2, \beta/\sigma^2)$ or equivalently for $\theta = (\alpha, \beta, \sigma^2)$ (1 to 1 correspondence).

Since the LSEs of α and β are unbiased and are functions of T, hence they must be UMVUE. Similarly, the UMVUE of σ^2 is given by $\hat{\sigma}^2 = SSE/(n-2)$.

Normal Linear Models

Rather than dealing with every single case, here we present some unified treatment for independent case, in particular, the "Normal Linear Models":

$$X_i \sim N(\xi_i, \sigma^2), \qquad i = 1, ..., n,$$

or equivalently

$$\mathbf{X} \sim N(\xi, \sigma^2 I),$$

where the X_i's are independent and $\xi = (\xi_1, ..., \xi_n) \in L_s$, an s-dimensional linear subspace of $R^n (s < n)$. Let

$$\mathbf{Y} = (Y_1, ..., Y_n) = (Y_1, ..., Y_n)(e_1, ..., e_n) = (X_1, ..., X_n)C = XC$$

where C is an orthogonal transformation. Denote $\eta_i = E(Y_i)$ then $\eta = \xi C$ or $\eta C^T = \xi$, and

$$\mathbf{Y} \sim N(\eta, \sigma^2 I),$$

Let the first s columns of C^T span L_s, then clearly,

$$\eta_{s+1} = ... = \eta_n = 0.$$

As a result, we have

$$\begin{aligned} f_Y(\mathbf{y}) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(\frac{-(y_i - \eta_i)^2}{2\sigma^2} \right) \\ &= C_1(\theta) \exp\left(-\frac{\sum_{i=1}^s (y_i - \eta_i)^2}{2\sigma^2} - \frac{\sum_{i=s+1}^n y_i^2}{2\sigma^2} \right) \\ &= C_1(\theta) \exp\left(-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2} - \frac{\sum_{i=1}^s \eta_i y_i}{2\sigma^2} \right) \end{aligned}$$

Hence, $T = (Y_1, ..., Y_s, \sum_{i=1}^n Y_i^2)$ or equivalently, $\tilde{T} = (Y_1, ..., Y_s, \sum_{i=1}^s Y_i^2)$, is C-S.

THEOREM **3.6** *UMVUE of $\sum_{i=1}^s \lambda_i\eta_i$ and σ^2 are, respectively,*

$$\sum_{i=1}^s \lambda_i Y_i, \qquad \textit{and} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^s Y_i^2}{n-s}$$

Proof. They are unbiased and functions of T, hence UMVUE by L-S theorem. ∎

THEOREM **3.7** *UMVUE of $\sum_{i=1}^s \xi_i\eta_i$ is $\sum_{i=1}^s \lambda_i\hat{\xi}_i$, where*

$$\hat{\xi} = (\hat{\xi}_1, ..., \hat{\xi}_n) = \arg\min \sum_{i=1}^n (X_i - \xi_i)^2$$

is an LSE of ξ.

Proof. From $\sum_{i=1}^n (X_i-\xi_i)^2 = \sum_{i=1}^n (Y_i-\eta_i)^2 = \sum_{i=1}^s (Y_i-\eta_i)^2 + \sum_{i=s+1}^n Y_i^2$, we get

$$\hat{\eta} = (Y_1, ..., Y_s, 0, ..., 0) = \hat{\xi}C.$$

Thus, $\hat{\xi} = \hat{\eta}C^T$, and hence $\sum_{i=1}^s \lambda_i\hat{\xi}_i$, is a function of $Y_1, ..., Y_s$. But $\eta = E\hat{\eta} = E\hat{\xi}C = \xi C$, thus $E(\sum_{i=1}^s \lambda_i\hat{\xi}_i) = \sum_{i=1}^s \lambda_i\xi_i$. ∎

EXAMPLE **3.10 One-way layout**. *Suppose*

$$X_{ij} \sim N(\xi_i, \sigma^2), \qquad j = 1, ..., n_i; \qquad i = 1, ..., s.$$

or equivalently,

$$X_{11}, ..., X_{1n_1} \sim N(\xi_1, \sigma^2);$$
$$..................$$
$$X_{s1}, ..., X_{sn_s} \sim N(\xi_s, \sigma^2);$$

Find UMVUEs for ξ_i and σ^2.

Solution. Let $X_{i\cdot}$ be the average of ith treatment. Note

$$\sum_{i=1}^s \sum_{j=1}^{n_i} (X_{ij}-\xi_i)^2 = \sum_{i=1}^s \left\{ \sum_{j=1}^{n_i} (X_{ij} - X_{i\cdot})^2 + n_i (X_{ij} - X_{i\cdot})^2 \right\}$$

So the LSE and hence the UMVUE of ξ_is are $\hat{\xi}_i = X_{i\cdot}$. And the UMVUE of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^s \sum_{j=1}^{n_i} (X_{ij} - X_{i\cdot})^2}{n-s}.$$

3.8 UMVUE may fail for small n

Let $X_1, \ldots, X_n \sim \textit{Poisson}(\lambda)$. Then $T = \sum X_i \sim \textit{Poisson}(n\lambda)$, and is complete and sufficient. To find a UMVUE for $e^{-a\lambda}$, we set $E\eta(T) = e^{-a\lambda}$, i.e.,

$$\sum \frac{\eta(t)(n\lambda)^t}{t!} = e^{(n-a)\lambda} = \sum \frac{(n-a)^t \lambda^t}{t!}$$

so $\eta(T) = (1 - a/n)^T$ is a UMVUE of $e^{-a\lambda}$.

- Big problems for small n.

  Suppose $a = 3$, and $n = 1$, then $\eta(T) = (-2)^{X_1}$. It goes from 1 to -2 to 4 to -8 etc as X goes from 0, 1, 2, ... This is crazy since it oscillates between negative and positive values.

- The above problem disappears for large n, e.g., as soon as $n > a$.

- The MLE is simply $e^{-a\bar{X}}$.

3.9 Exercises

1. $T = (X_1, \ldots, X_n) \overset{iid}{\sim} f(x)$, $n \geq 2$. Show that T is sufficient, but not complete.

2. Let T_i be a UMVUE of θ_i, $j = 1, ..., k$. Then

   (a) $T = \sum_{i=1}^k c_i T_i$ is a UMVUE of $\theta = \sum_{i=1}^k c_i \theta_i$, for constants c_i's.

   (b) $T = \prod_{i=1}^k T_i$ is a UMVUE of $\theta = ET = E\left(\prod_{i=1}^k T_i\right)$.

3. Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Find a UMVUE for μ^2/σ.

4. $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$. Find a UMVUE for $e^{-t\lambda}$ for $t > 0$.

5. $X_1, \ldots, X_n \overset{iid}{\sim} U(\theta_1, \theta_2)$, where $\theta_1 < \theta_2$. Find a UMVUE for $(\theta_1 + \theta_2)/2$.

6. Let $X_1, \ldots, X_n$ be iid $Bin(1, p)$, where $p \in (0, 1)$. Find the UMVUE of

   (a) p^m, where m is a positive integer $\leq n$;

   (b) $P(X_1 + ... + X_m = k)$, where m and k are positive integers $\leq n$;

   (c) $P(X_1 + ... + X_{n-1} > X_n)$.

7. $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x) = e^{-(x-\theta)}$, $x > \theta$. Find a UMVUE of θ^r, $r > 0$.

8. Let $X_1, \ldots, X_n \sim Bin(k, p)$. Find a UMVUE for $g(p) = P(X_1 = 1) = kp(1-p)^{k-1}$.

9. $X_1, \ldots, X_n \overset{iid}{\sim} N(0, \sigma^2)$. We wish to find a UMVUE of σ^2.

   (a) Find a complete and sufficient statistic for σ^2.

   (b) Find C_1, C_2 (which may depend on n) such that $W_1 = C_1(\bar{X})^2$ and $W_2 = C_2 \sum_{i=1}^n (X_i - \bar{X})^2$ are unbiased for σ^2.

   (c) Consider $W = aW_1 + (1-a)W_2$, where a is a constant. Find a which minimizes $var(W)$. Show that $var(W) \leq var(W_i)$, $i = 1, 2$

   (d) Find a UMVUE of σ^2.

Chapter 4

Maximum Likelihood Estimation

Likelihood-based methods lie in the heart of statistics. They have been widely used in practice, more so than you perhaps realise. Some methods, e.g., least square estimation (LSE), logistic regression, etc., which are seemingly unrelated to likelihood, can be cast into the likelihood framework. Others, e.g., Bayesian methods to be studied later, can be loosely regarded as the extension of likelihood methods.

The idea of the maximum likelihood estimate (MLE), first introduced by R.A. Fisher, is simple and yet very powerful. It enjoys many nice properties, such as invariance principle, asymptotic efficiency, existence principle. Of course, like every method, it also has drawbacks, e.g. strong assumptions, non-robust, etc.

4.1 Maximum Likelihood Estimate (MLE)

EXAMPLE **4.1** *Let* $X \sim P_\theta$, *where* $\theta = 0$ *or* 1. *Suppose*

$P_\theta(X = x)$	$x = 0$	$x = 1$	$x = 2$	$x = 3$
$\theta = 0$	*0.5*	*0.1*	*0.1*	*0.3*
$\theta = 1$	*0.2*	*0.3*	*0.2*	*0.3*

Let $\hat{\theta}$ *be an estimate of* θ.

- *If* $X = 0$, $P_0(0) > P_1(0)$, *so* $\theta = 0$ *is more likely. So we take* $\hat{\theta} = 0$.
- *Similarly, if* $X = 1$ *or* 2, *take* $\hat{\theta} = 1$.
- *If* $X = 3$, *then* $P_0(3) = P_1(3)$, *take* $\hat{\theta} = 0$ *or* 1.

Thus we obtained an MLE of θ*:*

$$\begin{aligned} \hat{\theta} &= 0 \qquad \textit{if } X = 0, 3 \\ &= 1 \qquad \textit{if } X = 1, 2, 3. \end{aligned}$$

Note immediately that $\hat{\theta}$ *may not be unique, e.g., when* $X = 3$, $\hat{\theta} = 0$ *or* 1. ∎

Definition **4.1** *Suppose* $\mathbf{X} \sim f_\theta(\mathbf{x})$, *where* $\theta \in \Theta \in R^k$. *Observe* $\mathbf{X} = \mathbf{x}$. *Let* $\overline{\Theta}$ *be the closure of* Θ.

- **likelihood function***: probability of observing* $\mathbf{x}$*, expressed as a function of* θ

$$L(\theta) \equiv L(\theta, \mathbf{x}) = f_\theta(\mathbf{x})$$

- **log-likelihood function***:*

$$l(\theta) \equiv l(\theta, \mathbf{x}) = \log L(\theta).$$

- *A* **maximum likelihood estimator** *(MLE) of* θ *is*

$$\hat{\theta} = \arg\sup_{\theta \in \Theta} L\left(\theta(\mathbf{x})\right) = \arg\sup_{\theta \in \Theta} l\left(\theta(\mathbf{x})\right).$$

  i.e., the principle of MLE: estimate θ *such that the probability of the observed data reaches the maximum.*

- $\hat{\theta}$ *always exists in* $\overline{\Theta}$ *(not necessarily in* Θ *though).*

4.2 Score function and information matrix

- Log-likelihood function:
$$l(\theta) = \log L(\theta)$$
- Score function:
$$s(\theta) = l'(\theta)$$
- Possible candidates for MLE's satisfy
$$l'(\theta) = s(\theta) = 0. \tag{2.1}$$
  - MLE may not be unique.
  - If $L(\beta)$ is concave function, $\hat{\theta}$ is unique.
    How to check concave function?
$$l''(\theta) < 0.$$
- Information matrix:
$$I(\theta) = -El''(\theta)$$

4.3 Invariance property of MLE

One very attractive property of the MLE is its invariance. To be more precise, let $\eta = g(\theta)$, where $g(\cdot)$ is a function from R^k to R^d. Basically, then the invariance property of MLE means: if $\hat{\theta}$ is an MLE of θ, then $g(\hat{\theta})$ is an MLE of $g(\theta)$.

4.3.1 $g(\cdot)$ is a 1-to-1 mapping

If $g(\cdot)$ is 1-to-1, estimating θ is equivalent to estimating $\eta = g(\theta)$ as a new parameter.

THEOREM **4.1** *Let* $\mathbf{X} \sim f_\theta(\mathbf{x})$ *and* $\eta = g(\theta)$ *where* $g(\cdot)$ *is 1-to-1. If* $\hat{\theta}$ *is an MLE for* θ*, then* $g(\hat{\theta})$ *is an MLE for* $g(\theta)$*.*

Proof. Since $g(\cdot)$ is 1-1, then $g^{-1}(\cdot)$ exists. Also

$$L(\theta) = f_\theta(\mathbf{x}) = f_{g^{-1}(\eta)}(\mathbf{x}) = L_1(\eta)$$

where L_1 is the likelihood function of η. Hence,

$$L(\hat{\theta}) = L_1(g(\hat{\theta}))$$

Since $\hat{\theta}$ is an MLE, hence $L(\theta)$ and $L_1(\eta)$ are maximized by $\hat{\theta}$ and $g(\hat{\theta})$, respectively. That is, $g(\hat{\theta})$ is an MLE of η. ∎

4.3.2 $g(\cdot)$ is not 1-to-1

If $g(\cdot)$ is NOT 1-to-1, $\eta = g(\theta)$ is no longer a parameter, since once knowing η, we could have more than one θ's, say, θ_1 and θ_2, such that $\eta = g(\theta_1) = g(\theta_2)$, i.e., we may not be able to completely specify θ in the parameter space Θ (non-identifiable). This is bad since we don't know whether $X \sim f_{\theta_1}(x)$ or $X \sim f_{\theta_2}(x)$.

For fixed η, Let $\Theta_\eta =: \{\theta \in \Theta : g(\theta) = \eta\}$. For example, we could have $\Theta_\eta =: \{\theta_1, \theta_2\}$. As we are not sure which θ to pick from Θ_η, it seems reasonable to choose the most probable one, i.e., the value of θ in Θ_η that maximizes the likelihood $L(\theta) = f_\theta(x)$. This is the so-called **induced likelihood**.

Definition **4.2** *Suppose $g(\cdot)$: $\Theta \to \Lambda \in R^p$, and let $\Theta_\eta =: \{\theta \in \Theta : g(\theta) = \eta\}$.*

- *The* **induced likelihood function** *for $\eta = g(\theta)$ is defined as*

$$\widetilde{L}(\eta) = \sup_{\{\theta\in\Theta : g(\theta)=\eta\}} L(\theta).$$

- *For fixed* $\mathbf{x}$, *let $\hat{\eta} = \hat{\eta}(\mathbf{x})$ satisfy $\widetilde{L}(\hat{\eta}) = \sup_{\eta\in g(\Theta)} \widetilde{L}(\eta)$ i.e.,*

$$\hat{\eta} = \arg \sup_{\eta\in g(\Theta)} \widetilde{L}(\eta).$$

Such a $\hat{\eta}$ always exists in $\overline{g(\Theta)}$ (not necessarily in $g(\Theta)$), and is still called a **maximum likelihood estimator** *(MLE) of θ.*

Theorem **4.2 (MLE Invariance Theorem)** *Let $\mathbf{X} \sim f_\theta(\mathbf{x})$ and $\eta = g(\theta)$. If $\hat{\theta}$ is an MLE for θ, then $g(\hat{\theta})$ is an MLE for $\eta = g(\theta)$.*

Proof. It suffices to show that $\widetilde{L}(\hat{\eta}) = \widetilde{L}(g(\hat{\theta}))$. By definition, we have

$$\widetilde{L}(\hat{\eta}) = \sup_{\eta\in g(\Theta)} \widetilde{L}(\eta) = \sup_{\eta} \sup_{\{\theta\in\Theta : g(\theta)=\eta\}} L(\theta) = \sup_{\theta\in\Theta} L(\theta) = L(\hat{\theta}).$$

On the other hand, from the definition of $\tilde{L}(\eta)$, we have

$$\widetilde{L}(g(\hat{\theta})) = \sup_{\{\theta\in\Theta : g(\theta)=g(\hat{\theta})\}} L(\theta) = L(\hat{\theta}). \quad \blacksquare$$

4.4 Examples

EXAMPLE **4.2** *Let $X_1, ..., X_n \sim Bin(1, p)$ with $p \in (0, 1)$. Find an MLE for p^2.*

Solution Note that

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i} = p^{n\bar{x}} (1-p)^{n(1-\bar{x})} \\ l(p) &= n[\bar{x} \log p + (1-\bar{x}) \log (1-p)] \end{aligned}$$

Setting

$$l'(p) = n\left(\frac{\bar{x}}{p} - \frac{1-\bar{x}}{1-p}\right) = \frac{(\bar{x}-p)}{p(1-p)} = 0,$$

we get $\hat{p} = \bar{x}$. Since

$$l''(p) = \left(-\frac{\bar{x}}{p^2} - \frac{1-\bar{x}}{(1-p)^2}\right) < 0,$$

$l(p)$ is a concave function. Then $\hat{p} = \bar{x}$ is an MLE.

Then, an MLE of p is $\hat{p} = \bar{X}$. By the invariance theorem, an MLE of p^2 is $(\bar{X})^2$.

REMARK **4.1** *(a) If $\bar{X} = 0$ or 1, then $\hat{p} = 0$ or $1 \notin \Theta = (0, 1)$. (b) Note that in this example, the MLE $\bar{X}$ is also the UMVUE.*

Example **4.3** *Let $X_1, ..., X_n \sim Poisson(\lambda)$ with $\lambda > 0$. Find an MLE for λ.*

Solution. Note that

$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{x_1!...x_n!}, \quad \text{and} \quad l(\lambda) = -n\lambda + (n\bar{x})\log\lambda - \log(x_1!...x_n!)$$

- If $\bar{x} > 0$, we set $l'(\lambda) = -n + \frac{n\bar{x}}{\lambda} = 0$ to get

  $$\hat{\lambda} = \bar{x}.$$

  Since $l''(\lambda) = -\frac{n\bar{x}}{\lambda^2} = -\frac{n}{\bar{x}} < 0$, so $l(\lambda)$ is concave. Hence, $\hat{\lambda} = \bar{x}$ is an MLE.
- If $\bar{x} = 0$, then $L(\lambda) = e^{-n\lambda}$, which has the supremum at $\hat{\lambda} = 0 = \bar{x}$.

Combining these results, an MLE of λ is thus $\hat{\lambda} = \bar{X}$. ∎

Remark **4.2** *(i) If $\bar{X} = 0$, then $\hat{\lambda} = 0 \notin \Theta = (0, \infty)$. (ii) Note that in this example, the MLE $\bar{X}$ is also the UMVUE.*

Example **4.4** *Let*

$$X_1, ..., X_n \sim_{i.i.d.} N(\mu, \sigma^2),$$

where $n \geq 2$, $\mu \in R$ and $\sigma^2 > 0$.

1. *Find an MLE for $\theta = (\mu, \sigma^2)$.*
2. *Do part 1 if $\mu \geq 0$.*

Solution. Note

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

and

$$l(\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi).$$

Setting

$$l'(\theta) = \begin{pmatrix} l'_\mu(\theta) \\ l'_{\sigma^2}(\theta) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2}\sum_{i=1}^n (x_i-\mu) \\ \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i-\mu)^2 - \frac{n}{2\sigma^2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

we get $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$, where

$$\hat{\mu} = \bar{x}, \qquad \text{and} \qquad \hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (x_i-\bar{x})^2.$$

1. It can be shown that $l''(\hat{\theta})$ is negative definite at $\hat{\theta}$ (exercise). Hence, the MLE of θ is

$$\hat{\mu} = \bar{X}, \qquad \text{and} \qquad \hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i-\bar{X})^2. \quad \blacksquare$$

2. The global MLE is $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (x_i-\bar{x})^2$.

Case I. If $\bar{x} \geq 0$, $\hat{\theta} \in \overline{\Theta} =: \{(\mu, \sigma^2) : \mu \geq 0, \sigma^2 \geq 0\}$. Hence, $\hat{\theta}$ is an MLE.

Case II. If $\bar{x} < 0$, $\hat{\theta} \notin \overline{\Theta}$ But as in the last example,

$$\begin{aligned} l(\theta) = l(\mu, \sigma^2) = &= -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi) \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\bar{x})^2 - \frac{n}{2\sigma^2}(\bar{x}-\mu)^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi). \end{aligned}$$

For each (fixed) σ^2, $l(\theta)$ attains its maximum when $\mu = 0$ since $\bar{x} \leq 0$. Then

$$l(0, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\bar{x})^2 - \frac{n}{2\sigma^2}(\bar{x})^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi),$$

which attains its maximum when $\sigma^2 = n^{-1}\sum x_i^2$. Therefore, an MLE of θ is

$$\hat{\mu} = 0, \qquad \hat{\sigma}^2 = n^{-1}\sum x_i^2.$$

In summary, an MLE of θ is

$$\begin{aligned} \hat{\theta} &= (0, n^{-1}\sum X_i^2), & \text{if } \bar{X} < 0 \\ &= (\bar{X}, n^{-1}\sum(X_i-\bar{X})^2), & \text{if } \bar{X} \geq 0. \quad \blacksquare \end{aligned}$$

Example **4.5 (Linear regression)**

EXAMPLE **4.6 (Logistic regression)**

4.5 Numerical solution to likelihood equations

Typically, there are no explicit solutions to likelihood equations, and some numerical methods have to be used to compute MLE's. We start with a couple of examples.

Example. Let $X_1, ..., X_n \sim Cauchy(\theta)$. Find an MLE of θ.

Solution. The pdf of Cauchy(θ) is $f_\theta(x) = \pi^{-1}\left(1+(x-\theta)^2\right)^{-1}$. Therefore, $L(\theta) = \pi^{n-1}\prod_{i=1}^n \left(1+(x_i-\theta)^2\right)^{-1}$ and

$$l(\theta) = -n\log\pi - \sum_{i=1}^n \log(1+(x_i-\theta)^2).$$

Setting

$$l'(\theta) = \sum_{i=1}^n \frac{2(x_i-\theta)}{1+(x_i-\theta)^2} = 0.$$

An MLE $\hat\theta$ can not be solved explicitly, but can be obtained by numerical method, e.g., Newton-Raphson.

Newton's algorithm.

Note that $\hat\theta \rightarrow_p \theta$, and $n^{-1}l(\theta)$, $n^{-1}l'(\theta)$ and $n^{-1}l''(\theta)$ are all sample means of i.i.d. r.v.'s. Thus

$$0 = \frac{1}{n}l'(\hat\theta) \quad\approx\quad \frac{1}{n}l'(\theta) + (\hat\theta-\theta)\frac{1}{n}l''(\theta) \approx \frac{1}{n}l'(\theta) + (\hat\theta-\theta)\frac{1}{n}El''(\theta)$$

Then

$$\hat\theta \quad\approx\quad \theta - l'(\theta)[l''(\theta)]^{-1} \approx \theta - l'(\theta)[El''(\theta)]^{-1}.$$

- *Newton-Raphson's algorithm:*

$$\theta_{j+1} = \theta_j - l'(\theta_j)\left[l''(\theta_j)\right]^{-1}, \qquad j = 0, 1, 2, ...,$$

- *Scoring method:*

$$\theta_{j+1} = \theta_j - l'(\theta_j)\left[El''(\theta_j)\right]^{-1}, \qquad j = 0, 1, 2, ...,$$

4.6 Certain "Issues" with MLE

4.6.1 MLE may not be unique

Example. Let $X_1, ..., X_n \sim U(\theta - 1/2, \theta + 1/2)$ with $\theta \in R$. Find an MLE of θ.

Solution. Note

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} I(\theta - 1/2 < x_i < \theta + 1/2) \\ &= I(\theta - 1/2 < x_{(1)} < x_{(n)} < \theta + 1/2) \\ &= I(x_{(n)} - 1/2 < \theta < x_{(1)} + 1/2). \end{aligned}$$

By plotting $L(\theta)$ against θ, we see that $\hat{\theta} \in (x_{(n)} - 1/2, x_{(1)} + 1/2)$. So we have infinite number of MLE's, and the range is $x_{(1)} + 1/2 - (x_{(n)} - 1/2) = 1 - (x_{(n)} - x_{(1)}) \to 0$ as $n \to \infty$. ∎

4.6.2 MLE may not be asymptotically normal

Example. Let $X_1, ..., X_n \sim U(0, \theta)$. Find an MLE for θ.

Solution. The likelihood function is

$$L(\theta, \mathbf{x}) = \frac{1}{\theta^n} I(0 \le x_{(1)} \le x_{(n)} \le \theta)$$

Draw a graph of $L(\theta, \mathbf{x})$ as a function of θ, we see that an MLE for θ is

$$\hat{\theta} = X_{(n)}.$$

Now

$$\begin{aligned} P(n(\theta - \hat{\theta})/\theta \le t) &= P(\hat{\theta} \ge \theta - t\theta/n) \\ &= 1 - P(X_{(n)} \le \theta - t\theta/n) \\ &= 1 - [P(X_1 \le \theta - t\theta/n)]^n \\ &= 1 - \left(1 - \frac{t}{n}\right)^n \\ &\to 1 - e^{-t} \qquad \text{as } n \to \infty. \end{aligned}$$

Hence, $\hat{\theta}$ is asymptotically $exp(1)$ (after normalization), instead of normal.

4.6.3 MLE can be inconsistent

When there are many nuisance parameters, then MLE's can behave very badly even for large samples, as shown below.

Example. We have p independent random samples

$$X_{11}, \cdots, X_{1n} \sim N(\mu_1, \sigma^2),$$

$$......$$

$$X_{p1}, \cdots, X_{pn} \sim N(\mu_p, \sigma^2).$$

Let $\theta = (\mu_1, \cdots, \mu_p, \sigma^2)$. Equivalently,

$$\mathbf{X}_i = (X_{1i}, ..., X_{pi}) \sim MN(\mu, \sigma^2 I), \qquad \text{where } \mu = (\mu_1, ..., \mu_p), \text{ and } i = 1, ..., n.$$

Show that

1. an MLE of θ is $\hat{\theta}$, where $\hat{\mu}_i = n^{-1}\sum_{j=1}^{n} x_{ij} = \bar{x}_{i\cdot}$, $\quad \hat{\sigma}^2 = (pn)^{-1}\sum_{i=1}^{p}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\cdot})^2$;
2. if p is fixed and $n \to \infty$, then $\hat{\theta}$ is consistent for θ;
3. if n is fixed and $p \to \infty$, then $\hat{\theta}$ is NOT consistent for θ.

Proof.

1. The log-likelihood function is

$$\begin{aligned} l(\theta) &= \log \prod_{i=1}^{p}\prod_{j=1}^{n} f_\theta(x_{ij}) = \log\left((2\pi\sigma^2)^{-pn/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{n}\frac{(x_{ij}-\mu_i)^2}{\sigma^2}\right\}\right) \\ &= -\frac{pn}{2}\log(2\pi) - \frac{pn}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{n}(x_{ij}-\mu_i)^2 \end{aligned}$$

The log-likelihood equations are

$$\frac{\partial l(\theta)}{\partial \mu_i} = \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_{ij}-\mu_i) = 0, \quad \text{and} \quad \frac{\partial l(\theta)}{\partial \sigma^2} = -\frac{pn}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{p}\sum_{j=1}^{n}(x_{ij}-\mu_i)^2 = 0$$

whose solution is $\hat{\theta}$ as required. Note

$$\hat{\sigma}^2 = \frac{\sigma^2}{pn}\sum_{i=1}^{p}\left(\sum_{j=1}^{n}\left(\frac{x_{ij}-\bar{x}_{i\cdot}}{\sigma}\right)^2\right) = \frac{\sigma^2}{pn}\sum_{i=1}^{p}\chi^2_{n-1} = \frac{\sigma^2}{pn}\chi^2_{p(n-1)}$$

Thus,

$$E\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{\sigma^2}{n}, \qquad Bias(\hat{\sigma}^2) = -\frac{\sigma^2}{n}, \qquad Var(\hat{\sigma}^2) = \frac{2(n-1)}{pn^2}\sigma^4$$

2. If p is fixed, $n \to \infty$, $\hat{\mu}_i$ is consistent for μ_i by WLLN. By Chebyshev's inequality,

$$\begin{aligned} P(|\hat{\sigma}^2 - \sigma^2| > \epsilon) &\leq \epsilon^{-2} E(\hat{\sigma}^2 - \sigma^2)^2 \\ &\leq \epsilon^{-2} \left[Bias^2(\hat{\sigma}^2) + Var(\hat{\sigma}^2) \right] \\ &\leq \epsilon^{-2} \left[\sigma^4/n^2 + Var(\hat{\sigma}^2) \right] \\ &\to 0. \end{aligned}$$

That is, $\hat{\sigma}^2$ is consistent for σ^2.

3. If n is fixed, $p \to \infty$, then

   - $\hat{\mu}_i$ is not consistent for μ_i since $E\hat{\mu}_i = \mu_i$, but $Var(\hat{\mu}_i) = \sigma^2/n > 0$;
   - $\hat{\sigma}^2$ is NOT consistent for σ^2 since the Bias($\hat{\sigma}^2$) $\neq 0$, even though $var(\hat{\sigma}^2) \to 0$.

The third part is often referred to as **"small n, large p"** problem in the literature ($p \gg n$). There are several ways to overcome this difficulty.

- Assume that both $n, p \to \infty$, but $n/p \to 0$.
- Assume sparsity: many μ_i's are zero's. We can impose L_1 penalty, e.g.

4.7 Exercises

1. Let $X_1, ..., X_n \sim_{iid} f_\theta(x)$. Find an MLE of θ if

   (a) $f_\theta(x) = \theta^{-1}$, where $x = 1, ..., \theta$ and θ is an integer between 1 and θ_0.

   (b) $f_\theta(x) = e^{-(x-\theta)}$, where $x \geq \theta$ and $\theta > 0$.

   (c) $f_\theta(x) = \theta(1-x)^{\theta-1} I(0 < x < 1)$, where $\theta > 1$.

   (d) $f_\theta(x)$ is the pdf of $N(\theta, \theta^2)$, where $\theta \in R$.

2. Let $\mathbf{X} \sim f_\theta(\mathbf{x})$ and $T = T(\mathbf{X})$ is sufficient. Show that if an MLE exists, it is a function of T but it may not be sufficient for θ.

3. $X_1, \cdots, X_n$ are i.i.d. r.v.s with $f_\theta(x) = (\theta+1)x^\theta,\ 0 \leq x \leq 1$. (i.e. Beta$(\theta+1, 1)$.)

   (1) Find the MLE of θ.

   (2) Find the asymptotic distribution of the MLE.

Chapter 5

Efficient Estimation

5.1 Information inequality (or Cramer-Rao lower bound)

Let $\mathbf{X} \sim f_\theta(\mathbf{x})$, and $l(\theta) = \log f_\theta(\mathbf{X})$, where $\theta \in R^k$. Define the Fisher Information of θ contained in $\mathbf{X}$ by

$$I_{\mathbf{X}}(\theta) = E\left[(l'(\theta))^\tau_{k\times 1}\,(l'(\theta))_{1\times k}\right]$$

Theorem **5.1 (Cramer-Rao (C-R) lower bound)** *Let $E\delta(\mathbf{X}) = g(\theta)$. Then*

$$Var(\delta(\mathbf{X})) \geq g'(\theta)_{1\times k}[I_{\mathbf{X}}(\theta)]^{-1}_{k\times k} g'(\theta)^\tau_{k\times 1},$$

where the equality holds iff $\delta(\mathbf{x})$ and $l'(\theta)$ are linearly dependent, provided

$$\frac{\partial}{\partial\theta}\int h(\mathbf{x}) f_\theta(\mathbf{x})dx = \int h(\mathbf{x})\frac{\partial}{\partial\theta} f_\theta(\mathbf{x})dx, \quad \textit{for } \theta \in \Theta$$

for $h(\mathbf{x}) = 1$ and $h(\mathbf{x}) = \delta(\mathbf{x})$,

Proof

(a) **Lemma.** For any r.v. δ and $Y = (Y_1, ..., Y_k)$, we have

$$Var(\delta) \geq Cov(\delta, Y)[Var(Y)]^{-1}[Cov(\delta, Y)]^\tau, \tag{1.1}$$

where the equality holds iff δ and Y are linearly dependent. Here,

$$\begin{aligned} Cov(\delta, Y) &= (Cov(\delta, Y_1), ..., Cov(\delta, Y_k))_{1\times k}, \\ Var(Y) &= (Cov(Y_i, Y_j))_{k\times k}. \end{aligned}$$

Proof. For any $\lambda = (\lambda_1, ..., \lambda_k) \in R^k$, we have

$$\begin{aligned} 0 &\leq Var(\delta - \lambda Y^\tau) = Cov(\delta - \lambda Y^\tau, \delta - \lambda Y^\tau) \\ &= Var(\delta) - 2Cov(\delta, Y\lambda^\tau) + Cov(\lambda Y^\tau, \lambda Y^\tau) \\ &= Var(\delta) - 2Cov(\delta, Y)\lambda^\tau + \lambda Var(Y)\lambda^\tau \\ &=: g(\lambda) \end{aligned}$$

Setting $g'(\lambda) = 2\lambda Var(Y) - 2Cov(\delta, Y) = 0$, we get the stationary point

$$\lambda_0 = Cov(\delta, Y)[Var(Y)]^{-1}.$$

Thus,

$$g(\lambda_0) = Var(\delta) - Cov(\delta, Y)[Var(Y)]^{-1}[Cov(\delta, Y)]^\tau \geq 0$$

where equality holds iff $g(\lambda_0) = 0 = Var(\delta - \lambda_0 Y^\tau)$ iff $\delta - \lambda_0 Y^\tau = C$.

(b) Take

$$Y = l'(\theta) = \frac{\partial}{\partial\theta} \log f_\theta(\mathbf{X}) = \frac{\frac{\partial}{\partial\theta} f_\theta(\mathbf{X})}{f_\theta(\mathbf{X})}.$$

Differentiating (w.r.t. θ)

$$1 = \int f_\theta(\mathbf{x})d\mathbf{x}, \quad \text{and} \quad g(\theta) = E(\delta(\mathbf{X})) = \int \delta(\mathbf{x}) f_\theta(\mathbf{x})d\mathbf{x}$$

we get

$$\begin{aligned}
0 &= \int \frac{\partial f_\theta(\mathbf{x})}{\partial\theta} d\mathbf{x} = \int l'(\theta) f_\theta(\mathbf{x})d\mathbf{x} = El'(\theta) = EY, \\
g'(\theta) &= \frac{\partial}{\partial\theta} E\delta(\mathbf{X}) = \int \delta(\mathbf{x}) \frac{\partial f_\theta(\mathbf{x})}{\partial\theta} d\mathbf{x} = \int \delta(\mathbf{x}) l'(\theta) f_\theta(\mathbf{x})d\mathbf{x} \\
&= E[\delta(\mathbf{x}) l'(\theta)] = cov\left(\delta(\mathbf{X}), l'(\theta)\right) = cov\left(\delta(\mathbf{X}), Y\right), \\
I_{\mathbf{X}}(\theta) &= E\left[\left(l'(\theta)\right)^\tau_{k\times 1} \left(l'(\theta)\right)_{1\times k}\right] = E(Y^\tau Y) = Var(Y).
\end{aligned}$$

Applying (1.1), we get

$$Var(\delta) \geq [E\delta(\mathbf{X})]'_\theta [I_{\mathbf{X}}(\theta)]^{-1} [[E\delta(\mathbf{X})]'_\theta]^\tau = g'(\theta)[I_{\mathbf{X}}(\theta)]^{-1}[g'(\theta)]^\tau,$$

where the equality holds iff $\delta - \lambda_0 l'(\theta)^\tau = C$ iff $\delta(\mathbf{X}) - E\delta(\mathbf{X}) = \lambda_0 l'(\theta)^\tau$.

5.2 Properties of Fisher information matrix

1. If $\mathbf{X} \perp \mathbf{Y}$, then $I_{(\mathbf{X},\mathbf{Y})}(\theta) = I_{\mathbf{X}}(\theta) + I_{\mathbf{Y}}(\theta)$.

2. If $\mathbf{X} = \{X_1, \ldots, X_n\} \sim_{i.i.d.} f_\theta(x)$, then $I_{\mathbf{X}}(\theta) = \sum_{i=1}^n I_{X_i}(\theta) = nI_{X_i}(\theta)$.

3. $I_{\mathbf{X}}(\theta) = -El''(\theta)$, where $\frac{\partial}{\partial\theta^\tau} \int \frac{\partial}{\partial\theta} f_\theta(\mathbf{x})dx = \int \frac{\partial^2}{\partial\theta\partial\theta^\tau} f_\theta(\mathbf{x})dx, \quad \text{for } \theta \in \Theta$.

   Proof. We only prove the case $k = 1$. Differentiating $1 = \int f_\theta(\mathbf{x})d\mathbf{x}$ w.r.t. θ, we get
$$0 = \int l'(\theta)f_\theta(\mathbf{x})d\mathbf{x} = E\left(l'(\theta)\right).$$
   Differentiating this again w.r.t. θ gives
$$0 = \int l''(\theta)f_\theta(\mathbf{x})d\mathbf{x} + \int \left(l'(\theta)\right)^2 f_\theta(\mathbf{x})d\mathbf{x} = E\left(l'(\theta)\right)^2 + El''(\theta).$$

4. Let $\theta = h(\eta)$ where $h : R^m \to R^k$ with $m, k \geq 1$. Then,
$$I_{\mathbf{X}}(\eta) = h'(\eta)^\tau_{m\times k} I_{\mathbf{X}}(\theta) h'(\eta)_{k\times m}.$$
   Proof. From the definition,
$$\begin{aligned} I_{\mathbf{X}}(\eta) &= E\left[\left(\frac{\partial l(\theta)}{\partial\eta}\right)^\tau \left(\frac{\partial l(\theta)}{\partial\eta}\right)\right] \\ &= E\left[\left(l'(\theta)h'(\eta)\right)^\tau \left(l'(\theta)h'(\eta)\right)\right] \\ &= \left(h'(\eta)\right)^\tau E\left[l'(\theta)^\tau l'(\theta)\right] h'(\eta) \\ &= h'(\eta)^\tau I_{\mathbf{X}}(\theta) h'(\eta) \end{aligned}$$

5. Let $\theta = \theta(\eta)$ is from R^m to R^k, where $m, k \geq 1$. Assume that $\theta(.)$ is differentiable. Then the Fisher information that $\mathbf{X}$ contains about η is
$$I_{\mathbf{X}}(\eta) = \left(\frac{\partial\theta}{\partial\eta^\tau}\right)_{m\times k} I_{\mathbf{X}}(\theta) \left(\frac{\partial\theta^\tau}{\partial\eta}\right)_{k\times m}.$$
   Proof. From the definition,
$$\begin{aligned} I_{\mathbf{X}}(\eta) &= E\left[\left(\frac{\partial l(\theta)}{\partial\eta}\right)^\tau \left(\frac{\partial l(\theta)}{\partial\eta}\right)\right] \\ &= E\left[\left(l'(\theta)\theta'(\eta)\right)^\tau \left(l'(\theta)\theta'(\eta)\right)\right] \\ &= \left(\theta'(\eta)\right)^\tau E\left[\left(l'(\theta)\right)^\tau \left(l'(\theta)\right)\right] \theta'(\eta) \\ &= \left(\theta'(\eta)\right)^\tau I_{\mathbf{X}}(\theta)\theta'(\eta) \end{aligned}$$

6. If $\theta = h(\eta)$ is 1-1 (from R^k to R^k), then the Cramer-Rao lower bound remains the same (i.e., reparametrization invariant).

   Proof. $X_1, ..., X_n \sim f_\theta(x)$, and $\theta = h(\eta)$ is 1-1. Now $E\delta(X) = g(\theta) = g(h(\eta)) = r(\eta)$. By C-R lower bound theorem,
$$Var(\delta(\mathbf{X})) \geq \left(g'(\theta)\right)_{1\times k} \left[I_X(\theta)\right]^{-1}_{k\times k} \left(g'(\theta)\right)^\tau_{k\times 1} =: A,$$
   and also
$$Var(\delta(\mathbf{X})) \geq \left(r'(\eta)\right)_{1\times k} \left[I_X(\eta)\right]^{-1}_{k\times k} \left(r'(\eta)\right)^\tau_{k\times 1} =: B.$$
   We shall show that $A = B$. Note that
$$l'(\eta) = \frac{\partial l(\eta)}{\partial\eta} = \frac{\partial l(\eta)}{\partial\theta}\frac{\partial\theta^\tau}{\partial\eta} = \frac{\partial g(\theta)}{\partial\theta} h'(\eta)^\tau = g'(\theta)h'(\eta)^\tau,$$
   and that $\frac{\partial\theta^\tau}{\partial\eta}$ is $k \times k$ matrix of full rank. So from this and (4) above, we get
$$B = \left(g'(\theta)h'(\eta)^\tau\right) \left(h'(\eta)^\tau\right)^{-1} \left[I_{\mathbf{X}}(\theta)\right]^{-1} \left(h'(\eta)\right)^{-1} \left(g'(\theta)h'(\eta)^\tau\right)^\tau = \left(g'(\theta)\right) \left[I_X(\theta)\right]^{-1} \left(g'(\theta)\right)^\tau = A.$$

7. If θ is a scalar (i.e., $k = 1$), then the Cramer-Rao lower bound becomes

$$Var(\delta(\mathbf{X})) \geq \frac{[g'(\theta)]^2}{I_X(\theta)}, \qquad \text{where} \quad I_X(\theta) = E\left(l'(\theta)\right)^2 = -E\left(l''(\theta)\right).$$

5.3 Efficiency, sub-efficiency, and super-efficiency

An unbiased estimator $\delta(X)$ is

- efficient if $Var(\delta) = I_X^{-1}(\theta)$
- sub-efficient if $Var(\delta) > I_X^{-1}(\theta)$
- super-efficient if $Var(\delta) < I_X^{-1}(\theta)$

We have used the Cramer-Rao lower bound as a benchmark to define efficiency.

5.3.1 Examples of efficient and sub-efficient estimators

A necessary and sufficient condition for reaching Cramer-Rao lower bound is that $\delta(\mathbf{x})$ and $l'(\theta)$ are linearly dependent, i.e.

$$\delta(\mathbf{x}) - E\delta(\mathbf{X}) = \lambda_0 \cdot l'(\theta).$$

UMVUE may or may not be efficient.

Example **5.1** *Let $X_1, ..., X_n$ be iid from the $Bin(1, p)$. Show that the Cramer-Rao lower bound can be reached in estimating p but not for p^2.*

Proof. First $f_p(\mathbf{x}) = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$, $x_i = 0, 1$. So $l(p) = \log f_p(\mathbf{x}) = \sum_i x_i \log p + (n - \sum_i x_i)\log(1-p)$. Then

$$l'(p) = \frac{\sum_i x_i}{p} - \frac{n - \sum_i x_i}{1-p} = \frac{n(\bar{x} - p)}{p(1-p)}, \qquad x_i = 0, 1. \tag{3.2}$$

Therefore,

$$I_{\mathbf{X}}(p) = \frac{nE(X_1 - p)^2}{p^2(1-p)^2} = \frac{n}{p(1-p)}, \qquad x = 0, 1.$$

If $E\delta(\mathbf{X}) = g(p)$, then by the Cramer-Rao inequality,

$$Var(\delta(\mathbf{X})) \geq \frac{[g'(p)]^2}{I_{\mathbf{X}}(p)} = \frac{[g'(p)]^2 p(1-p)}{n}.$$

From (3.2), the equality only holds for $g(p) = c_0 + c_1 p$, a linear function of p.

1. If $g(p) = p$, then $Var(\delta(\mathbf{X})) \geq p(1-p)/n$, which is indeed reached by $\delta(\mathbf{X}) = \bar{X}$ (UMVUE of p).

2. If $g(p) = p^2$, then $Var(\delta(\mathbf{X})) \geq 4p^3(1-p)/n = CRLB$.

   Since this lower bound can never be reached, we are interested in how big the gap is. Take $\eta(T) = T(T-1)/(n(n-1))$ (the UMVUE of p^2). After some tedious algebra, we get

   $$Var(\eta(T)) = \frac{E[T^2(T-1)^2]}{[n^2(n-1)^2]} - p^4 = \frac{4p^3(1-p)}{n} + \frac{2p^2(1-p)^2}{n(n-1)} = CRLB + O(n^{-2}). \quad \blacksquare$$

   Note: To find out $Var(\eta(T))$, e.g., the m.g.f. of T is $M_n(t) = (q + pe^t)^n$. So

   $$\begin{aligned} M_n'(t) &= n(q+pe^t)^{n-1}pe^t = npe^t M_{n-1}(t), \\ M_n''(t) &= ne^t M_{n-1}'(t) + ne^t M_{n-1}(t), \\ &= n(n-1)e^{2t}M_{n-2}(t) + ne^t M_{n-1}(t), \\ M_n^{(3)}(t) &= .... \\ M_n^{(4)}(t) &= .... \end{aligned}$$

   And then setting $t = 0$, we get the first four moments of T.

5.3.2 Examples of super-efficient estimator

5.3.3 On regularity conditions

EXAMPLE **5.2** *Let $X_1, \ldots, X_n \sim U(0, \theta)$. Show that*

1. *the assumption in the Cramer-Rao theorem does not hold (the range of the pdf depends on the unknown parameter).*
2. *the information inequality does not apply to the UMVUE of θ.*

Solution.

1. Use Leibnitz's rule, we have

$$\begin{aligned}
\frac{\partial}{\partial\theta}\int h(x)f_\theta(x)dx &= \frac{\partial}{\partial\theta}\int_0^\theta \frac{h(x)}{\theta}dx \\
&= \frac{h(\theta)}{\theta} + \int_0^\theta h(x)\frac{\partial}{\partial\theta}\frac{1}{\theta}dx \\
&= \frac{h(\theta)}{\theta} + \int h(x)\frac{\partial}{\partial\theta}f_\theta(x)dx \\
&\neq \int h(x)\frac{\partial}{\partial\theta}f_\theta(x)dx
\end{aligned}$$

unless $h(\theta)/\theta = 0$ for all θ. Therefore, the assumption fails here.

2. Here $f_\theta(x) = 1/\theta$, $0 < x < \theta$. So $\frac{\partial}{\partial\theta}\log f_\theta(x) = -1/\theta$, and

$$I_{\mathbf{X}}(\theta) = nE\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2 = n\int_0^\theta \frac{1}{\theta^2}\frac{1}{\theta}dx = \frac{n}{\theta^2}$$

By Cramer-Rao theorem, we get, if W is any unbiased estimator, then

$$Var(W) \geq \frac{\theta^2}{n}.$$

Take $\hat\theta = (n+1)/nX_{(n)}$ (the UMVUE). Recall $f_{X_{(n)}}(x) = \frac{nx^{n-1}}{\theta^n}I(0 < x < \theta)$, then

$$\begin{aligned}
EX_{(n)}^2 &= \int_0^\theta \frac{nx^{n+1}}{\theta^n}dx = \frac{n}{n+2}\theta^2 \\
var\left(\hat\theta\right) &= \left(\frac{n+1}{n}\right)^2 EX_{(n)}^2 - \theta^2 = \left(\frac{n+1}{n}\right)^2\frac{n}{n+2}\theta^2 - \theta^2 = \frac{\theta^2}{n(n+2)} \\
&< \frac{\theta^2}{n} = \frac{1}{nI(\theta)}.
\end{aligned}$$

Note that the $Var(X_{(n)})$ is of size $O(n^{-2})$, as opposed to the usual size $O(n^{-1})$.

Remark **5.1** *In this example, the UMVUE has variance below the C-R lower bound, which is a good thing. In fact, the variance of $\hat{\theta}$ is of order n^{-2}, much smaller than the usual n^{-1}. More precisely, we have*

$$n(\theta - \hat{\theta})/\theta \sim_{approx} Exp(\lambda = 1)$$

Therefore, $\hat{\theta}$ is asymptotically $Exp(\lambda = 1)$, but not asymptotically normal.

Proof. For any $y \geq 0$, we have

$$\begin{aligned} P(n(\theta - X_{(n)}) \leq y) &= P(X_{(n)} \geq \theta - y/n) \\ &= 1 - P(X_{(n)} \leq \theta - y/n) \\ &= 1 - [P(X_1 \leq \theta - y/n)]^n \\ &= 1 - \left(1 - \frac{y}{n\theta}\right)^n . \end{aligned}$$

Taking $y = t\theta$, we get

$$P\left(n(\theta - X_{(n)})/\theta \leq t\right) = 1 - (1 - t/n)^n \to 1 - e^{-t} \qquad \text{as } n \to \infty.$$

By Slutsky's theorem, we get $n(\theta - \hat{\theta})/\theta \sim_{approx} \text{Exp}(\lambda = 1)$.

5.4 Asymptotic efficiency

5.4.1 Asymptotic Cramer-Rao lower bound

The Cramer-Rao theorem (under appropriate regularity conditions) states

$$Var_\theta(\delta_n) \geq \frac{([E\delta_n]'_\theta)^2}{nI_{X_1}(\theta)} \qquad \Longleftrightarrow \qquad Var_\theta(\sqrt{n}\delta_n) \geq \frac{([E\delta_n]'_\theta)^2}{I_{X_1}(\theta)} \qquad \forall \theta \in \Theta, \tag{4.3}$$

The above inequality holds for fixed n. What happens as $n \to \infty$?

Assume

$$\sqrt{n}[\delta_n - g(\theta)] \to_d N(0, v(\theta)), \qquad v(\theta) > 0 \tag{4.4}$$

which holds if $bias(\delta_n) = E\delta_n - g(\theta) = o(n^{-1/2})$ (i.e., δ_n is asymptotically unbiased) and $Var(\sqrt{n}\delta_n) \to v(\theta)$. Let $n \to \infty$ in (4.3), we would get

$$v(\theta) \geq \frac{[g'(\theta)]^2}{I(\theta)} \qquad \text{for ALL } \theta \in \Theta. \tag{4.5}$$

However, (4.5) may not hold uniformly for ALL $\theta \in \Theta$, even under nice regularity conditions. (The example when regularity conditions fails is $Unif(0, \theta)$.)

Example **5.3 (Hodges)** *Let $X_1, ..., X_n \sim N(\theta, 1)$, $\theta \in R$. Then $I_{X_1}(\theta) = 1$. The MLE of θ is $\hat{\theta} = \bar{X} \sim N(\theta, 1/n)$. Now let us define*

$$\begin{aligned} \tilde{\theta} &= \bar{X} && \textit{if } |\bar{X}| \geq n^{-1/4} \\ & \quad t\bar{X} && \textit{if } |\bar{X}| < n^{-1/4}, \end{aligned}$$

where $0 \leq t < 1$. Show that

$$\begin{aligned} \sqrt{n}\left(\tilde{\theta} - \theta\right) &\to_d N(0, t^2) && \textit{if } \theta = 0, \\ & \quad\; N(0, 1) && \textit{if } \theta \neq 0. \end{aligned}$$

Remark **5.2** *$\tilde{\theta}$ shrinks $\bar{X}$ around 0 to $t\bar{X}$. Hence it is more efficient at 0* **if the true parameter is 0**. *The trouble is that we don't know where the true θ is.*

Solution. Note that $\tilde{\theta} = \bar{X}I_n + t\bar{X}(1 - I_n)$, where $I_n = I(|\bar{X}| \geq n^{-1/4}) = I(|\sqrt{n}\bar{X}| \geq n^{1/4})$. This is a shrinkage estimator, where shrinkage occurs near 0. Rewrite

$$\begin{aligned} \sqrt{n}(\tilde{\theta} - \theta) &= t\sqrt{n}\bar{X}(1 - I_n) + \sqrt{n}\bar{X}I_n && (4.6) \\ &= \sqrt{n}(\bar{X} - \theta) + (t - 1)\sqrt{n}\bar{X}(1 - I_n). && (4.7) \end{aligned}$$

(1) Suppose $\theta = 0$. Then we have $\sqrt{n}\bar{X} \sim N(0, 1)$, and $I_n \to_p 0$ as $P(I_n \geq \epsilon) = P(|\sqrt{n}\bar{X}| \geq n^{1/4}) = P(|N(0, 1)| \geq n^{1/4}) \to 0$. Hence, applying Slutsky theorem to (4.6), we have

$$\sqrt{n}(\tilde{\theta} - \theta) = t\sqrt{n}\bar{X}(1 - I_n) + \sqrt{n}\bar{X}I_n \to_d N(0, t^2).$$

(2) Suppose $\theta \neq 0$. Then $\bar{X} \to_p \mu$, and $\sqrt{n}(\bar{X}-\theta) \to_d N(0,1)$. The proof follows naturally from Slutsky Theorem to (4.7) if we can show $\sqrt{n}(1-I_n) \to_p 0$. Now

$$\begin{aligned}
P(\sqrt{n}(1-I_n) \geq \epsilon) &= P(I_n \leq 1-\epsilon/\sqrt{n}) = P(I_n = 0) = P\left(|\sqrt{n}\bar{X}| < n^{1/4}\right) \\
&= P(-n^{1/4} < \sqrt{n}\bar{X} < n^{1/4}) \\
&= P\left(-\sqrt{n}\theta - n^{1/4} < \sqrt{n}(\bar{X}-\theta) < n^{1/4} - \sqrt{n}\theta\right) \\
&= P\left(-\sqrt{n}\theta - n^{1/4} < N(0,1) < -\sqrt{n}\theta + n^{1/4}\right) \\
&= \Phi\left(-\sqrt{n}\theta + n^{1/4}\right) - \Phi\left(-\sqrt{n}\theta - n^{1/4}\right) \\
&\to 0. \quad \blacksquare
\end{aligned}$$

From the example, we see that in the case of $\theta = 0$, the asymptotic Cramer-Rao theorem does not hold, since

$$t^2 < 1 = [I(\theta)]^{-1}. \quad \blacksquare$$

Remark **5.3**

- *The set of all points violating (4.5) is called points of* **superefficiency**. *Superefficiency in fact is a very good property.*
- *Even for the normal family (the best possible family), (4.5) may not hold. However, it will hold for a more restricted class (i.e., "regular" estimators).*
- *The example can be easily extended to have countable number of super-efficiency points. For instance, we can define, for $k = 0, \pm 1, \pm 2, \ldots\ldots$,*

$$\begin{aligned}
\hat{\theta} &= \bar{X} && \text{if } |\bar{X}-k| \geq n^{-1/4}, \\
& \quad k + t_k(\bar{X}-k) && \text{if } |\bar{X}-k| < n^{-1/4}
\end{aligned}$$

where we choose $0 \leq t_k < 1$ for all k. Then it can be shown that this estimator will be super-efficient at all $k = 0, \pm 1, \pm 2, \ldots\ldots.$

- *Although (4.5) is not always true, however, it holds a.e., i.e., the points of super-efficiency has Lebesgue measure 0. See the next theorem (Le Cam (1953), Bahadur (1964)).*

Theorem **5.2** *If $\delta_n = \delta_n(X_1, \ldots, X_n)$ is any estimator satisfying (4.4), then $v(\theta)$ satisfies (4.5) except on a set of Lebesgue measure 0, given the following conditions hold:*

1. *The set $A = \{x : f_\theta(x) > 0\}$ is independent of θ. [i.e., they have common support, and the boundary of Θ does not depend on θ.]*
2. *For every $x \in A$, $l''(\theta)$ exists and is continuous in θ.*
3. *$\int f_\theta(x)dx$ can be twice differentiated under the integral sign.*
4. *$0 < I_{X_1}(\theta) < \infty$, where $I_{X_1}(\theta) = E\left(l'(\theta)\right)^2$.*
5. *For any $\theta_0 \in \Theta$, there exists $c > 0$ and a function $M(x)$ (both may depend on θ_0) such that for all $x \in A$ and $|\theta - \theta_0| < c$, we have*

$$|l''(\theta)| \leq M(x), \qquad \text{and} \qquad E_{\theta_0}[M(X_1)] < \infty.$$

5.5 Consistency of the MLE

Assume throughout the section:

(A1) $X = (X_1, \cdots, X_n) \sim_{iid} f(x|\theta)$ with respect to μ with common support.

(A2) The true parameter value θ_0 is an interior point of the parameter space Ω.

THEOREM **5.3** *For any fixed $\theta \neq \theta_0$, we have*

$$P_{\theta_0}[L(\theta_0|\mathbf{X}) > L(\theta|\mathbf{X}), \textit{for large } n] = P_{\theta_0}[l(\theta_0) > l(\theta), \textit{for large } n] = 1.$$

Proof. By SLLN and Jensen inequality, we have

$$\frac{1}{n}[l(\theta) - l(\theta_0)] = \frac{1}{n}\sum \log \frac{f(X_i|\theta)}{f(X_i|\theta_0)} \longrightarrow_{a.s.} E_{\theta_0} \log \frac{f(X_i|\theta)}{f(X_i|\theta_0)} < \log E_{\theta_0} \frac{f(X_i|\theta)}{f(X_i|\theta_0)} = 0. \quad \blacksquare$$

Although for a single point x, we don't usually have $l(\theta_0) = f(x_1|\theta_0) > f(x_1|\theta) = l(\theta)$. However, for large enough sample size n, we have $l(\theta_0) = \prod_{i=1}^n f(x_i|\theta_0) > \prod_{i=1}^n f(x_i|\theta) = l(\theta)$ with high probability. We do not know θ_0, but we can determine the value $\hat{\theta}$ of θ which maximizes $l(\theta) = \prod_{i=1}^n f(x_i|\theta)$. The above theorem suggests that, if $f_\theta(x)$ varies smoothly with θ, the MLE $\hat{\theta}$ typically should be close to the true value θ_0, and hence be a reasonable estimator.

In fact, we illustrate this for the simple case in which Θ is finite, say, $\Theta = \{\theta_0, \theta_1, \cdots, \theta_k\}$. By the last theorem, for any θ_i, $i = 1, ..., k$, we have

$$P_{\theta_0}[l(\theta_0) > l(\theta_i), \text{for large } n] = 1.$$

Hence,

$$P_{\theta_0}\left(l(\theta_0) > \max_{1 \leq i \leq n} l(\theta_i), \text{for large } n\right) = P_{\theta_0}[l(\theta_0) > l(\theta_i), \text{for large } n] = 1.$$

5.6 MLE is asymptotically efficient

5.6.1 Asymptotically efficiency

Suppose that $\{\delta_n\}$ is a sequence of estimators $g(\theta)$. Then typically,

1. **(Consistency):** $\delta_n \to_p g(\theta)$.
2. **(Asymptotic normality):** $\sqrt{n}[\delta_n - g(\theta)] \to_d N(0, v(\theta))$.

If the lower bound in (4.5) is attained, δ_n is called *Asymptotic Efficient Estimators.*

Definition. A sequence $\{\delta_n\}$ is said to be **asymptotically efficient (AE)** if

$$\sqrt{n}[\delta_n - g(\theta)] \to_d N\left(0, \frac{[g'(\theta)]^2}{I(\theta)}\right).$$

Two questions immediately arise:

- Do AE estimators exist? Under what conditions?

  AE estimators are not unique. If δ_n is AE, so is $\delta_n + R_n$ if $\sqrt{n}R_n \to_p 0$, since

  $$\sqrt{n}\left(\delta_n + R_n - g(\theta)\right) = \sqrt{n}\left(\delta_n - g(\theta)\right) + \sqrt{n}R_n \to_d N\left(0, \frac{[g'(\theta)]^2}{I(\theta)}\right).$$

- How do we find such estimators if they exist?

  AE estimators can be obtained by different methods, like MLE and Bayes.

5.6.2 MLE is asymptotically efficient

Denote $\mathbf{x} = (x_1, ..., x_n)$, then

$$l(\theta) = \log f_\theta(\mathbf{x}) = \sum_{i=1}^n \log f_\theta(x_i), \qquad l'(\theta) = \partial \log f_\theta(\mathbf{x})/\partial\theta = \sum_{i=1}^n \partial \log f_\theta(x_i)/\partial\theta.$$

Theorem **5.4** *Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, where Θ is an open set in R. Assume that*

1. *For each $\theta \in \Theta$, $l'''(\theta)$ exists for all x, and also there exists a $H(x) \geq 0$ (possibly depending on θ) such that for $\theta \in N(\theta_0, \epsilon) = \{\theta : |\theta - \theta_0| < \epsilon\}$,*

   $$|l'''(\theta)| \leq H(x), \qquad E_\theta H(X_1) < \infty.$$

2. *For $g_\theta(x) = f_\theta(x)$ or $l'''(\theta)$, we have $\frac{\partial}{\partial\theta}\int g_\theta(x)dx = \int \frac{\partial g_\theta(x)}{\partial\theta}dx$.*
3. *For each $\theta \in \Theta$, we have $0 < I_{X_1}(\theta) = E_\theta\left(l'(\theta)\right)^2 < \infty$.*

Let $X_1, ..., X_n \sim F_\theta$. Then with probability 1, the likelihood equations, $l'(\theta) = 0$, admit a sequence of solutions $\{\hat{\theta}_n\}$ satisfying

1. *(Strong consistency)* $\quad \hat{\theta}_n \to_{a.s.} \theta_0$
2. *(Asymptotic efficiency)* $\quad \sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to_d N\left(0, I_{X_1}^{-1}(\theta)\right)$.

Proof.

1. Let $s(\theta) =: s(\mathbf{X}, \theta) = l'(\theta)/n$. Then

$$|s''(\theta)| \le \frac{1}{n}\sum_{i=1}^n \left|\frac{\partial^3 \log f_\theta(X_i)}{\partial\theta^3}\right| \le \frac{1}{n}\sum_{i=1}^n |H(X_i)| =: \bar{H}(\mathbf{X})$$

where $\bar{H}(\mathbf{X}) = n^{-1}\sum_{i=1}^n H(X_i)$. By Taylor's expansion,

$$\begin{aligned} s(\theta) &= s(\theta_0) + s'(\theta_0)(\theta - \theta_0) + \frac{1}{2}s''(\xi)(\theta - \theta_0)^2 \\ &= s(\theta_0) + s'(\theta_0)(\theta - \theta_0) + \frac{1}{2}\bar{H}(\mathbf{X})\eta^*(\theta - \theta_0)^2 \end{aligned}$$

where $|\eta^*| = |s''(\xi)|/\bar{H}(\mathbf{X}) \le 1$ by assumption. By the SLLN, we have, a.s.

$$\begin{aligned} &s(\theta_0) \to Es(\theta_0) = 0, \\ &s'(\theta_0) \to Es'(\theta_0) = -I_{X_1}(\theta_0), \\ &\bar{H}(\mathbf{X}) \to E_\theta H(X_i) < \infty. \end{aligned}$$

Remark **5.4** *Heuristically, we have, for large n and $\theta = \theta_0 \pm \epsilon$,*

$$s(\theta_0 \pm \epsilon) \approx \mp I_{X_1}(\theta_0)\epsilon + \frac{1}{2}EH(X_1)\eta^*\epsilon^2 \approx \mp I_{X_1}(\theta_0)\epsilon$$

which is negative at $\theta_0 + \epsilon$ and positive at $\theta_0 - \epsilon$. Thus there is a change of sign for $s(\theta)$ near θ_0, and hence must have at least one root. By choosing ϵ smaller and smaller, we get a sequence of solutions converging to θ_0. Let us make this precise below.

Fix $\epsilon > 0$, as n is large enough, say $n \ge n_\epsilon$ (depending on ϵ), we have, a.s.

$$\begin{aligned} &|s(\theta_0)| \le \epsilon^2, \\ &|s'(\theta_0) + I_{X_1}(\theta_0)| \le \epsilon^2, \\ &|\bar{H}(\mathbf{X}) - E_\theta H(X_i)| \le \epsilon^2. \end{aligned}$$

Consequently, if ϵ is chosen to be small enough, we have

$$\begin{aligned} s(\theta_0 + \epsilon) &= s(\theta_0) + s'(\theta_0)\epsilon + \frac{1}{2}\bar{H}(\mathbf{X})\eta^*\epsilon^2 \\ &\le \epsilon^2 + \left(-I_{X_1}(\theta_0) + \epsilon^2\right)\epsilon + \frac{1}{2}\left(E_{\theta_0}H(X_i) + \epsilon^2\right)\epsilon^2 \\ &= -I_{X_1}(\theta_0)\epsilon + \epsilon^2 + \frac{1}{2}\left(E_{\theta_0}H(X_i) + \epsilon^2\right)\epsilon^2 + \epsilon^3 \\ &< 0, \end{aligned}$$

Similarly, as n is large enough, we have, a.s.

$$s(\theta_0 - \epsilon) = s(\theta_0) - s'(\theta_0)\epsilon + \frac{1}{2}\bar{H}(\mathbf{X})\eta^*\epsilon^2 > 0.$$

Then, by continuity of $s(\theta)$, the interval $[\theta_0 - \epsilon, \theta_0 + \epsilon]$ contains a solution of $s(\theta) = 0$. In particular, it contains the solution

$$\hat{\theta}_{n_\epsilon,\epsilon} = \inf\left\{\theta : \theta_0 - \epsilon \le \theta \le \theta_0 + \epsilon, \quad \text{and} \quad s(\theta) = 0\right\}.$$

We now show that $\hat{\theta}_{n_\epsilon,\epsilon}$ is a proper random variable, that is, $\hat{\theta}_{n_\epsilon,\epsilon}$ is measureable. Note that, for all $t \ge \theta_0 - \epsilon$, we have

$$\begin{aligned} \{\hat{\theta}_{n_\epsilon,\epsilon} > t\} &= \left\{\inf_{\theta_0-\epsilon\le\theta\le t} s(\theta) > 0\right\} \\ &= \left\{\inf_{\theta_0-\epsilon\le\theta\le t, \theta \text{ is rational}} s(\theta) > 0\right\}, \end{aligned}$$

which is clearly a measurable set, since $s(\theta)$ is continuous in θ.

Take $\epsilon = 1/k$ for $k \ge 1$. For each $k \ge 1$, we define a sequence of r.v.'s by

$$\hat{\theta}_k = \hat{\theta}_{n_{1/k},1/k}, \qquad k = 1, 2, ...$$

Clearly, $\hat{\theta}_k \to \theta_0$ with probability one.

2. For large n, we have seen that

$$0 = s(\hat{\theta}_n) = s(\theta_0) + s'(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}\bar{H}(\mathbf{X})\eta^*(\hat{\theta}_n - \theta_0)^2$$

Thus,

$$\sqrt{n}s(\theta_0) = \sqrt{n}(\hat{\theta}_n - \theta_0)\left(-s'(\theta_0) - \frac{1}{2}\bar{H}(\mathbf{X})\eta^*(\hat{\theta}_n - \theta_0)\right)$$

Since $\sqrt{n}s(\theta_0) \to N(0, I(\theta_0))$ in distribution by CLT, and $-s'(\theta_0) - \frac{1}{2}\bar{H}(\mathbf{X})\eta^*(\hat{\theta}_0 - \theta_0) \to I(\theta_0)$ a.s., then it follows from Slutsky's theorem that

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta_0) &= \frac{\sqrt{n}s(\theta_0)}{-s'(\theta_0) - \frac{1}{2}\bar{H}(\mathbf{X})\eta^*(\hat{\theta}_0 - \theta_0)} \\ &\to_d \frac{N(0, I(\theta_0))}{I(\theta_0)} = N\left(0, I^{-1}(\theta_0)\right). \quad \blacksquare \end{aligned}$$

Remark **5.5** *To estimate $g(\theta)$ (g is smooth), we can use $g(\hat{\theta}_n)$. Consequently, we can apply the continuous mapping theorem and δ-method to obtain*

- *Strong consistency: $g(\hat{\theta}_n) \to g(\theta_0)$ with probability 1.*
- *Asymptotic efficiency: $\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta_0)\right) \to_d N\left(0, \frac{[g'(\theta)]^2}{I_{X_1}(\theta)}\right)$.*

5.6.3 An example: asymptotic efficiency holds when the assumptions are voilated

EXAMPLE **5.4** *Let $X_1, \ldots, X_n \sim_{iid} f_\theta(x) = \frac{1}{2}\exp\{-|x-\theta|\}$ (the double exponential). Despite the violation of regularity conditions, the asymptotic version of the Cramer-Rao theorem still holds for the MLE.*

Proof. First of all, for EVERY x, $f_\theta(x)$ is not differentable w.r.t θ at $\theta = x$. So the regularity conditions are not satisfied. Secondly,

$$l(\theta) = \log \prod_{i=1}^{n} f_\theta(x_i) = \log \frac{1}{2^n} \exp\left\{ -\sum_{i=1}^{n} |x_i - \theta| \right\} = C - \sum_{i=1}^{n} |x_i - \theta|.$$

So an MLE of θ is $\hat{\theta} = \arg\max_{\theta \in R} l(\theta) = \arg\min_{\theta \in R} \sum_{i=1}^{n} |x_i - \theta| =$ sample median, as seen in an earlier chapter.

Next $l'(\theta) = -\sum_{i=1}^{n} |x_i - \theta|'$, where $|x_i - \theta|' = \frac{\partial}{\partial \theta}|x_i - \theta| = \pm 1$ except at $\theta = x_i$. Thus $I(\theta) = E[l'(\theta)]^2 = 1$. Thus $\hat{\theta}$ is an asymptotic efficient estimator if

$$\sqrt{n}(\hat{\theta} - \theta) \to_d N\left(0, 1\right).$$

Let us assume that n is odd at this moment, then $\hat{\theta} = \hat{m}$. It is well known that

$$\sqrt{n}(\hat{\theta} - \theta) \sim_{asymptotic} N\left(0, [4f^2(\theta)]^{-1}\right).$$

In our current example, we have $[4f^2(\theta)]^{-1} = [4 \times (1/4)]^{-1} = 1$. Hence, the asymptotic Cramer-Rao bound does hold. ∎

5.7 Some comparisons about the UMVUE and MLE's

1. It can be shown that UMVUE's are always consistent and MLE's usually are. (Bickel and Doksum, 1977., p133.)
2. In the exponential family, both estimates are functions of the natural sufficient statistics $T_i(X)$.
3. When n is large, both give essentially the same answer. In that case, use is governed by ease of computation.
4. MLEs always exist while UMVUEs may not exist, and MLEs are usually easier to calculate than UMVUE's, due partially to the invariance property of MLE's.
5. Neither MLE's nor UMVUE's are satisfactory in general if one takes a Bayesian or minimax point of view. Nor are they necessarily robust.

5.8 Exercises

1. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. In the lecture, we have seen that the information matrix about $\theta = (\mu, \sigma)$ contained in $X_1, \ldots, X_n$ is

$$I_{\mathbf{X}}(\theta) \quad = \quad \frac{n}{\sigma^2} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

   Using this or otherwise, find the information matrix about $\theta = (\mu, \sigma^2)$ contained in $X_1, \ldots, X_n$.

2. Suppose that $X_1, \ldots, X_n$ is a random sample from a population with density $f_\theta(x) = 2x\theta \exp(-\theta x^2)$, for $x > 0$ and $\theta > 0$. Let $T = \sum_{i=1}^n X_i^2/n$.

   (a) Find ET and $Var(T)$.

   (b) Find the Fisher information $I_{\mathbf{X}}(\theta)$.

Chapter 6

Transformation

In statistics, data transformation is the application of a deterministic mathematical function to each point in a data set - that is, each data point x_i is replaced with the transformed value $z_i = f(x_i)$. A transformation changes the shape of a distribution or relationship. It has important applications.

Transforms are usually applied so that the data appear to more closely meet the assumptions of a statistical inference procedure that is to be applied, or to improve the interpretability or appearance of graphs.

In fact, many familiar measured scales are really transformed scales: decibels, pH and the Richter scale of earthquake magnitude are all logarithmic. Furthermore, transformations are needed because there is no guarantee that the world works on the scales it happens to be measured on, such as music.

6.1 Why transformations?

- Convenience

  An example is standardisation: $z = \dfrac{x-a}{b}$. It puts all variables on the same scale.

  Standardisation makes no difference to the shape of a distribution.

- Reducing skewness

  A (nearly) symmetric distribution is often easier to handle and interpret than a skewed one. More specifically, a normal distribution is often regarded as ideal as it is assumed by many statistical methods.

  - To reduce right skewness, try $\sqrt{x}$, $\log x$ or $1/x$.
  - To reduce left skewness, take x^2, x^3, ...

- Equal spreads/variability (homoscedasticity)

  A transformation may be used to produce approximately equal spreads, despite marked variations in level, which again makes data easier to handle and interpret. This is called homoscedasticity, as opposed to heteroscedasticity.

- Linear relationships

  When looking at relationships between variables, it is often far easier to think about patterns that are approximately linear than about patterns that are highly curved. This is vitally important when using linear regression, which amounts to fitting such patterns to data.

  For example, a plot of logarithms of a series of values against time has the property that periods with constant rates of change (growth or decline) plot as straight lines.

- Additive relationships

  Relationships are often easier to analyse for an additive model

  $$y = a + bx$$

  rather than an multiplicative one

  $$y = ax^b$$

  Additivity is a vital issue in analysis of variance.

- Easier to visualize.

  For example, suppose we have a scatterplot in which the points are the countries of the world, and the data values being plotted are the land area and population of each country. If the plot is made using untransformed data (e.g. square kilometers for area and the number of people for population), most of the countries would be plotted in tight cluster of points in the lower left corner of the graph. The few countries with very large areas and/or populations would be spread thinly around most of the graph's area. Simply rescaling units (e.g., to thousand square kilometers, or to millions of people) will not change this. However, following logarithmic transformations of both area and population, the points will be spread more uniformly in the graph.

In practice, a transformation often works, serendipitously, to do several of these at once, particularly to reduce skewness, to produce nearly equal spreads and to produce a nearly linear or additive relationship. But this is not guaranteed.

6.2 Box-Cox transform

6.3 Review of most common transformations

The most useful transformations are the reciprocal, logarithm, cube root, square root, and square.

Reciprocal: $1/x$**,** $x \neq 0$

- This has a drastic effect on distribution shape.
- The reciprocal of a ratio may often be interpreted as easily as the ratio itself: e.g.
  - population density (people per unit area) becomes area per person;
  - persons per doctor becomes doctors per person;
  - rates of erosion become time to erode a unit depth.

Logarithm $\log x$**,** $x > 0$

- This has a major effect on distribution shape. It is commonly used for reducing right skewness.
- Exponential growth or decline $y = a \exp(bx)$ is made linear by

  $$\ln y = \ln a + bx.$$

  Exponential function does not go through origin, as at $x = 0$, we have $y = a \neq 0$.
- Power function $y = ax^b$ is made linear by

  $$\log y = \log a + b \log x.$$

  The power function for $b > 0$ **goes through the origin**, as at $x = 0$ $(b > 0)$, we have $y = 0$. This often makes physical or biological or economic sense.
- Consider ratios $y = p/q \in (0, \infty)$ where $p, q > 0$. Examples are
  - males/females;
  - dependants/workers;
  - downstream length/downvalley length

  which are usually skewed, because there is a clear lower limit and no clear upper limit. The logarithm

  $$\log y = \log p/q = \log p - \log q \in (-\infty, \infty),$$

  which is likely to be more symmetrically distributed.

Square root $x^{1/2}$ $(x \geq 0)$

- The square root $x^{1/2}$ has a moderate effect on distribution shape.
- It is weaker than the logarithm and the cube root.
- It is also used for reducing right skewness, and also has the advantage that it can be applied to zero values.
- Note that the square root of an area has the units of a length. It is commonly applied to counted data, especially if the values are mostly rather small.

Cube root $x^{1/3}$

- The cube root $x^{1/3}$ is a fairly strong transformation with a substantial effect on distribution shape: it is weaker than the logarithm.
- It is also used for reducing right skewness, and has the advantage that it can be applied to zero and negative values.
- Note that the cube root of a volume has the units of a length. It is commonly applied to rainfall data.

Square x^2

- The square x^2 has a moderate effect on distribution shape and it could be used to reduce left skewness.
- In practice, the main reason for using it is to fit a response by a quadratic function

$$y = a + bx + cx^2.$$

****

Test

QQ plots

Box plot

6.4 Preliminaries

Many properties in statistics are preserved under smooth transformations.

Theorem **6.1 (Continuous Mapping Theorem)** *$g : \mathcal{R}^k \to \mathcal{R}$ is continuous. Then*

- $X_n \to_{a.s.} X \implies g(X_n) \to_{a.s.} g(X)$.
- $X_n \to_p X \implies g(X_n) \to_p g(X)$.
- $X_n \to_d X \implies g(X_n) \to_d g(X)$.

Now let us focus on the convergence in distribution (useful in statistical applications). We will see that more could be said: if the limiting distribution of X_n is Normal (or Gamma, Cauchy, etc), so is the limiting distribution of $g(X_n)$. That is, the form of the limiting distribution (such as asymptotic normality) is also preserved under smooth transformation.

Theorem **6.2 (Slutsky's Theorem)** *Let $X_n \to_d X$, $Y_n \to_p C$. Then*

- $X_n + Y_n \to_d X + C$.
- $X_n Y_n \to_d CX$.
- $X_n / Y_n \to_d X/C$ *if* $C \neq 0$.

6.4.1 Stochastic Order relations

Order relations could greatly simplify writings in a proof. Let $\{X_n\}$'s and $\{Y_n\}$'s be two sequences of r.v.'s.

(a) $X_n = O_p(1)$

$\iff \forall \epsilon > 0$, $\exists M_\epsilon$ and N_ϵ such that $P(|X_n| > M_\epsilon) < \epsilon$, $\forall n \geq N_\epsilon$.

$(P(|X_n| \leq M_\epsilon) \geq 1 - \epsilon)$

$\iff \forall \epsilon > 0$, $\exists M_\epsilon$ such that $\sup_n P(|X_n| > M_\epsilon) < \epsilon$.

$\iff \lim_{M\to\infty} \limsup_n P(|X_n| > M) = 0$. ("tightness".)

We say that $\{X_n\}$ is "stochastically bounded", or "tight." It means that no mass will escape to ∞ with positive probability.

(b) $X_n = O_p(Y_n)$ if $X_n/Y_n = O_p(1)$.

(c) $X_n = o_p(1)$ if $X_n \to_p 0$.

(d) $X_n = o_p(Y_n)$, if $X_n/Y_n = o_p(1)$.

Here are some useful relations.

Theorem **6.3**

1. *If $X_n \to_d X$, then $X_n = O_p(1)$.*
2. *If $X_n = o_p(1)$ then $X_n = O_p(1)$.*
3. *$O_p(1)o_p(1) = o_p(1)$.*
4. *$O_p(1)O_p(1) = O_p(1)$.*
5. *$R_n = o(Y_n)$, $Y_n = O_p(1)$, $\Longrightarrow R_n = o_p(1)$.*
6. *$R_n = \frac{r_n}{\sqrt{n}}$, $r_n = O_p(1)$ or $o_p(1)$, $\Longrightarrow R_n = O_p(n^{-1/2})$, or $o_p(n^{-1/2})$.*

Proof.

1. $X_n \to_d X \Longrightarrow |X_n| \to_d |X|$. Note that the set of discontinuity points of $F_{|X|}$, $D(F_{|X|})$, is countable. For every $\epsilon > 0$, we can always find a large $M_\epsilon \in C(F_{|X|})$ (the set of continuity points of $F_{|X|}$) such that
$$P(|X_n| > M_\epsilon) = 1 - F_{|X_n|}(M_\epsilon) \longrightarrow 1 - F_{|X|}(M_\epsilon) = P(|X| > M_\epsilon) < \epsilon.$$

2. $X_n = o_p(1) \iff X_n \to_p 0 \Longrightarrow P(|X_n| > 1) \to 0 \Longrightarrow P(|X_n| > 1) < \epsilon$, $\forall n \geq N_\epsilon$. $\Longrightarrow$ $X_n = O_p(1)$.

   Alternatively, $X_n = o_p(1) \Longrightarrow \forall \epsilon > 0 : \lim_n P(|X_n| > \epsilon) = 0 \Longrightarrow$
$$\lim_{M\to\infty} \limsup_n P(|X_n| > M) = \lim_{M\to\infty} 0 = 0. \quad \blacksquare$$

3. Denote $W_n = O_p(1)$ and $V_n = o_p(1)$.

   $\forall \epsilon > 0$, $\exists M_\epsilon$, such that for all large n, we have

$$\begin{aligned} P(|W_nV_n| > \delta) &= P(|W_nV_n| > \delta, |W_n| \le M_\epsilon) + P(|W_nV_n| > \delta, |W_n| > M_\epsilon) \\ &\le P(|V_n| > \delta/M_\epsilon) + P(|W_n| > M_\epsilon) \\ &\le \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

4. Denote $W_n = O_p(1)$ and $U_n = O_p(1)$.

   $\forall \epsilon > 0$, $\exists M_{1\epsilon}$ and $M_{2\epsilon}$ such that for all large n, we have $P(|W_n| > M_{1\epsilon}) \le \epsilon/2$ and $P(|U_n| \ge M_{2\epsilon}) \le \epsilon/2$. Thus,

$$\begin{aligned} P(|W_nU_n| > M_{1\epsilon}M_{2\epsilon}) &\le P(|W_n| > M_{1\epsilon},\ \textit{or}\ |U_n| \ge M_{2\epsilon}) \\ &\le P(|W_n| > M_{1\epsilon}) + P(|U_n| \ge M_{2\epsilon}) \\ &\le \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

5. Proof. $\forall \epsilon > 0$, note that

   $Y_n = O_p(1) \Longrightarrow \forall \eta > 0 : \limsup_n P(|Y_n| > M_\epsilon) < \epsilon.$

   $R_n = o(Y_n) \Longrightarrow \forall \delta > 0 : \lim_n P(|R_n/Y_n| > \delta) = 0.$

   Hence,

$$\begin{aligned} P(|R_n| > \eta) &\le P(|R_n| > \eta, |Y_n| < M_\epsilon) + P(|Y_n| > M_\epsilon) \\ &\le P(|R_n/Y_n| > \eta/M_\epsilon) + P(|Y_n| > M_\epsilon) \\ &\le \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

6. $R_n/n^{-1/2} = \sqrt{n}R_n = r_n = O_p(1)$ or $o_p(1)$ by assumption. ∎

6.5 The δ-method

The δ-method is a very useful tool in proving limit theorems and Slutsky's theorems are extensively employed. We illustrate the idea with smooth function of means.

6.5.1 The univariate case

THEOREM **6.4** *Let $X_1, ..., X_n$ be i.i.d. with $\mu = EX_1$ and $\sigma^2 = Var(X_1) \in (0, \infty)$, and $g'(\mu) \neq 0$. Then,*

$$g(\bar{X}) \to_d N\left(g(\mu), n^{-1}[g'(\mu)]^2\sigma^2\right), \quad \Longleftrightarrow \quad \frac{\sqrt{n}(g(\bar{X}) - g(\mu))}{g'(\mu)\sigma} \to_d N(0,1). \tag{5.1}$$

Proof. From Young's Taylor's expansion,

$$g(\bar{X}) = g(\mu) + g'(\mu)(\bar{X} - \mu) + o(|\bar{X} - \mu|)$$

Then,

$$\begin{aligned} \frac{\sqrt{n}(g(\bar{X}) - g(\mu))}{g'(\mu)\sigma} &= \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} + o(|\sqrt{n}(\bar{X} - \mu)|) \\ &= \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} + o_p(1) \\ &\quad (\text{as } \sqrt{n}(\bar{X} - \mu)/\sigma = O_p(1)) \\ &\longrightarrow_d N(0,1), \qquad (\text{Slutsky's Theorem}). \quad \blacksquare \end{aligned}$$

The following generalization is straightforward.

THEOREM **6.5** *If $T_n \to_d N(\mu, \sigma_n^2)$ with $\sigma_n \to 0$, and $g'(\mu) \neq 0$, then,*

$$g(T_n) \longrightarrow_d N\left(g(\mu), [g'(\mu)]^2\sigma_n^2\right), \Longleftrightarrow \frac{g(T_n) - g(\mu)}{g'(\mu)\sigma_n} \longrightarrow_d N(0,1). \quad \blacksquare$$

Now consider the case that g is differentiable but $g'(\mu) = 0$.

THEOREM **6.6** *Suppose T_n is $AN(\mu, \sigma_n^2)$ with $\sigma_n \to 0$. If $g(\cdot)$ is m times at μ differentiable with $g^{(j)}(\mu) = 0$ for $1 \le j \le m-1$ and $g^{(m)}(\mu) \neq 0$, then,*

$$\frac{g(T_n) - g(\mu)}{\frac{1}{m!}g^{(m)}(\mu)\sigma_n^m} \longrightarrow_d [N(0,1)]^m.$$

Proof. We shall prove this for $m = 2$ only.

$$\begin{aligned} g(T_n) &= g(\mu) + g'(\mu)(T_n - \mu) + \frac{g''(\mu)}{2}(T_n - \mu)^2 + o[(T_n - \mu)^2] \\ &= g(\mu) + \frac{g''(\mu)\sigma_n^2}{2}\left(\frac{T_n - \mu}{\sigma_n}\right)^2 + o_p(\sigma_n^2) \qquad (\text{as } (T_n - \mu)/\sigma_n = O_p(1)). \end{aligned}$$

Applying Slutsky's Theorem, we get

$$\frac{[g(T_n) - g(\mu)]}{\frac{g''(\mu)\sigma_n^2}{2}} = \left(\frac{T_n - \mu}{\sigma_n}\right)^2 + o_p(1) \longrightarrow_d [N(0,1)]^2.$$

6.5.2 The multivariate case

THEOREM **6.7** *If $X, X_1, ..., X_n$ are i.i.d. random k-vectors with $\mu = EX$ and $\Sigma = Cov(X, X)$. Assume that $g(\cdot) : R^k \to R^1$ is differentiable with $\nabla g(\mu) \neq 0$, where $\nabla g(x)_{k\times 1} = (\partial g(x)/\partial x_1, ..., \partial g(x)/\partial x_k)^\tau$. Then,*

$$g(\bar{X}) \longrightarrow_d N\left(g(\mu), v_n\right).$$

where $v_n = n^{-1}\nabla g(\mu)^\tau_{1\times k}\Sigma_{k\times k}\nabla g(\mu)_{k\times 1}$. ∎

Proof. From Taylor's expansion, $g(\bar{X}) = g(\mu) + \bigtriangledown^\tau g(\mu)(\bar{X} - \mu) + R$, hence

$$\sqrt{n}[g(\bar{X}) - g(\mu)] = \bigtriangledown^\tau g(\mu)\sqrt{n}(\bar{X} - \mu) + \sqrt{n}R.$$

By the multivariate CLT, we get $\sqrt{n}(\bar{X}-\mu) \to_d N(0, \Sigma)$. Furthermore, $\sqrt{n}R = o(\sqrt{n}||\bar{X} - \mu||) = o(O_p(1)) = o_p(1)$. Then apply Slutsky's Theorem. ∎

6.5.3 Examples

EXAMPLE **6.1** *Find the limiting d.f.'s of* $(\bar{X})^2, 1/\bar{X}, S^2, S, S/\bar{X}$.

Solution.

1. Take $g(x) = x^2$, so $g'(x) = 2x$.

   If $\mu \neq 0$, $(\bar{X})^2 \sim AN\left(\mu^2, \dfrac{4\mu^2\sigma^2}{n}\right)$. If $\mu = 0$, $\dfrac{n(\bar{X})^2}{\sigma^2} \to_d \chi_1^2$.

2. Take $g(x) = 1/x$, so $g'(x) = -1/x^2$.

   If $\mu \neq 0$, $1/\bar{X} \sim AN(1/\mu, n^{-1}(\sigma^2/\mu^4))$. If $\mu = 0$, $\dfrac{\sigma}{\sqrt{n}\bar{X}} \to_d \dfrac{1}{N(0,1)}$. ∎

3. Let $Y_i = X_i - \mu$. Take $g(x,y) = y - x^2$, then

   $$S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n}\sum_{i=1}^n Y_i^2 - (\bar{Y})^2 = g\left(\bar{Y}, \overline{Y^2}\right).$$

   Now $\tilde{\mu} = E\left(\bar{Y}, \overline{Y^2}\right) = (0, \sigma^2)$, $\bigtriangledown g(\tilde{\mu}) = (-2x, 1)^\tau|_{\tilde{\mu}} = (0,1)^\tau$, and

   $$\Sigma = \begin{pmatrix} Var(Y) & Cov(Y, Y^2) \\ Cov(Y, Y^2) & Var(Y^2) \end{pmatrix} = \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}$$

   Then, $g(\tilde{\mu}) = \sigma^2, \quad \bigtriangledown^\tau g(\tilde{\mu})\Sigma \bigtriangledown g(\tilde{\mu}) = \mu_4 - \sigma^4$. Therefore,

   $$S^2 \longrightarrow_d N(\sigma^2, n^{-1}(\mu_4 - \sigma^4)).$$

4. For $S = \sqrt{S^2} = h(S^2)$, where $h(x) = \sqrt{x}$. So $h'(x) = 1/(2\sqrt{x})$. From the above,

   $$S \longrightarrow_d N(h(\sigma^2), n^{-1}[h'(\sigma^2)]^2(\mu_4 - \sigma^4)) = N\left(\sigma, \frac{\mu_4 - \sigma^4}{4\sigma^2 n}\right). \quad ∎$$

5. If $\mu \neq 0$, take $g(x,y) = (y - x^2)^{1/2}/x$. Then

   $$S/\bar{X} = \left(\frac{1}{n}\sum_{i=1}^n X_i^2 - (\bar{X})^2\right)^{1/2} / \bar{X} := g(x,y).$$

   After some routine but tedious calculation, we get

   $$\frac{S}{\bar{X}} \longrightarrow_d N\left(\frac{\sigma}{\mu}, \frac{1}{n}\left(\frac{\sigma^2\mu_2}{\mu^4} - \frac{\mu_3}{\mu^3} + \frac{\mu_4 - \mu_2^2}{4\mu^2\sigma^2}\right)\right).$$

   If $\mu = 0$, we take $T_n = \sqrt{n}\bar{X}/S \longrightarrow_d N(0,1) := Z$ and $g(t) = 1/t$, then $g(T_n) \longrightarrow_d g(Z)$ (continuous mapping theorem), i.e.,

   $$\frac{S}{\sqrt{n}\bar{X}} \longrightarrow_d \frac{1}{N(0,1)}.$$

6.6 Variance-stabilizing transformations

We often have

$$T_n \to_d N(\mu, \sigma^2(\mu))$$

i.e., the variance changes with the mean.

We now seek some transformation g so that $g(T_n)$ has stable variance. Since

$$g(T_n) \longrightarrow_d N\left(g(\mu), [g'(\mu)]^2\sigma(\mu)^2\right),$$

we would like to have

$$[g'(\mu)]^2\sigma^2(\mu) = C^2, \qquad i.e., \qquad g'(\mu) = \frac{C}{\sigma(\mu)}.$$

Thus

$$g(\mu) = \int \frac{C}{\sigma(\mu)} d\mu.$$

Example. $X_1, ..., X_n \sim \exp(\lambda)$, i.e., $f(x) = \lambda^{-1}\exp\{-x/\lambda\}$, $x \geq 0$. Then,

$$\mu = \lambda, \qquad \sigma^2 = \lambda^2 = \mu^2.$$

Here, $\sigma^2(\mu) = \mu^2$, hence,

$$g(\mu) = \int \frac{1}{\sigma(\mu)} d\mu = \int \frac{1}{\mu} d\mu = \ln \mu. \quad \blacksquare$$

In fact, if $X_1, ..., X_n \sim Gamma(k, \lambda)$ with k known, then $EX = k\lambda$, $var(X) = k\lambda^2$. Similarly to the exponential distribution, ln transformation will stabilize variance.

Example. $X_1, ..., X_n \sim Poisson(\lambda)$. The CLT implies $\bar{X} \sim AN(\lambda, \lambda/n)$, hence,

$$1 - \alpha \quad \approx \quad P\left(\Phi^{-1}(\alpha/2) < \frac{\sqrt{n}(\bar{X} - \lambda)}{\sqrt{\lambda}} < \Phi^{-1}(1 - \alpha/2)\right)$$

from which one can get an approximate confidence interval for λ.

Alternatively, we can use the transformation approach. Here, $\sigma^2(\lambda) = \lambda$, hence,

$$g(\theta) = \int \frac{C}{\sigma(\theta)} d\theta = \int \frac{C}{\sqrt{\theta}} d\theta = 2C\sqrt{\theta}.$$

Choose $C = 1/2$. Then we get $g(\theta) = \sqrt{\theta}$, hence

$$\sqrt{\bar{X}} \sim AN\left(\sqrt{\lambda}, \frac{1}{4n}\right), \qquad \text{or} \qquad 2\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda}) \sim AN(0, 1)$$

Hence, an approximate two-sided confidence interval for $\sqrt{\lambda}$ at $(1 - \alpha)$-level is

$$(L, U) = \sqrt{\bar{X}} \pm \frac{1}{2\sqrt{n}} \Phi^{-1}(1 - \alpha/2).$$

Hence, an approximate two-sided confidence interval for λ at $(1 - \alpha)$-level is given by (L^2, U^2) if $L > 0$, or $(0, U^2)$ if $L \leq 0$. $\blacksquare$

6.6.1 An example: the sample correlation coefficient

Example 1. Let $(X, Y), (X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. random vectors. For simplicity, we first assume that $\mu_x = \mu_y = 0$. Then the correlation coefficient is

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{E(X-\mu_x)(Y-\mu_y)}{\sqrt{V(X)V(Y)}} = \frac{EXY}{\sqrt{EX^2EY^2}}.$$

Correspondingly, we consider a simpler sample correlation coefficient given by

$$\hat{\rho} = \frac{\frac{1}{n}\sum_{i=1}^n X_iY_i}{\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^{1/2}}.$$

(**Note that this is different from the usual definition of $\hat{\rho}$; see the next example.**)

(i) Find the limiting d.f. of $\hat{\rho}$.

(ii) Further assume (X, Y) is bivariate normal with pdf

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_1^2(1-\rho^2)} + \frac{\rho(x-\mu_x)(y-\mu_2)}{\sigma_1\sigma_2(1-\rho^2)} - \frac{(y-\mu_2)^2}{2\sigma_2^2(1-\rho^2)}\right\}$$

(a) Show that $\hat{\rho} \to_{a.s.} \rho$, and $\sqrt{n}(\hat{\rho}-\rho) \to_d N(0,(1-\rho^2)^2)$. Further,

$$\frac{\sqrt{n}(\hat{\rho}-\rho)}{(1-\hat{\rho}^2)} \to_d N(0,1).$$

(b) Find a variance stabilizing transformation (Fisher transformation).

(**Hint:** X and $Y - \beta X$ are independent for $\beta = \sigma_{xy}/\sigma_x^2$, since $(X, Y - \beta X)$ is bivariate Normal with $Cov(X, Y - \beta X) = EXY - \beta EX^2 = 0$.)

Solution. (i) Let $\theta = (EX^2, EY^2, EXY)$ and $\hat{\theta} = \left(\frac{1}{n}\sum_{i=1}^n X_i^2, \frac{1}{n}\sum_{i=1}^n Y_i^2, \frac{1}{n}\sum_{i=1}^n X_iY_i\right)$, and

$$g(z_1, z_2, z_3) = \frac{z_3}{(z_1z_2)^{1/2}}.$$

Then, $\rho = g(\theta)$ and $\hat{\rho} = g(\hat{\theta})$. Note that $E\hat{\theta} = \theta$. Assuming that $EX^4 < \infty$ and $EY^4 < \infty$, from the multivariate CLT, we have

$$\hat{\theta} \sim AN(\theta, \Sigma/n),$$

where $\Sigma_{3\times3}$ is the covariance matrix of (X^2, Y^2, XY):

$$\Sigma = \begin{pmatrix} V(X^2) & Cov(X^2,Y^2) & Cov(X^2,XY) \\ Cov(X^2,Y^2) & V(Y^2) & Cov(Y^2,XY) \\ Cov(X^2,XY) & Cov(Y^2,XY) & V(XY) \end{pmatrix}.$$

Therefore, from the theorem in this section, we have

$$\hat{\rho} = g(\hat{\theta}) \sim AN(\rho, n^{-1}[\bigtriangledown g(\theta)]^\tau \Sigma \bigtriangledown g(\theta)),$$

From

$$[\bigtriangledown g(z)]^\tau = \left(\frac{\partial g(z)}{\partial z_1}, \frac{\partial g(z)}{\partial z_2}, \frac{\partial g(z)}{\partial z_3}\right) = \left(-\frac{z_3}{2z_1^{3/2}z_2^{1/2}}, -\frac{z_3}{2z_1^{1/2}z_2^{3/2}}, \frac{1}{z_1^{1/2}z_2^{1/2}}\right),$$

we have

$$[\bigtriangledown g(\theta)]^\tau = \left(-\frac{\sigma_{xy}}{2\sigma_x^3\sigma_y}, -\frac{\sigma_{xy}}{2\sigma_x\sigma_y^3}, \frac{1}{\sigma_x\sigma_y}\right).$$

(ii) (a) From $\hat{\theta} \to_{a.s.} \theta$, we get $\hat{\rho} = g(\hat{\theta}) \to_{a.s.} g(\theta) = \theta$.

For simplicity, we shall also assume that $\sigma_x^2 = \sigma_y^2 = 1$. Then $X \sim N(0,1)$, $X^2 \sim \chi_1^2$, hence $EX^2 = 1$, $V(X^2) = 2$, $EX^3 = 0$, $EX^4 = 3$. Similar results hold for Y as well. Note that X and $Y - \beta X$ are independent for $\beta = \sigma_{xy}/\sigma_x^2 = \rho$. Therefore

$$\begin{aligned}
E(X^2Y^2) &= E(X^2(Y-\beta X+\beta X)^2) \\
&= EX^2(Y-\beta X)^2 + 2\beta EX^3(Y-\beta X) + \beta^2 EX^4 \\
&= EX^2E(Y-\beta X)^2 + 2\beta EX^3E(Y-\beta X) + \beta^2 EX^4 \\
&= EX^2E(Y-\beta X)^2 + \beta^2 EX^4 \\
&= EX^2\left(EY^2 - 2\beta EXY + \beta^2 EX^2\right) + \beta^2 EX^4 \\
&= \left(1 - 2\rho^2 + \rho^2\right) + 3\rho^2 \\
&= 1 + 2\rho^2 \\
V(XY) &= E(X^2Y^2) - (EXY)^2 = 1 + 2\rho^2 - \rho^2 = 1 + \rho^2 \\
Cov(X^2, Y^2) &= E(X^2Y^2) - EX^2EY^2 = 1 + 2\rho^2 - 1 = 2\rho^2 \\
Cov(X^2, XY) &= E(X^3Y) - EX^2EXY = EX^3(Y-\beta X) + \beta EX^4 - \rho \\
&= EX^3E(Y-\beta X) + 3\rho - \rho = 2\rho \\
Cov(Y^2, XY) &= 2\rho \qquad \text{(similar to above)}
\end{aligned}$$

Hence,

$$\Sigma = \begin{pmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1+\rho^2 \end{pmatrix} = 2\begin{pmatrix} 1 & \rho^2 & \rho \\ \rho^2 & 1 & \rho \\ \rho & \rho & (1+\rho^2)/2 \end{pmatrix}.$$

and $[\bigtriangledown g(\theta)]^\tau$ can be simplified to

$$[\bigtriangledown g(\theta)]^\tau = \left(-\frac{\sigma_{xy}}{2\sigma_x^3\sigma_y}, -\frac{\sigma_{xy}}{2\sigma_x\sigma_y^3}, \frac{1}{\sigma_x\sigma_y}\right) = (-\rho/2, -\rho/2, 1) = \frac{1}{2}(-\rho, -\rho, 2).$$

Finally,

$$\begin{aligned}
[\bigtriangledown g(\theta)]^\tau \Sigma \bigtriangledown g(\theta) &= \frac{1}{2}(-\rho,-\rho,2)\begin{pmatrix} 1 & \rho^2 & \rho \\ \rho^2 & 1 & \rho \\ \rho & \rho & (1+\rho^2)/2 \end{pmatrix}\begin{pmatrix} -\rho \\ -\rho \\ 2 \end{pmatrix} \\
&= \frac{1}{2}\left(\rho(1-\rho^2), \rho(1-\rho^2), (1-\rho^2)\right)\begin{pmatrix} -\rho \\ -\rho \\ 2 \end{pmatrix} \\
&= \frac{1}{2}\left(-2\rho^2(1-\rho^2) + 2(1-\rho^2)\right) \\
&= (1-\rho^2)^2.
\end{aligned}$$

Therefore, $\sqrt{n}(\hat{\rho} - \rho) \to_d N(0, (1-\rho^2)^2)$. This, $\hat{\rho} \to_{a.s.} \rho$ and Slutsky's theorem imply

$$\frac{\sqrt{n}(\hat{\rho} - \rho)}{(1 - \hat{\rho}^2)} \to_d N(0, 1).$$

(b) Let us find a variance stabilizing transform:

$$h(t) = \int \frac{1}{\sigma(t)} dt = \int \frac{1}{(1-t^2)} dt = \int \frac{1}{(1-t)(1+t)} d\rho = \frac{1}{2} \ln \frac{1+t}{1-t},$$

Hence, $\sqrt{n}\,(h(\hat{\rho}) - h(\rho)) \to_d N(0,1)$. ∎

Example 2. (Continuation of the last example.) Let $(X,Y), (X_1,Y_1), ..., (X_n,Y_n)$ be i.i.d. random vectors. The general definition of correlation coefficient is

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{E(X-\mu_x)(Y-\mu_y)}{\sqrt{V(X)V(Y)}}.$$

The sample correlation coefficient is

$$\hat{\rho} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2\right)^{1/2}}.$$

Find the limiting d.f. of $\hat{\rho}$.

Solution. Let $\theta = (EX, EY, EX^2, EY^2, EXY)$ and $\hat{\theta} = \left(\bar{X}, \bar{Y}, \frac{1}{n}\sum_{i=1}^n X_i^2, \frac{1}{n}\sum_{i=1}^n Y_i^2, \frac{1}{n}\sum_{i=1}^n X_i Y_i\right)$ and

$$g(z_1, ..., z_5) = \frac{z_5 - z_1 z_2}{(z_3 - z_1^2)^{1/2}(z_4 - z_2^2)^{1/2}}.$$

Then, $\rho = g(\theta)$ and $\hat{\rho} = g(\hat{\theta})$. Note that $E\hat{\theta} = \theta$. Assuming that $EX^4 < \infty$ and $EX^4 < \infty$, from the multivariate CLT, we have

$$\hat{\theta} \sim AN(\theta, \Sigma/n),$$

where $\Sigma_{5\times 5}$ is the covariance matrix of (X, Y, X^2, Y^2, XY). (Why?) Therefore, from the theorem in this section, we have

$$\hat{\rho} = g(\hat{\theta}) \sim AN(g(\theta) = \rho, n^{-1}[\triangledown g(\theta)]^\tau \Sigma \triangledown g(\theta)),$$

Some routine calculations yield

$$[\triangledown g(\theta)]^\tau = \left(\frac{\rho\mu_x}{\sigma_x^2} - \frac{\mu_y}{\sigma_x\sigma_y}, \frac{\rho\mu_y}{\sigma_y^2} - \frac{\mu_x}{\sigma_x\sigma_y}, -\frac{\rho}{2\sigma_x^2}, -\frac{\rho}{2\sigma_y^2}, \frac{1}{\sigma_x\sigma_y}\right).$$

Finish the rest as an exercise. ∎

6.7 Exercises

Below $X, X_1, ..., X_n$ are i.i.d. r.v.'s with mean μ, variance σ^2, and $\mu_k = E(X-\mu)^k$.

1. Prove the equivalence of several definitions for $X_n = O_p(1)$ given in the notes.
2. (a) If $X_n = O_p(Z_n)$, and $P(Y_n = 0) = 0$ for all n, then $X_n Y_n = O_p(Y_n Z_n)$.
   (b) If $X_n = O_p(Z_n)$ and $Y_n = O_p(Z_n)$, then $X_n + Y_n = O_p(Z_n)$.
   (c) If $E|X_n| = O(a_n)$ for a sequence $a_n > 0$, then $X_n = O_p(a_n)$.
   (d) If $X_n \to_{a.s.} X$, then $\sup_n |X_n| = O_p(1)$.
3. Find the limiting d.f.'s of $e^{\bar{X}}, \ln|\bar{X}|, \cos(\bar{X})$.
4. Let $(X, Y), (X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. with mean $\mu = (\mu_x, \mu_y)^\tau$ and covariance matrix Σ. Let $r = \mu_x/\mu_y$ with $\mu_y \neq 0$, and an estimate of r is $\hat{r} = \bar{X}/\bar{Y}$. Investigate the limiting behavior of $\hat{r}$ (i.e. both consistency and asymptotic d.f.).
5. Let $U_n \sim_{iid}$ Uniform$[0,1]$, and $X_n = (\prod_{i=1}^{n} U_i)^{-1}$. Show that $\sqrt{n}(X_n^{1/n} - e) \to_d N(0, e^2)$.
6. If $X_n \sim Bin(n,p)$, then $X_n \to_d N(np, np(1-p))$. Find a variance stabilizing transformation.

Chapter 7

Quantiles

We are often interested in a special portion of a population distribution, other than the centre. For instance, we might be interested in

- underweight new-born babies in medical studies;
- losses in financial companies (e.g., VAR) [risk management];
- the low and high water levels of a reservoir.

These are all examples of quantiles. The layout of the chapter is as follows:

- definitions of quantiles
- consistency
- asymptotic normality
- Bahadur theorem
- quantile regression

7.1 Definition

The pth quantile of a d.f. F is

$$\xi_p := F^{-1}(p) = \inf\{t : F(t) \geq p\}.$$

A plot of $F^{-1}(p)$ reveals that

- F^{-1} jumps when F is flat. F^{-1} is flat when F jumps.
- In fact, $F^{-1}(p)$ is a mirror image of $F(t)$ along the line $p = t$.

THEOREM **7.1** *$\forall p \in (0,1)$, we have*

1. *$F^{-1}(p)$ is non-decreasing and left continuous.*
2. *$F^{-1}(F(x)) \leq x$, $\forall x \in R$.*
3. *$F(\xi_p-) \leq p \leq F(\xi_p)$.*
4. *$\xi_p = F^{-1}(p) \leq t \iff \quad p \leq F(t)$.*
5. *If F is continuous, then $F(F^{-1}(p)) = F(\xi_p) = p$.*

Proof.

1. **Monotonicity.** If $p_1 < p_2$, then $\{t : F(t) \geq p_2\} \subset \{t : F(t) \geq p_1\}$. Hence,

$$F^{-1}(p_1) = \inf\{t : F(t) \geq p_1\} \leq \inf\{t : F(t) \geq p_2\} = F^{-1}(p_2).$$

   Left-continuity. Let $p_n \nearrow p$.
   $\Longrightarrow F^{-1}(p_n) \leq F^{-1}(p)$ and $F^{-1}(p_n) \nearrow b$, say.
   (Monotonicity shown above)
   $\Longrightarrow F^{-1}(p_n) \leq b \leq F^{-1}(p)$ for all n.
   $\Longrightarrow p_n \leq F(b)$ from (4) (proof given below).
   $\Longrightarrow F(b) \geq \lim_n p_n = p$.
   $\Longrightarrow b \in \{t : F(t) \geq p\}$, hence, $b \geq F^{-1}(p)$.
   $\Longrightarrow \lim_n F^{-1}(p_n) = b = F^{-1}(p)$.

2. $F^{-1}(F(x)) = \inf\{t : F(t) \geq F(x)\} \leq x$ as $x \in \{t : F(t) \geq F(x)\}$.

3. By definition of ξ_p, we have $F(\xi_p - \epsilon) < p \leq F(\xi_p + \epsilon)$. Then let $\epsilon \to 0$.

4. - If $\xi_p := F^{-1}(p) \leq t$, then $F(t) \geq F[\xi_p] \geq p$ from (3).
   - If $p \leq F(t)$, then $F^{-1}(p) \leq F^{-1}[F(t)] \leq t$ from (2).

5. It follows from (3).

Theorem **7.2 (Quantile transformation.)** *F is a d.f., $U \sim Uniform(0,1)$. Then*

$$X := F^{-1}(U) \sim F.$$

Proof. From Theorem 7.1 (4): $P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$. ∎

Theorem 7.2 forms the basis for sampling from d.f. F, at least in principle. Feasibility of this technique depends on either having F^{-1} available in closed form, or being able to approximate it numerically, or using some other techniques.

7.2 Consistency of sample quantiles

Given $X_1, ..., X_n \sim_{iid} F$, a natural estimate of ξ_p is the sample quantile of F_n:

$$\hat{\xi}_p := F_n^{-1}(p) = \inf\{t : F_n(t) \geq p\}.$$

Theorem **7.3** *If $F(\xi_p - \delta) < p < F(\xi_p + \delta)$ for all $\delta > 0$ (i.e., F is not flat at ξ_p), then for all $\epsilon > 0$ and n,*

$$P(|\hat{\xi}_p - \xi_p| > \epsilon) \leq 2Ce^{-2n\delta_\epsilon^2},$$

where $\delta_\epsilon = \min\{F(\xi_p + \epsilon) - p,\ p - F(\xi_p - \epsilon)\}$, and $C > 0$ is an absolute constant.

Proof. Fix $\epsilon > 0$. Recall $F^{-1}(t) \leq x \iff t \leq F(x)$. Hence,

$$\begin{aligned} & P(\hat{\xi}_p > \xi_p + \epsilon) = P(F_n^{-1}(p) > \xi_p + \epsilon) = P(p > F_n[\xi_p + \epsilon]) \\ = \ & P\left(F_n[\xi_p + \epsilon] - F[\xi_p + \epsilon] < p - F[\xi_p + \epsilon]\right) \\ = \ & P\left(F_n[\xi_p + \epsilon] - F[\xi_p + \epsilon] < -\{F[\xi_p + \epsilon] - p\}\right) \\ \leq \ & P\left(|F_n[\xi_p + \epsilon] - F[\xi_p + \epsilon]| > \{F[\xi_p + \epsilon] - p\}\right) \\ \leq \ & P(\sup_{x \in R} |F_n(x) - F(x)| > \delta_\epsilon) \\ \leq \ & Ce^{-2n\delta_\epsilon^2} \qquad \text{from DWK inequality.} \end{aligned}$$

Similarly, $P(\hat{\xi}_p < \xi_p - \epsilon) \leq Ce^{-2n\delta_\epsilon^2}$. ∎

Corollary **7.1** *Let $F(\xi_p - \epsilon) < p < F(\xi_p + \epsilon)$ for any $\epsilon > 0$. Then $\hat{\xi}_p \longrightarrow_{a.s.} \xi_p$.*

Proof. $\sum_{n=1}^{\infty} P(|\hat{\xi}_p - \xi_p| > \epsilon) \leq \sum_{n=1}^{\infty} 2Ce^{-2n\delta_\epsilon^2} < \infty$. ∎

7.3 Asymptotic normality of the sample qunatiles

Asymptotic normality can be proved in several different ways, e.g. Bahadur representation. For simplicity, let $m = F^{-1}(1/2)$ and $\hat{m} = F_n^{-1}(1/2)$ be the population and sample medians, respectively. General quantiles can be treated similarly.

THEOREM **7.4** *Assume $f(x) = F'(x)$ is continuous near m and $f(m) > 0$. Then*

$$\hat{m} \sim_{asymp.} N\left(m, \frac{1}{4f^2(m)n}\right).$$

Proof. WLOG, assume n is odd. Let $Y_{ni} = I\{X_i \le m + x/\sqrt{n}\}$. Then $\sum Y_{ni} \sim Bin(n, p_n)$, where $p_n = P(X_i \le m + x/\sqrt{n})$. Now

$$\begin{aligned}
P\left(\sqrt{n}(\hat{m} - m) \le x\right) &= P\left(\hat{m} \le m + x/\sqrt{n}\right) \\
&= P\left(\sum_{i=1}^{n} I\{X_i \le m + x/\sqrt{n}\} \ge \frac{n+1}{2}\right) \\
&= P\left(\sum_{i=1}^{n} Y_{ni} \ge \frac{n+1}{2}\right) \\
&= P\left(\frac{\sum_{i=1}^{n} Y_{ni} - np_n}{\sqrt{np_n(1-p_n)}} \ge \frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}}\right)
\end{aligned}$$

Since F is differentiable (hence continuous) at m, thus $F(m) = 1/2$. So

$$p_n = F(m + x/\sqrt{n}) = \frac{1}{2} + \frac{x}{\sqrt{n}} f\left(m + \frac{\delta x}{\sqrt{n}}\right) \to \frac{1}{2},$$

and $\sqrt{n}(p_n - 1/2) \to xf(m)$. Thus,

$$\frac{\frac{1}{2}(n+1) - np_n}{\sqrt{np_n(1-p_n)}} = \frac{\frac{1}{2\sqrt{n}} + \sqrt{n}(\frac{1}{2} - p_n)}{\sqrt{p_n(1-p_n)}} \to \frac{-xf(m)}{1/2} = -2xf(m)$$

Applying the CLT to the triangular array Y_{ni} and Slutsky's theorem, we have

$$P\left(\sqrt{n}(\hat{m} - m) \le x\right) \to 1 - \Phi(-2xf(m)) = \Phi(2xf(m)) = \Phi(x/(2f(m))^{-1}). \quad \blacksquare$$

7.4 Bahadur representation

Sample quantile $\hat{\xi}_p$ is a nonlinear statistic (i.e., not an average of iid r.v.s). However, Bahadur representation says that it can be approximately by a linear statistic, which is much easier to handle.

Heuristically, by Taylor's expansion, $F(\hat{\xi}_p) - F(\xi_p) = f(\xi_p)(\hat{\xi}_p - \xi_p) + ...$ Thus,

$$\hat{\xi}_p - \xi_p \approx \frac{F(\hat{\xi}_p) - F(\xi_p)}{f(\xi_p)} \approx \frac{F_n(\hat{\xi}_p) - F_n(\xi_p)}{f(\xi_p)} \approx \frac{p - F_n(\xi_p)}{f(\xi_p)},$$

where we have used $F \approx F_n$ and $F_n(\hat{\xi}_p) \approx p$ (in fact $|F_n(\hat{\xi}_p) - p| \le 1/n$.)

More precisely, we have

THEOREM **7.5 (Bahadur representation)** *If* $f(\xi_p) = F'(\xi_p) > 0$, *then*

$$\hat{\xi}_p = \xi_p + \frac{F(\xi_p) - F_n(\xi_p)}{f(\xi_p)} + o_p(n^{-1/2}).$$

Or equivalently,

$$\sqrt{n}\left(\hat{\xi}_p - \xi_p\right) = \frac{\sqrt{n}[F(\xi_p) - F_n(\xi_p)]}{f(\xi_p)} + o_p(1). \quad \blacksquare$$

We state the following useful lemma, which will be shown later.

LEMMA **7.1 (Ghosh, 1971)** *If* $X_n = O_p(1)$ *and* $P(X_n \le t, Y_n \ge t+\delta) + P(X_n \ge t+\delta, Y_n \le t) = o(1)$, *for any fixed* $t \in \mathcal{R}$ *and* $\delta > 0$, *then*

$$X_n - Y_n = o_p(1). \quad \blacksquare$$

Proof. It is equivalent to show $X_n - Y_n = o_p(1)$, where

$$X_n := \sqrt{n}(\hat{\xi}_p - \xi_p), \qquad Y_n = \frac{\sqrt{n}[F(\xi_p) - F_n(\xi_p)]}{f(\xi_p)}.$$

To apply the lemma, let $t \in R$ and $\xi_{nt} = \xi_p + \dfrac{t}{\sqrt{n}}$. We have

$$\begin{aligned}
P(X_n \le t, Y_n \ge t+\epsilon) &= P\left(\hat{\xi}_p \le \xi_{nt}, Y_n \ge t+\epsilon\right) \\
&= P\left(F_n(\hat{\xi}_p) \le F_n\left(\xi_{nt}\right), Y_n \ge t+\epsilon\right) \\
&= P\left(F\left(\xi_{nt}\right) - F_n\left(\xi_{nt}\right) \le F\left(\xi_{nt}\right) - F_n(\hat{\xi}_p), Y_n \ge t+\epsilon\right) \\
&= P\left(\frac{\sqrt{n}[F\left(\xi_{nt}\right) - F_n\left(\xi_{nt}\right)]}{f(\xi_p)} \le \frac{\sqrt{n}[F\left(\xi_{nt}\right) - F_n(\hat{\xi}_p)]}{f(\xi_p)},\right. \\
&\qquad \left.\frac{\sqrt{n}[F\left(\xi_p\right) - F_n\left(\xi_p\right)]}{f(\xi_p)} \ge t+\epsilon\right) \\
&:= P\left(Z_n(t) \le U_n(t), Z_n(0) \ge t+\epsilon\right),
\end{aligned}$$

where

$$Z_n(t) = \frac{\sqrt{n}[F\left(\xi_{nt}\right) - F_n\left(\xi_{nt}\right)]}{f(\xi_p)}, \qquad U_n(t) = \frac{\sqrt{n}[F\left(\xi_{nt}\right) - F_n(\hat{\xi}_p)]}{f(\xi_p)}.$$

We can show (later), as $n \to \infty$,

(a) $U_n(t) = t + o(1)$, (b) $Z_n(t) - Z_n(0) = o_p(1)$.

For any $t \in \mathcal{R}$ and $\epsilon > 0$, from (a) and (b), we have

$$\begin{aligned} P(X_n \le t, Y_n \ge t+\epsilon) &= P\left(Z_n(t) \le U_n(t), Z_n(0) \ge t+\epsilon\right) \\ &\le P\left(Z_n(t) \le U_n(t), Z_n(0) \ge t+\epsilon, |U_n(t)-t| \le \epsilon/2\right) \\ &\qquad +P\left(|U_n(t)-t| \ge \epsilon/2\right) \\ &\le P\left(|Z_n(t)-Z_n(0)| \ge \epsilon/2\right) + P\left(|U_n(t)-t| \ge \epsilon/2\right) \\ &\longrightarrow 0. \end{aligned}$$

Proof of (a): First we prove the following.

(a1) $\left|F_n(\hat{\xi}_p) - p\right| \le n^{-1}$.

Proof. Let $\hat{\xi}_p = F_n^{-1}(p)$. Note that $F_n(\hat{\xi}_p-) \le p \le F_n(\hat{\xi}_p)$. Thus $|F_n(\hat{\xi}_p) - p| \le |F_n(\hat{\xi}_p) - F_n(\hat{\xi}_p-)| = n^{-1}$.

(a2) Note $\dfrac{\sqrt{n}[F(\xi_{nt}) - p]}{f(\xi_p)} \to t$, as $\dfrac{F\left(\xi_p + \frac{t}{\sqrt{n}}\right) - F(\xi_p)}{\frac{t}{\sqrt{n}}} \to f(\xi_p)$.

Then, from (a1) and (a2), we have, as $n \to \infty$,

$$U_n(t) = \frac{\sqrt{n}[F(\xi_{nt}) - p]}{f(\xi_p)} + \frac{\sqrt{n}[p - F_n\left(\hat{\xi}_p\right)]}{f(\xi_p)} \to t.$$

Proof of (b): Rewrite

$$\begin{aligned} Z_n(t) - Z_n(0) &= \frac{\sqrt{n}[F(\xi_{nt}) - F_n(\xi_{nt})]}{f(\xi_p)} - \frac{\sqrt{n}[F(\xi_p) - F_n(\xi_p)]}{f(\xi_p)} \\ &= \frac{\sqrt{n}[F(\xi_{nt}) - F(\xi_p)]}{f(\xi_p)} - \frac{\sqrt{n}[F_n(\xi_{nt}) - F_n(\xi_p)]}{f(\xi_p)} \\ &= -\frac{\sum_{i=1}^n \left(I\{\xi_p < X_i \le \xi_{nt}\} - EI\{\xi_p < X_i \le \xi_{nt}\}\right)}{\sqrt{n} f(\xi_p)} \\ &= -\frac{\sum_{i=1}^n (Y_{ni} - EY_{ni})}{\sqrt{n} f(\xi_p)} \qquad (4.1) \end{aligned}$$

where $Y_{ni} = I\{\xi_p < X_i \le \xi_{nt}\}$. Note that $EY_{ni} = P(\xi_p < X_1 \le \xi_{nt}) = F(\xi_{nt}) - F(\xi_p) = f(\xi_p)t/\sqrt{n} + o(n^{-1/2})$ [see the proof of (a2) above]. Thus,

$$\begin{aligned} Var\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_{ni} - EY_{ni})\right) &= \frac{1}{n}\sum_{i=1}^n Var(Y_{ni}) \le Var(Y_{n1}) \\ &\le E(Y_{n1}^2) = EY_{n1} = f(\xi_p)t/\sqrt{n} + o(n^{-1/2}) \\ &\longrightarrow 0, \end{aligned}$$

which implies that the first term in (4.1) is $o_p(1)$. Thus $Z_n(t) - Z_n(0) = o_p(1)$.

Similarly, we can show that $P(X_n \ge t+\epsilon, Y_n \le t) \to 0$.

Applying Lemma 7.1, we get $X_n - Y_n = o_p(1)$. ∎

An useful application of Bahadur representation

THEOREM **7.6** *Let F have positive derivatives at ξ_{p_j}, where $0 < p_1 < ... < p_m < 1$. Then*

$$\sqrt{n}\left[\left(\hat{\xi}_{p_1}, ..., \hat{\xi}_{p_m}\right) - (\xi_{p_1}, ..., \xi_{p_m})\right] \longrightarrow_d N_m(0, \Sigma),$$

where $\Sigma = (\sigma_{ij})_{m\times m}$ with

$$\sigma_{ij} = \frac{p_i(1-p_j)}{f(\xi_{p_i})f(\xi_{p_j})}. \quad ∎$$

The theorem is useful in deriving the (joint) d.f. of the sample range, inter-quartile range, etc. The proof is left as an exercise.

Proof of Lemma 7.1

Fix $\epsilon > 0$ any $\delta > 0$. Since $X_n = O_p(1)$, there exists M_ϵ and N_ϵ such that

$$P(|X_n| > M_\epsilon) < \epsilon, \qquad \forall n \geq N_\epsilon.$$

Divide the intervals $[-M_\epsilon, M_\epsilon]$ into m sub-intervals with the length of each sub-interval $\leq \delta/2$, e.g., we can take $m = 2[2M_\epsilon/(\delta/2)]$. Write $[-M_\epsilon, M_\epsilon] \subset \sum_{k=1}^m [t_k, t_{k+1}]$ with $t_{k+1} - t_k \leq \delta/2$. Then for all $n \geq N_\epsilon$, we have

$$\begin{aligned}
P(|X_n - Y_n| \geq \delta) &\leq P(|X_n - Y_n| \geq \delta,\ |X_n| \leq M_\epsilon) + P(|X_n| > M_\epsilon) \\
&\leq P(|X_n - Y_n| \geq \delta,\ X_n \in [-M_\epsilon, M_\epsilon]) + \epsilon \\
&\leq \sum_{k=1}^m P(|X_n - Y_n| \geq \delta,\ X_n \in [t_k, t_{k+1}]) + \epsilon \\
&\quad \text{(since } X_n \in [t_k, t_{t+1}]\text{, a very small interval,} \\
&\quad \text{essentially we have fixed the value of } X_n\text{)} \\
&= \sum_{k=1}^m P(X_n - Y_n \geq \delta,\ X_n \in [t_k, t_{k+1}]) \\
&\quad + \sum_{k=1}^m P(X_n - Y_n \leq -\delta,\ X_n \in [t_k, t_{k+1}]) + \epsilon \\
&:= \sum_{k=1}^m A_k + \sum_{k=1}^m B_k + \epsilon.
\end{aligned}$$

Now for every k and as $n \to \infty$, we have (draw a diagram to illustrate)

$$\begin{aligned}
A_k &= P(X_n - Y_n \geq \delta,\ X_n \in [t_k, t_{k+1}]) \\
&= P(Y_n \leq X_n - \delta \leq t_{k+1} - \delta,\quad X_n \in [t_k, t_{k+1}]) \\
&\leq P(Y_n \leq t_k - \delta/2,\ X_n \geq t_k) \\
&= o(1), \\
B_k &= P(X_n - Y_n \leq -\delta,\ X_n \in [t_k, t_{k+1}]) \\
&= P(Y_n \geq X_n + \delta \geq t_k + \delta,\quad X_n \in [t_k, t_{k+1}]) \\
&\leq P(Y_n \geq t_{k+1} + \delta/2,\ X_n \leq t_{k+1}) \\
&= o(1).
\end{aligned}$$

Since ϵ is arbitrary, we have $\lim_n P(|X_n - Y_n| \geq \delta) = 0$. ∎

7.5 A more general definition of quantiles

Population quantiles

The earlier definition $\xi_p := F^{-1}(p) = \inf\{t \in R : F(t) \geq p\}$ is uniquely defined. It satisfies $F(\xi_p-) \leq p \leq F(\xi_p)$, which could be used as an alternative definition.

Definition **7.1** *A p-th quantlle is any value satisfying*

$$F(\xi_p-) \leq p \leq F(\xi_p). \quad ∎$$

For simplicity, we first consider medians (i.e. $p = 1/2$) and denote $m = \xi_{1/2}$, i.e.

$$F(m-) \leq 1/2 \leq F(m).$$

Theorem **7.7**

1. *Medians form a closed interval:* $m_0 \leq m \leq m_1$.
   (If $m_0 = m_1$*, the interval reduces to a single point.)*
2. *If* F *is continuous at* m*, then* $F(m) = 1/2$.
3. *We have*
   $$m = \arg\min_c E|X - c|,$$
   i.e. for any c*, we have*
   $$E_F|X - m| \leq E_F|X - c|.$$

Proof.

1.
   - Medians form an interval.
     Proof. Let a and b be two medians. For any $c \in (a, b)$, i.e. $a < c < b$, we have
     $$F(c-) \leq F(b-) \leq 1/2 \leq F(a) \leq F(c).$$
     Hence c is also a median.
   - Medians are closed, i.e. $\underline{m} = \inf\{m : m \text{ is a median}\}$ and $\overline{m} = \sup\{m : m \text{ is a median}\}$ are also medians.
     Proof. WLOG, we only show that $\underline{m}$ is a median. Clearly, there exist a decreasing subsequence of medians, denoted by $\{m_k, k \geq 1\}$ such that
     - $m_k \downarrow \underline{m}$.
     - From $F(m_k-) \leq 1/2 \leq F(m_k)$

     Hence,
     - $F(\underline{m}-) \leq F(m_k-) \leq 1/2$,
     - $F(\underline{m}) = \lim_k F(m_k) \geq 1/2$.

   Similarly, we can show that $\overline{m}$ is a median too. ∎
2. If F_X is continuous at m, then $1/2 \leq F_X(m) = F_X(m-) \leq 1/2$.

3. WLOG, assume that $c > m$, then

$$\begin{aligned}
E|X-c| &= \int_{(-\infty,c]} + \int_{(c,\infty)} |x-c| dF(x) \\
&= \int_{(-\infty,c]} (c-x)dF(x) + \int_{(c,\infty)} (x-c)dF(x) \\
&= \int_{(-\infty,m]} (c-x)dF(x) + \int_{(m,c]} (c-x)dF(x) \\
&\quad + \int_{(m,\infty)} (x-c)dF(x) - \int_{(m,c]} (x-c)dF(x) \\
&= \int_{(-\infty,m]} (c-m)dF(x) + \int_{(-\infty,m]} (m-x)dF(x) \\
&\quad + \int_{(m,c]} (c-x)dF(x) \\
&\quad + \int_{(m,\infty)} (x-m)dF(x) + \int_{(m,\infty)} (m-c)dF(x) \\
&\quad + \int_{(m,c]} (c-x)dF(x) \\
&= \int_{(-\infty,m]} (c-m)dF(x) + \int_{(-\infty,m]} (m-x)dF(x) \\
&\quad + \int_{(m,\infty)} (x-m)dF(x) + \int_{(m,\infty)} (m-c)dF(x) \\
&\quad +2\int_{(m,c]} (c-x)dF(x) \\
&= (c-m)F(m) + \int_{(-\infty,m]} (m-x)dF(x) \\
&\quad + \int_{(m,\infty)} (x-m)dF(x) + (m-c)[1-F(m)] \\
&\quad +2\int_{(m,c]} (c-x)dF(x) \\
&= \int_{(-\infty,m]} (m-x)dF(x) + \int_{(m,\infty)} (x-m)dF(x) \\
&\quad +(c-m)[2F(m)-1] + 2\int_{(m,c]} (c-x)dF(x) \\
&= \int_{(-\infty,m]} + \int_{(m,\infty)} |x-m| dF(x) \\
&\quad +(c-m)[2F(m)-1] + 2\int_{(m,c]} (c-x)dF(x) \\
&\geq E|X-m|,
\end{aligned}$$

where the last inequality follows from $F(m) \geq 1/2$ and the assumption $c > m$. ∎

Sample quantiles

Definition. Given a sample $X_1, ..., X_n$ from a continuous d.f. F, the sample median is any value $\hat{m}$ satisfying

$$F_n(\hat{m}-) \le 1/2 \le F_n(\hat{m})$$

where F_n is the empirical distribution function (e.d.f.):

$$F_n(x) = n^{-1} \sum I(X_i \le x).$$

THEOREM **7.8**

(1) We have

$$\begin{aligned} \hat{m} \;&=\; X_{\left(\frac{n+1}{2}\right)} && \textit{if } n \textit{ is odd} \\ & \quad \textit{any value in } \left[X_{\left(\frac{n}{2}\right)}, X_{\left(\frac{n}{2}+1\right)}\right] && \textit{if } n \textit{ is even} \end{aligned}$$

(2) We have

$$\hat{m} = \arg\min_c E_{F_n}|X - c| = \arg\min_c \sum_{i=1}^n |X_i - c|,$$

i.e. for any c, we have

$$\sum_{i=1}^n |X_i - \hat{m}| \le \sum_{i=1}^n |X_i - c|$$

(3) For n pairs (X_i, Y_i) $(1 \le i \le n)$, we have

$$\arg\min_b \sum_{i=1}^n |Y_i - bX_i| = Z_{(k)},$$

where k satisfies $\sum_{i=1}^k p_{(i)} \ge 1/2$ and $\sum_{i=k}^n p_{(i)} \ge 1/2$ with $p_i = \frac{|X_i|}{\sum |X_j|}$, and $Z_{(1)}, ..., Z_{(n)}$ are order statistics of $Y_1/X_1, ..., Y_n/X_n$, and $p_{(i)}$ is the value corresponding to $Z_{(i)}$.

Proof.

(1) It suffices to show that $\hat{m}$ satisfies $nF_n(\hat{m}-) \le n/2 \le nF_n(\hat{m})$, or equivalently

$$LHS =: \sum_{i=1}^n I\{X_i < \hat{m}\} \le n/2 \le \sum_{i=1}^n I\{X_i \le \hat{m}\} =: RHS.$$

- If n is odd, then for $\hat{m} = X_{\left(\frac{n+1}{2}\right)}$, we have

$$\begin{aligned} RHS &= \#\left\{X_i \le \hat{X}_{\left(\frac{n+1}{2}\right)}\right\} = \frac{n+1}{2} \ge \frac{n}{2}, \\ LHS &= \#\left\{X_i < \hat{X}_{\left(\frac{n+1}{2}\right)}\right\} = \frac{n+1}{2} - 1 = \frac{n-1}{2} \le \frac{n}{2}. \end{aligned}$$

- If n is even, then for $\hat{m} \in [X_{(\frac{n}{2})}, X_{(\frac{n}{2}+1)}]$, we have

$$\begin{aligned} RHS &= \#\{X_i \le \hat{m}\} = \left\{\frac{n}{2} \textit{ or } \frac{n}{2} + 1\right\} \ge n/2, \\ LHS &= \#\{X_i < \hat{m}\} = \left\{\frac{n}{2} - 1 \textit{ or } \frac{n}{2}\right\} \le n/2. \end{aligned}$$

(2) Let $X \sim F_n(x)$, then $E_{F_n}|X-a| = n^{-1}\sum |X_i - a|$.

(3) Note that

$$\begin{aligned}
\sum |Y_i - X_i b| &= \sum \left|\frac{Y_i}{X_i} - b\right| \ |X_i| = \left(\sum |X_i|\right)\sum \left|\frac{Y_i}{X_i} - b\right| \ \frac{|X_i|}{\sum |X_i|} \\
&= \left(\sum |X_i|\right)\sum |Z_i - b| p_i \\
&= \left(\sum |X_i|\right) E_G|Z-b|,
\end{aligned}$$

where $Z \sim G(x) = \sum p_i I(Y_i/X_i \le x)$, where $p_i = \frac{|X_i|}{\sum |X_j|}$. From (2), $\sum |Y_i - X_i b|$ is minimized when $b = Z_{(k)}$, the median of $F_{np}(x)$, where k satisfies $\sum_{i=1}^{k-1} p_{(i)} \le 1/2 \le \sum_{i=1}^{k} p_{(i)}$. Note here $p_{(i)}$ corresponds to $Z_{(i)}$. ∎

7.5.1 Alternative definition of quantiles

THEOREM **7.9** *A* **p-th quantile** *of X is any value ξ_p such that (Lehmann, p58.)*

$$F_X(\xi_p-) \le p \le F_X(\xi_p).$$

Then

1. *The set of quantiles forms a closed interval: $\xi_{p0} \le \xi_p \le \xi_{p1}$.*
   If $\xi_{p0} = \xi_{p1}$, then the interval reduces to a single point.
2. *If F is continuous at ξ_p, then $F(\xi_p) = p$.*
3. *$pE(X-c)^+ + (1-p)E(X-c)^-$ is minimized by any pth quantile ξ_p, i.e., for any c,*
   $$pE(X-c)^+ + (1-p)E(X-c)^- \ge pE(X-\xi_p)^+ + (1-p)E(X-\xi_p)^-,$$
   i.e., $\xi_p = \arg\min_c\{pE(X-c)^+ + (1-p)E(X-c)^-\}$.

   REMARK ***7.1*** *Take $p = 0.05$ for example, $\xi_{0.05}$ minimizes the loss function $0.05E(X-c)^+ + 0.95E(X-c)^-$. Here we notice that it puts more weight (0.95) on the negative loss $E(X-c)^-$. This has some bearings on the value-at-risk (VaR).*

Proof. We only prove the last below since the proofs for the first two are similar to those of medians. WLOG, assume that $c > \xi_p$, then

$$\begin{aligned}
& pE(X-c)^+ + (1-p)E(X-c)^- \\
= \;& p\int (x-c)I\{x-c>0\}dF(x) + (1-p)\int[-(x-c)]I\{x-c\le 0\}dF(x) \\
= \;& (1-p)\int_{(-\infty,c]}(c-x)dF(x) + p\int_{(c,\infty)}(x-c)dF(x) \\
= \;& (1-p)\int_{(-\infty,\xi_p]}(c-x)dF(x) + (1-p)\int_{(\xi_p,c]}(c-x)dF(x) \\
& +p\int_{(\xi_p,\infty)}(x-c)dF(x) - p\int_{(\xi_p,c]}(x-c)dF(x) \\
= \;& (1-p)\int_{(-\infty,\xi_p]}(c-\xi_p)dF(x) + (1-p)\int_{(-\infty,\xi_p]}(\xi_p-x)dF(x) \\
& +p\int_{(\xi_p,\infty)}(x-\xi_p)dF(x) + p\int_{(\xi_p,\infty)}(\xi_p-c)dF(x) \\
& +(1-p)\int_{(\xi_p,c]}(c-x)dF(x) - p\int_{(\xi_p,c]}(x-c)dF(x) \\
= \;& (1-p)\int_{(-\infty,\xi_p]}(\xi_p-x)dF(x) + p\int_{(\xi_p,\infty)}(x-\xi_p)dF(x) \\
& +(1-p)\int_{(-\infty,\xi_p]}(c-\xi_p)dF(x) + p\int_{(\xi_p,\infty)}(\xi_p-c)dF(x) \\
& +(1-p)\int_{(\xi_p,c]}(c-x)dF(x) + p\int_{(\xi_p,c]}(c-x)dF(x) \\
= \;& pE(X-\xi_p)^+ + (1-p)E(X-\xi_p)^- \\
& +(c-\xi_p)[(1-p)F(\xi_p) - p(1-F(\xi_p))] + \int_{(\xi_p,c]}(c-x)dF(x) \\
= \;& pE(X-\xi_p)^+ + (1-p)E(X-\xi_p)^- \\
& +(c-\xi_p)[F(\xi_p)-p] + \int_{(\xi_p,c]}(c-x)dF(x) \\
\ge \;& pE(X-\xi_p)^+ + (1-p)E(X-\xi_p)^-.
\end{aligned}$$

where the last inequality holds since $F(\xi_p) \geq p$ (as ξ_p is a p-th quantile), and also the assumption that $c > \xi_p$. The statement is proved. ∎

Remark **7.2** *As we have seen earlier, our quantiles may not be unique. In fact, we have shown that it can be any value from the closed interval $[a, b]$, where*

$$a := \inf\{t : F(t) \geq p\}, \qquad b := \sup\{t : F(t) \geq p\}$$

Definition. Let $X_1, ..., X_n$ be a set of distinct real numbers, its order statistics are $X_{(1)}, ..., X_{(n)}$. Then we define the sample p-th quantile to be

$$\begin{aligned} \hat{\xi}_p &= \text{any value in } \left[X_{([np])}, X_{([np]+1)}\right] \quad \text{if } np = [np] \text{ (i.e., } np \text{ is an integer)} \\ &= X_{([np]+1)} \quad \text{if } np \neq [np] \text{ (i.e., } np \text{ is NOT an integer).} \end{aligned}$$

(Compare this with the median where $p = 1/2$.)

Theorem **7.10** *Let $X_1, ..., X_n$ be a set of distinct real numbers.*

(1) $\hat{\xi}_p$ is also a p-th quantile of $F_n(x) = n^{-1}\sum_{i=1}^n I(X_i \leq x)$.

(2) Any p-th quantile, $\hat{\xi}_p$, minimizes

$$\begin{aligned} & pE_{F_n}(X - c)^+ + (1-p)E_{F_n}(X - c)^- \\ = \; & n^{-1}\left(\sum_{i \in \{i : X_i \geq c\}} p|X_i - c| + \sum_{i \in \{i : X_i < c\}} (1-p)|X_i - c|\right) \\ = \; & n^{-1}\left(\sum_{i=1}^n p(X_i - c)^+ + \sum_{i=1}^n (1-p)(X_i - c)^-\right) \\ = \; & n^{-1}\sum_{i=1}^n \left\{p(X_i - c)^+ + (1-p)(X_i - c)^-\right\}. \end{aligned}$$

(3) For n given points (X_i, Y_i), $i = 1, ..., n$,

$$b := \arg\min\left(\sum_{i \in \{i : Y_i - cX_i \geq 0\}} p|Y_i - cX_i| + \sum_{i \in \{i : Y_i - cX_i < 0\}} (1-p)|Y_i - cX_i|\right)$$

is called the p-th quantile regression estimator of $Y_i = cX_i + \epsilon_i$.

If further $X_i > 0$, there is a simple algorithm to calculate b, which is given by the p-th sample quantile of the weighted empirical d.f. $G(x) = \sum p_i I(Y_i/X_i \leq x)$, where $p_i = \frac{|X_i|}{\sum |X_j|}$. The details are similar to the medians.

Proof.

(1) It suffices to show that $nF_n(\hat{\xi}_p-) \leq np \leq nF_n(\hat{\xi}_p)$, i.e.,

$$LHS =: \sum_{i=1}^n I\{X_i < \hat{\xi}_p\} \leq np \leq \sum_{i=1}^n I\{X_i \leq \hat{\xi}_p\} =: RHS.$$

- If np is NOT an integer, then $\hat{\xi}_p = X_{([np]+1)}$, therefore,

$$\begin{aligned} RHS &= \#\left\{X_i \le \hat{X}_{([np]+1)}\right\} = [np]+1 \ge np, \\ LHS &= \#\left\{X_i < X_{([np]+1)}\right\} = [np] \le np. \end{aligned}$$

- If np is an integer, then $\hat{\xi}_p$ could be any value in the close interval $\left[X_{([np])}, X_{([np]+1)}\right] = \left[X_{(np)}, X_{(np+1)}\right]$, therefore,

$$\begin{aligned} RHS &= \#\left\{X_i \le \hat{\xi}_p\right\} = \{np \ or \ np+1\} \ge [np], \\ LHS &= \#\left\{X_i < \hat{\xi}_p\right\} = \{np-1 \ or \ np\} \le np = [np]. \end{aligned}$$

(2) Let $X \sim F_n(x)$. Since

$$\begin{aligned} & pE_{F_n}(X-c)^+ + (1-p)E_{F_n}(X-c)^- \\ = \ & n^{-1}\left(\sum_{i\in\{i:X_i\ge c\}} p|X_i-c| + \sum_{i\in\{i:X_i<c\}} (1-p)|X_i-c|\right), \end{aligned}$$

by (1), is minimized when $a =$ a p-th quantile of $F_n(x)$, or a p-th quantile of $X_1, ..., X_n$.

(3) If we assume that $X_i > 0$ for all i. Then, letting $Z_i = Y_i/X_i$ and $p_i = |X_i|/\sum|X_i|$, we have

$$\begin{aligned} & \sum_{i\in\{i:Y_i-cX_i\ge 0\}} p|Y_i-cX_i| + \sum_{i\in\{i:Y_i-cX_i<0\}} (1-p)|Y_i-cX_i| \\ = \ & \left(\sum|X_i|\right)\left(\sum_{i\in\{i:Y_i/X_i-c\ge 0\}} p|Y_i/X_i-c|p_i + \sum_{i\in\{i:Y_i/X_i-c<0\}} (1-p)|Y_i/X_i-c|p_i\right) \\ = \ & \left(\sum|X_i|\right)\left(\sum_{i\in\{i:Z_i-c\ge 0\}} p|Z_i-c|p_i + \sum_{i\in\{i:Z_i-c<0\}} (1-p)|Z_i-c|p_i\right) \end{aligned}$$

The rest is left as an exercise. ∎

7.6 Median or Quantile (Robust) Regression

Suppose that (x_i, Y_i) follow the model

$$Y_i = \alpha + \beta x_i^\tau + \epsilon_i, \qquad i = 1, ..., n,$$

where $\alpha \in R$, $\beta \in R^k$, and ϵ_i are errors satisfying $E\epsilon_i = 0$.

- The least square estimates (LSE) of (α, β) are

$$\min_{a,b} \left\{ \sum_{i=1}^n [y_i - (a + bx_i^\tau)]^2 \right\}.$$

- The least absolute deviation (LAD) estimates [or **median regression**] of (α, β) are

$$\min_{a,b} \left\{ \sum_{i=1}^n |y_i - (a + bx_i^\tau)| \right\}.$$

- The p**-th quantile regression estimates** of (α, β) are

$$\min_{a,b} \left\{ \sum_{i=1}^n p\left(y_i - (a + bx_i^\tau)\right)^+ + (1-p)\left(y_i - (a + bx_i^\tau)\right)^- \right\}.$$

Remark **7.3**

- *LAD is more robust than the LSE. If the d.f. of ϵ is heavy tailed, one can use LAD. There are many other robust estimates, balancing robustness and efficiency.*
- *LSE is easier to compute than LAD. Theorem 7.10 shows how to compute one unknown for LAD. For more than one unknowns, we could fix one value given all others, and repeat for others.*

Remark **7.4** *It can be shown (see the chapter on M-functionals) that, the mean, the median and the p-th quantile satisfy*

$$\begin{aligned} \mu &= \arg\min_c E(X-c)^2 \\ m &= \arg\min_c E|X-c| \\ \xi_p &= \arg\min_c \; E(p(X-c)^+ + (1-p)(X-c)^-), \end{aligned}$$

with respective influence functions given by

$$\begin{aligned} \psi(x;\mu) &= x \\ \psi(x;m) &= I\{x>0\} - I\{x<0\} \\ \psi(x;\xi_p) &= pI\{x>0\} - (1-p)I\{x<0\} \end{aligned}$$

The above three loss functions, $(X-c)^2$, $|X-c|$ and $(p(X-c)^+ + (1-p)(X-c)^-)$, are all convex functions. Furthermore, the first two are symmetric functions while the last one is asymmetric. If we plot $px^+ + (1-p)x^-$ against x with $p = 5\%$ for instance, clearly we penalize more if X move further away toward the left tail than toward the right. This makes sense since we are interested in the quantiles in the left tail ($p = 5\%$ e.g.).

7.7 Exercises

1. Suppose $X \sim F$, and $m = F^{-1}(1/2)$, $\mu = EX$, $\sigma^2 = Var(x)$. Show that

$$|m - \mu| \leq \sigma.$$

   (Hint: use the L_1-optimal property of the median.)

2. Let $F(x) = |x|$, which is convex and has derivatives everywhere except 0. Denote its left derivative by $D(x) =: f(x-) = I\{x > 0\} - I\{x \leq 0\}$. Show that the one-term Taylor expansion satisfies

$$\begin{aligned} |F(x+t) - F(x) - f(x-)t| &=: ||x+t| - |x| - D(x)t| \\ &\leq 2|t| I\{|x| \leq |t|\}. \end{aligned}$$

Chapter 8

M-Estimation

Loosely speaking, an M-Estimation refers to any procedure which involves **maximization or minimization** of certain quantities. For example, in practice, we might like to minimize a loss function $E_F\rho(X,t)$, say, over a range of values for t. Many estimation procedures in statistics are in the form of M-estimation. Other related procedures in the literature are general estimating equations (GEE) or zero-estimation (Z-estimation).

8.1 Definition

$M-$**functional** is usually defined by minimizing an expected loss function:

$$T(F) = \arg\min_{t_0} E_F\rho(X,t_0) = \arg\min_{t_0}\int \rho(x,t_0)dF(x),$$

which can be solved (under certain conditions) from

$$\frac{\partial}{\partial t}\int \rho(x,t)dF(x) = \int \psi(x,t)dF(x) = 0, \qquad \text{where } \psi(x,t) = \partial\rho(x,t)/\partial t$$

Definition: Let $X_1,\ldots,X_n \sim F$. F_n is the empirical d.f. Define

$$\lambda_F(t) = E_F\psi(X,t) = \int \psi(x,t)dF(x).$$

(1). An $M-$**functional** $T(F)$ w.r.t. ψ is a solution t_0 to

$$\lambda_F\left(t_0\right) = 0.$$

(2). An M-estimator of $T(F)$ is $\hat{t} = T(F_n)$, or the solution of

$$\lambda_{F_n}(\hat{t}) = E_{F_n}\psi(X,\hat{t}) = n^{-1}\sum_{i=1}^{n}\psi(X_i,\hat{t}) = 0.$$

(3). If $\psi(x,t) = \tilde{\psi}(x-t)$, $T(F)$ is called a location parameter.

Remark **8.1** *Although the M-functional and M-estimators are usually derived by* **Minimization** *or* **Maximization** *of certain quantities $E_F\rho(X,t)$ (hence the name M), we actually define those items without any mention of $E_F\rho(X,t)$.*

8.2 Examples of M-estimators

Example 1 (MLE). Let $\mathcal{F} = \{F_\theta(\cdot) : \theta \in \Theta\}$. Assume that $F'_\theta(x) = f_\theta(x)$ exists. Maximizing log-likelihood $\log f_\theta(x)$ is equivalent to minimizing $\rho(x, \theta) = -\log f_\theta(x)$ Choose

$$\psi(x, \theta) = -\frac{\partial}{\partial \theta} \log f_\theta(x).$$

The resulting M-estimator $T(F_n)$ is an MLE.

Example 2 (Location parameter estimation). Let M-functional $T(F)$ be a solution of

$$E_F \psi(X - t_0) = \int \psi(x - t_0) dF(x) = 0,$$

and the corresponding M-estimators $T(F_n)$ to the location parameter. Different choices of ψ lead to different estimators. Here are some examples.

1. **The least square estimate (LSE).** If $\rho(x) = x^2/2$, then $\psi(x) = x$. Then the solution to $E_F\psi(X - t_0) = E_F(X - t_0) = \mu_F - t_0$ is $t_0 = \mu_F$. Therefore,

   $$T(F) = \int x dF(x) = EX, \qquad \text{mean functional,}$$

   and

   $$T(F_n) = \int x dF_n(x) = \bar{X}, \qquad \text{sample mean.}$$

2. **The least absolute deviation (LAD) estimate.** If $\rho(x) = |x|$, then

   $$\psi(x) \;\; = \;\; I\{x > 0\} - I\{x < 0\} = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

   (In fact, $\rho(x)$ is not differentiable at 0, but we still define $\psi(0) = 0$).

   Then we claim that any solution to

   $$\begin{aligned} E_F\psi(X - t_0) &= E_F I\{X - t_0 > 0\} - E_F I\{X - t_0 < 0\} \\ &= P(X > t_0) - P(X < t_0) \\ &= P(X \geq t_0) - P(X \leq t_0) \\ &= 0 \end{aligned}$$

   must be a median m.

   > **Proof.** We have $P(X \geq t_0) = P(X \leq t_0)$, and hence, $P(X < t_0) = P(X > t_0)$. Therefore,
   >
   > $$\begin{aligned} 1 &= P(X > t_0) + P(X < t_0) + P(X = t_0) = 2P(X < t_0) + P(X = t_0) \\ &= P(X \geq t_0) + P(X \leq t_0) - P(X = t_0) = 2P(X \leq t_0) - P(X = t_0), \end{aligned}$$
   >
   > implying that $P(X < t_0) = \frac{1}{2}(1 - P(X = t_0)) \leq \frac{1}{2}$, and $P(X \leq t_0) = \frac{1}{2}(1 + P(X = t_0)) \geq \frac{1}{2}$. Therefore, t_0 must be a median. ∎

   Therefore, $T(F) = m$ (median functional) and $T(F_n) = \hat{m}$ (sample median).

3. ***p*th LAD estimator (LAD)**. More generally, we can take $\rho(x) = |x|^p/p$, where $p \in [1,2]$, then

$$\psi(x) = \begin{cases} -|x|^{p-1} & x < 0 \\ 0 & x = 0 \\ |x|^{p-1} & x > 0. \end{cases}$$

The M-estimator $T(F_n)$ is called the ***p*th least absolute deviation estimator** (LAD) or the minimum L_p distance estimator. (Again, $\rho(x)$ may not be differentiable at 0, but we still define $\psi(0) = 0$).

4. **A type of trimmed mean.** Huber (1964) considers

$$\rho(x) = \begin{cases} x^2/2 & |x| \le C \\ C^2/2 & |x| > C. \end{cases}$$

with

$$\begin{aligned} \psi(x) &= \begin{cases} x & |x| \le C \\ 0 & |x| > C \end{cases} \\ &= xI\{|x| \le C\}. \end{aligned}$$

Then the solution to

$$\begin{aligned} E_F\psi(X - t_0) &= E_F(X - t_0)I\{|X - t_0| \le C\} \\ &= E_F XI\{|X - t_0| \le C\} - t_0 P(|X - t_0| \le C) = 0 \end{aligned}$$

satisfies

$$t_0 = \frac{E_F XI\{|X - t_0| \le C\}}{P(|X - t_0| \le C)} = E_F\left(X|\{|X - t_0| \le C\}\right) := T(F).$$

Intuitively, $t_0 = T(F)$ is the mean for the restricted range $[t_0 - C, t_0 + C]$.

Now the solution $\hat{t}$ to

$$\begin{aligned} E_{F_n}\psi(X - \hat{t}) &= E_{F_n}(X - \hat{t})I\{|X - \hat{t}| \le C\} \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{t})I\{|X_i - \hat{t}| \le C\} \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} X_i I\{|X_i - \hat{t}| \le C\} - \hat{t}\sum_{i=1}^{n} I\{|X_i - \hat{t}| \le C\}\right) \\ &= 0 \end{aligned}$$

satisfies

$$\hat{t} = \frac{\sum_{i=1}^n X_i I\{|X_i - \hat{t}| \le C\}}{\sum_{i=1}^n I\{|X_i - \hat{t}| \le C\}} = E_{F_n}\left(X|\{|X - \hat{t}| \le C\}\right) := T(F_n).$$

Intuitively, $\hat{t} = T(F_n)$ is the mean for the restricted range $[\hat{t} - C, \hat{t} + C]$. Notice that both sides in the above equation involves $\hat{t}$, but we can solve $\hat{t}$ iteratively with the initial choice of $\hat{t}_0$ being the sample mean or any other reasonable guess.

Clearly, $T(F_n)$ is a type of **trimmed mean.** It completely eliminates the influence of the extreme values (outliers).

5. **Huber's estimate (or winsorized mean).** Huber (1964) considers

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & |x| \le C \\ C|x| - \frac{1}{2}C^2 & |x| > C. \end{cases}$$

with

$$\begin{aligned} \psi(x) &= \begin{cases} C & x > C \\ x & |x| \le C \\ -C & x < -C. \end{cases} \\ &= xI\{|x| \le C\} + CI\{|x| > C\}. \end{aligned}$$

Then the solution to

$$\begin{aligned} E_F\psi(X - t_0) &= E_F(X - t_0)I\{|X - t_0| \le C\} + CP(|X - t_0| > C) \\ &= E_F XI\{|X - t_0| \le C\} - t_0 P(|X - t_0| \le C) + CP(|X - t_0| > C) = 0 \end{aligned}$$

satisfies

$$t_0 = \frac{E_F XI\{|X - t_0| \le C\} + CP(|X - t_0| > C)}{P(|X - t_0| \le C)} := T(F).$$

Intuitively, $t_0 = T(F)$ is the mean for the restricted range $[t_0 - C, t_0 + C]$.

Now the solution $\hat{t}$ to

$$\begin{aligned} E_{F_n}\psi(X - \hat{t}) &= E_{F_n}\left((X - \hat{t})I\{|X - \hat{t}| \le C\} + CI\{|x| > C\}\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^n (X_i - \hat{t})I\{|X_i - \hat{t}| \le C\} + C\sum_{i=1}^n I\{|X_i - \hat{t}| > C\}\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^n X_i I\{|X_i - \hat{t}| \le C\} - \hat{t}\sum_{i=1}^n I\{|X_i - \hat{t}| \le C\} + C\sum_{i=1}^n I\{|X_i - \hat{t}| > C\}\right) \\ &= 0 \end{aligned}$$

satisfies

$$\hat{t} = \frac{\sum_{i=1}^n X_i I\{|X_i - \hat{t}| \le C\} + C\sum_{i=1}^n I\{|X_i - \hat{t}| > C\}}{\sum_{i=1}^n I\{|X_i - \hat{t}| \le C\}} := T(F_n).$$

Here, $T(F_n)$ is a type of **winsorized mean**. It limits, but does not entirely eliminate, the influence of outliers.

6. **Hampel's estimate.** Hampel (1964) considers $\psi(-x) = -\psi(-x)$ (i.e. an odd function), and

$$\psi(x) = \begin{cases} x & 0 \le x \le a, \\ a & a < x \le b, \\ a\left(\frac{c - x}{c - b}\right) & b < x \le c, \\ 0 & x > c. \end{cases}$$

Then $T(F_n)$ is called **Hampel's estimate.** It is a combination of the trimmed mean and the Winsorized mean. It down-weights the influence of extreme values and completely removes the influence of the most extreme values.

7. **Smoothed Hampel's estimate.** Hampel (1964) considers $\psi(-x) = -\psi(-x)$ (i.e. an odd function), and

$$\psi(x) = \begin{cases} \sin(ax) & 0 \le x \le \pi/a \\ 0 & x > \pi/a. \end{cases}$$

Then $T(F_n)$ is called **smoothed Hampel's estimate.**

8.3 Consistency and asymptotic normality of M-estimates

We consider the following cases separately.

(a) $\psi(x,t)$ is monotone in t. (e.g. sample mean, median, Huber's estimates.)

(b) $\psi(x,t)$ is continuous in t and bounded. (e.g. Hampel's estimates.)

Recall $\lambda_F(t) = E_F\psi(X,t) = \int \psi(x,t)dF(x)$ and $\lambda_{F_n}(t) = E_{F_n}\psi(X,t) = \frac{1}{n}\sum \psi(X_i,t)$.

Case I: $\psi(x,t)$ is monotone in t

Theorem **8.1 (Consistency)** *Assume that*

(a) $\psi(x,t)$ is monotone (say, non-increasing) in t.

(b) Either $t_0 = T(F)$ is an **isolated** *root of $\lambda_F(t) = 0$, (or $\lambda_F(t)$ changes sign uniquely in a neighborhood of t_0).*

(c) $\{T_n\}$ is a root of $\lambda_{F_n}(t) = 0$, (or $\lambda_{F_n}(t)$ changes sign in a neighborhood of T_n).

Then t_0 is unique and $T_n \to t_0$ a.s.

Proof. Since $\psi(x,t)$ is non-increasing in t, so are $\lambda_F(t)$ and $\lambda_{F_n}(t)$.

If t_0 is an isolated root to $\lambda_F(t_0) = 0$, t_0 is the unique root, then for any $\epsilon > 0$,

$$\lambda_F(t_0+\epsilon) < \lambda_F(t_0) = 0 < \lambda_F(t_0-\epsilon).$$

If $\lambda_F(t)$ changes sign uniquely in a neighborhood of t_0, then for any $\epsilon > 0$,

$$\lambda_F(t_0+\epsilon) < 0 < \lambda_F(t_0-\epsilon).$$

By SLLN, $\lambda_{F_n}(t) \to \lambda_F(t)$ a.s. for each t. Then as $n \to \infty$,

$$\lambda_{F_n}(t_0+\epsilon) < 0 < \lambda_{F_n}(t_0-\epsilon), \qquad a.s. \tag{3.1}$$

Recall that $\lambda_{F_n}(t)$ is non-increasing in t. Then from assumption (c), we must have

$$t_0 - \epsilon \le T_n \le t_0 + \epsilon, \qquad a.s.$$

(Note that T_n does not depend on ϵ and may not be unique, but they all satisfy the above inequality.) Letting $\epsilon \to 0$, we see that $T_n \to t_0$ a.s. ∎

Theorem **8.2 (Asymptotic Normality)** *Let t_0 be an isolated root of $\lambda_F(t) = E_F\psi(X,t) = 0$. Suppose that*

(a) $\psi(x,t)$ is monotone (say, non-increasing) in t.

(b) $\lambda_F(t)$ is differentiable at $t = t_0$ with $\lambda'_F(t_0) \neq 0$.

(c) $E\psi^2(X,t) < \infty$ for t in a neighborhood of t_0 and is continuous at $t = t_0$.

Suppose that $\{T_n\}$ is a root of $\lambda_{F_n}(t) = 0$, or $\lambda_{F_n}(t)$ changes sign in a neighborhood of T_n. Then,

(i) $T_n \longrightarrow_{a.s.} t_0$;

(ii) $\sqrt{n}\,(T_n - t_0) \to_d N(0, \sigma^2)$, where $\sigma^2 = \dfrac{E_F\psi^2(X,t_0)}{[\lambda'_F(t_0)]^2}$.

Proof. (i) This is guaranteed by Theorem 8.1.

(ii) Since $\psi(x,t)$ non-increasing in t, so is $\lambda_{F_n}(t)$. Then we have

$$\{\lambda_{F_n}(t) < 0\} \subset \{T_n \le t\} \subset \{\lambda_{F_n}(\hat{t}) \le 0\}. \tag{3.2}$$

Thus, $P(\lambda_{F_n}(t) < 0) \le P(T_n \le t) \le P(\lambda_{F_n}(t) \le 0)$. Note that

$$P\left(\frac{\sqrt{n}(T_n - t_0)}{\sigma} \le x\right) = P\left(T_n \le t_0 + \frac{\sigma x}{\sqrt{n}}\right) = P\left(T_n \le t_{nx}\right), \quad \text{where } t_{nx} = t_0 + \frac{\sigma x}{\sqrt{n}}.$$

Hence it suffices to show that

$$\lim_n P(\lambda_{F_n}(t_{nx}) < 0) = \lim_n P(\lambda_{F_n}(t_{nx}) \le 0) = \Phi(x).$$

Now

$$\begin{aligned} P\left(\lambda_{F_n}(t_{nx}) \le 0\right) &= P\left(\sum \psi(X_i, t_{nx}) \le 0\right) \\ &= P\left(\frac{\sum[\psi(X_i,t_{nx}) - E\psi(X_i,t_{nx})]}{\sqrt{n}\sigma_{nx}} \le \frac{-\sqrt{n}E\psi(X_i,t_{nx})}{\sigma_{nx}}\right), \\ &=: P\left(n^{-1/2}\sum Y_{ni} \le \frac{-\sqrt{n}\lambda_F(t_{nx})}{\sigma_{nx}}\right) \end{aligned}$$

where $\sigma_{nx}^2 = Var(\psi(X_i, t_{nx}))$ and $Y_{ni} = \dfrac{\psi(X_i,t_{nx}) - E\psi(X_i,t_{nx})}{\sigma_{nx}}$. However, as $n \to \infty$, we have

$$\begin{aligned} \sqrt{n}\lambda_F(t_{nx}) &= \sqrt{n}[\lambda_F(t_{nx}) - \lambda_F(t_0)] \\ &= \sqrt{n}[t_{nx} - t_0]\left(\frac{\lambda_F(t_{nx}) - \lambda_F(t_0)}{t_{nx} - t_0}\right) \\ &= \sigma x[\lambda'_F(t_0) + o(1)] \quad \text{as } n \to \infty \\ &\to \lambda'_F(t_0)\sigma x \end{aligned}$$

and, since $E_F\psi^2(X,t)$ are $E_F\psi(X,t) = \lambda_F(t)$ are both continuous at t_0, we have

$$\begin{aligned} \sigma_{nx}^2 &= E_F\psi^2(X,t_{nx}) - [E_F\psi(X,t_{nx})]^2 \\ &= E_F\psi^2(X,t_0) + o(1) - [E_F\psi(X,t_0) + o(1)]^2 \\ &= E_F\psi^2(X,t_0) + o(1) \\ &= \sigma^2[\lambda'_F(t_0)]^2 + o(1) \\ &\to \sigma^2[\lambda'_F(t_0)]^2. \end{aligned}$$

Therefore, $\sigma_{nx} \to -\lambda_F'(t_0)\sigma$ as $\lambda_F'(t_0) \le 0$. Hence, as $n \to \infty$,

$$\frac{-\sqrt{n}\lambda_F(t_{nx})}{\sigma_{nx}} \longrightarrow x.$$

Since Y_{ni}, $1 \le i \le n$, are i.i.d. with mean 0 and variance 1, we may apply the double array CLT (Theorem 8.6) with $B_n^2 = nVar(Y_{n1}) = n$. Applying Slutsky's theorem and **the CLT for double arrays** (see Theorem 8.6 in Section 8.6 below), we get

$$\lim_n P(\lambda_{F_n}(t_{nx}) \le 0) = \Phi(x).$$

By the following remark, we have $\lim_n P(\lambda_{F_n}(t_{nx}) < 0) = \Phi(x)$. The proof is complete. ∎

Remark **8.2** *There are some equivalent definitions of weak convergence, referred to as Portmanteau Theorem:*

(a) $X_n \Longrightarrow X$ *or* $X_n \to_d X$*.*

(b) $\liminf_{n\to\infty} P(X_n \in G) \ge P(X \in G)$ *for all open sets* G*.*

(c) $\limsup_{n\to\infty} P(X_n \in K) \le P(X \in K)$ *for all closed sets* K*.*

(d) $\lim_{n\to\infty} P(X_n \in A) = P(X \in A)$ *for all sets* A *with* $P(X \in \partial A) = 0$*, where* ∂A *is the boundary of* A*.*

(e) $Eg(X_n) \to Eg(X)$ *for all bounded continuous function* g*.*

(f) $Eg(X_n) \to Eg(X)$ *for all functions* g *of the form* $g(x) = h(x)I_{[a,b]}(x)$*, where* h *is continuous on* $[a,b]$ *and* $a, b \in C(F)$*.*

(g) $\lim_n \psi_n(t) = \psi(t)$*, where* $\psi_n(t)$ *and* $\psi(t)$ *are the c.f.s of* X_n *and* X*, respectively.*

Remark **8.3** *In the quantile example below, the estimating equation* $\lambda_{F_n}(t) = 0$ *reduces to* $F_n(t) = p$*, which may have no solution or an infinite number of solutions, since* $F_n(t)$ *has jump sizes* n^{-1} *at each observation. If* $\lambda_{F_n}(t) = 0$ *has no solution, then* T_n *is defined to be a value which changes sign of* $\lambda_{F_n}(t)$*.*

Example 1. (Sample pth Quantile). F is continuous, and $0 < p < 1$. Let ξ_p be the unique root of $F(t) = p$, and let $\hat{\xi}_p$ be a root of $F_n(t) = p$ or $F_n(t) = p$ changes signs at $\hat{\xi}_p$. Further assume that $F'(\xi_p) = f(\xi_p)$. Then

$$\sqrt{n}\left(\hat{\xi}_p - \xi_p\right) \longrightarrow_d N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right)$$

Proof. Take $\psi(x,t) = \psi(x-t)$ (somewhat abuse of notation), where

$$\begin{aligned}\psi(x) &= \begin{cases} -1 & x < 0, \\ 0 & x = 0, \\ \frac{p}{1-p} & x > 0 \end{cases} \\ &= \frac{p}{1-p} I\{x > 0\} - I\{x < 0\} \\ &= \frac{1}{1-p}\left(pI\{x > 0\} - (1-p)I\{x < 0\}\right).\end{aligned}$$

Then

$$\begin{aligned}\lambda_F(t) &= E_F\psi(X-t) = (-1)P(X-t<0) + \frac{p}{1-p}P(X-t>0) \\ &= \frac{p}{1-p}[1-F(t)] - P(X<t) = \frac{p}{1-p}[1-F(t)] - F(t) \\ &= \frac{p-F(t)}{1-p}.\end{aligned}$$

Clearly, $\lambda_F(\xi_p) = 0$. Also,

$$\begin{aligned}E\psi^2(X,\xi_p) &= (-1)^2P(X-\xi_p<0) + \frac{p^2}{(1-p)^2}P(X-\xi_p>0) \\ &= p + \frac{p^2}{(1-p)^2}(1-p) = \frac{p}{1-p}.\end{aligned}$$

and

$$\lambda'_F(\xi_p) = \frac{-F'(\xi_p)}{1-p} = \frac{-f(\xi_p)}{1-p}.$$

It is easy to check that all conditions in Theorem 8.2 are satisfied with $t_0 = \xi_p$. Therefore, $\sqrt{n}\left(\hat{\xi}_p - \xi_p\right) \to_d N\left(0,\sigma^2\right)$, where

$$\sigma^2 = \frac{E\psi^2(X,\xi_p)}{[\lambda'_F(\xi_p)]^2} = \frac{p/(1-p)}{f^2(\xi_p)/(1-p)^2} = \frac{p(1-p)}{f^2(\xi_p)}. \quad \blacksquare$$

REMARK **8.4** *In this example, we integrate $\psi(x)$ to get*

$$\begin{aligned}\rho(x) &= \int_0^x \psi(x)dx \\ &= \begin{cases} -x & x \le 0, \\ \frac{p}{1-p}x & x>0 \end{cases} \\ &= \frac{1}{1-p}\left(pxI\{x>0\} - (1-p)xI\{x<0\}\right) \\ &= \frac{1}{1-p}\left(px_+ + (1-p)x_-\right).\end{aligned}$$

Ignoring the constant term $\frac{1}{1-p}$, we notice that the p-th quantile estimate minimizes the loss function weighted L_1 loss, $EL(c)$, where

$$L(c) = p(X-c)_+ + (1-p)(X-c)_-.$$

Take $p = 5\%$ for instance, if we plot $L(c)$ against c, we see that we penalize more if X move further away toward the left tail than it toward the right tail. This makes sense since we are interested in the quantiles in the left tail ($p = 5\%$ e.g.).

Case II: $\psi(x,t)$ is differentiable in t (e.g. MLE)

The next theorem trades monotonicity of $\psi(x,\cdot)$ for smoothness restrictions. The method has been most used for MLE.

Lemma **8.1** *Let $g(x,t)$ be continuous at t_0 uniformly in x. Let F be a d.f. such that $E_F|g(X,t_0)| < \infty$. Let $\{X_i\}$ be i.i.d. with d.f. F and suppose that $T_n \longrightarrow_p t_0$. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i, T_n) \longrightarrow_p \; E_F g(X, t_0).$$

The convergence in probability can be replaced by a.s. convergence throughout.

Proof. Write

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i,T_n) - E_F g(X,t_0)\right| &\le \left|\frac{1}{n}\sum_{i=1}^{n}[g(X_i,T_n) - g(X_i,t_0)]\right| \\ &\quad + \left|\frac{1}{n}\sum_{i=1}^{n} g(X_i,t_0) - E_F g(X,t_0)\right|\end{aligned}$$

The second term on the RHS $\longrightarrow_p 0$, while the first term is bounded by

$$\frac{1}{n}\sum_{i=1}^{n} |g(X_i,T_n) - g(X_i,t_0)| \longrightarrow_p 0,$$

since $|g(x,T_n) - g(x,t_0)| \to_p 0$ for all x uniformly. ∎

Case III: $\psi(x,t)$ is continuous in t and bounded

Theorem **8.3 (Consistency)** *Assume that*

(*a*) *$\psi(x,t)$ is continuous in t and bounded.*

(*b*) *$t_0 = T(F)$ is an* **isolated** *root of $\lambda_F(t) = 0$,* **and** *$\lambda_F(t)$ changes sign uniquely in a neighborhood of t_0. (so it does not behaves like $y = x^2$.)*

Then $\lambda_{F_n}(t) = 0$ admits a strongly consistent estimation sequence T_n for t_0.

Proof. Since $\lambda_F(t)$ is continuous in t, assumption (b) implies that $\exists\epsilon > 0$ such that

$$\lambda_F(t_0+\epsilon) < \lambda_F(t_0) = 0 < \lambda_F(t_0-\epsilon),$$

or

$$\lambda_F(t_0+\epsilon) > \lambda_F(t_0) = 0 > \lambda_F(t_0-\epsilon).$$

For simplicity, we assume that first inequality holds. By the Strong Law of Large Numbers (SLLN), $\lambda_{F_n}(t) \to \lambda_F(t)$ a.s. for each t. Then as $n \to \infty$,

$$\lambda_{F_n}(t_0+\epsilon) < 0 < \lambda_{F_n}(t_0-\epsilon), \qquad a.s. \tag{3.3}$$

Since $\lambda_{F_n}(t)$ is also continuous in t, then there **EXISTS** a solution sequence, $\{T_{n\epsilon}\}$, of $\lambda_{F_n}(t) = 0$ in the interval $[t_0-\epsilon, t_0+\epsilon]$. That is,

$$t_0 - \epsilon \le T_{n\epsilon} \le t_0 + \epsilon, \qquad a.s.$$

In particular, choosing $\epsilon = 1/n$ and $T_n = T_{n,1/n}$, we have

$$t_0 - 1/n \le T_n = T_{n,1/n} \le t_0 + 1/n, \qquad a.s.$$

Letting $n \to \infty$, we get $T_n \to t_0$ a.s. ∎

Remark **8.5** *Note that $\lambda_{F_n}(t) = 0$ in Theorem 8.3 may have multiple solutions. In that case, one needs to identify a consistent solution sequence in practice. Therefore, for a given solution sequence, e.g., one obtained by a specified algorithm, one needs to CHECK if it is consistent or not.*

Theorem **8.4 (Asymptotic Normality)** *Let $t_0 = T(F)$ be an M-functional which is a root of $\lambda_F(t) = E_F\psi(X,t) = 0$. Suppose that*

(a) $\psi(x,t)$ be bounded and continuous in t.

(b) $\lambda_F(t)$ is continuously differentiable at t_0 with $\lambda_F'(t_0) \neq 0$.

Let $\{T_n\}$ be a strongly consistent solution sequence of $\lambda_{F_n}(t) = 0$ (The existence is established in Theorem 8.3). Then,

$$\sqrt{n}\,(T_n - t_0) \to_d N(0,\sigma^2), \quad \textit{where } \sigma^2 = Var\,(\phi_F(X_1)) = \frac{E\psi^2(X,t_0)}{[\lambda_F'(t_0)]^2}.$$

Proof. We provide two methods.

- **Method 1. This is a special case of the next theorem.**
- **Method 2.** It suffices to show that T is d_∞-Hadamard differentiable at F with

$$\phi_F(x) = \frac{\psi(x,t_0)}{\lambda_F'(t_0)} = \frac{\psi(x,T(F))}{\lambda_F'(T(F))}. \tag{3.4}$$

The derivation of (3.4) will be given in the next theorem. Let us prove d_∞-Hadamard differentiability next.

We now show Hadamard differentiability. For $G_j = F + t_j\Delta_j$, we have

$$T_F'(G_j - F) = -\frac{\lambda_{G_j}(T(F))}{\lambda_F'(T(F))} = -\frac{\lambda_{G_j - F}(T(F))}{\lambda_F'(T(F))} = -\frac{t_j\lambda_{\Delta_j}(T(F))}{\lambda_F'(T(F))}$$

Then,

$$\begin{aligned}
0 &= \lambda_{G_j}[T(G_j)] - \lambda_F[T(F)] \\
&= \lambda_{G_j}[T(G_j)] - \lambda_{G_j}[T(F)] + \lambda_{G_j}[T(F)] - \lambda_F[T(F)] \\
&= [T(G_j) - T(F)]\frac{\lambda_{G_j}[T(G_j)] - \lambda_{G_j}[T(F)]}{T(G_j) - T(F)} + t_j\lambda_{\Delta_j}[T(F)] \\
&= [T(G_j) - T(F)]\left(\frac{\lambda_{G_j}[T(G_j)] - \lambda_{G_j}[T(F)]}{T(G_j) - T(F)}\right) - T_F'(G_j - F)\lambda_F'(T(F))
\end{aligned}$$

So

$$\begin{aligned}
T(G_j) - T(F) &= T_F'(G_j - F)\frac{\lambda_F'[T(F)]}{\frac{\lambda_{G_j}[T(G_j)] - \lambda_{G_j}[T(F)]}{T(G_j) - T(F)}} \\
&= T_F'(G_j - F) + T_F'(G_j - F)\left(\frac{\lambda_F'[T(F)]}{\frac{\lambda_{G_j}[T(G_j)] - \lambda_{G_j}[T(F)]}{T(G_j) - T(F)}} - 1\right)
\end{aligned}$$

$$
\begin{aligned}
&= T'_F(G_j - F) + t_j \lambda_{\Delta_j}[T(F)] \left(\frac{\lambda'_F[T(F)]}{\frac{\lambda_{G_j}[T(G_j)] - \lambda_{G_j}[T(F)]}{T(G_j) - T(F)}} - 1 \right) \\
&:= T'_F(G_j - F) + t_j \lambda_{\Delta_j}[T(F)]\Xi
\end{aligned}
$$

It suffices to show that $|\lambda_{\Delta_j}[T(F)]| < C$ and $\Xi \to 0$. The first relation is easy to prove. The second can also be proved as well.

Case IV: $\psi(x,t)$ is continuous in t (but not necessarily bounded)

Denote $\|\cdot\|_V$ be the **variation norm**, $\|h\|_V = \lim_{a\to-\infty, b\to\infty} V_{a,b}(h)$, where

$$V_{a,b}(h) = \sup \sum_{i=1}^{k} |h(x_i) - h(x_{i-1})|,$$

the supremum being taken over all partitions $a = x_0 < ... < x_k = b$ of $[a,b]$.

LEMMA **8.2** *Let the function h be continuous with $\|h\|_V < \infty$ and the function K be right-continuous with $\|K\|_\infty < \infty$ and $K(\infty) = K(-\infty) = 0$. Then,*

$$\left| \int h dK \right| \leq \|h\|_V \|K\|_\infty.$$

Proof. Applying integration by parts, we get

$$\int h dK = \lim_{x\to\infty} h(x)K(x) - \lim_{x\to-\infty} h(x)K(x) - \int K dh = -\int K dh$$

using the fact that $h(\pm\infty) < \infty$ (Why? Please check) and the assumption $K(\infty) = K(-\infty) = 0$. Hence

$$\left| \int h dK \right| = \left| \int K dh \right| \leq \int |K(x)| d|h(x)| \leq \sup_x |K(x)| \int d|h(x)| \leq \|K(x)\|_\infty \|h(x)\|_V.$$

For another proof, see Natanson (1961), page 232. ∎

THEOREM **8.5 (Asymptotic Normality)** *Let $t_0 = T(F)$ be an M-functional which is an isolated root of $\lambda_F(t) = E_F \psi(X,t) = 0$. Suppose that*

- *(a) $\psi(x,t)$ is continuous in x and satisfy* $\lim_{t\to t_0} \|\psi(\cdot,t) - \psi(\cdot,t)\|_V = 0$.
- *(b) $\lambda_F(t)$ is differentiable at t_0 with $\lambda'_F(t_0) \neq 0$.*
- *(c) $E\psi^2(X,t_0) < \infty$.*

Let $\{T_n\}$ be a solution sequence of $\lambda_{F_n}(t) = 0$ satisfying $T_n \to t_0$. Then,

$$\sqrt{n}\,(T_n - t_0) \to_d N(0,\sigma^2), \quad \textit{where} \;\; \sigma^2 = Var\,(\phi_F(X_1)) = \frac{E\psi^2(X,t_0)}{[\lambda'_F(t_0)]^2}.$$

Proof. We shall use the von Mises differential approach here.

- Let us first find out $\phi_F(x)$. Let $F_t = F + t(\delta_x - F)$ and $\lambda_F(t) = E_F\psi(X,t) = \int \psi(x,t)dF(x)$. Then,

$$\begin{aligned}
\lambda_{F_t}(T(F_t)) &= E_{F_t}\psi(X, T(F_t)) \\
&= E_F\psi(X, T(F_t)) + t[E_G\psi(X, T(F_t)) - E_F\psi(X, T(F_t))] \\
&= \lambda_F(T(F_t)) + t[\lambda_G(T(F_t)) - \lambda_F(T(F_t))] \\
&= 0.
\end{aligned}$$

Differentiating w.r.t. t, we get

$$\lambda'_F(T(F_t))\frac{dT(F_t)}{dt} + [\lambda_G(T(F_t)) - \lambda_F(T(F_t))] + t\,\frac{d}{dt}[\lambda_G(T(F_t)) - \lambda_F(T(F_t))] = 0.$$

Setting $t = 0$, we get

$$\lambda'_F(t_0)\,\frac{dT(F_t)}{dt}\bigg|_{t=0} + [\lambda_G(t_0) - \lambda_F(t_0)] = 0.$$

Note that $\lambda_F(t_0) = 0$. Therefore,

$$T'_F(G - F) = \frac{dT(F_t)}{dt}\bigg|_{t=0} = -\frac{\lambda_G(t_0)}{\lambda'_F(t_0)} = -\frac{E_G\psi(X, t_0)}{\lambda'_F(t_0)}.$$

From this, we get

$$\phi_F(x) = T'_F(\delta_x - F) = -\frac{E_{\delta_x}\psi(X, t_0)}{\lambda'_F(t_0)} = -\frac{\psi(x, t_0)}{\lambda'_F(t_0)}.$$

- In order to deal with $R_n = T(G) - T(F) - T'_F(G - F)$, we need to have some explicit expression for $T(G) - T(F)$. Note that

$$\begin{aligned}
T(G) - T(F) &= \begin{cases} \dfrac{\lambda_F(T(G)) - \lambda_F(t_0)}{\dfrac{\lambda_F(T(G)) - \lambda_F(t_0)}{T(G) - t_0}} & \text{if } T(G) \neq t_0 \\ 0 & \text{if } T(G) = t_0 \end{cases} \\
&= \frac{\lambda_F(T(G)) - \lambda_F(t_0)}{h[T(G)]},
\end{aligned}$$

where

$$h[T(G)] = \begin{cases} \dfrac{\lambda_F(T(G)) - \lambda_F(t_0)}{T(G) - t_0} & \text{if } T(G) \neq t_0 \\ \lambda'_F(t_0) & \text{if } T(G) = t_0 \end{cases}$$

It is easy to see that $h(t) \to h(t_0) = \lambda'_F(t_0)$. Therefore,

$$R_n(G, F) = \frac{\lambda_F(T(G)) - \lambda_F(t_0)}{h[T(G)]} + \frac{\lambda_G(t_0)}{\lambda'_F(t_0)}.$$

As it turns out, this expression is not especially manageable. (Try this yourself.) However, if the denomenatior in the second term of R_n is replaced by $h[T(G)]$, things becomes a lot easier. In other words, we can study

$$\begin{aligned}
\widetilde{R}_n(G, F) &= T(G) - T(F) - \frac{\lambda'_F(t_0)}{h[T(G)]}T'_F(G - F) \qquad (3.5) \\
&= \frac{\lambda_F(T(G)) - \lambda_F(t_0)}{h[T(G)]} + \frac{\lambda_G(t_0)}{h[T(G)]} \\
&= \frac{\lambda_F(T(G)) + \lambda_G(t_0)}{h[T(G)]} \\
&= \frac{-\int[\psi(x, T(G)) - \psi(x, t_0)]d[G(x) - F(x)]}{h[T(G)]}
\end{aligned}$$

We now specialize to $G = F_n$ and $T(G) = T(F_n) = T_n$. The assumption $T_n \longrightarrow_p t_0$ yields that $h[T(F_n)] \longrightarrow_p h[t_0] = \lambda'_F(t_0)$. Check that Lemma 8.2 is applicable. We thus have

$$\begin{aligned}
|\widetilde{R}_n(F_n, F)| &= \left|\frac{-\int[\psi(x,T_n) - \psi(x,t_0)]d[F_n(x) - F(x)]}{h[T_n]}\right| \\
&\leq \frac{\|\psi(\cdot,T_n) - \psi(\cdot,t_0)\|_V \|F_n - F\|_\infty}{|h[T_n]|} \\
&= o_p(1)\,\|F_n - F\|_\infty = o_p\left(n^{-1/2}\right)[n^{1/2}\|F_n - F\|_\infty] \\
&= o_p\left(n^{-1/2}\right) O_p(1) \\
&\quad \text{(recall that } E\left(n^{1/2}\|F_n - F\|_\infty\right)^s < \infty \text{ for any } s > 0 \text{ from before)} \\
&= o_p\left(n^{-1/2}\right).
\end{aligned}$$

From this, (3.5) and $h[T_n] \longrightarrow_p \lambda_F(t_0)$, we get

$$\sqrt{n}[T_n - t_0] \longrightarrow_d N\left(E\phi_F(X), n^{-1}\sigma^2\right) = N\left(0, \frac{E\psi^2(X,t_0)}{n[\lambda'_F(t_0)]^2}\right). \quad \blacksquare$$

Remark **8.6** *Consider M-estimation of a location parameter with $\psi(x,t) = \psi(x-t)$ (a slight abuse of notation). Then, the condition (a) of Theorem 8.5 is satisfied by*

(*a*) *least pth power estimates, i.e., $\psi(x) = |x|^{p-1}sgn(x)$ $(1 < p \leq 2)$.*

(*b*) *Huber's estimates*

(*c*) *Hampel's (smoothed or non-smoothed) estimates.*

Appendix: CLT for double arrays of r.v.'s

Consider a double array of r.v.'s:

$$X_{11}, X_{12}, ..., X_{1k_1};$$

$$X_{21}, X_{22}, ..., X_{2k_2};$$

$$........................$$

$$X_{n1}, X_{n2}, ..., X_{nk_n};$$

For each $n \geq 1$, there are k_n r.v.'s in that row. Assume that $k_n \to \infty$. The case $k_n = n$ is called a "triangular array".

Write $F_{nj}(x) = P(X_{nj} \leq x)$, and

$$\mu_{nj} = EX_{nj}, \qquad A_n = \sum_{j=1}^{k_n} \mu_{nj}, \qquad B_n^2 = Var\left(\sum_{j=1}^{k_n} X_{nj}\right).$$

The Lindeberg-Feller-Levy Theorem for a double array of r.v.'s become

Theorem **8.6** *Let $\{X_{nj} : 1 \leq j \leq k_n; n = 1, 2, ...\}$ be a double array with independent r.v.'s within rows. Then* *(i) the uniform asymptotic negligibility condition*

$$\max_{1\le j\le k_n} P\left(|X_{nj}-\mu_{nj}|>\tau B_n\right)\to 0, \qquad n\to\infty, \qquad \textit{each} \quad \tau>0,$$

(ii) the asymptotic normality condition $\sum_{j=1}^{k_n} X_{nj} \sim_{asy} N(A_n, B_n^2)$

together holds if and only if the following Lindeberg condition holds:

$$B_n^{-2}\sum_{j=1}^{k_n} EX_{nj}^2 I\{|X_{nj}-\mu_{nj}|>\tau B_n\}\to 0, \qquad n\to\infty, \qquad \textit{each} \quad \tau>0.$$

8.4 Asymptotic Relative Efficiency (ARE)

Definition. Let T_{n1} and T_{n2} be two estimators of θ. We say that

- T_{n1} is more efficient than T_{n2} if

$$RE(T_{n1},T_{n2}) := \frac{MSE(T_{n2})}{MSE(T_{n1})}\ge 1,$$

where RE := Relative Efficiency.

- T_{n1} is asymptotically more efficient than T_{n2} if

$$ARE(T_{n1},T_{n2}) := \lim_{n\to\infty}\frac{MSE(T_{n2})}{MSE(T_{n1})}\ge 1,$$

where ARE := Asymptotic Relative Efficiency.

- If both estimators are asymptotically unbiased, i.e., $MSE(T_{ni})\sim h(n)\sigma_i^2$, then

$$ARE(T_{n1},T_{n2}) = \frac{\sigma_2^2}{\sigma_1^2}.$$

Comparison between $\hat{m}$ and $\bar{X}$.

Let $X_1,...,X_n \sim F(x)$ with **symmetric** pdf $f(x)$. Then, $\mu = m$. Here we could use either the sample mean $\bar{X}$ or sample median $\hat{m}$ to estimator of the center.

1). $\bar{X} \sim_{asymp.} N(\mu, \sigma^2/n)$.
2). $\hat{m} \sim_{asymp.} N\left(\mu, [4f^2(\mu)]^{-1}/n\right)$.

Therefore,

$$ARE(\hat{m},\bar{X}) = \frac{\sigma^2/n}{[4f^2(\mu)]^{-1}/n} = 4\sigma^2 f^2(\mu).$$

1. Take $f(x) = (2\pi\sigma^2)^{-1/2}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ (i.e., $N(\mu,\sigma^2)$). Then $f^2(\mu) = 1/(2\pi\sigma^2)$. So

$$ARE(\hat{m},\bar{X}) = 2/\pi = 0.64.$$

2. If $f(x) = \frac{e^{-x}}{(1+e^{-x})^2}$ (logistic distribution), then $\mu = EX = 0$ and $\sigma^2 = var(X) = \pi^2/3$ (see the appendix below). $f(\mu) = f(0) = 1/4$. So

$$ARE(\hat{m}, \bar{X}) = 4\sigma^2 f^2(\mu) = \pi^2/12 = 0.82.$$

3. If $f(x) = \frac{1}{2}e^{-|x|}$ (double exponential), then $\mu = 0$, $\sigma^2 = \int x^2 f(x)dx = \int_0^\infty x^2 e^{-x}dx = \Gamma(3) = 2! = 2$. So

$$ARE(\hat{m}, \bar{X}) = 4\sigma^2 f^2(\mu) = 8 \times (1/4) = 2.$$

   So it is better to use the median in this case.

   Note that the MLE is the sample median in this case, which is known to be asymptotically efficient. So one can not find a more efficient one than the sample median asymptotically.

4. If $f(x)$ is the pdf of Cauchy distribution, then $\sigma^2 = \infty$. So

$$ARE(\hat{m}, \bar{X}) = \infty.$$

Remark: This example illustrates that as the tail of pdf gets thicker, the median becomes more efficient.

Appendix: proof of the logistic distribution

Logistic distribution. $X \sim f(x) = \dfrac{e^{-x}}{(1+e^{-x})^2}$. Find $\mu = EX$ and $\sigma^2 = var(X)$.

Solution. Clearly, $f(-x) = \dfrac{e^{x}}{(1+e^{x})^2} = \dfrac{e^{-x}}{(1+e^{-x})^2} = f(x)$, i.e., $f(x)$ is even and symmetric around the origin. Thus $\mu = EX = 0$. And

$$\begin{aligned}
\sigma^2 &= E(X-\mu)^2 = EX^2 = \int_{-\infty}^{\infty} x^2 f(x)dx = 2\int_0^\infty x^2 f(x)dx \\
&= 2\int_0^\infty \frac{x^2 e^x}{(1+e^x)^2}dx \\
&= -2\int_0^\infty x^2 d\left(\frac{1}{1+e^x}\right) \\
&= -2\frac{x^2}{1+e^x}\bigg|_0^\infty + 4\int_0^\infty \frac{x}{1+e^x}dx \\
&= 4\int_0^\infty \frac{xe^{-x}}{1+e^{-x}}dx \\
&= 4\int_0^\infty xe^{-x}\left(1 - e^{-x} + e^{-2x} - e^{-3x} + e^{-4x} + ...\right)dx \\
&= 4\int_0^\infty xe^{-x}\sum_{m=0}^{\infty}(-1)^m e^{-mx}dx \\
&= 4\int_0^\infty \sum_{m=0}^{\infty}(-1)^m xe^{-(m+1)x}dx \\
&= 4\sum_{m=0}^{\infty}(-1)^m \int_0^\infty xe^{-(m+1)x}dx \qquad \text{(interchanging integration and summation)}
\end{aligned}$$

$$
\begin{aligned}
&= 4\sum_{m=0}^{\infty}\frac{(-1)^m}{(m+1)^2}\int_0^{\infty} ye^{-y}dy \qquad [\text{take } y=(m+1)x] \\
&= 4\sum_{m=0}^{\infty}\frac{(-1)^m}{(m+1)^2} = 4\left(1-\frac{1}{2^2}+\frac{1}{3^2}-\frac{1}{4^2}+\frac{1}{5^2}-\frac{1}{6^2}+......\right)
\end{aligned}
$$

But it is known that $1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\frac{1}{5^2}+\frac{1}{6^2}+......=\frac{\pi^2}{6}$ So

$$
2\left(\frac{1}{2^2}+\frac{1}{4^2}+\frac{1}{6^2}+\frac{1}{8^2}+......\right)=2\times\frac{1}{4}\left(1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+......\right)=\frac{\pi^2}{12}
$$

The difference of the above two relations results in

$$
1-\frac{1}{2^2}+\frac{1}{3^2}-\frac{1}{4^2}+\frac{1}{5^2}-\frac{1}{6^2}+......=\frac{\pi^2}{12}
$$

Therefore, $\sigma^2=4\times\frac{\pi^2}{12}=\frac{\pi^2}{3}$.

8.5 Robustness v.s. Efficiency

Assuming that $\psi(x,\theta)=\psi(x-\theta)$, then

$$
\sigma_M^2=\frac{E\psi^2(X-\theta_0)}{[E\psi'(X-\theta_0)]^2}=\frac{A}{B^2}.
$$

Assume $E|\psi(X-\theta_0)|<\infty$, then $f_{\theta_0}(x)\psi(x-\theta_0)\to 0$ as $|x|\to\infty$. Thus

$$
\begin{aligned}
B &= \int \psi'(x-\theta_0)f_{\theta_0}(x)dx \\
&= \int f_{\theta_0}(x)d\psi(x-\theta_0) \\
&= f_{\theta_0}(x)\psi(x-\theta_0)|_{-\infty}^{\infty}-\int \psi(x-\theta_0)f'_{\theta_0}(x)dx \\
&= -\int \psi(x-\theta_0)f'_{\theta_0}(x)dx, \\
&= -E\left(\psi(X-\theta_0)\frac{f'_{\theta_0}(X)}{f_{\theta_0}(X)}\right), \\
&= -E\left(\psi(X-\theta_0)\frac{\partial \ln f_{\theta_0}(X)}{\partial\theta}\right).
\end{aligned}
$$

Therefore, by C-S inequality,

$$
B^2\le E\left(\psi^2(X-\theta_0)\right)E\left(\frac{\partial \ln f_{\theta_0}(X)}{\partial\theta}\right)^2=AI(\theta),
$$

where $I(\theta)$ is the Fisher information. Combining all these, we get

$$
\sigma_M^2=\frac{A}{B^2}\ge\frac{1}{I(\theta)},
$$

where the equality holds iff

$$
\psi(X-\theta_0)=a\frac{\partial \ln f_{\theta_0}(X)}{\partial\theta}, \quad \text{for some } a.
$$

That is, M-estimators are not efficient unless they are MLE's.

8.6 Generalized Estimating Equations (GEE)

The method of GEE is a very powerful tool and general method of deriving point estimators, which includes many other methods as special cases. We omit the details here.

8.7 Key figures

8.7.1 Frank Hampel, 1941-2018

Frank Rudolf Hampel was well known for his fundamental contributions to robust statistics, in particular for the introduction of the basic concepts of influence function and breakdown point. The influence function – ?perhaps the most useful heuristic tool of robust statistics,? according to Peter Huber (Robust Statistics, Wiley 1981, pp.13?14)?describes the approximate effect on an estimate when inserting, deleting or modifying a single observation. Moreover, the asymptotic variance of an estimator is given by the expected value of the squared influence function. This connection allowed Frank to formulate and solve a central optimality problem in robust statistics, namely to minimize the asymptotic variance under a bound on the influence of a single observation (Lemma 5 in his thesis). In contrast to the infinitesimal description provided by the influence function, the breakdown point is a global measure that gives the largest percentage of arbitrary bad observations an estimator can tolerate without diverging. His book Robust statistics: The approach based on the influence function, written together with Elvezio Ronchetti, Peter Rousseeuw and Werner Stahel (Wiley, 1986), contains a systematic exposition of the area. It served as a key reference for more than two decades and was highly influential.

In addition to deviations from an assumed marginal distribution, Frank also considered deviations from independence, advocating the use of long-range dependence models as the most relevant type of unsuspected dependence. Another important contribution by Frank is what he called *small sample asymptotics*, a variant of saddle-point approximations for the distribution of estimators, based on a different derivation. They provide an excellent agreement with the exact distribution even for very small samples.

In his later years, Frank focused on the philosophical foundations of statistics. He argued for describing epistemic uncertainty by upper and lower probabilities, corresponding to one-sided bets. In his approach, total ignorance about an event means that one refuses to bet on either the event or its complement. It remains to be seen if these ideas will be recognized in the future as a fundamental new approach.

Frank grew up in Germany during World War II; his father died when he was one year old. His mother then moved to the house of his grandfather in Upper Silesia. Because this region became Polish at the end of the war, the family was forced to leave and ended up near Gttingen. After high school, Frank studied physics, mathematics and philosophy in Munich and Gttingen. His professor in Gttingen, Konrad Jacobs, who worked in ergodic theory, showed him the seminal 1964 Annals of Mathematical Statistics paper by Peter Huber, and encouraged him to go to Berkeley with a one-year exchange scholarship. He decided to stay there and completed his PhD in 1968. Officially, Erich Lehmann was his advisor, but Erich wrote in Reminiscences of a Statistician (Springer, 2008, p.158) that "...in fact I had essentially no input. My contribution consisted of my immediate realization of the importance and maturity of this work ... and my task was to encourage, smooth the process and otherwise stay out of the way." After his PhD, Frank accepted an offer by Volker Strassen (famous for proving an invariance principle for the law of the iterated logarithm) to move with him from Berkeley to the University of Zurich and to take a position as "Oberassistent", being in charge of the statistical consulting service. In 1970-71, Frank was invited together with Peter Bickel and Peter Huber to join John Tukey during the Princeton robustness year, which had a big impact on the further development of robust statistics. In 1974, he was elected as associate professor at ETH Zurich, thus becoming a colleague of Peter Huber. He was soon promoted to full professor and stayed at ETH until his retirement in 2006. In 2007 he received an honorary doctorate from the University of Dortmund for his "scientific achievements in the area of modern statistics and data analysis."

8.8 Exercises

1. *Huber's M-estimate.* Take $\psi(x,\theta) = \psi_0(x-\theta)$, where

$$\begin{aligned} \psi_0(x) &= -k, \qquad x < -k, \\ &= x, \qquad |x| \le k, \\ &= k, \qquad x > k, \end{aligned}$$

Using the theorem given in the lecture, show that any solution sequence T_n of $\sum_{i=1}^n \psi(X_i, T_n) = 0$ satisfies asymptotically

$$\sqrt{n}\,(T_n - \theta) \longrightarrow N(0,\sigma^2), \qquad \text{in distribution,}$$

where

$$\sigma^2 = \frac{\int_{\theta-k}^{\theta+k}(x-\theta)^2 dF(x) + k^2\int_{-\infty}^{\theta-k} dF(x) + k^2\int_{\theta+k}^{\infty} dF(x)}{\left(\int_{\theta-k}^{\theta+k} dF(x)\right)^2}$$

2. Let $X_1,\ldots,X_n$ be i.i.d. from the Cauchy distribution with pdf given by

$$f(x) = \frac{C}{1+x^2}.$$

   (a) Show that $E|X_{(k)}|^r < \infty$ iff $r < k < n-r+1$.

   (Hint: the pdf of $X_{(k)}$ is $f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}[1-F(x)]^{n-k}f(x)$.)

   (b) Cauchy d.f. does not have expectations. However, from the above,

      i. $E|X_{(k)}| < \infty$ iff $2 \le k \le n-1$,
      i.e., only the smallest/largest order statistics do not have finite first moments.

      ii. $EX_{(k)}^2 < \infty$ iff $3 \le k \le n-2$,
      i.e., only the smallest/largest two order statistics do not have finite second moments.

   (c) Check if the trimmed mean by removing the smallest and largest two observations is a consistent estimate of the center of the d.f. consistently.

   (Hint: find the mean/variance of the trimmed mean.)

Chapter 9

U-Statistics

Hoeffding (1948) introduced U-statistics and proved their central limit theorem. There are several factors that are unique to the class of U-statistics.

- Many commonly used statistics are or can be approximated by U-statistics, so they are much more applicable than it first appears.
- U-statistics are generalisations of sums of independent r.v.s, thus these developments on the theory of U-statistics are strongly influenced by the classical theory on sums of independent random variables.
- Many of the techniques developed for U-statistics are also essential in studying even more general classes of statistics, such as the symmetrical statistics.
- The special dependence structure in the U-statistics allows one to employ the martingale properties extensively.

9.1 U-statistics

Definition. Let $X_1, ..., X_n \sim_{i.i.d.} F$ and $h(x_1, ..., x_m)$ be a symmetric function in its arguments. Consider

$$\theta = T(F) = E_F h(X_1, ..., X_m) = \int ... \int h(x_1, ..., x_m) dF(x_1) ... dF(x_m).$$

The U- and V-statistics of order m with kernel h are defined, respectively, by

$$\begin{aligned} V_n &= T(F_n) = \frac{1}{n^m} \sum_{i_1=1}^{n} ... \sum_{i_m=1}^{n} h(X_{i_1}, ..., X_{i_m}), \\ U_n &= \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < ... < i_m \le n}^{n} h(X_{i_1}, ..., X_{i_m}) \\ &= \frac{1}{n(n-1)...(n-m+1)} \sum_{1 \le i_1 \neq ... \neq i_m \le n}^{n} h(X_{i_1}, ..., X_{i_m}). \quad \blacksquare \end{aligned}$$

Remark **9.1** *U-statistics are always unbiased ("U" = unbiased) while V-statistics ("V" comes from "von Mises", who studied this type of statistics first) are typically biased for $m \ge 2$, since usually $Eh(X_1, ..., X_1) \neq Eh(X_1, ..., X_n)$ for diagonal elements.*

9.2 Examples

- If $h(x) = x$, the U-statistic
$$U_n = \bar{x}_n = (x_1 + \cdots + x_n)/n$$
is the sample mean.

- If $h(x_1, x_2) = |x_1 - x_2|$, the U-statistic
$$U_n = \frac{1}{n(n-1)} \sum_{i \neq j} |x_i - x_j|, \qquad n \geq 2$$
is the mean pairwise deviation.

- If $h(x_1, x_2) = (x_1 - x_2)^2/2$, the U-statistic is
$$U_n = \frac{1}{n-1} \sum (x_i - \bar{x}_n)^2, \qquad n \geq 2$$
the sample variance.

- The third k-statistic
$$k_{3,n}(x) = \frac{n}{(n-1)(n-2)} \sum (x_i - \bar{x}_n)^3$$
i.e., the sample skewness defined for $n \geq 3$, is a U-statistic.

9.3 Hoeffding-decomposition

Given $T := f(X_1, ..., X_n)$, we have the Hoeffding-decomposition (or projection, or ANOVA decomposition):

$$T = \sum_{i=1}^{n} T_i + \sum_{i<j} T_{ij} + \sum_{i<j<k} T_{ijk} + \cdots + \sum_{1 \leq i_1 < \ldots < i_n} T_{i_1 \ldots i_n},$$

where

$$\begin{aligned}
T_i &= E(T|X_i) \\
T_{ij} &= E(T|X_i, X_j) - T_i - T_j \\
T_{ijk} &= E(T|X_i, X_j, X_k) - E(T|X_i, X_j) - E(T|X_i, X_k) - E(T|X_j, X_k) \\
&\qquad\qquad\qquad + E(T|X_i) + E(T|X_j) + E(T|X_k) \\
&= E(T|X_i, X_j, X_k) \\
&\quad -\{E(T|X_i, X_j) - E(T|X_i) - E(T|X_j)\} \\
&\quad -\{E(T|X_i, X_k) - E(T|X_i) - E(T|X_k)\} \\
&\quad -\{E(T|X_i, X_k) - E(T|X_j) - E(T|X_k)\} \\
&\quad -E(T|X_i) - E(T|X_j) - E(T|X_k) \\
&= E(T|X_i, X_j, X_k) - (T_{ij} + T_{ik} + T_{jk}) - (T_i + T_j + T_k).
\end{aligned}$$

Clearly,

- T_i is the main effect of X_i on T,
- T_{ij} is the 2-way interaction effect of X_i and X_j on T, after removing the main effects.
- T_{ijk} is the 3-way interaction effect of X_i, X_j, X_k on T, after removing the main and two-way interaction effects.

Heuristic proof

For $T = f(X_1, ..., X_n)$, note $E(T|X_1, ..., X_n) = T$.

- For $n = 1$,

$$T \;:=\; f(X_1) = E(T|X_1)$$

- For $n = 2$,

$$\begin{aligned} T &:= f(X_1, X_2) \\ &= E(T|X_1) + E(T|X_2) + \{E(T|X_1, X_2) - E(T|X_1) - E(T|X_2)\} \\ &= \sum_{i=1}^{2} E(T|X_i) + \sum_{1 \le i < j \le 2} [E(T|X_i, X_j) - E(T|X_i) - E(T|X_j)] \end{aligned}$$

- For $n = 3$,

$$\begin{aligned} T &:= f(X_1, X_2, X_3) \\ &= E(T|X_1) + E(T|X_2) + E(T|X_3) \\ &\quad +\{E(T|X_1, X_2) - E(T|X_1) - E(T|X_2)\} \\ &\quad +\{E(T|X_1, X_3) - E(T|X_1) - E(T|X_3)\} \\ &\quad +\{E(T|X_2, X_3) - E(T|X_2) - E(T|X_3)\} \\ &\quad +\{E(T|X_1, X_2, X_3) - E(T|X_1, X_2) - E(T|X_1, X_3) - E(T|X_2, X_3) \\ &\qquad\qquad +E(T|X_1) + E(T|X_2) + E(T|X_3)\} \\ &= \sum_{i=1}^{3} E(T|X_i) + \sum_{1 \le i < j \le 3} [E(T|X_i, X_j) - E(T|X_i) - E(T|X_j)] \\ &\quad + \sum_{1 \le i < j < k \le 3} \{E(T|X_i, X_j, X_k) - E(T|X_i, X_j) - E(T|X_i, X_k) - E(T|X_j, X_k) \\ &\qquad\qquad +E(T|X_i) + E(T|X_j) + E(T|X_k)\}. \end{aligned}$$

- The general n can be shown similarly. ∎

REMARK **9.2 Hoeffding-decomposition** *is an identity. Given T, its overall effect is ET, giving*

$$T = ET + R_1.$$

For R_1, the main effect (or contribution) from X_i is given by T_i, resulting in

$$T - ET = \sum_{i=1}^{n} T_i + R_2.$$

For R_2, the main effect of 2-way interactions from (X_i, X_j) is $T_{i,j}$, giving

$$T = ET + \sum_{i=1}^{n} T_i + \sum_{1 \le i_1 < i_2 \le 2}^{n} T_{i_1, i_2} + R_3.$$

Continuing on until all interaction effects are taken care of, giving the above ANOVE decomposition.

THEOREM **9.1** *The Hoeffding-decomposition to 2-order U-statistics is*

$$U_n - \theta = \frac{2}{n}\sum_{i=1}^{n} g(X_i) + \frac{2}{n(n-1)}\sum_{i<j}\psi(X_i, X_j),$$

where $h_{ij} = h(X_i, X_j)$, $g_i = g(X_i) = E\left(h_{ij} \mid X_i\right) - \theta$ *and* $\psi(X_i, X_j) = h_{ij} - g_i - g_j + \theta$ *and all the terms on the RHS of* U_n *are uncorrelated.*

Proof. WLOG, assume that $\theta = 0$.

$$\begin{aligned} T_1 &= E(U_n|X_1) = \frac{2}{n(n-1)}\sum_{i<j} E[h(X_i, X_j)|X_1] \\ &= \frac{2}{n(n-1)}(n-1)E[h(X_1, X_j)|X_1] \\ &= \frac{2}{n}g_1, \\ E(U_n|X_1, X_2) &= \frac{2}{n(n-1)}\sum_{i<j} E[h(X_i, X_j)|X_1, X_2] \\ &= \frac{2}{n(n-1)}[h_{12} + (n-2)g_1 + (n-2)g_2] \end{aligned}$$

Hence,

$$\begin{aligned} T_{12} &= E(U_n|X_1, X_2) - E(U_n|X_1) - E(U_n|X_2) \\ &= \frac{2}{n(n-1)}\{h_{12} + (n-2)[g_1 + g_2] - (n-1)[g_1 + g_2]\} \\ &= \frac{2}{n(n-1)}\left(h_{12} - g_1 - g_2\right) \\ &= \frac{2}{n(n-1)}\psi(X_1, X_2). \end{aligned}$$

To prove uncorrelatedness, we will show a few examples. First we note that $E[\psi(X_1, X_2)|X_1]\} = E[h(X_1, X_2) - g(X_1) - g(X_2)|X_1]\} = g(X_1) - g(X_1) - Eg(X_2) = 0$. Hence,

$$Eg(X_1)\psi(X_1, X_2) = EE[g(X_1)\psi(X_1, X_2)|X_1] = E\{g(X_1)E[\psi(X_1, X_2)|X_1]\}$$

and

$$\begin{aligned} &E\psi(X_1, X_2)\psi(X_1, X_3) = EE[\psi(X_1, X_2)\psi(X_1, X_3)|X_1] \\ &= E\{E[\psi(X_1, X_2)|X_1]\}E\{E[\psi(X_1, X_3)|X_1]\} = 0. \quad \blacksquare \end{aligned}$$

9.3.1 Asymptotic Normality

THEOREM **9.2** *If* $Eh^2(X_1, X_2) < \infty$ *and* $\sigma_g^2 > 0$, *then*

$$\frac{\sqrt{n}(U_n - \theta)}{2\sigma_g} \longrightarrow_d N(0, 1).$$

Proof. Note $\sqrt{n}(U_n - \theta) = \frac{2}{\sqrt{n}}\sum_{i=1}^{n} g(X_i) + \sqrt{n}R_n$, where $R_n = \frac{2}{n(n-1)}\sum_{i<j}\psi(X_i, X_j)$. Now

$$Var(\sqrt{n}R_n) = n\frac{4}{n^2(n-1)^2}\sum_{i<j} Var(\psi(X_i, X_j)) = \frac{2n}{n(n-1)}Var(\psi(X_i, X_j)) \to 0,$$

implying $\sqrt{n}R_n \to_p 0$. Then apply the CLT and Slutsky theorem. $\blacksquare$

9.4 Jackknife variance estimates for U-statistics

Note that $Var(U_n) = 4Var[g(X_1)]/n + O(n^{-2})$ which is approximated by

$$\widetilde{Var}(U_n) \simeq \frac{4}{n}\frac{1}{n}\sum_{i=1}^{n}[g(X_i)]^2 = \frac{4}{n^2}\sum_{i=1}^{n} g^2(X_i).$$

Here $g(X_i) = E\{h(X_i, X_j)|X_i\}$ can be approximated by $g(X_i) \simeq \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} h_{ij} - U_n$. Then an approximation to the variance of U_n is

$$\widehat{Var}(U_n) = \frac{4}{n^2}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1,j\neq i}^{n} h_{ij} - U_n\right)^2.$$

This is almost the same as the jackknife variance estimator, $\widehat{Var}_{Jack}(U_n)$, given in the next theorem. In effect, $\widehat{Var}_{Jack}(U_n)$ estimates the variance of the dominating term in the H-decomposition of U_n.

Theorem **9.3** *If U_n is a U-statistics of order 2 with kernel $h(x, y)$, then*

$$\widehat{Var}_{Jack}(U_n) = \frac{4(n-1)}{n(n-2)^2}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1,i\neq j}^{n} h_{ij} - U_n\right)^2. \tag{4.1}$$

Furthermore, $\widehat{Var}_{Jack}(U_n)$ is a consistent estimator of $Var(U_n)$ in the sense that

$$\frac{\widehat{Var}_{Jack}(U_n)}{Var(U_n)} \longrightarrow_p 1, \qquad \text{as } n \to \infty. \tag{4.2}$$

Proof. We only derive (3.1) below (the proof for (3.2) is more involved and hence omitted). From Theorem 10.1, we have

$$\begin{aligned}
U_n &= \frac{1}{n(n-1)}\sum_{i\neq j} h_{ij} \\
U_{n-1}^{(-k)} &= \frac{1}{(n-1)(n-2)}\left(\sum_{i\neq j} h_{ij} - 2\sum_{j=1,j\neq k}^{n} h_{kj}\right) \\
U_{n-1}^{(\cdot)} &= \frac{1}{n}\sum_{k=1}^{n} U_{n-1}^{(-k)} \\
&= \frac{1}{(n-1)(n-2)}\left(\sum_{i\neq j}^{n} h_{ij} - \frac{2}{n}\sum_{k=1}^{n}\sum_{j=1}^{n} h_{kj} + \frac{2}{n}\sum_{k=1}^{n} h_{kk}\right) \\
&= \frac{1}{(n-1)(n-2)}\left(\sum_{i\neq j} h_{ij} - \frac{2}{n}\sum_{i\neq j} h_{ij}\right) \\
&= \frac{1}{(n-1)(n-2)}\left(\sum_{i\neq j} h_{ij} - 2(n-1)U_n\right) \\
\widehat{Var}_{Jack}(U_n) &= \frac{n-1}{n}\sum_{k=1}^{n}\left(U_{n-1}^{(-k)} - U_{n-1}^{(\cdot)}\right)^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{n-1}{n}\frac{1}{(n-1)^2(n-2)^2}\sum_{k=1}^{n}\left(2\sum_{j=1,j\neq k}^{n} h_{kj} - 2(n-1)U_n\right)^2 \\
&= \frac{n-1}{n}\frac{4(n-1)^2}{(n-1)^2(n-2)^2}\sum_{k=1}^{n}\left(\frac{1}{n-1}\sum_{j=1,j\neq k}^{n} h_{kj} - U_n\right)^2 \\
&= \frac{4(n-1)}{n(n-2)^2}\sum_{k=1}^{n}\left(\frac{1}{n-1}\sum_{j=1,j\neq k}^{n} h_{kj} - U_n\right)^2 . \quad \blacksquare
\end{aligned}
$$

From Slutsky theorem, we get

THEOREM **9.4** *If* $Eh^2(X_1, X_2) < \infty$ *and* $\sigma_g^2 > 0$, *then*

$$
\frac{\sqrt{n}(U_n - \theta)}{S_n} \longrightarrow_d N(0,1).
$$

Chapter 10

Jackknife

The jackknife and bootstrap are two popular nonparametric methods in statistical inference. The first idea related to the bootstrap was von-Mises, who used the plug-in principle in the 1930's, (although it can be argued that Gauss invented the bootstrap in 1800's). Then in the late 40's Quenouille found a way of correcting the bias for estimators whose bias was known to be of the form. The method is called the jackknife.

Both methods provide several advantages over the traditional parametric approach: the methods are easy to describe and they apply to arbitrarily complicated situations; distribution assumptions, such as normality, are never made.

The jackknife preceded the bootstrap. It is simple to use. The original work on the "delete-one" jackknife is due to Quenouille (1949) and Tukey (1958).

10.1 The traditional approach

Let $\mathbf{X} = X_1, \ldots, X_n \sim_{i.i.d.} F$ (typically unknown). The parameter of interest is $\theta \equiv T(F)$. Let $T_n \equiv T(X_1, \ldots, X_n)$ be an estimator of θ. For T_n, we might be interested in

1. its bias: $Bias(T_n) = E(T_n) - \theta$ (for correcting for bias of T_n).
2. its variance: $\sigma^2_{T_n} = Var(T_n)$ (for measuring the accuracy of T_n).
3. its sampling distribution: $P\left((T_n - \theta)/\sigma_{T_n} \le x\right)$, or $P\left((T_n - \theta)/\hat{\sigma}_{T_n} \le x\right)$? (for constructing confidence intervals for θ).

Example 1. (Sample mean.) Let $\theta = \mu \equiv EX_1$, and $T_n = \bar{X}$, then

1. $Bias(\bar{X}) = E(\bar{X}) - \mu = 0$, unbiased.
2. $Var(\bar{X}) = \sigma^2/n$, where $\sigma^2 = Var(X_1)$. So an estimate of $Var(\bar{X})$ is
$$\hat{\sigma}^2_{T_n} =: \widehat{Var}(\bar{X}) = \widetilde{S}^2/n, \qquad \text{where} \quad \widetilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$
3. By the CLT and Slutsky Theorem, $P\left((T_n - \theta)/\hat{\sigma}_{T_n} \le x\right) \longrightarrow \Phi(x)$. So an approximate $1 - \alpha$ C.I. for μ is
$$\bar{X} \mp z_{\alpha/2}\frac{\widetilde{S}}{\sqrt{n}}.$$

Example 2. Let $\theta = \mu^2$, and $T_n = (\bar{X})^2$, then

1. $Bias[(\bar{X})^2] = E[(\bar{X})^2] - \mu^2 = Var(\bar{X}) + [E(\bar{X})]^2 - \mu^2 = \dfrac{\sigma^2}{n}$. Hence, a biased-corrected estimator of μ^2 is given by
$$\widetilde{T}_n = T_n - \frac{\widetilde{S}^2}{n} = (\bar{X})^2 - \frac{\widetilde{S}^2}{n},$$
which is in fact unbiased for μ^2. (But its variance may increase as a result, so one need to use the MSE criterion, e.g., to choose a better one.)
2. Let us find out $Var(T_n)$ next. Let $Y = X - \mu$. Note that $T_n = (\bar{X} - \mu + \mu)^2 = (\bar{Y} + \mu)^2 = (\bar{Y})^2 + 2\mu\bar{Y} + \mu^2$. Hence
$$\begin{aligned} Var(T_n) &= Var[(\bar{Y})^2] + 4Cov(\bar{Y}, (\bar{Y})^2) + 4\mu^2 Var(\bar{Y}) \\ &= E[(\bar{Y})^4] - \{E[(\bar{Y})^2]\}^2 + 4E[(\bar{Y})^3] + 4\mu^2\sigma^2/n \\ &= \end{aligned}$$
(Please finish the rest of the calculations.)

Alternatively, as we have seen, by using the transformation approach, the asymptotic variance of T_n is $4\mu^2\sigma^2/n$, which can be estimated by $4(\bar{X})^2\widetilde{S}^2/n$.

3. By the transformation approach, $\dfrac{T_n - \mu^2}{2|\mu|\sigma} \to_d N(0,1)$, or $\dfrac{T_n - \mu^2}{2|\bar{X}|\widetilde{S}} \to_d N(0,1)$, which can then be used to construct a C.I. for μ^2. ∎

Exercise. Let (X_i, Y_i) be i.i.d. r.v.s. Let $\theta = \rho$ and $T_n = \hat{\rho}$ be the population and sample correlation coefficients, respectively. Repeat the last example. ∎

The weakness of the traditional approaches

There are several inherent weaknesses associated with traditional approaches.

(1) The theoretical formula for bias, variance, etc can be very difficult or even impossible to obtain. It often requires the data analysts to have strong mathematical and statistical background.

(2) The theorectical formula may to be too complicated to be useful.

(3) The formula are problem-specific. For a new problem, we need to repeat the whole process all over again.

(4) The formula are often valid only for large sample size n (e.g. CLT). The performance for small n can often be poor.

In the following two chapters, we shall introduce alternative methods such as bootstrap and jackknife and other resampling methods to overcome the above-mentioned difficulties. Some of the advantages are listed below.

(1) The methods are highly automatic, easy to use.

(2) The methods require little mathematical background on the user's part.

(3)) The methods are general, not problem-specific.

(4) The methods have good performances for small sample size, as compared to some asymptotic methods.

10.2 Jackknife

Definition. Suppose that $\hat{\theta} = T(F_n) := T(X_1, ..., X_n)$ estimates $\theta = T(F)$. Let

$$T_{n-1}^{(-i)} \equiv T(F_{n-1}^{(-i)}) = T(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n),$$

where

$$F_{n-1}^{(-i)}(x) \equiv \frac{1}{n-1} \sum_{j \neq i} I_{\{X_j \leq x\}}.$$

That is, $T_{n-1}^{(-i)}$ is the estimator of $\theta = T(F)$ based on the data after deleting X_i.

Remark **10.1** *Tukey (1958) defined the i-th pseudo-value as*

$$T_i^{pseudo} = nT_n - (n-1)T_{n-1}^{(-i)}.$$

and conjectured that

- T_i^{pseudo} *$(1 \leq i \leq n)$ may be treated as i.i.d. r.v.'s, which have been verified in many cases such as U-statistics and functions of means.*
- T_i^{pseudo} *has approximately the same variance as $\sqrt{n}T_n$.*

We shall now use pseudo-values T_i^{pseudo} to estimate bias, variance, etc. Let

$$T_{n-1}^{(\cdot)} = \frac{1}{n} \sum_{i=1}^{n} T_{n-1}^{(-i)}.$$

1. The jackknife estimate of $\theta = T(F)$ (biased-corrected) is

$$T_n^{Jack} = \frac{1}{n} \sum_{i=1}^{n} T_i^{pseudo} = nT_n - (n-1)T_{n-1}^{(\cdot)}.$$

2. The **jackknife bias estimate** is

$$\widehat{bias}(T_n) \equiv T_n - T_n^{Jack} = (n-1)[T_{n-1}^{(\cdot)} - T_n].$$

   (By definition, $T_n^{Jack} = T_n - \widehat{bias}(T_n)$ should hopefully have smaller bias than T_n.)

3. From Remark 10.1, we estimate $var(\sqrt{n}T_n)$ by the sample variance based on $\widetilde{T}_{n(i)}^{pseudo}$. Hence, the **jackknife estimate of** $Var(T_n)$ is

$$\begin{aligned} \hat{V}_{Jack}(T_n) &= \frac{1}{n}\left[\frac{1}{(n-1)} \sum_{i=1}^{n} \left(T_i^{pseudo} - \frac{1}{n}\sum_{i=1}^{n} T_i^{pseudo}\right)^2\right] \\ &= \frac{n-1}{n} \sum_{i=1}^{n} \left(T_{n-1}^{(-i)} - T_{n-1}^{(\cdot)}\right)^2. \end{aligned}$$

4. The jackknife can also be used to estimate d.f.'s. Details are omitted.

Remark **10.2** *The delete-1 jackknife estimate of $P(T_n \leq t)$ can be taken as $\hat{P}_{Jack}(T_n \leq t) = \frac{1}{n}\sum_{i=1}^{n} I\{T_{n-1}^{(-i)} \leq t\}$ or $\hat{P}_{Jack}(T_n \leq t) = \frac{1}{n}\sum_{i=1}^{n} I\{T_i^{pseudo} \leq t\}$. This has poor properties for i.i.d. sample. But for dependent sample, one could use the delete-1 jackknife estimate, which will be mentioned later.*

10.2.1 Heuristic justification

Suppose that we have the following expansion,

$$E(T_n) = T(F) + \frac{a}{n} + \frac{b}{n^2} + O(n^{-3}).$$

Then,

$$\begin{aligned} E[T_{n-1}^{(-i)}] &= E[T_{n(.)}] \\ &= T(F) + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O(n^{-3}). \end{aligned}$$

Hence

$$\begin{aligned} E[T_n^{Jack}] &= nET_n - (n-1)ET_{n-1}^{(\cdot)} \\ &= n\left(T(F) + \frac{a}{n} + \frac{b}{n^2} + O(n^{-3})\right) - (n-1)\left(T(F) + \frac{a}{n-1} + \frac{b}{(n-1)^2} + O(n^{-3})\right) \\ &= \left(nT(F) + a + \frac{b}{n} + O(n^{-2})\right) - \left((n-1)T(F) + a + \frac{b}{(n-1)} + O(n^{-2})\right) \\ &= T(F) - \frac{b}{n(n-1)} + O(n^{-2}). \end{aligned}$$

So T_n^{Jack} has bias of order $O(n^{-2})$, one order smaller than the bias of T_n (of $O(n^{-1})$).

One can repeat the jackknife again and again to keep on reducing the bias. **But be aware**: the jackknife might introduce more variability. One could use the Mean Square Error (MSE) to see if one- or more-step jackknife is worthwhile or not. Usually, the jackknife is rarely used beyond one-step.

10.2.2 The examples revisited

Example 1. (Sample mean.) Let $\theta = T(F) = \mu \equiv EX_1$, and $T_n = \bar{X}$. Then,

$$T_{n-1}^{(-i)} = T(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n) = \frac{X_1, +... + X_{i-1} + X_{i+1} ... + X_n}{n-1} = \frac{n\bar{X} - X_i}{n-1}$$

$$T_{n-1}^{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} T_{n-1}^{(-i)} = \frac{n\bar{X} - \bar{X}}{n-1} = \bar{X}.$$

So we have

1. The i-th pseudo-value is
$$T_i^{pseudo} = nT_n - (n-1)T_{n-1}^{(-i)} = \sum_{i=1}^{n} X_i - (n\bar{X} - X_i) = X_i.$$

2. The jackknife estimate of $\theta = T(F)$ (biased-corrected) is
$$T_n^{Jack} = \frac{1}{n}\sum_{i=1}^{n} T_i^{pseudo} = \bar{X}.$$

3. The jackknife bias estimate is
$$\widehat{bias}(T_n) \equiv T_n - T_n^{Jack} = \bar{X} - \bar{X} = 0.$$

4. The jackknife estimate of $Var(\hat{\theta})$ is
$$\begin{aligned} \widehat{Var}_{Jack}(\hat{\theta}) &= \frac{n-1}{n}\sum_{i=1}^{n}\left(T_{n-1}^{(-i)} - T_{n-1}^{(\cdot)}\right)^2 = \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{n\bar{X} - X_i}{n-1} - \bar{X}\right)^2 \\ &= \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} = \frac{\widetilde{S}^2}{n}, \end{aligned}$$

   which is unbiased for σ^2/n. ∎

10.2.3 Definition

Example 2. Let $\theta = T(F) = \mu^2$, and $T_n = T(F_n) = (\bar{X})^2$. Then,

$$T_{n-1}^{(-i)} = \left(\frac{n\bar{X} - X_i}{n-1}\right)^2, \qquad T_{n-1}^{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} T_{n-1}^{(-i)}.$$

The i-th pseudo-value is $T_i^{pseudo} = nT_n - (n-1)T_{n-1}^{(-i)}$.

1. The jackknife bias estimate is

$$\begin{aligned}
\widehat{bias}(T_n) &\equiv (n-1)[T_{n-1}^{(\cdot)} - T_n] \\
&= \frac{(n-1)}{n}\sum_{i=1}^{n}\left(\left(\frac{n\bar{X} - X_i}{n-1}\right)^2 - (\bar{X})^2\right) \\
&= \frac{(n-1)}{n}\sum_{i=1}^{n}\left(\frac{n\bar{X} - X_i}{n-1} + \bar{X}\right)\left(\frac{n\bar{X} - X_i}{n-1} - \bar{X}\right) \\
&= \frac{(n-1)}{n}\sum_{i=1}^{n}\left(\frac{(2n-1)\bar{X} - X_i}{n-1}\right)\left(\frac{\bar{X} - X_i}{n-1}\right) \\
&= \frac{1}{(n-1)n}\sum_{i=1}^{n}(-X_i)\left(\bar{X} - X_i\right) \\
&= \frac{1}{(n-1)n}\sum_{i=1}^{n}\left(\bar{X} - X_i\right)\left(\bar{X} - X_i\right) \\
&= \frac{1}{(n-1)n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{\widetilde{S}^2}{n}.
\end{aligned}$$

2. The jackknife estimate of $\theta = T(F)$ (biased-corrected) is

$$T_n^{Jack} \equiv T_n - \widehat{bias}(T_n) = (\bar{X})^2 - \frac{\widetilde{S}^2}{n},$$

which is unbiased as $ET_n^{Jack} - \mu^2 = E(\bar{X})^2 - \frac{E\widetilde{S}^2}{n} - \mu^2 = \frac{\sigma^2}{n} - \mu^2 - \frac{\sigma^2}{n} - \mu^2 = 0.$

3. The jackknife estimate of $Var(\hat{\theta})$ is left as an exercise.
4. As an exercise, compare the MSEs of T_n and T_n^{Jack}. ∎

Example 3. Let $\theta = T(F) = \sigma^2 = Var(X)$, and $T_n = T(F_n) = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$, which is biased estimator of σ^2. Show that the jackknife estimate of $\theta = T(F)$ (biased-corrected) is

$$T_n^{Jack} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{(unbiased)}.$$

Proof. Exercise. ∎

Remark: As we shall see, $T(F_n)$ is a V-statistic of order 2, whose jackknife estimator is a U-statistic of order 2, hence unbiased.

Example 4. Let $T_n = X_{(n)}$ be the largest order statistic of a random sample $X_1, ..., X_n$. (T_n is often used to estimate the boundary point of a d.f., such as Unif$[0, \theta]$). Find its jackknife estimator T_n^{Jack}. Compare their MSEs.

Proof. It is easy to find that $T_n^{Jack} = \dfrac{2n-1}{n}X_{(n)} - \dfrac{n-1}{n}X_{(n-1)}$. Finding the MSEs is more tedious. ∎

10.3 Application of the Jackknife to U- and V-statistics

Definition. Let $X_1, ..., X_n \sim_{i.i.d.} F$ and $h(x_1, ..., x_m)$ be a symmetric function in its arguments. Consider

$$\theta = T(F) = Eh(X_1, ..., X_m) = \int ... \int h(x_1, ..., x_m) dF(x_1)...dF(x_m).$$

The U- and V-statistics of order m with kernel h are defined, respectively, by

$$\begin{aligned}
V_n &= T(F_n) = \frac{1}{n^m} \sum_{i_1=1}^{n} ... \sum_{i_m=1}^{n} h(X_{i_1}, ..., X_{i_m}), \\
U_n &= \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < ... < i_m \le n}^{n} h(X_{i_1}, ..., X_{i_m}) \\
&= \frac{1}{n(n-1)...(n-m+1)} \sum_{1 \le i_1 \ne ... \ne i_m \le n}^{n} h(X_{i_1}, ..., X_{i_m}). \quad \blacksquare
\end{aligned}$$

Remark **10.3** *U-statistics are always unbiased ("U" = unbiased) while V-statistics ("V" comes from "von Mises", who studied this type of statistics first) are typically biased for $m \ge 2$, since usually $Eh(X_1, ..., X_1) \ne Eh(X_1, ..., X_n)$ for diagonal elements.*

10.3.1 Reduce bias for V-statistics by jackknife

Examples.

1. $\theta = \mu = T(F)$, here $h(x) = x$ and $m = 1$. Then, $V_n = T(F_n) = \bar{X}$ and $U_n = \bar{X}$.

2. $\theta = \mu^2 = T(F)$, here $h(x, y) = xy$ and $m = 2$. Then, $V_n = T(F_n) = (\bar{X})^2$, and

$$\begin{aligned}
U_n &= \frac{1}{n(n-1)} \sum_{i \neq j} X_i X_j = \frac{1}{n(n-1)} \left(\sum_{i,j} X_i X_j - \sum_{i=j} X_i X_j \right) \\
&= \frac{1}{n(n-1)} \left(\left(\sum_{i=1}^{n} X_i \right)^2 - \sum_{i=1}^{n} X_i^2 \right) \\
&= \frac{1}{n(n-1)} \left(n^2 \left(\bar{X}\right)^2 - \left(\sum_{i=1}^{n} (X_i - \bar{X})^2 + n \left(\bar{X}\right)^2 \right) \right) \\
&= \left(\bar{X}\right)^2 - \frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2 := \left(\bar{X}\right)^2 - \frac{\widetilde{S}^2}{n},
\end{aligned}$$

which is unbiased for μ^2.

3. $\theta = \sigma^2 = T(F) = Var_F(X) = \frac{1}{2} Var_F(X - Y) = \frac{1}{2} E_F \left(X - Y\right)^2$, where $X, Y \sim F$ and independent, hence $h(x, y) = \frac{(x-y)^2}{2}$. Hence,

$$\begin{aligned}
V_n &= Var_{F_n}(X) = \int (x - \int x dF_n(x))^2 dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2, \\
U_n &= \frac{1}{n(n-1)} \sum_{i \neq j} \frac{(X_i - X_j)^2}{2} = \frac{1}{n(n-1)} \sum_{i,j} \frac{(X_i - X_j)^2}{2} \\
&= \frac{1}{n(n-1)} n^2 V_n = \frac{1}{n(n-1)} n^2 \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \\
&= \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2 := \widetilde{S}^2,
\end{aligned}$$

which is unbiased for σ^2.

The above three examples are all special cases of the next theorem.

THEOREM **10.1** *If V_n is a V-statistics of order 2, then*

$$V_{n,Jack} = U_n.$$

That is, jackknifing V_n (of order 2) leads to U_n.

Proof. Recall that $V_{n,Jack} = nV_n - (n-1)V_{n-1}^{(\cdot)}$. We first calculate $V_{n-1}^{(\cdot)}$. Denote $h_{ij} := h(X_i, X_j)$, then

$$V_n = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n h_{ij}.$$

Hence, (easy to see from a diagram)

$$\begin{aligned}
V_{n-1}^{(-k)} &= \frac{1}{(n-1)^2}\left(\sum_{i=1}^n\sum_{j=1}^n h_{ij} - \sum_{i=1}^n h_{ik} - \sum_{j=1}^n h_{kj} + h_{kk}\right) \\
&= \frac{1}{(n-1)^2}\left(\sum_{i=1}^n\sum_{j=1}^n h_{ij} - 2\sum_{j=1}^n h_{kj} + h_{kk}\right) \\
V_{n-1}^{(\cdot)} &= \frac{1}{n}\sum_{k=1}^n V_{n-1}^{(-k)} \\
&= \frac{1}{(n-1)^2}\left(\sum_{i=1}^n\sum_{j=1}^n h_{ij} - \frac{2}{n}\sum_{k=1}^n\sum_{j=1}^n h_{kj} + \frac{1}{n}\sum_{k=1}^n h_{kk}\right) \\
&= \frac{1}{(n-1)^2}\left(\frac{n-2}{n}\sum_{i=1}^n\sum_{j=1}^n h_{ij} + \frac{1}{n}\sum_{k=1}^n h_{kk}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
V_{n,Jack} &= nV_n - (n-1)V_{n-1}^{(\cdot)} \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n h_{ij} - \frac{1}{(n-1)}\left(\frac{n-2}{n}\sum_{i=1}^n\sum_{j=1}^n h_{ij} + \frac{1}{n}\sum_{k=1}^n h_{kk}\right) \\
&= \frac{1}{n}\left(1 - \frac{n-2}{n-1}\right)\sum_{i=1}^n\sum_{j=1}^n h_{ij} - \frac{1}{n(n-1)}\sum_{k=1}^n h_{kk} \\
&= \frac{1}{n(n-1)}\left(\sum_{i=1}^n\sum_{j=1}^n h_{ij} - \sum_{i=j}^n h_{ij}\right) \\
&= \frac{1}{n(n-1)}\sum_{i\neq j}^n h_{ij} \\
&= U_n. \quad \blacksquare
\end{aligned}$$

REMARK **10.4** *The above theorem is NOT true if $m > 2$ in general.*

10.3.2 Use the jackknife to estimate variance for U-statistics

Jackknife's variance estimation for U-statistics has an explicit expression.

Theorem **10.2** *If U_n is a U-statistics of order 2 with kernel $h(x,y)$, then*

$$\widehat{Var}_{Jack}(U_n) = \frac{4(n-1)}{n(n-2)^2}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1,i\neq j}^{n} h_{ij} - U_n\right)^2. \tag{3.1}$$

Furthermore, $\widehat{Var}_{Jack}(U_n)$ is a consistent estimator of $Var(U_n)$ in the sense that

$$\frac{\widehat{Var}_{Jack}(U_n)}{Var(U_n)} \longrightarrow_p 1, \qquad as\ n\to\infty. \tag{3.2}$$

Proof. We only derive (3.1) below (the proof for (3.2) is more involved and hence omitted). From Theorem 10.1, we have

$$\begin{aligned}
U_n &= \frac{1}{n(n-1)}\sum_{i\neq j} h_{ij} \\
U_{n-1}^{(-k)} &= \frac{1}{(n-1)(n-2)}\left(\sum_{i\neq j} h_{ij} - 2\sum_{j=1,j\neq k}^{n} h_{kj}\right) \\
U_{n-1}^{(\cdot)} &= \frac{1}{n}\sum_{k=1}^{n} U_{n-1}^{(-k)} \\
&= \frac{1}{(n-1)(n-2)}\left(\sum_{i\neq j}^{n} h_{ij} - \frac{2}{n}\sum_{k=1}^{n}\sum_{j=1}^{n} h_{kj} + \frac{2}{n}\sum_{k=1}^{n} h_{kk}\right) \\
&= \frac{1}{(n-1)(n-2)}\left(\sum_{i\neq j} h_{ij} - \frac{2}{n}\sum_{i\neq j} h_{ij}\right) \\
&= \frac{1}{(n-1)(n-2)}\left(\sum_{i\neq j} h_{ij} - 2(n-1)U_n\right) \\
\widehat{Var}_{Jack}(U_n) &= \frac{n-1}{n}\sum_{k=1}^{n}\left(U_{n-1}^{(-k)} - U_{n-1}^{(\cdot)}\right)^2 \\
&= \frac{n-1}{n}\frac{1}{(n-1)^2(n-2)^2}\sum_{k=1}^{n}\left(2\sum_{j=1,j\neq k}^{n} h_{kj} - 2(n-1)U_n\right)^2 \\
&= \frac{n-1}{n}\frac{4(n-1)^2}{(n-1)^2(n-2)^2}\sum_{k=1}^{n}\left(\frac{1}{n-1}\sum_{j=1,j\neq k}^{n} h_{kj} - U_n\right)^2 \\
&= \frac{4(n-1)}{n(n-2)^2}\sum_{k=1}^{n}\left(\frac{1}{n-1}\sum_{j=1,j\neq k}^{n} h_{kj} - U_n\right)^2. \quad \blacksquare
\end{aligned}$$

Remark **10.5 Hoeffding-decomposition**

- *The Hoeffding-decomposition (or projection method, or ANOVA decomposition) is*
$$T - ET = \sum_{i=1}^{n} T_i + \sum_{1\le i_1<i_2\le 2}^{n} T_{i_1,i_2} + \cdots + \sum_{1\le i_1<\ldots<i_n\le n}^{n} T_{i_1,\ldots,i_n}.$$
where $T_i = E(T-ET|X_i)$, $T_{i_1,i_2} = E(T-ET|X_{i_1},X_{i_2}) - T_{i_1} - T_{i_2}$, etc.

- *We decompose T into the main effects T_i, 2-way interactions, 3-way interactions, etc. This is where the name "ANOVA decomposition" comes from.*

- **Hoeffding-decomposition** *is in fact an identity, which can be thought as follows. Given T, its overall effect is ET, and we write*
$$T = ET + R_1.$$
From R_1, we take the main effects or contributions from each X_i, given by T_i, resulting in
$$T = ET + \sum_{i=1}^{n} T_i + R_2.$$
From R_2, we take the 2-way interactions from each pair (X_i, X_j) (but without the), given by $T_{i,j}$, resulting in
$$T = ET + \sum_{i=1}^{n} T_i + \sum_{1\le i_1<i_2\le 2}^{n} T_{i_1,i_2} + R_3.$$
We can continue until all the interaction effects have been taken care of, giving the above ANOVE decomposition.

- *Applying the Hoeffding-decomposition to U-statistics, it is easy to check that*
$$U_n - \theta = \frac{2}{n}\sum_{i=1}^{n} g(X_i) + \frac{2}{n(n-1)}\sum_{i<j}\psi(X_i,X_j),$$
where $g(X_i) = E\left(h_{ij} \mid X_i\right) - \theta$ and $\psi(X_i,X_j) = h_{ij} - g(X_i) - g(X_j) + \theta$ and all the terms on the RHS of U_n are uncorrelated.

Proof. *WLOG, assume that $\theta = 0$.*
$$\begin{aligned} T_1 &= E(U_n|X_1) = \frac{1}{n(n-1)}\sum_{i\ne j} E[h(X_i,X_j)|X_1] \\ &= \frac{1}{n(n-1)}2(n-1)E[h(X_1,X_j)|X_1] = \frac{1}{n(n-1)}2(n-1)g_1 \\ &= \frac{2}{n}g_1, \\ E(U_n|X_1,X_2) &= \frac{1}{n(n-1)}\sum_{i\ne j} E[h(X_i,X_j)|X_1,X_2] \\ &= \frac{1}{n(n-1)}\left(h_{12} + 2(n-2)g_1 + 2(n-2)g_2\right) \end{aligned}$$
Hence,
$$\begin{aligned} T_{12} &= E(U_n|X_1,X_2) - E(U_n|X_1) - E(U_n|X_2) \\ &= \frac{1}{n(n-1)}\left(h_{12} + 2(n-2)[g_1+g_2] - 2(n-1)[g_1+g_2]\right) \\ &= \frac{1}{n(n-1)}\left(h_{12} - g_1 - g_2\right) \\ &= \psi(X_1,X_2). \end{aligned}$$

Remark **10.6** *From the above, we see that*

$$Var(U_n) = \frac{4Var[g(X_1)]}{n} + O(n^{-2})$$

which can be approximated by

$$\widetilde{Var}(U_n) \simeq \frac{4}{n}\frac{1}{n}\sum_{i=1}^{n}[g(X_i)]^2 = \frac{4}{n^2}\sum_{i=1}^{n} g^2(X_i).$$

But $g(X_i)$ are not statistics, and can be further approximated by

$$g(X_i) \simeq \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} h_{ij} - U_n.$$

Combining the above, we get an approximation to the variance of U_n:

$$\widehat{Var}(U_n) = \frac{4}{n^2}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1, j\neq i}^{n} h_{ij} - U_n\right)^2.$$

This is almost the same as the jackknife variance estimator $\widehat{Var}_{Jack}(U_n)$. **Therefore, $\widehat{Var}_{Jack}(U_n)$ estimates the variance of the dominating term in the H-decomposition of U_n.**

Remark **10.7** *Asymptotic properties such as the CLT for U-statistics, Berry-Esseen bounds, and so on will be described in a later chapter.*

10.4 Application of the jackknife to the smooth function of means models

Let $X, X_1, ..., X_n$ be i.i.d. k-random vectors with $\mu = EX$ and $\Sigma = Cov(X, X)$. It is known that

$$g(\bar{X}) \longrightarrow_d N\left(g(\mu), \frac{1}{n}\sigma_g^2\right),$$

where

$$\sigma_g^2 = g'(\mu)^\tau \Sigma g'(\mu).$$

Here $g(\cdot) : R^k \to R^1$ is assumed to be differentiable with $g'(\mu) \neq 0$, and $g'(x)$ is the gradient of $g(x)$, given by

$$g'(x)_{k\times 1} = \left(\frac{\partial g(x)}{\partial x_1}, ..., \frac{\partial g(x)}{\partial x_k}\right)^\tau.$$

Let $\bar{X}$ and $\hat{\Sigma} = n^{-1}\sum_{i=1}^n \left(\bar{X}_i - \bar{X}\right)\left(\bar{X}_i - \bar{X}\right)^\tau$ denote sample mean and sample covariance matrix. One obvious estimate of σ_g^2 is the plug-in (i.e., bootstrap) estimator:

$$\hat{\sigma}_g^2 \quad = \quad g'(\bar{X})^\tau \hat{\Sigma} g'(\bar{X}) \longrightarrow_{a.s} g'(\mu)^\tau \Sigma g'(\mu) = \sigma_g^2.$$

However, the formula requires the user to carry out much analytical work, such as calculating $g'(\bar{X})$ etc. By comparion, the jackknife provides a very simple alternative, which requires no differentiation at all.

THEOREM **10.3** *If $g'(x)$ is continuous in a neighborhood of μ and $g'(\mu) \neq 0$. Then, the jackknife variance estimator of $g(\bar{X})$ is strongly consistent, i.e.,*

$$\frac{\widehat{Var}_{Jack}[g(\bar{X})]}{\sigma_g^2/n} = \frac{n\widehat{Var}_{Jack}[g(\bar{X})}{\sigma_g^2} \longrightarrow_{a.s.} 1.$$

Proof. Let $\theta = g(\mu) = T(F)$, and $\hat{\theta} = T(F_n) = g(\bar{X})$. The jackknife estimate of $Var(\hat{\theta})$ is

$$\begin{aligned}
n\widehat{Var}_{Jack}[g(\bar{X})] &= n\widehat{Var}_{Jack}(T_n) \\
&= (n-1)\sum_{i=1}^{n}\left(T_{n-1}^{(-i)} - T_{n-1}^{(\cdot)}\right)^2 \\
&= (n-1)\sum_{i=1}^{n}\left(g(\bar{X}^{(-i)}) - \frac{1}{n}\sum_{i=1}^{n} g(\bar{X}^{(-i)})\right)^2 \\
&= (n-1)\sum_{i=1}^{n}\left([g(\bar{X}^{(-i)}) - g(\bar{X})] - \frac{1}{n}\sum_{i=1}^{n}[g(\bar{X}^{(-i)}) - g(\bar{X})]\right)^2 .
\end{aligned}$$

Now

$$\begin{aligned}
\bar{X}^{(-i)} &= \frac{1}{n-1}\left(X_1 + ... + X_{i-1} + X_{i+1} + ... + X_n\right) = \frac{n\bar{X} - X_i}{n-1}, \\
\bar{X}^{(-i)} - \bar{X} &= \frac{n\bar{X} - X_i - (n-1)\bar{X}}{n-1} = \frac{\bar{X} - X_i}{n-1}, \qquad \frac{1}{n}\sum_{i=1}^{n}\left(\bar{X}^{(-i)} - \bar{X}\right) = 0.
\end{aligned}$$

Furthermore, by the mean-value theorem, we have

$$g(\bar{X}^{(-i)}) - g(\bar{X}) = g'(\xi_i)^\tau\left(\bar{X}^{(-i)} - \bar{X}\right) = g'(\bar{X})^\tau\left(\bar{X}^{(-i)} - \bar{X}\right) + R_i,$$

where ξ_i is a point on the segment between $X^{(-i)}$ and $\bar{X}$ and

$$R_i = [g'(\xi_i) - g'(\bar{X})]^\tau\left(\bar{X}^{(-i)} - \bar{X}\right) = \frac{1}{n-1}[g'(\xi_i) - g'(\bar{X})]^\tau\left(\bar{X} - X_i\right)$$

Therefore,

$$\begin{aligned}
n\widehat{Var}_{n,Jack}(T_n) &= (n-1)\sum_{i=1}^{n}[g'(\bar{X})^\tau\left(\bar{X}^{(-i)} - \bar{X}\right) + (R_i - \bar{R})]^2 \\
&:= A_n + B_n + 2C_n,
\end{aligned}$$

where

$$\begin{aligned}
A_n &= (n-1)g'(\bar{X})^\tau\sum_{i=1}^{n}\left(\bar{X}^{(-i)} - \bar{X}\right)\left(\bar{X}^{(-i)} - \bar{X}\right)^\tau g'(\bar{X})^\tau \\
B_n &= (n-1)\sum_{i=1}^{n}(R_i - \bar{R})^2 \le (n-1)\sum_{i=1}^{n} R_i^2 \\
C_n^2 &\le |A_n|\,|B_n|
\end{aligned}$$

For A_n, since $g'(x)$ is continuous at μ, we have

$$\begin{aligned}
A_n &= \frac{1}{(n-1)}g'(\bar{X})^\tau\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^\tau g'(\bar{X})^\tau \\
&= \frac{n}{(n-1)}g'(\bar{X})^\tau\hat{\Sigma}g'(\bar{X})^\tau \\
&\longrightarrow_{a.s.} \sigma_g^2.
\end{aligned}$$

For B_n, let $u_n = \max_{1\le i\le n} \|g'(\xi_i)^\tau - g'(\bar{X})^\tau\|$, we have

$$\begin{aligned}
B_n &\le \frac{(n-1)}{(n-1)^2}\sum_{i=1}^n \left([g'(\xi_i) - g'(\bar{X})]^\tau\,(\bar{X} - X_i)\right)^2 \\
&\le \frac{1}{(n-1)}\sum_{i=1}^n \left(\|g'(\xi_i) - g'(\bar{X})\|^2\|\bar{X} - X_i\|^2\right) \\
&\qquad \text{(by Cauchy-Schwarze inequality)} \\
&\le \frac{u_n^2}{(n-1)}\sum_{i=1}^n \|\bar{X} - X_i\|^2 \\
&\le \frac{n}{(n-1)}u_n^2 \times (\text{trace of } \hat{\Sigma}).
\end{aligned}$$

Note that (trace of $\hat{\Sigma}$) $\longrightarrow_{a.s.}$ (trace of Σ). Since ξ_i is between $\bar{X}^{(-i)}$ and $\bar{X}$, we have

$$\begin{aligned}
\left(\max_{1\le i\le n}\|\xi_i - \bar{X}\|\right)^2 &= \max_{1\le i\le n}\|\xi_i - \bar{X}\|^2 \le \max_{1\le i\le n}\|\bar{X}^{(-i)} - \bar{X}\|^2 \le \frac{1}{(n-1)^2}\max_{1\le i\le n}\|X_i - \bar{X}\|^2 \\
&\le \frac{1}{(n-1)^2}\sum_{i=1}^n \|X_i - \bar{X}\|^2 \le \frac{n}{(n-1)^2}(\text{trace of } \hat{\Sigma}) \longrightarrow_{a.s.} 0,
\end{aligned}$$

That is, $\max_{1\le i\le n}\|\xi_i - \bar{X}\| \longrightarrow_{a.s.} 0$. This, together with $\bar{X} \longrightarrow_{a.s.} \mu$, implies that $u_n \longrightarrow_{a.s.} 0$. Hence, $B_n \longrightarrow_{a.s.} 0$.

For $C_n^2 \le |A_n|\,|B_n|$, since $A_n \longrightarrow_{a.s.} \sigma_g^2$ and $B_n \longrightarrow_{a.s.} 0$, we have $C_n \longrightarrow_{a.s.} 0$. ∎

Remark **10.8** *Theorem 10.3 implies that* $[g(\bar{X}) - g(\mu)]/\sqrt{\widehat{Var}_{n,Jack}[g(\bar{X})]} \longrightarrow_d N(0,1)$, *which could be used to construct a C.I. for* $g(\mu)$.

Remark **10.9** *From Theorem 10.3,* $\widehat{Var}_{Jack}[g(\bar{X})]$ *is consistent if* $g'(x)$ *is continuous in a neighborhood of* μ. *On the other hand,* $\hat{\sigma}_g^2$ *is consistent if* $g'(x)$ *is continuous only at* μ. *So the jackknife needs stronger condition.*

10.5 Failure of the jackknife

The jackknife is good, but it can fail miserably if the statistic $\hat{\theta}$ is not "smooth" (i.e., small changes in the data can cause big changes in the statistics). One example is the sample median (or quantile), whose influence curve $IC(x, F, T)$ has a jump (i.e., not smooth) near the sample median. Note that the sample median only depends one or two values in the middle of the data.

10.5.1 An example

Consider the following $n = 10$ ordered values

$$10, \quad 27, \quad 31, \quad 38, \quad 40, \quad 50, \quad 52, \quad 104, \quad 146, \quad 150.$$

Then, we get

$$\begin{aligned} T_{n-1}^{(-i)} = &= \begin{cases} 50 & 1 \le i \le 5 \\ 40 & 1 \le 6 \le 10. \end{cases} \\ T_{n-1}^{(\cdot)} &= 45 \\ T_{n-1}^{(-i)} - T_{n-1}^{(\cdot)} &= \begin{cases} 5 & 1 \le i \le 5 \\ -5 & 6 \le i \le 10. \end{cases} \end{aligned}$$

Thus, $\widehat{Var}_{Jack}(T_n)$ tried to estimate the variance from just these two values, 40 and 50, resulting in

$$\widehat{Var}_{Jack}(T_n) \quad = \quad \frac{10-1}{10} \times 10 \times 5^2 = 9 \times 25.$$

Note that the estimate depends solely on two points 40 and 50, which makes it highly unstable. (It is impossible to estimate from one data value, and near impossible to estimate from two data values. The variance estimate should be based on "many" data points.)

The jackknife described so far deletes one observation at a time. If we try to delete 2 observations at a time to the above quantile example, the total number of ways one can do this is $\binom{n}{2} = \binom{10}{2} = 10 \times 9/2 = 45$. You can try to list the pseudo-values here. Clearly, we have more different values now. More generally, one can do delete-d jackknife, to be discussed below.

10.5.2 Main results

THEOREM **10.4** *Let $\theta = F^{-1}(1/2)$ and $T_n = F_n^{-1}(1/2)$ and $n = 2m$. Assume that $f = F'$ exists and continuous at θ, and $f(\theta) > 0$. Then*

1. *The jackknife variance estimate is*

$$\widehat{Var}_{Jack}(T_n) = \frac{n-1}{4}\left(X_{(m+1)} - X_{(m)}\right)^2.$$

(Hence, the jackknife estimates the variance from just two pseudo-values, not a very smart idea.)

2. *The jackknife variance estimate fails to be consistent, i.e.,*

$$\frac{\widehat{Var}_{Jack}(T_n)}{v_n} \longrightarrow_d [exponential(1)]^2 =_d \left(\frac{\chi_2^2}{2}\right)^2 =_d Weibull\left(1, \frac{1}{2}\right),$$

where $v_n = \dfrac{1}{4nf^2(\theta)}$.

(Hence, no matter how large n is, the ratio on the RHS does not go to 1.)

Proof.

1. The order statistics are $X_{(1)}, ..., X_{(m)}, X_{(m+1)}, ..., X_{(n)}$. Now

$$\begin{aligned}
T_{n-1}^{(-i)} &= \begin{cases} X_{(m+1)} & 1 \le i \le m \\ X_{(m)} & 1 \le m+1 \le n. \end{cases} \\
T_{n-1}^{(\cdot)} &= \frac{mX_{(m+1)} + mX_{(m)}}{n} = \frac{X_{(m+1)} + X_{(m)}}{2} \\
T_{n-1}^{(-i)} - T_{n-1}^{(\cdot)} &= \begin{cases} \dfrac{X_{(m+1)} - X_{(m)}}{2} & 1 \le i \le m \\ -\dfrac{X_{(m+1)} - X_{(m)}}{2} & m+1 \le i \le n. \end{cases} \\
\widehat{Var}_{Jack}(T_n) &= \frac{n-1}{n}\sum_{i=1}^{n}\left(T_{n-1}^{(-i)} - T_{n-1}^{(\cdot)}\right)^2 = \frac{n-1}{4}\left(X_{(m+1)} - X_{(m)}\right)^2.
\end{aligned}$$

2. **Proof.** The assumptions in the theorem implies that F has local inverse at θ. WLOG, we might assume that F has inverse everywhere. By Lemma 10.1, we have

$$X_k =: F^{-1}(1 - e^{-Y_k}) := H(Y_k) \sim F, \qquad 1 \le k \le n,$$

where $H(t) = F^{-1}(1 - e^{-t})$, and

$$X_k \sim_{i.i.d.} F, \qquad F(X_k) \sim Uniform[0,1], \qquad Y_k =: -\ln[1 - U_k] \sim Exp(1).$$

Since $H(t) = F^{-1}(1 - e^{-t})$ is increasing in t, from Lemma 10.1, we have

$$X_{(k)} = F^{-1}(1 - e^{-Y_{(k)}}) := H(Y_{(k)}), \qquad 1 \le k \le n,$$

where $Y_{(1)}, ..., Y_{(n)}$ are order statistics of $Y_1, ..., Y_n \sim_{iii} Exp(1)$.

Note that $H'(t) = \dfrac{e^{-t}}{f(H(t))}$. By (5.3), we have $Y_{(m+1)} - Y_{(m)} = Z_m/m = O_p(n^{-1})$, and

$$\begin{aligned}
EY_{(m)} &= \sum_{k=m}^{n} \frac{1}{k} = \sum_{k=1}^{n} \frac{1}{k} - \sum_{k=1}^{m-1} \frac{1}{k} = (C + \ln n + \epsilon_n) - (C + \ln(m-1) + \epsilon_{m-1}) \\
&= \ln\left(\frac{n}{m-1}\right) + (\epsilon_n - \epsilon_{j-1}), \\
&= \ln 2 + o(1) \\
Var(Y_{(m)}) &= \sum_{k=m}^{n} \frac{Var(Z_k)}{k^2} = \frac{1}{m^2} + ... + \frac{1}{n^2} = O(n^{-1}) \\
Y_{(m)} &= EY_{(m)} + O_p(n^{-1/2}) \\
H(EY_{(m)}) &= F^{-1}(e^{-EY_{(m)}}) = F^{-1}(e^{-\ln 2 + o(1)}) = F^{-1}(1/2 + o(1)) = \theta + o(1) \\
H'(EY_{(m)}) &= \frac{e^{-EY_{(m)}}}{f(H(EY_{(m)}))} = \frac{e^{-\ln 2 + o(1)}}{f(\theta + o(1))} = \frac{1}{2f(\theta)} + o(1) \\
H'(Y_{(m)}) &= H'(EY_{(m)}) + o(Y_{(m)} - EY_{(m)}).
\end{aligned}$$

By the mean value theorem,

$$\begin{aligned}
n[X_{(m+1)} - X_{(m)}] &= n[H(Y_{(m+1)}) - H(Y_{(m)})] \\
&= H'\left(Y_{(m)} + \theta[Y_{(m+1)} - Y_{(m)}]\right)\ n[Y_{(m+1)} - Y_{(m)}], \qquad |\theta| \le 1 \\
&= H'\left(\ln 2 + \frac{1}{\sqrt{n}}\sqrt{n}[Y_{(m)} - \ln 2] + \frac{1}{n}\theta n[Y_{(m+1)} - Y_{(m)}]\right)\ n[Y_{(m+1)} - Y_{(m)}] \\
&= H'\left(\ln 2 + o_p(1)\right)\ n[Y_{(m+1)} - Y_{(m)}] \\
&\to_d H'\left(\ln 2\right)\ Exp(1)
\end{aligned}$$

by Slutsky's theorem, continuity of g' and $\sqrt{n}[U_{(m)} - 1/2] \longrightarrow_d N(0, \sigma^2)$. Hence

$$\begin{aligned}
n\widehat{Var}_{Jack}(T_n) &= \frac{n(n-1)}{4}\left(X_{n,m+1} - X_{n,m}\right)^2 \\
&\longrightarrow_d \frac{1}{4}[H'\left(\ln 2\right)]^2\ [Exp(1)]^2. \quad \blacksquare
\end{aligned}$$

10.5.3 Delete-d jackknife

Let $\hat{\theta}$ estimate θ. Let $\hat{\theta}_{(s)}$ denote $\hat{\theta}$ applied to the data with subset s removed. Then the delete-d jackknife variance estimator is

$$\widehat{Var}_{Jack}(T_n)[d] = \frac{n-d}{d\begin{pmatrix} n \\ d \end{pmatrix}} \sum_d \left(\hat{\theta}_{(s)} - \hat{\theta}_{(s\cdot)}\right)^2,$$

where $\sum_d$ is the sum over all subsets s of size $n-d$ chosen without replacement from $\{X_1, ..., X_n\}$ and $\hat{\theta}_{(s\cdot)} = \sum_d \hat{\theta}_{(s)}/\binom{n}{d}$.

THEOREM **10.5** *Under the same conditions of Theorem 10.4, and $\sqrt{n}/d \to 0$ and $n-d \to \infty$, then*

$$\frac{\widehat{Var}_{Jack}(T_n)[d]}{v_n} \longrightarrow_p 1.$$

REMARK **10.10**

1. *The conditions require d to go to ∞, but not too fast. One such choice is $d = n^{2/3}$.*
2. *If n is large, and $\sqrt{n} < d < n$, then $\binom{n}{d}$ is large. In this case, one can use random sampling (or Monte-Carlo simulations) to approximate $\widehat{Var}_{Jack}(T_n)[d]$. If this is too much trouble, one might use the bootstrap in the first place.*
3. *The delete-d jackknife can be shown to be consistent for the median if $\sqrt{n}/d \to 0$ and $n - d \to 0$. Roughly speaking, it is preferable to choose a d such that $\sqrt{n} < d < n$ for the delete-d jackknife estimation of standard error*
4. *d plays the role of smoothing parameter, similar to that of bandwidth in curve estimation.*
5. *The delete-d jackknife can also be used to estimate d.f.'s for t-statistics, for example. But its convergence rate is slower than the bootstrap for the independent data. For dependent data though, the delete-d jackknife works beautifully while "naive" bootstrap or delete-1 jackknife all fails here.*

10.5.4 A useful lemma

LEMMA **10.1** *Let $Y_1, ..., Y_n$ be iid from Exp(1), and $Y_{(1)}, ..., Y_{(n)}$ its order statistics.*

1. *(i)* $e^{-Y_k} \sim Uniform[0, 1]$, *(ii)* $Y_k^2 \sim Weibull(1, 1/2)$.
2. *Let* $Z_i := (n - i + 1)[Y_{(i)} - Y_{(i-1)}]$, $1 \le i \le n.$, *where* $Y_{(0)} = 0$. *Then,*
$$Z_1, ..., Z_n \text{ are i.i.d. from Exp(1).}$$
Consequently, $Y_{(i)}$ *can be represented as*
$$Y_{(i)} = \frac{Z_1}{n} + ... + \frac{Z_i}{n-i+1} = \sum_{k=1}^{i} \frac{Z_k}{n-k+1}, \qquad 1 \le i \le n. \tag{5.3}$$
3. *Let* $U_{(1)}, ..., U_{(n)}$ *be order statistics from* $Uniform(0, 1)$. *Let* $U_{(0)} = 0$, $U_{(n+1)} = 1$. *Then,*
$$V_i := n[U_{(i)} - U_{(i-1)}] \longrightarrow_d Exp(1), \qquad 1 \le i \le n+1.$$

Proof.

1. We prove (ii) only. $f_Y(y) = e^{-y} I\{y > 0\}$. Let $w = y^2$, so $y = \sqrt{w}$, $dy/dw = 1/(2\sqrt{w})$. Hence,
$$f_W(w) = f_Y(x)|dy/dw| = e^{-\sqrt{w}}/(2\sqrt{w}) = \frac{1}{2} w^{-1/2} e^{-w^{1/2}}, \qquad w > 0.$$
(Recall that $Weibull(\alpha, \beta)$ has density $f(w) = \alpha\beta w^{\beta-1} e^{-\alpha w^\beta}$, where $\alpha > 0$, $\beta > 0$, $w > 0$.)
2. The joint pdf of $(Y_{(1)}, ..., Y_{(n)})$ is
$$f_{Y_{(1)},...,Y_{(n)}}(y_1, ..., y_n) = n! f_Y(y_1)...f_Y(y_n) = n! e^{-y_1}...e^{-y_n}, \quad \textit{where } y_1 < ... < y_n.$$
Let
$$\begin{aligned} Z_1 &= nY_{(1)} \\ Z_2 &= (n-1)[Y_{(2)} - Y_{(1)}] \\ Z_3 &= (n-2)[Y_{(3)} - Y_{(2)}] \\ ... & \quad ...... \\ Z_n &= 1 \times [Y_{(n)} - Y_{(n-1)}], \end{aligned} \tag{5.4}$$
whose jacobian is $\left|\frac{dz}{dy}\right| = n!$. Therefore,
$$f_{Z_1,...,Z_n}(z_1, ..., z_n) = f_{Y_{(1)},...,Y_{(n)}}(y_1, ..., y_n) \left|\frac{dy}{dz}\right| = n! e^{-y_1 - ... - y_n} \frac{1}{n!} = e^{-z_1 - ... - z_n} = e^{-z_1}...e^{-z_n},$$
which implies that $Z_1, ..., Z_n$ are i.i.d. Exp(1) r.v.s.
The second part follows from the transformation in (5.4):
$$\begin{aligned} Y_{(1)} &= \frac{Z_1}{n} \\ Y_{(2)} &= \frac{Z_1}{n} + \frac{Z_2}{n-1} \\ ... & \quad ...... \\ Y_{(n)} &= \frac{Z_1}{n} + \frac{Z_2}{n-1} + ... + \frac{Z_n}{1}. \end{aligned}$$

3. The joint d.f. of $(U_{(k)}, U_{(k+1)})$ is

$$f_{U_{(k)},U_{(k+1)}}(u,v) = \frac{n!}{(k-1)!(n-k-1)!}u^{k-1}[1-v]^{n-k-1}, \qquad\qquad 0 \le u < v \le 1.$$

Let

$$\begin{aligned} X &= U_{(k)} \\ Y &= n[U_{(k+1)} - U_{(k)}]. \end{aligned}$$

So

$$\begin{aligned} U_{(k)} &= X \\ U_{(k+1)} &= X + \frac{Y}{n}. \end{aligned}$$

Therefore,

$$\begin{aligned} f_{(X,Y)}(x,y) &= f_{(U_{(k)},U_{(k+1)})}(u,v)\left|\frac{dU}{dY}\right| \\ &= \frac{(n-1)!}{(k-1)!(n-k-1)!}u^{k-1}[1-v]^{n-k-1} \\ &= \frac{(n-1)!}{(k-1)!(n-k-1)!}x^{k-1}[1-y/n-x]^{n-k-1}, \quad \text{where } 0 \le x \le 1-y/n. \end{aligned}$$

(Note that $0 \le u < v \le 1 \iff 0 \le x < x+y/n \le 1 \iff 0 \le x \le 1-y/n$.) Thus,

$$\begin{aligned} f_Y(y) &= \frac{(n-1)!}{(k-1)!(n-k-1)!}\int_0^{1-y/n} x^{k-1}[1-y/n-x]^{n-k-1}dx \\ &= \left(1-\frac{y}{n}\right)^{n-1}\frac{(n-1)!}{(k-1)!(n-k-1)!}\int_0^1 t^{k-1}\left(1-t\right)^{n-k-1}dt, \quad \text{where } x = t\left(1-\frac{y}{n}\right) \\ &= \left(1-\frac{y}{n}\right)^{n-1} \\ &\longrightarrow e^{-y}, \quad y > 0. \ \blacksquare \end{aligned}$$

10.6 Application of jackknife in a real example

In the study of income shares, sometimes we need to estimate the poverty line (or low income cutoff point). Consider the model

$$\log Z_i = \gamma_1 + \gamma_2 \log Y_i + \sum_{t=1}^{m} \beta_t x_{it} + \epsilon_i, \qquad i = 1, ..., n,$$

where

Z_i = expenditure on "necessities",

Y_i = total income,

x_{it} = urbanization category, family size, etc.

ϵ_i = independent errors (also assumed to be independent of all other covariates)

Let γ_0 be the overall proportion of income spent on "necessities". Then the poverty line θ can be defined to be the solution of

$$\log[(\gamma_0 + 0.2)\theta] = \gamma_1 + \gamma_2 \log \theta + \sum_{t=1}^{m} \beta_t x_{0t}$$

for a particular set $x_{01}, ..., x_{0m}$. Our purpose is to estimate θ and construct a $(1-\alpha)$ confidence interval for θ.

Solution. First of all, it is reasonable to estimate γ_0 by its sample counterpart:

$$\hat{\gamma}_0 = \frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n Y_i} = \frac{\bar{Z}}{\bar{Y}}.$$

On the other hand, γ_1, γ_2, β_t's can be estimated by the LSE, denoted by $\hat{\gamma}_1, \hat{\gamma}_2$, $\hat{\beta}_t$'s. As a consequence, θ can be estimated (implicitly) from

$$\log[(\hat{\gamma}_0 + 0.2)\hat{\theta}] = \hat{\gamma}_1 + \hat{\gamma}_2 \log \hat{\theta} + \sum_{t=1}^{m} \hat{\beta}_t x_{0t},$$

which can be solved numerically by standard software.

Note that

$$\hat{\theta} = g\left(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2, \hat{\beta}_1, ..., \hat{\beta}_m\right) := g(\bar{X}),$$

where

$$X_i = \left(Z_i, Y_i, \log Z_i, \log Y_i, x_{i1}, ..., x_{im}, \text{all other cross product terms}\right).$$

Therefore, we can express $\hat{\theta}$ into a very smooth function of multivariate sample means of independent (not necessarily identically distributed) r.v.'s. Hence, $\hat{\theta}$ is asymptotically normal. We need to get a variance estimate.

If one uses the traditional variance estimator, we have

$$v_n = \bigtriangledown g(\bar{X})^\tau \widehat{Var}(\bar{X}) \bigtriangledown g(\bar{X}).$$

Since g is not explicit, it is very tedious to calculate $\bigtriangledown g(\bar{X})$ in this case. On the other hand, the jackknife can provide an easy way to get a consistent variance estimator:

- Step I: Run OLS.
- Step II: Apply Newton algorithm to solve $\hat{\theta}_{(i)}$.
- Step III: Use Jackknife estimation of variance. $\blacksquare$

10.7 A quick summary

1. The jackknife replaces the theoretical derivations by numerical computations. Asymptotically, it is equivalent to the traditional approach.

2. The jackknife variance estimator $\hat{v}_{Jack}$ is consistent for many types of statistics under mild conditions, such as sample means, function of sample means, U-, V-, L-statistics, M-estimates, etc.

3. If delete-1 jackknife fails, one can try delete-d jackknife, or bootstrap.

4. The jackknife can be easily extended to multivariate case.

10.8 Exercises

1. Let $X_1, \ldots, X_n$ be i.i.d with variance σ^2. Take $\widehat{\theta} = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ as an estimator of σ^2.

   (a) Show that $\widehat{\theta}$ has bias equal to $-\sigma^2/n$.

   (b) Find $\widehat{\theta}_J$, the jackknife estimator of σ^2.

   (b) Find $\widehat{v}_J(\widehat{\theta})$, the jackknife estimator of $Var(\widehat{\theta})$.

2. Suppose $X_1, \ldots, X_n$ are i.i.d with mean μ. We take $\widehat{\theta} = (\bar{X})^3$ as an estimator of μ^3.

   (a) Find $\widehat{\theta}_J$, the jackknife estimator of μ^3.

   (b) Show that $\widehat{\theta} = (\bar{X})^3$ is a V-statistic of order 3 with the kernel $h(x, y, z) = xyz$.

   (c) Derive the U-statistic of order 3 with the kernel $h(x, y, z) = xyz$.

   (d) From (a)-(c), check whether the jackknifed V-statistic in (b) produces the U-statistic in (c).

3. (Use a package to do this question.) The following data (X_i, Y_i) are English scores of 6 people before and after attending an intensive study program.

   (80, 90), (85, 78), (65, 75), (88, 92), (77, 86), (95,90).

   (a) Calculate $\hat{\theta}_1 = \bar{X}/\bar{Y}$ and $\hat{\theta}_2 = \hat{\rho}$ (sample correlation coefficient).

   (b) Find $\widehat{bias}(\hat{\theta})$, $\widehat{\theta}_J$ and $\widehat{v}_J(\hat{\theta})$ for $\hat{\theta} = \hat{\theta}_1$ and $\hat{\theta}_2$. Hence find a 95% confidence intervals for $\theta_1 = \mu_1/\mu_2$ and $\theta_2 = \rho$.

   (c) Assuming the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. The parameter of interest here is $\theta_0 = -\beta_0/\beta_1$. Take $\hat{\theta}_0 = -\hat{\beta}_0/\hat{\beta}_1$ where $(\hat{\beta}_0, \hat{\beta}_1)$ are LSE of (β_0, β_1). Use the jackknife to find the variance estimator of $\hat{\theta}_0$ and hence find a 95% confidence interval for θ_0.

   (d) [**Do this after the chapter on the bootstrap.**] Repeat these calculations by using the bootstrap method and compare the results by the two methods.

4. Show that for linear statistics, the jackknife and bootstrap estimates of bias are 0. (Note that $\theta = T(F) = \int \psi dF$ is linear with $T_n = T(F_n)$ is the corresponding linear statistics.)

5. (a) Given n distinct data items, show that the probability that a given data item does not appear in a bootstrap resample is $e_n = (1 - 1/n)^n \to e^{-1} \simeq 0.368$ as $n \to \infty$.

   (b) Show that the probability that each of B bootstrap samples contains an item i is $(1 - e^n)^B$.

6. Let $X_1, ..., X_n \sim_{i.i.d.}$ from F. Let $\theta = T(F)$ and $T_n = T(F_n)$. Suppose that

   $$\sqrt{n}\,(T_n - \theta) \longrightarrow_d N(0, \sigma_F^2),$$

   where $\sigma_{T_n}^2 = E_F\psi_F^2(X) = EIC(X, F, T_n)^2$. An estimate of $\sigma_{T_n}^2$ can be given by

   $$\widehat{\sigma}_{T_n}^2 = \frac{1}{n-1}\sum_{i=1}^{n} IC(X, F_n, T_n)^2.$$

   The factor $n - 1$ (instead of n) was substituted to preserve equivalence with the classical formula for the sample mean. Check if this variance estimate is consistent in the case of the sample median, $T_n = F_n^{-1}(1/2)$.

7. Let $\theta = Var(X)$, and $T_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Show that $\widehat{Var}_{Jack} = \frac{n^2}{(n-1)(n-2)^2}(\hat{\mu}_4 - \hat{\mu}_2^2)$. Recall that the delta method gives $\widehat{Var}_{delta} = \frac{1}{n}(\hat{\mu}_4 - \hat{\mu}_2^2)$.

8. Let $X_1, ..., X_n \sim_{i.i.d.}$ from F. Let $\theta = \mu^2$ and $T_n = (\bar{X})^2$. Let $\hat{\mu}_j$ be the jth sample central moments. Show that the variance estimators using jackknife and bootstrap are

$$\widehat{Var}_{Jack}(T_n) = \frac{4\bar{X}^2\hat{\mu}_2}{n-1} - \frac{4\bar{X}\hat{\mu}_3}{(n-1)^2} + \frac{\hat{\mu}_4 - \hat{\mu}_2^2}{(n-1)^3},$$

$$\widehat{Var}_{Boot}(T_n) = \frac{4\bar{X}^2\hat{\mu}_2^2}{n} + \frac{4\bar{X}\hat{\mu}_3}{n^2} + \frac{\hat{\mu}_4 - \hat{\mu}_2^2}{n^3}.$$

Remark: The asymptotic variance of T_n *is* $\frac{4\mu^2\mu_2}{n}$ *whose plug-in estimate is* $\frac{4\bar{X}^2\hat{\mu}_2}{n}$. *Therefore, all three approaches are asymptotically equivalent in the sense that the ratio goes to 1 a.s.*

9. Let $X_1, ..., X_n \sim_{i.i.d.}$ from F. Let $\theta = T(F)$ and $T_n = T(F_n)$. Suppose that

$$\sqrt{n}\,(T_n - \theta) \longrightarrow_d N(0, \sigma_F^2),$$

where $\sigma_{T_n}^2 = E_F\psi_F^2(X) = \mathbf{E}IC(X, F, T_n)^2$. An estimate of $\sigma_{T_n}^2$ can be given by

$$\widehat{\sigma}_{T_n}^2 = \frac{1}{n}\sum_{i=1}^{n} IC(X, F_n, T_n)^2.$$

Check if this variance estimate is consistent in the case of the sample median, $T_n = F_n^{-1}(1/2)$.

10. Let $g(x,t)$ be continuous at t_0 uniformly in x. Let F be a distribution function for which $E_F|g(X, t_0)| < \infty$. Let $\{X_i\}$ be i.i.d. with d.f. F and suppose that $T_n \longrightarrow_p t_0$. Then,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i, T_n) \longrightarrow_p \; E_F g(X, t_0).$$

Further, the convergence in probability can be replaced by a.s. convergence throughout.

Chapter 11

Bootstrap

The bootstrap was first introduced by Efron (1979). It is a powerful nonparametric technique, like the jackknife. In a sense, jackknife can be considered as a first-order approximation to the bootstrap.

11.1 The bootstrap (or plug-in, or substitution method)

Suppose

- $X_1, \ldots, X_n \sim_{iid} F$ (typically unknown),
- $\theta \equiv \theta(F) = T(F)$ is the parameter of interest,
- An estimator of θ is
$$T_n \equiv T_n(X_1, \ldots, X_n) \equiv T_n(F_n).$$

Again, we are interested in

- the bias of T_n: $Bias(T_n) = E(T_n) - \theta = E_F(T(X_1, \ldots, X_n)) - T(F)$.
- the variance of T_n: $Var(T_n) = E\left(T_n - E(T_n)\right)^2 = E_F\left(E_F(T(X_1, \ldots, X_n)) - T(F)\right)^2$.
- the sampling distribution of
$$H(x) = P\left(\frac{T_n - \theta}{\sqrt{Var(T_n)}} \le x\right), \qquad or \qquad K(x) = P\left(\frac{T_n - \theta}{\sqrt{\widehat{Var}(T_n)}} \le x\right)?$$

Let $X_1^*, ..., X_n^* \sim_{i.i.d.} F_n$. Then, the bootstrap approximations of the above quantities are

(1) $\widehat{Bias}_{Boot}(T_n) = E_{F_n}[T(X_1^*, ..., X_n^*) - T(F_n)]$.
(2) $\widehat{Var}_{Boot}(T_n) = E_{F_n}\left(T(X_1^*, ..., X_n^*) - E_{F_n}T(X_1^*, ..., X_n^*)\right)^2$
(3) $\widehat{H}_{Boot}(x) = P_{F_n}\left(\frac{\theta(F_n^*) - \theta(F_n)}{\sigma_n(F_n)} \le x\right)$, and $\widehat{K}_{Boot}(x) = P_{F_n}\left(\frac{\theta(F_n^*) - \theta(F_n)}{\sigma_n(F_n^*)} \le x\right)$?

Example Let $\theta = \mu^2$ and $T = (\bar{X})^2$. Find the bootstrap estimate of the bias.

Solution. Let $X_1^*, \cdots, X_n^* \sim_{i.i.d.} F_n$,

$$\begin{aligned} E_{F_n}\left(\bar{X}^*\right)^2 &= V_{F_n}\left(\bar{X}^*\right) + \left[E_{F_n}\left(\bar{X}^*\right)\right]^2 \\ &= \frac{V_{F_n}\left(X_1^*\right)}{n} + \left(EX_1^*\right)^2 \\ &= \frac{1}{n}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + \left(\bar{X}\right)^2 \\ &= \frac{1}{n}S_n^2 + \left(\bar{X}\right)^2. \end{aligned}$$

Hence, the bootstrap bias estimate is

$$\widehat{Bias}_{Boot}\left[\left(\bar{X}\right)^2\right] = E_{F_n}\left(\bar{X}^*\right)^2 - \left(\bar{X}\right)^2 = \frac{1}{n}S_n^2.$$

Recall that the true bias is

$$Bias\left[\left(\bar{X}\right)^2\right] = E_F\left(\bar{X}\right)^2 - \mu^2 = \frac{\sigma^2}{n} + \mu^2 - \mu^2 = \frac{\sigma^2}{n}. \quad \blacksquare$$

Definition. Let $\eta = E_F\ T(X_1, \ldots, X_n, F)$, where $X_1, \ldots, X_n \sim_{iid} F$. Its bootstrap approximation is

$$\hat{\eta}_{Boot} = E_{F_n}\ T(X_1^*, ..., X_n^*, F_n), \quad \text{where } X_1^*, ..., X_n^* \sim_{iid} F_n.$$

That is, the generalized function η is approximated by the generalized statistical functional $\hat{\eta}_{Boot}$. $\blacksquare$

11.2 Monte Carlo simulations

Generally speaking, there is no closed form expression for the bootstrap estimate $\hat{\eta}_{Boot}$. On the other hand, for each k, we note

$$P\left(X_k^* = X_j\right) = \frac{1}{n}, \qquad j = 1, \cdots, n.$$

Hence, one can find the exact value $\hat{\eta}_{Boot}$ by exhausting all possible bootstrap samples (all together n^n) and calculate

$$\hat{\eta}_{Boot} = \frac{1}{n^n} \sum_{k=1}^{n^n} T(X_1^*(k), ..., X_n^*(k), F_n)$$

If T is symmetric in $X_1, \ldots, X_n$, then the total number of different samples is $\binom{2n-1}{n}$ (Why?). If $n = 10$, then $n^n = 10,000,000,000$, and $\binom{2n-1}{n} = \binom{19}{10}$, a huge number indeed. Therefore, the exact calculation is not practical.

One alternative is to resort to Monte Carlo simulations:

1. Draw B independent random samples $\mathcal{X}_k^* = \{X_1^*(k), ..., X_n^*(k)\}$ from F_n, $k = 1, ..., B$. (This can be done by drawing from F_n with replacement).
2. The Monte Carlo approximation to $\hat{\eta}_{Boot}$ is

   $$\hat{\eta}_{MC} = \frac{1}{B} \sum_{k=1}^{B} T(X_1^*(k), ..., X_n^*(k), F_n).$$

   By the SLLN, we have $\hat{\eta}_{MC} \to \hat{\eta}_{Boot}$ a.s. as $B \to \infty$.

11.3 Application of the bootstrap to quantile

Closed expressions

Example. Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s from d.f. F and F_n be its e.d.f. Let $\xi_p = F^{-1}(p) = \inf\{t : F(t) \geq p\}$ and

$$\begin{aligned}\xi_p &= F^{-1}(p) = \inf\{t : F(t) \geq p\} \\ \hat{\xi}_p &= F_n^{-1}(p) = \inf\{t : nF_n(t) \geq np\} := X_{(r)},\end{aligned}$$

where $r = \lceil np \rceil$, giving the smallest integer $\geq np$ (namely $= [np]$ if integer, $[np] + 1$ otherwise). Find the bootstrap estimates of

(1) the mean $\mu_\xi = E_F(\hat{\xi}_p)$, and the bias $bias_\xi = E_F(\hat{\xi}_p - \xi_p)$

(2) the variance $\sigma_\xi^2 = Var_F(\hat{\xi}_p)$, and the Mean Square Error $MSE_\xi = E_F(\hat{\xi}_p - \xi_p)^2$

(3) the d.f.'s of standardized and studentized sample quantiles

$$H(t) = P_F\left(\frac{\sqrt{n}(\hat{\xi}_p - \xi_p)}{\sigma_\xi} \leq t\right), \qquad K(t) = P_F\left(\frac{\sqrt{n}(\hat{\xi}_p - \xi_p)}{\hat{\sigma}_\xi} \leq t\right),$$

where $\sigma_\xi^2 = \sigma_\xi^2(F) = Var_F(\hat{\xi}_p)$ and $\hat{\sigma}_\xi^2$ is an consistent estimate of σ_ξ^2.

Solution. Let $F_{(r)}(x) = P(X_{(r)} \leq x)$, then

$$\begin{aligned}dF_{(r)}(x) &= \frac{n!}{(r-1)!1!(n-r)!}F^{r-1}(x)[1-F(x)]^{n-r}dF(x) \\ &= r\binom{n}{r}F^{r-1}(x)[1-F(x)]^{n-r}dF(x).\end{aligned}$$

Note that $X_{(r)} =_d F^{-1}(U_{(r)})$, where $U_{(r)}$ is the rth order statistic from $U(0,1)$ with pdf given by

$$dG_{(r)}(u) = r\binom{n}{r}u^{r-1}[1-u]^{n-r}, \qquad 0 < x < 1.$$

Therefore,

$$\begin{aligned}E_F(\hat{\xi}_p - c)^m &= E[X_{(r)} - c]^m = \int (x-c)^m dF_{(r)}(x) \\ &= \int_{-\infty}^{\infty} (x-c)^m r\binom{n}{r}F^{r-1}(x)[1-F(x)]^{n-r}dF(x) \\ &= r\binom{n}{r}\int_0^1 [F^{-1}(u) - c]^m u^{r-1}[1-u]^{n-r}du \\ &= E[F^{-1}(U_{(r)}) - c]^m.\end{aligned}$$

For the bootstrap,

$$\hat{\xi}_p^* = F_n^{*-1}(p) = \inf\{t : nF_n^*(t) \geq np\} := X_{(r)}^*,$$

Therefore, the bootstrap estimate of $E_F(\hat{\xi}_p)^m$ is given by

$$\begin{aligned}
E_{F_n}(\hat{\xi}_p^* - c)^m &= E[X_{(r)}^* - c]^m = r\binom{n}{r}\int_0^1 [F_n^{-1}(u) - c]^m u^{r-1}[1-u]^{n-r}du \\
&= r\binom{n}{r}\sum_{k=1}^n \int_{(\frac{k-1}{n},\frac{k}{n}]} [F_n^{-1}(u) - c]^m u^{r-1}[1-u]^{n-r}du \\
&= r\binom{n}{r}\sum_{k=1}^n \int_{(\frac{k-1}{n},\frac{k}{n}]} [X_{(k)} - c]^m u^{r-1}[1-u]^{n-r}du \\
&= \sum_{k=1}^n \left(r\binom{n}{r} \int_{(\frac{k-1}{n},\frac{k}{n}]} u^{r-1}[1-u]^{n-r}du \right) [X_{(k)} - c]^m \\
&= \sum_{k=1}^n w_{nk}[X_{(k)} - c]^m,
\end{aligned}$$

where

$$w_{nk} = r\binom{n}{r}\int_{(\frac{k-1}{n},\frac{k}{n}]} u^{r-1}[1-u]^{n-r}du, \qquad k = 1, ..., n.$$

(1) the bootstrap estimates of the mean and bias are

$$\begin{aligned}
\hat{\mu}_\xi &= E_{F_n}(\hat{\xi}_p^*) = \sum_{k=1}^n w_{nk}X_{(k)}, \\
\hat{bias}_\xi &= \widehat{bias}_{Boot}(\hat{\xi}_p) = E_{F_n}(\hat{\xi}_p^* - X_{(r)}) = \sum_{k=1}^n w_{nk}\left(X_{(k)} - X_{(r)}\right).
\end{aligned}$$

(2) the bootstrap estimates of the variance and MSE are

$$\begin{aligned}
\hat{\sigma}_\xi^2 &= \widehat{var}_{Boot}(\hat{\xi}_p) = \sum_{k=1}^n w_{nk}X_{(k)}^2 - \left(\sum_{k=1}^n w_{nk}X_{(k)}\right)^2, \\
\widehat{MSE}_{Boot}(\hat{\xi}_p) &= E_{F_n}(\hat{\xi}_p^* - \hat{\xi}_p)^2 = \sum_{k=1}^n w_{nk}[X_{(k)} - X_{(r)}]^2.
\end{aligned}$$

(3) First we note that

$$\begin{aligned}
H(t) &= P\left(\hat{\xi}_p \le s\right) = P\left(X_{(r)} \le s\right) \quad \text{where } s = \xi_p + t\sigma_\xi/\sqrt{n} \\
&= P\left(\{\text{ \# of } X_i \le s \ \} \ge r\right) \\
&= \sum_{k=r}^n \binom{n}{k} F^k(s)[1 - F(s)]^{n-k}.
\end{aligned}$$

Thus, its bootstrap approximation is given by

$$\hat{H}_{Boot}(t) = P_{F_n}\left(\frac{\hat{\xi}_p^* - \hat{\xi}_p}{\sigma_\xi^2(F_n)} \le t\right) = \sum_{k=r}^n \binom{n}{k} F_n^k(s)[1 - F_n(s)]^{n-k}.$$

Comparisons with other methods

(a) the kernel estimate $\dfrac{p(1-p)}{\hat{f}^2(\hat{\xi}_p)}$, where $\hat{f}(x) = \dfrac{1}{nh}\sum_{i=1}^{n} K\left(\dfrac{x - X_i}{h}\right)$.

(b) the bootstrap variance estimate $\hat{\sigma}^2_{\xi,boot} = \sigma^2_{\xi}(F_n) = \sum_{k=1}^{n} w_{nk} X^2_{(k)} - \left(\sum_{k=1}^{n} w_{nk} X_{(k)}\right)^2$.

(c) Recall that delete-1 jackknife variance estimate is inconsistent.

For illustration, we take the bootstrap variance estimate given by

$$\hat{K}_{Boot}(t) = P_{F_n}\left(\frac{\hat{\xi}^*_p - \hat{\xi}_p}{\sigma^2_{\xi}(F^*_n)} \le t\right)$$

where

$$\sigma^2_{\xi}(F^*_n) = \sum_{k=1}^{n} w_{nk} {X^*_{(k)}}^2 - \left(\sum_{k=1}^{n} w_{nk} X^*_{(k)}\right)^2.$$

Note that there is no close form expression for $\hat{K}_{Boot}(t)$. Monte Carlo simulations have be carried out here. ∎

Remark **11.1** *For more materials see Hall (1992), page 319.*

Remark **11.2** *Bootstrap approximations of the mean, bias, and variance are all weighted averages of the ordered statistics. Note that $u^{r-1}[1-u]^{n-r}$ is maximized at $u = (r-1)/(n-1) \approx p$. Hence, all the bootstrap estimates place more weights around $\hat{\theta} = F_n^{-1}(p) = X_{(r)}$, and less weights if the data are far away from $\hat{\theta}$, thus producing a consistent estimator. By comparison, the jackknife variance estimator, $\widehat{Var}_{Jack}(\hat{\theta})$ only uses two values near $\hat{\theta}$ to approximate the variance, which fails to be consistent.*

11.4 Bootstrap approximations to d.f.'s

Definition (k-th order consistency (or accuracy) of the bootstrap). Let $R_n = R(X_1, \dots, X_n, F)$, $R_n^* = R(X_1^*, ..., X_n^*, F_n)$. Let $H(x) = P_F(R_n \le x)$ and $\hat{H}_B(x) = P_{F_n}(R_n^* \le x)$. Then, $\hat{H}_B(x)$ is said to be k-th order strongly or weakly consistenct or accurate if

$$\sup_x \left|\hat{H}_B(x) - H(x)\right| = o(n^{-(k-1)/2}) \quad \textit{or} \quad O(n^{-k/2}), \quad \textit{a.s.} \quad \textit{or} \quad \textit{in prob.}$$

In particular,

(a) $\hat{H}_B(x)$ is (first-order) consistent if $\sup_x \left|\hat{H}_B(x) - H(x)\right| = o(1) \quad \textit{or} \quad O(n^{-1/2})$.

(b) $\hat{H}_B(x)$ is second-order accurate if $\sup_x \left|\hat{H}_B(x) - H(x)\right| = o(n^{-1/2}) \quad \textit{or} \quad O(n^{-1})$.

(c) $\hat{H}_B(x)$ is third-order accurate if $\sup_x \left|\hat{H}_B(x) - H(x)\right| = o(n^{-1}) \quad \textit{or} \quad O(n^{-3/2})$.

11.4.1 First- and second-order accurate or consistent

Under certain regularity conditions, the bootstrap provides second-order accurate approximation in a number of familiar situations:

1. smooth functions of means
2. U-statistics
3. L-statistics (including quantiles)
4. symmetric statistics

By comparison, normal approximations only provide first-order accurate approximations.

11.5 Exercises

1. Consider an artificial data set consisting of the 8 numbers

$$1, \quad 2, \quad 3.5, \quad 4, \quad 7, \quad 7.3, \quad 8.6, \quad 12.4, \quad 13.8, \quad 18.1.$$

Define the α trimmed mean by $\hat{\theta} = \dfrac{1}{n-2[n\alpha]} \displaystyle\sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{n,i}$.

(a) Take $\alpha = 25\%$, calculate the bootstrap variance estimate $\hat{v}_B(\hat{\theta})$ by Monte-Carlo simulations for $B = 20, 100, 200, 500, 1000, 2000$. Do they become more stable as B gets larger?

(b) Find the jackknife variance estimate $\hat{v}_J(\hat{\theta})$.

2. (Generation of bivariate normal random variables.) Suppose that we would like to generate a bivariate normal random vector $(X, Y)^T$ with mean $(\mu_x, \mu_y)^T$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

The following procedure could be used.

Step 1. Generate R_1 and R_2, where $R_i \sim N(0,1)$ for $i = 1, 2$, and R_1 and R_2 are independent.

Step 2. Calculate

$$X = \mu_x + \sigma_x R_1 \qquad Y = \mu_y + \frac{\sigma_y}{\sqrt{1+c^2}}(R_1 + cR_2)$$

where $c = \sqrt{\rho^{-2} - 1}$.

Show that $(X, Y)^T$ have the required bivariate normal distribution.

3. Show that for function of means, one-sided (lower) confidence interval by the hybrid method is only first-order accurate.

4. The following are the numbers of minutes that a doctor kept 10 patients waiting beyond their appointment times,

$$12.1, \quad 9.8, \quad 10.5, \quad 5.6, \quad 8.2, \quad 0.5, \quad 6.8, \quad 10.1, \quad 17.2, \quad 4.2.$$

find a two-sided 90% confidence interval for the mean waiting time by the normal approximation, percentile-t, percentile, ABC method.

5. The following data are the number of hours 10 students studied for a certain achievement test and their scores on the test.

Number of hours (x): 8, 6, 10, 9, 14, 6, 18, 10, 15, 24,

Score on the test (Y): 80, 62, 91, 77, 89, 65, 96, 85, 94, 91.

(i) Plot Y against x.

(ii) Fit a simple linear regression.

(iii) Find a two-sided 90% confidence interval for the slope parameter by the normal approximation method and hybrid method.

6. "Are statistical tables obsolete?" In this example, we shall use Monte Carlo simulations to obtain quantiles from any given distribution. We shall illustrate this with the standard normal distribution $\Phi(x)$. Suppose we wish to find the 90% and 95% quantile, i.e. $q_1 = \Phi^{-1}(0.90)$ and $q_2 = \Phi^{-1}(0.95)$.

   (1) Generate B random numbers, $X_1, \cdots, X_B$, from the standard normal distribution, where B can be arbitrarily big, say 10,000.

   (2) Plot the histogram for the generated data. Is it belled shaped?

   (3) Rank the random numbers in increasing order, say, $X_{B,1}, \cdots, X_{B,B}$. So the Monte Carlo approximations for q_1 and q_2 are

   $$q_1^{MC} = X_{B,[B*0.90]}, \qquad\qquad q_1^{MC} = X_{B,[B*0.95]}.$$

   (4) Compare and comment on those Monte Carlo approximations and the exact values.

   (5) Let $X \sim N(0,1)$. Suppose that we are interested in calculate $P(X \geq 2)$, $\mu = EX$ and $\sigma^2 = var(X)$. How do we use the Monte Carlo simulations to approximate these? Compare your approximations with the exact ones.

Chapter 12

Empirical likelihood method

The empirical likelihood (EL) method was first introduced by Owen (1988, 1990) as a nonparametric approach to statistical inference, e.g., in doing hypothesis testing and constructing confidence regions. Hall and La Scala (1990) has summarized some advantages of the empirical likelihood (over its competitors such as the bootstrap):

- the shape of confidence regions "automatically" determined by the sample,
- Bartlett correctable,
- range preserving,
- transformation respecting.

For these reasons, the empirical likelihood has found many applications such as in smooth functions of means, in nonparametric density and regression function estimation, in quantile related estimation, and so on. For a more thorough review of the empirical likelihood method and its applications, the reader is referred to the recent monograph by Owen (2001).

12.1 Nonparametric MLE (NMLE) or Maximimum EL estimates (MELE)

Let $X_1, ..., X_n$ be a random sample from a d.f. $F \in \mathcal{F}$, the set of all d.f.'s. Define the nonparametric likelihood function to be

$$L(G) = \prod_{i=1}^{n} P_G(\{X_i\}), \qquad G \in \mathcal{F}.$$

To maximize $L(G)$ over $\mathcal{F}$, it suffices to consider those distributions which have non-zero probabilities at all $\{X_i\}$'s. Denote $p_i = P_G(\{X_i\})$, $i = 1, ..., n$.

THEOREM **12.1** *Let $X_1, ..., X_n$ be i.i.d. with $F \in \mathcal{F}$. Then, the empirical d.f. F_n maximize $L(G)$ over $G \in \mathcal{F}$. That is, F_n is an NMLE or MELE.*

Proof. We only need to consider $G \in \mathcal{F}$ such that $L(G) > 0$. Let $c \in (0, 1]$ and $\mathcal{F}(c) \subset \mathcal{F}$ containing G's satisfying $p_i > 0$ and $\sum p_i = c$ (the rest of the mass may be distributed somewhere else). Then, by the Lagrange multiplier method, the empirical likelihood

$$L(\theta) = \max_{\sum p_i = c} \prod_{i=1}^{n} p_i$$

attains its maximum $\left(\frac{c}{n}\right)^n$ at $p_i = \frac{c}{n}$. Maximizing over $c \in (0, 1]$, we get the $p_i = \frac{1}{n}$. ∎

12.2 EL for the univariate mean

We shall illustrate the idea of the empirical likelihood by a simple example of the sample mean.

Let $X_1, ..., X_n$ be a random sample from a d.f. F (typically unkonwn). Suppose that we are interested in the mean functional $\theta = \theta(F) = \int x dF(x)$. Consider the following hypothesis test:

$$H_0 : \theta = \theta_0 \qquad vs \qquad H_0 : \theta \neq \theta_0.$$

Let $p = (p_1, \cdots, p_n)$ be a probability vector, i.e., $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$ for $1 \leq i \leq n$. Let G_p be the distribution function which assigns probability p_i at the ith observation X_i, namely

$$G_p(x) = \sum_{i=1}^n p_i I\{X_i \leq x\},$$

and hence $\theta(G_p) = \sum_{i=1}^n p_i X_i$. Then, the empirical likelihood, evaluated at θ, is given by

$$L(\theta) = \max\left\{\prod_{i=1}^n p_i : \sum p_i = 1, \sum_{i=1}^n p_i X_i = \theta\right\}.$$

Note that $\prod_{i=1}^n p_i$, subject to $\sum_{i=1}^n p_i = 1$, attains its maximum n^{-n} at $p_i = n^{-1}$. So we define the empirical likelihood ratio at θ by

$$\begin{aligned} R(\theta) &= \frac{\max\{\prod_{i=1}^n p_i : \sum p_i = 1, \sum_{i=1}^n p_i X_i = \theta\}}{\max\{\prod_{i=1}^n p_i : \sum p_i = 1\}} = \frac{L(\theta)}{n^{-n}} \\ &= \max\left\{\prod_{i=1}^n (np_i) : \sum p_i = 1, \sum_{i=1}^n p_i X_i = \theta\right\}. \qquad (2.1) \end{aligned}$$

Using Lagrange multipliers, when $X_{(1)} \leq \theta \leq X_{(n)}$, which has probability going to one, we have (why???)

$$p_i = \frac{1}{n}\frac{1}{1 + \lambda(X_i - \theta)}, \qquad (2.2)$$

where λ satisfies

$$g(\lambda) \equiv \frac{1}{n}\sum_{i=1}^n \frac{X_i - \theta}{1 + \lambda(X_i - \theta)} = 0. \qquad (2.3)$$

Proof. Let

$$h(p_1, ..., p_n, \lambda_0, \lambda) = \sum_{i=1}^n \ln(np_i) - n\lambda_0\left(\sum_{i=1}^n p_i - 1\right) - n\lambda\left(\sum_{i=1}^n p_i X_i - \theta\right).$$

Set

$$\begin{aligned} \frac{\partial h}{\partial p_i} &= \frac{1}{p_i} - n\lambda_0 - n\lambda X_i = 0, \\ \frac{\partial h}{\partial \lambda_0} &= -n\left(\sum_{i=1}^n p_i - 1\right) = 0, \\ \frac{\partial h}{\partial \lambda} &= -n\left(\sum_{i=1}^n p_i X_i - \theta\right) = 0. \end{aligned}$$

From the first equation, we get $1 = n\lambda_0 p_i + n\lambda p_i X_i$. Summing over i and using the second and third relations, we get $n = n\lambda_0 + n\lambda\theta$, i.e., $1 = \lambda_0 + \lambda\theta$, i.e., $\lambda_0 = 1 - \lambda\theta$. Therefore, from the first equation, we have

$$p_i = \frac{1}{n\lambda_0 + n\lambda X_i} = \frac{1}{n}\frac{1}{\lambda_0 + \lambda X_i} = \frac{1}{n}\frac{1}{1 + \lambda(X_i - \theta)},$$

where λ satisfies

$$\sum_{i=1}^{n} p_i X_i - \theta = \sum_{i=1}^{n} p_i (X_i - \theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i - \theta}{1 + \lambda(X_i - \theta)} = 0. \quad \blacksquare$$

After plugging the p_i's back into (2.1) and taking the logarithm of $R(\theta)$, we get the nonparametric log-likelihood ratio

$$\log R(\theta) = -\sum_{i=1}^{n} \log\{1 + \lambda(X_i - \theta)\}.$$

The next theorem shows that Wilks's theorem holds here under a mild condition.

Theorem **12.2** *Assume that $EX_1^2 < \infty$, then*

$$-2\log R(\theta) \to_d \chi_1^2. \quad \blacksquare$$

Remark **12.1** *An approximate $1 - \alpha$ level confidence interval for θ can be defined as*

$$C.I.(1 - \alpha) = \{\theta : R(\theta) \geq c_0\} = \{\theta : -2\log R(\theta) \leq c\},$$

where c satisfies $P(\chi_1^2 \geq c) = \alpha$. From Theorem 12.2, we have

$$\lim_{n\to\infty} P\{\theta \in C.R.(1 - \alpha)\} = P(\chi_1^2 \leq c) = 1 - \alpha.$$

In other words, the interval $\Re_c$ gives asymptotic correct coverage probability.

Remark **12.2** *Under stronger moment conditions and some smoothness condition, we can describe coverage accuracy more precisely. For example, suppose $EX_1^4 < \infty$ and X_1 satisfies Cramér condition, then*

$$P\{\theta \in \Re_c\} = \alpha + O(n^{-1}).$$

Proof of Theorem 12.2.

Lemma **12.1**

1. *Let* $Y_n = \max\limits_{1\le i\le n} |X_i|$. *If* $EX_1^2 < \infty$, *then* $Y_n = o(n^{1/2})$ *a.s.*
2. *Let* $Z_n = \max_{1\le i\le n} |X_i - \theta|$, *if* $EX_1^2 < \infty$, *then*

$$Z_n = o(n^{1/2}) \qquad a.s., \qquad and \qquad \frac{1}{n}\sum_{i=1}^{n} |X_i-\theta|^3 = o(n^{1/2}) \qquad a.s.$$

3. *Show that the root of (2.3) satisfies* $|\lambda| = O_p(n^{-1/2})$.

Proof.

1. Since $EX_1^2 < \infty$, using $\sum_{n=1}^{\infty} P(|\xi| \ge n) \le E|\xi| \le 1 + \sum_{n=1}^{\infty} P(|\xi|\ge n)$, we have

$$\sum_{n=1}^{\infty} P\left(X_1^2 > n\right) = \sum_{n=1}^{\infty} P\left(|X_1| > \sqrt{n}\right) = \sum_{n=1}^{\infty} P\left(|X_n| > \sqrt{n}\right) < \infty.$$

By Borel-Cantelli Lemma, with probability 1,

$$|X_n| > n^{1/2}, \qquad \text{finitely often (for every } i \ge 1)$$

Thus with probability 1, $Y_n =: \max_{1\le i\le n} |X_i| > n^{1/2}$ occurs finitely often. By the same argument $Y_n > \epsilon n^{1/2}$ finitely often with probability 1 for any $\epsilon > 0$. Consequently,

$$\limsup_{n\to\infty} \frac{Y_n}{n^{1/2}} \le \epsilon \qquad a.s. \tag{2.4}$$

Inequality (2.4) holds simultaneously with probability 1 for any countable set of values for ϵ. Therefore $Y_n = o(n^{1/2})$ *a.s.* ∎

2. Note that $|X_i - \theta| \le Y_n + |\theta|$ for any $1 \le i \le n$. By Lemma 12.1, $Z_n = o(n^{1/2})$ *a.s.*

For the second assertion, with probability 1

$$\frac{1}{n}\sum_{i=1}^{n} |X_i-\theta|^3 \quad\le\quad Z_n \times \frac{1}{n}\sum_{i=1}^{n}(X_i-\theta)^2 = o(n^{1/2}) \times (\sigma^2 + o(1)) = o(n^{1/2}) \quad ∎$$

3. Write $\tilde{X}_t = X_t - \theta$. Note that

$$0 = \frac{1}{n}\sum_{t=1}^{n} \frac{\widetilde{X}_t}{1+\lambda\widetilde{X}_t} = \frac{1}{n}\sum_{t=1}^{n}\widetilde{X}_t - \frac{\lambda}{n}\sum_{t=1}^{n}\frac{\widetilde{X}_t^2}{1+\lambda\widetilde{X}_t}.$$

Thus, we have

$$|\frac{1}{n}\sum_{t=1}^{n}\widetilde{X}_t| = \frac{|\lambda|}{n}\left|\sum_{t=1}^{n}\frac{\widetilde{X}_t^2}{1+\lambda\widetilde{X}_t}\right| \ge \frac{|\lambda|}{1+|\lambda|\max_t|\widetilde{X}_t|}\frac{1}{n}\sum_{t=1}^{n}\widetilde{X}_t^2$$

Hence,

$$\left|\frac{1}{n}\sum_{t=1}^{n}\widetilde{X}_t\right|\left(1+|\lambda|\max_t|\widetilde{X}_t|\right) \ge |\lambda|\frac{1}{n}\sum_{t=1}^{n}\widetilde{X}_t^2$$

It then follows that

$$|\lambda| \leq \frac{|\frac{1}{n}\sum_{t=1}^{n}\widetilde{X}_t|}{\frac{1}{n}\sum_{t=1}^{n}\widetilde{X}_t^2 - |\frac{1}{n}\sum_{t=1}^{n}\widetilde{X}_t| \max_t |\widetilde{X}_t|}$$

From (a)-(c) of the present lemma, we have

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_t = O_p(\frac{1}{\sqrt{n}}), \qquad \frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_t^2 \to_{a.s.} \sigma^2, \qquad \max_{1\leq t\leq n}|\widetilde{X}_t| = o_p(n^{\frac{1}{2}}).$$

Therefore, $\lambda = O_p(n^{-1/2})$. ∎

Lemma **12.2** *We have* $\sum_{i=1}^{n}\lambda\widetilde{X}_i = \sum_{i=1}^{n}(\lambda\widetilde{X}_i)^2 + o_p(1)$.

Proof: From Lemma 12.1, we have $\lambda\widetilde{X}_i = o_p(1)$. By Taylor expansion, we get

$$\begin{aligned} 0 &= \lambda\sum_{i=1}^{n}\frac{\widetilde{X}_i}{1+\lambda\widetilde{X}_i} = \lambda\sum_{i=1}^{n}\widetilde{X}_i(1-\lambda\widetilde{X}_i + O_p((\lambda\widetilde{X}_i)^2) \\ &= \sum_{i=1}^{n}(\lambda\widetilde{X}_i) - \sum_{i=1}^{n}(\lambda\widetilde{X}_i)^2 + \lambda^3 O_p(\sum_{i=1}^{n}\widetilde{X}_i^3). \end{aligned}$$

Note that $\lambda^3|\sum_{i=1}^{n}\widetilde{X}_i^3| \leq \lambda^3\max_{1\leq t\leq n}|\widetilde{X}_i|\sum_{i=1}^{n}\widetilde{X}_i^2 = O_p(n^{-3/2})o_p(n^{\frac{1}{2}})O_p(n) = o_p(1)$. So have $\sum_{i=1}^{n}\lambda\widetilde{X}_i = \sum_{i=1}^{n}(\lambda\widetilde{X}_i)^2 + o_p(1)$. ∎

Completion of the proof of Theorem 12.2. From Lemma 12.2, we have $\sum_{i=1}^{n}\widetilde{X}_i = \lambda\sum_{i=1}^{n}\widetilde{X}_i^2 + o_p(\lambda^{-1})$. So

$$\begin{aligned} \lambda &= \frac{\sum_{i=1}^{n}\widetilde{X}_i}{\sum_{i=1}^{n}\widetilde{X}_i^2} + o_p\left(\lambda^{-1}(\sum_{i=1}^{n}\widetilde{X}_i^2)^{-1}\right) = \frac{\sum_{i=1}^{n}\widetilde{X}_i}{\sum_{i=1}^{n}\widetilde{X}_i^2} + o_p(O_p(n^{\frac{1}{2}})O_p(n^{-1})) \\ &= \frac{\sum_{i=1}^{n}\widetilde{X}_i}{\sum_{i=1}^{n}\widetilde{X}_i^2} + o_p(\frac{1}{\sqrt{n}}). \end{aligned}$$

Therefore, we have

$$\begin{aligned} 2\sum_{i=1}^{n}\log(1+\lambda\widetilde{X}_i) &= 2\sum_{i=1}^{n}\lambda\widetilde{X}_i - \sum_{i=1}^{n}(\lambda\widetilde{X}_i)^2 + 2\sum_{i=1}^{n}o_P((\lambda\widetilde{X}_i)^2) = \lambda\sum_{i=1}^{n}\widetilde{X}_i + o_p(1) \\ &= \frac{(n^{-1/2}\sum_{i=1}^{n}\widetilde{X}_i)^2}{n^{-1}\sum_{i=1}^{n}\widetilde{X}_i^2} + o_p(1) \longrightarrow_d \chi_1^2. \end{aligned}$$ ∎

12.3 EL for the multivariate mean

Let $X_1, ..., X_n$ be an i.i.d. random d-vectors from a d.f. F (typically unkonwn). Suppose that we are interested in the vector mean $\theta = \theta(F) = EX_1$. Similarly to the case of the univariate mean, we can show that

THEOREM **12.3** *Assume that $EX_1^2 < \infty$, then*

$$-2 \log R(\theta) \rightarrow_d \chi_d^2.$$

REMARK **12.3** *An approximate $1 - \alpha$ level confidence region for θ can be defined as*

$$C.R.(1 - \alpha) = \{\theta : -2 \log R(\theta) \leq c\},$$

where c satisfies $P(\chi_d^2 \geq c) = \alpha$. From Theorem 12.2, we have

$$\lim_{n\to\infty} P\{\theta \in C.R.(1 - \alpha)\} = P(\chi_1^2 \leq c) = 1 - \alpha.$$

Namely, the confidence region $C.R.(1 - \alpha)$ gives asymptotic correct coverage probability.

REMARK **12.4** *The confidence region defined above can be shown to form a convex set. The shape of the region is, however, completely determined by the data. This is one of the major advantages over its competitor, the bootstrap method, which requires the user to specify the shape of the confidence region.*

REMARK **12.5** *Other nice properties of the empirical likelihood method is that it's range preserving and transformation respecting.*

REMARK **12.6** *The above remarks apply more generally to other quantities other than the population means.*

Chapter 13

Bayes Rules and Empirical Bayes

13.1 Introduction

It is impossible to find estimators which minimize the risk $R(\theta, \delta)$ uniformly at every value of θ (see the example below) unless we restrict ourselves to a smaller class. We shall now drop such restrictions, admitting all estimators into competition, but shall then have to be satisfied with a weaker optimality property than uniformly minimum risk. We shall look for estimators that make the risk function $R(\theta, \delta)$ small in some overall sense. Two such optimality properties will be considered:

- Average Risk Optimality (minimizing the weighted average risk);
- Minimax Optimality (minimizing the maximum risk)

We consider the first approach now while the second will be studied in the next chapter.

An illustrative example

Suppose we are interested in the average income μ of an area A. If we have no data for the area, a natural guess would be μ_0, the average income from the last national census; whereas if we have a random sample $X_1, X_2, ..., X_n$ from the area A, we may want to combine μ_0 and $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ into an estimator, for instance,

$$\hat{\mu} \;=\; (0.2)\mu_0 + (0.8)\bar{X}.$$

For comparison, we calculate their $MSE(\hat{\mu}) = E(\hat{\mu} - \mu)^2 = Var(\hat{\mu}) + bias^2(\mu)$. Hence,

$$\begin{aligned} R(\mu, \mu_0) &= MSE(\mu_0) = (\mu_0 - \mu)^2, \\ R(\mu, \hat{\mu}) &= MSE(\hat{\mu}) = .04(\mu_0 - \mu)^2 + (.64)\sigma^2/n, \\ R(\mu, \bar{X}) &= MSE(\bar{X}) = \sigma^2/n. \end{aligned}$$

We can plot these risks as functions of μ.

Here are some observations:

1. Since μ is unknown, none of the estimators can be proclaimed as the best.
2. We can place a prior $\mu \sim N(\mu_0 + \delta, A)$.
   (a) If $\delta \approx 0$ and $A \approx 0$, $MSE(\mu_0)$ will be the smallest.
   (b) If μ is close to μ_0, then $R(\mu, \hat{\mu}) < R(\mu, \bar{X})$ with the minimum relative risk

$$\inf\{R(\mu, \hat{\mu})/R(\mu, \bar{X}); \mu \in R\} = 0.64, \text{ when } \mu = \mu_0$$

3. If we use as our criteria the maximum (over μ) of the MSE (called the minimax criteria), then $\bar{X}$ is optimal (shown later).

13.2 Bayes estimator

Assume that

$$\begin{aligned} \mathbf{X}|\theta &\sim f(\mathbf{x}|\theta), \\ \theta &\sim \pi. \end{aligned}$$

Let $\delta = \delta(\mathbf{X})$ be an estimator of $g(\theta)$. (Sometimes we use $d \approx \delta$ (decision), or a (action)). Define

- Loss function: $l(\theta, \delta)$.

  Examples:

  (1) L_2-Loss:
  $$l(\theta, \delta) = (\delta(x) - g(\theta))^2.$$

  (2) Weighted L_2-Loss:
  $$l(\theta, \delta) = w(\theta)(\delta(x) - g(\theta))^2.$$

  (3) L_1-Loss:
  $$l(\theta, \delta) = |\delta(x) - g(\theta)|.$$

  (4) Supremum-Loss:
  $$l(\theta, \delta) = \sup_x |\delta(x) - g(\theta)|.$$

  (5) $0 - 1$ Loss:
  $$l(\theta, \delta) = I\{\delta \neq g(\theta)\}.$$

- A priori risk (a.k.a. frequentist risk)

$$R(\theta, \delta) \equiv E[l(\theta, \delta)|\theta] \equiv E_\theta l(\theta, \delta) = \int l(\theta, \delta(\mathbf{x}))f(\mathbf{x}|\theta)dx$$

- A posterior risk:

$$r(\delta|\mathbf{x}) \equiv E[l(\theta, \delta(\mathbf{x}))|\mathbf{X} = \mathbf{x}] = \int l(\theta, \delta(\mathbf{x}))f(\theta|\mathbf{x})d\theta$$

- Bayes risk (or average risk w.r.t. θ):

$$r(\pi, \delta) \equiv El(\theta, \delta) = Er(\delta|\mathbf{X}) = ER(\theta, \delta) = \int \int l(\theta, \delta(\mathbf{x}))f(\theta, \mathbf{x})d\theta d\mathbf{x}.$$

Definition **13.1** *An estimator δ^* is called a Bayes estimator w.r.t. π if*

$$r(\pi, \delta^*) \equiv El(\theta, \delta^*) \leq El(\theta, \delta) \equiv r(\pi, \delta).$$

i.e.,

$$\delta^* = \arg\inf_\delta r(\pi, \delta) = \arg\inf_\delta El(\theta, \delta).$$

Why Bayes estimators?

Bayes estimators are important in a number of different contexts.

1. As Mathematical Tools

   Bayes estimators play a central role in Wald's decision theory. One of its main results is that in any given statistical problem, attention can be restricted to Bayes solutions and suitable limits of Bayes solutions; given any other procedure δ, there exists a procedure δ' in this class such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all θ. (In view of this result, it is not surprising that Bayes estimators provide a tool for solving minimax problems, as will be seen in the next chapter.)

2. As a Way of Utilizing Past Experience

   It is frequently reasonable to treat θ as the realization of a r.v. with known d.f. rather than as a constant. For example, in estimating the probability of a penny showing heads when spun on a flat surface, we would have considered n spins, and found the sample proportion of heads. Suppose, however, that we have had considerable experience before, it might be reasonable to represent this past knowledge as a prior probability distribution for p.

3. As a Description of a State of Mind

   A formally similar approach is adopted by the so-called Bayesian school, which interprets π as expressing the subjective feeling about the likelihood of different θ values. The subjective Bayesian uses the observations $\mathbf{X}$ to modify prior beliefs. After $\mathbf{X} = \mathbf{x}$ has been observed, the belief about θ is expressed by the posterior (i.e., conditional) distribution of θ given $\mathbf{x}$.

Finding a Bayes estimator is, in principle, quite simple

Before any observations are taken, the Bayes estimator δ^* minimizes the Bayes risk $\mathbf{E}l(\theta, \delta)$ by definition.

Since we observe $\mathbf{X} = \mathbf{x}$, the Bayes estimator should minimize the posterior risk $E\{L[\theta, \delta(\mathbf{x})]|\mathbf{x}\}$. The following is a precise statement of this result.

Theorem **13.1** *If $\delta^*(\mathbf{x})$ minimizes the posterior risk*

$$r(\delta|\mathbf{x}) \equiv E\{l[\theta, \delta(\mathbf{x})]|\mathbf{X} = \mathbf{x}\}, \qquad (a.s.),$$

then $\delta^*(\mathbf{X})$ is a Bayes estimator.

Proof. $E\{l[\theta, \delta(\mathbf{x})]|\mathbf{X} = \mathbf{x}\} \geq E\{l[\theta, \delta^*(\mathbf{x})]|\mathbf{X} = \mathbf{x}\}$ a.e. Then take expectation. ∎

Corollary **13.1** *For L_2-Loss: $l(\theta, \delta) = (\delta - g(\theta))^2$, then*

$$\delta^*(\mathbf{x}) = E[g(\theta)|\mathbf{x}] = \int g(\theta) f(\theta|x) d\theta = \frac{\int g(\theta) f(x|\theta)\pi(\theta) d\theta}{f(x)} = \frac{\int g(\theta) f(x|\theta)\pi(\theta) d\theta}{\int f(x|\theta)\pi(\theta) d\theta}.$$

More generally, for weighted L_2-Loss: $l(\theta, \delta) = w(\theta)(\delta - g(\theta))^2$, then

$$\delta^*(\mathbf{x}) = \frac{E[w(\theta)g(\theta)|\mathbf{x}]}{E[w(\theta)|\mathbf{x}]} = \frac{\int w(\theta) g(\theta) f(x|\theta)\pi(\theta) d\theta}{\int w(\theta) f(x|\theta)\pi(\theta) d\theta}$$

Proof. For weighted L_2-loss,

$$r(\delta|\mathbf{x}) = E[w(\theta)(\delta - g(\theta))^2|\mathbf{x}] = \delta^2 E[w(\theta)|\mathbf{x}] - 2\delta E[w(\theta)g(\theta)|\mathbf{x}] + E[w(\theta)g^2(\theta)|\mathbf{x}].$$

Setting $r'(\delta|\mathbf{x}) = 0$, we get the desired result. ∎

Determininination of Bayes estimates may be daunting and the use of conjugate priors might facilitate its calculations, and reduce computational complexity in $\delta^*(\mathbf{x})$. Conjugate priors are natural parametric families of priors such that the posterior distributions also belong to this family. Such families are called conjugate.

Examples

Example **13.1 (Normal mean with known variance)**

Suppose

$$\begin{aligned} X_i|\mu &\sim N(\mu, \tilde{\sigma}_0^2), \qquad i = 1, ..., n \\ \mu &\sim N(M, A), \qquad \text{(conjugate prior)} \end{aligned}$$

Denote $\mathbf{x} = (x_1, ..., x_n)'$. The posterior pdf is

$$\begin{aligned} f(\mu|\mathbf{x}) &= f(\mathbf{x}, \mu)/f(\mathbf{x}) \propto f(\mathbf{x}, \mu) \\ &= \frac{1}{(2\pi\tilde{\sigma}_0^2)^{n/2}} \exp\left\{-\frac{1}{2\tilde{\sigma}_0^2}\sum(x_i - \mu)^2\right\} \\ &\quad \times \frac{1}{(2\pi A)^{1/2}} \exp\left\{-\frac{1}{2A}(\mu - M)^2\right\} \\ &\propto \exp\left\{-\frac{n}{2\tilde{\sigma}_0^2}(-2\bar{x}\mu + \mu^2)\right\} \exp\left\{-\frac{1}{2A}(\mu^2 - 2M\mu)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\tilde{\sigma}_0^2} + \frac{1}{A}\right)\left(\mu^2 - 2\mu\frac{n\bar{x}/\tilde{\sigma}_0^2 + M/A}{n/\tilde{\sigma}_0^2 + 1/A}\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_0^2} + \frac{1}{A}\right)\left(\mu - \frac{\bar{x}/\sigma_0^2 + M/A}{1/\sigma_0^2 + 1/A}\right)^2\right\} \\ &\sim N\left(E[\mu|\mathbf{x}], var[\mu|\mathbf{x}]\right), \end{aligned}$$

where $\sigma_0^2 = \tilde{\sigma}_0^2/n$, and

$$\begin{aligned} E[\mu|\mathbf{x}] &= \frac{\bar{x}/\sigma_0^2 + M/A}{1/\sigma_0^2 + 1/A} = \frac{A\bar{x} + \sigma_0^2 M}{A + \sigma_0^2}, \\ var[\mu|\mathbf{x}] &= \frac{1}{1/\sigma_0^2 + 1/A} = \frac{A\sigma_0^2}{A + \sigma_0^2}. \end{aligned}$$

Under L_2-loss, the Bayes estimator is

$$\delta^* = E[\mu|\mathbf{x}] = \frac{A}{A + \sigma_0^2}\bar{x} + \frac{\sigma_0^2}{A + \sigma_0^2}M. \quad \blacksquare$$

Question: The posterior mean is now biased, but the posterior variance is shrunk. It would be interesting to know if and when the MSE will be smaller than the frequentist mean $\bar{X}$.

Remark **13.1**

1. $\delta^* = w\bar{x} + (1-w)M$, a weighted average of the standard estimator $\bar{X}$ and the mean M of the prior distribution.
2. As $n \to \infty$ with M and A fixed, $\delta^* \approx \bar{X} \to_p \theta$. Hence, choice of priors is of little import when n is large.
3. When n is fixed, choice of prior is important. For example,
   (a) As $A \to 0$, then $\delta^* \to_p M$. (Very informative prior.)
   (b) As $A \to \infty$, $\delta^* \approx \bar{X}$. (Non-informative (flat) prior.)
4. δ^* is a shrinkage estimator; e.g., if $M = 0$, $\delta^* = \frac{A}{A+\sigma_0^2}\bar{x}$ shrinks $\bar{x}$ toward 0.

Remark **13.2** *Inference can be based on sufficient statistics $z = \bar{x}$ only. That is,*

$$\begin{aligned} z|\mu &\sim N(\mu, \sigma_0^2), \qquad i = 1, ..., n \\ \mu &\sim N(M, A), \qquad (\textit{conjugate prior}). \end{aligned}$$

It can be easily seen that, under L_2-loss, the Bayes estimator is

$$\delta^* = E[\mu|z] = \frac{A}{A+\sigma_0^2}z + \frac{\sigma_0^2}{A+\sigma_0^2}M. \quad ▮$$

Example **13.2 (Normal variance with known mean)**

Suppose

$$\begin{aligned} X_i|\sigma^2 &\sim N(0, \sigma^2), \\ \tau \equiv \frac{1}{\sigma^2} &\sim Gamma(\alpha, \beta) \qquad \text{(conjugate prior)} \end{aligned}$$

where $\pi(\tau) = (\beta^\alpha/\Gamma(\alpha))\tau^{\alpha-1}e^{-\beta\tau}$ with

$$E\tau = \frac{\alpha}{\beta}, \quad E\tau^2 = \frac{\alpha(\alpha+1)}{\beta^2},$$

$$E\tau^{-1} = E\sigma^2 = \frac{\beta}{\alpha-1}, \quad E\tau^{-2} = E\sigma^4 = \frac{\beta^2}{(\alpha-1)(\alpha-2)}.$$

So the posterior distribution is

$$\begin{aligned} f(\tau|\mathbf{x}) &\propto f(\mathbf{x}|\mu,\tau)\pi(\tau) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}\right) \times \frac{\beta^\alpha}{\Gamma(\alpha)}\tau^{\alpha-1}e^{-\beta\tau} \\ &\propto \tau^{n/2}\exp\left(-\frac{\tau}{2}\sum_{i=1}^n x_i^2\right)\tau^{\alpha-1}e^{-\beta\tau} \\ &\propto \tau^{n/2+\alpha-1}\exp\left\{-\tau\left(\frac{1}{2}\sum_{i=1}^n x_i^2 + \beta\right)\right\} \\ &\sim Gamma\left(\frac{n}{2}+\alpha, \frac{1}{2}\sum_{i=1}^n x_i^2 + \beta\right). \end{aligned}$$

Under L_2-loss, the Bayes estimator of $\sigma^2 = 1/\tau$ is

$$\begin{aligned} \delta^* &= E[\tau^{-1}|\mathbf{x}] = \frac{\sum_{i=1}^n x_i^2/2 + \beta}{n/2+\alpha-1} = \frac{\sum_{i=1}^n x_i^2 + 2\beta}{n+2\alpha-2} \\ &= \left(\frac{n}{n+2\alpha-2}\right)\frac{\sum_{i=1}^n x_i^2}{n} + \left(\frac{2\alpha-2}{n+2\alpha-2}\right)\frac{\beta}{\alpha-1} \\ &=: wS_n^2 + (1-w)\,E\sigma^2. \end{aligned}$$

which is a weighted average of sample variance $S_n^2 = \sum_{i=1}^n x_i^2/n$ and prior variance $E\sigma^2 = \beta/(\alpha-1)$.

Example **13.3 (Normal variance with unknown mean)**

Suppose

- $X_i|(\mu, \sigma^2) \sim_{iid} N(\mu, \sigma^2)$
- $\pi(\mu) = 1$, or $\mu \sim N(M, \infty)$, (improper uniform prior)
- $\tau \equiv \frac{1}{\sigma^2} \sim Gamma(\alpha, \beta)$, i.e., $\pi(\tau) \propto \tau^{a-1} e^{-b\tau}$, (conjugate prior)
- $\mu \perp \tau$

Now

$$\begin{aligned} f(\mathbf{x}|\mu, \tau) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right) \\ &\propto \tau^{n/2} \exp\left(-\frac{\tau}{2}\sum_{i=1}^n (x_i - \mu)^2\right). \end{aligned}$$

So the posterior distribution is

$$\begin{aligned} f(\mu, \tau|\mathbf{x}) &\propto f(\mathbf{x}|\mu, \tau)\pi(\mu)\pi(\tau) \\ &\propto \tau^{n/2+\alpha-1} \exp\left(-\frac{\tau}{2}\left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 + 2\beta\right)\right). \end{aligned}$$

By integrating out μ, we get

$$\begin{aligned} f(\tau|\mathbf{x}) &= \int f(\mu, \tau|\mathbf{x}) d\mu \\ &\propto \tau^{n/2+\alpha-3/2} \exp\left(-\frac{\tau}{2}\left(\sum_{i=1}^n (x_i - \bar{x})^2 + 2\beta\right)\right) \\ &\quad \times \int \tau^{1/2} \exp\left(-\tau n(\mu - \bar{x})^2\right) d\mu \\ &\propto \tau^{n/2+\alpha-3/2} \exp\left(-\frac{\tau}{2}\left(\sum_{i=1}^n (x_i - \bar{x})^2 + 2\beta\right)\right) \\ &\sim Gamma\left(\frac{n}{2} + \alpha - \frac{1}{2},\ \frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2 + \beta\right). \end{aligned}$$

Under L_2-loss, the Bayes estimator of $\sigma^2 = 1/\tau$ is

$$\begin{aligned} \delta^* &= E(\tau^{-1}|\mathbf{x}) = \frac{\frac{1}{2}\sum_{i=1}^n (x_i - \bar{x})^2 + \beta}{n/2 + \alpha - 3/2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 + 2\beta}{n + 2\alpha - 3} \\ &= \left(\frac{n-1}{n+2\alpha-3}\right)\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} + \left(\frac{2\alpha-2}{n+2\alpha-3}\right)\frac{\beta}{(\alpha-1)} \\ &=: wS^2 + (1-w)\, E\sigma^2, \end{aligned}$$

which is a weighted average of sample variance $S^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$ and prior variance $E\sigma^2 = \beta/(\alpha - 1)$. ∎

Example **13.4 (Binomial mean)**

Suppose

$$\begin{aligned} X|p &\sim Binomial(n, p), \\ p &\sim Beta(a, b), \qquad \text{(conjugate prior)} \end{aligned}$$

The joint pdf is

$$f(\mathbf{x}, p) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{x+a-1}(1-p)^{n-x+b-1}$$

and the posterior pdf is

$$\begin{aligned} f(p|\mathbf{x}) &= f(\mathbf{x}, \theta)/f(\mathbf{x}) \propto f(\mathbf{x}, \theta) \\ &\propto p^{x+a-1}(1-p)^{n-x+b-1} \\ &\sim Beta(x+a, n-x+b), \end{aligned}$$

Recall, for $p \sim Beta(a, b)$, we have

$$E(p) = \frac{a}{a+b} \qquad \text{and} \qquad var(p) = \frac{ab}{(a+b)^2}(a+b+1)$$

So the Bayes estimator under L_2-loss is

$$\delta_k^*(\mathbf{x}) = E[p|\mathbf{x}] = \frac{a+x}{a+b+n} = \left(\frac{a+b}{a+b+n}\right)\frac{a}{a+b} + \left(\frac{n}{a+b+n}\right)\frac{X}{n},$$

which is a weighted average of standard estimate X/n and prior mean $a/(a+b)$. ∎

Example **13.5 (Poisson)**

Suppose

$$\begin{aligned} X_i|\theta &\sim Poisson(\theta),, \\ \theta &\sim Gamma(\alpha,\beta), \qquad \text{(conjugate prior)} \end{aligned}$$

The posterior distribution is

$$f(\theta|\mathbf{x}) = Gamma\left(\alpha+\bar{x}, \frac{\beta}{1+\beta}\right)$$

Note that $EX = var(X) = \theta$. Although $L_0(\theta,\delta) = (\theta-\delta)^2$ is often preferred for the estimation of a mean, some type of scaled squared error loss, e.g., $L_k(\theta,\delta) = (\theta-\delta)^2/\theta^k$, may be more appropriate for the estimation of a variance. The Bayes estimator under L_k-loss is (by taking $w(\theta) = \theta^{-k}$ in Corollary 13.1)

$$\begin{aligned} \delta_k^*(\mathbf{x}) &= \frac{E(\theta^{1-k}|\mathbf{x})}{E(\theta^{-k}|\mathbf{x})} = \frac{\beta}{\beta+1}\left(\bar{x}+\alpha-k\right), \\ &= \left(\frac{\beta}{\beta+1}\right)\bar{x} + \left(\frac{1}{\beta+1}\right)(\alpha-k)\beta, \qquad \text{for } \alpha-k>0 \end{aligned}$$

which is a weighted average of sample mean $\bar{x}$ and $(\alpha-k)\beta$ (the prior mean is $\alpha\beta$). ▮

Note that the choice of loss function can have a large effect on the resulting Bayes estimator.

Example 13.6 (Regression)

Consider generalized least squares regression

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N(0, V).$$

1. The MLE of β is $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y := AY$.
   (In particular, if $V = \sigma^2 I$, then $\hat{\beta} = (X^T X)^{-1} X^T Y$.)
2. Assume we have prior $\beta \sim N(\beta_P, P)$, then

   $$\hat{\beta}_{Bayes} = (X^T V^{-1} X + P^{-1})^{-1} (X^T V^{-1} Y + P^{-1} \beta_P).$$

   Remarks:

   - This is simply a ridge estimator.
   - If $V = \sigma^2 I$ and $\beta_P = 0$, we have $\hat{\beta}_{Bayes} = (X^T X + \sigma^{-2} I)^{-1} X^T Y$.
3. $\hat{\beta}_{Bayes}$ satisfies

   $$\hat{\beta}_{Bayes} = \beta_P + K(Y - X\beta_P) = (I - KX)\beta_P + KY$$

   where $K = (X^T V^{-1} X + P^{-1})^{-1} X^T V^{-1}$.

Proof. The likelihood of β is

$$\begin{aligned} l(\beta) = f(\epsilon) &= (2\pi)^{n/2} |V|^{-1/2} \exp\{-\frac{1}{2}(Y - X\beta)^T V^{-1} (Y - X\beta)\} \\ &\propto \exp\{-g(\beta)\}, \end{aligned}$$

where $g(\beta) = \frac{1}{2}(Y - X\beta)^T V^{-1} (Y - X\beta)$.

1. The MLE of β is

   $$\hat{\beta} = \arg\max_{\beta} l(\beta) = \arg\min_{\beta} g(\epsilon).$$

   That is, $\hat{\beta}$ satisfies $g'(\beta) = -X^T V^{-1}(Y - X\beta) = -X^T V^{-1} Y + X^T V^{-1} X\beta = 0$, resulting in

   $$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y := AY.$$

   It is easy to see

   $$\begin{aligned} E\hat{\beta} &= (X^T V^{-1} X)^{-1} X^T V^{-1} X\beta = \beta, \\ Var(\hat{\beta}) &= AVar(Y)A^T = (X^T V^{-1} X)^{-1} X^T V^{-1} V V^{-1} X (X^T V^{-1} X)^{-1} = (X^T V^{-1} X)^{-1}. \end{aligned}$$

2. The posterior of β is

   $$\begin{aligned} f(\epsilon, \beta) &= f(\epsilon|\beta)\pi(\beta) \\ &= (2\pi)^{n/2} |V|^{-1/2} \exp\{-\frac{1}{2}(Y - X\beta)^T V^{-1} (Y - X\beta)\} \\ &\quad \times (2\pi)^{p/2} |P|^{-1/2} \exp\{-\frac{1}{2}(\beta - \beta_P)^T P^{-1} (\beta - \beta_P)\} \\ &\propto \exp\{-\frac{1}{2}(Y - X\beta)^T V^{-1} (Y - X\beta) - \frac{1}{2}(\beta - \beta_P)^T P^{-1} (\beta - \beta_P)\} \\ &:= \exp\{-h(\beta)\}, \end{aligned}$$

where $h(\beta) = \frac{1}{2}(Y - X\beta)^T V^{-1}(Y - X\beta) - \frac{1}{2}(\beta - \beta_P)^T P^{-1}(\beta - \beta_P)$.

The maximum a posteriori (MAP) estimate is

$$\hat{\beta}^{Bayes} = \arg\min_{\beta} f(\epsilon, \beta) = \arg\max_{\beta} h(\epsilon).$$

That is, $\hat{\beta}^{Bayes}$ satisfies

$$\begin{aligned}
0 &= h'(\beta) \\
&= -X^T V^{-1}(Y - X\beta) + P^{-1}(\beta - \beta_P) \\
&= -X^T V^{-1} Y + X^T V^{-1} X\beta + P^{-1}\beta - P^{-1}\beta_P \\
&= (X^T V^{-1} X + P^{-1})\beta - (X^T V^{-1} Y + P^{-1}\beta_P)
\end{aligned}$$

resulting in

$$\hat{\beta}^{Bayes} = (X^T V^{-1} X + P^{-1})^{-1}(X^T V^{-1} Y + P^{-1}\beta_P).$$

3. $\hat{\beta}_{Bayes}$ can be rewritten as

$$\begin{aligned}
\hat{\beta}^{Bayes} &= (X^T V^{-1} X + P^{-1})^{-1}[X^T V^{-1}(Y - X\beta_P) + (X^T V^{-1} X + P^{-1})\beta_P] \\
&= \beta_P + (X^T V^{-1} X + P^{-1})^{-1} X^T V^{-1}(Y - X\beta_P) \\
&:= \beta_P + K(Y - X\beta_P).
\end{aligned}$$

It is easy to see

$$\begin{aligned}
E\hat{\beta}^{Bayes} &= \beta_P + K(X\beta - X\beta_P) = (1 - KX)\beta_P + KX\beta \\
Var(\hat{\beta}^{Bayes}) &= KVar(Y)K^T = (X^T V^{-1} X + P^{-1})^{-1} X^T V^{-1} X (X^T V^{-1} X + P^{-1})^{-1}.
\end{aligned}$$

Clearly, as with other Bayes estimator, $\hat{\beta}^{Bayes}$ is a linear combination of prior estimator and frequentist estimator. It is biased with reduced variance.

13.3 Properties of Bayes Estimators

Bayes estimators: balance between evidence and belief

In many cases, a Bayes estimator is a weighted average of frequentist estimate $\hat{\theta}_{freq}$ and prior estimate θ_{prior}:

$$\delta^* = w\hat{\theta}_{freq} + (1-w)\theta_{prior},$$

where the weight w strikes balances between evidence from data and prior belief.

1. As $n \to \infty$ with all hyper-parameters fixed, δ^* becomes essentially $\hat{\theta}_{freq}$. Hence, choice of priors is of little import when n is large.
2. When n is fixed, choice of prior is important. For example,
   (a) As the prior $\pi(\cdot)$ becomes more concentrated (degenerate), we have $w \to 0$. Then $\delta^* \approx \hat{\theta}_{prior}$.
   (b) As the prior $\pi(\cdot)$ becomes non-informative, we have $w \to 1$. Then $\delta^* \approx \hat{\theta}_{freq}$.

Bayes estimators can also be regarded as shrinkage estimator. Without priors, the usual estimator is $\hat{\theta}_{freq}$. Once we place priors, our new estimator shrinks toward θ_{prior}. More discussion will be given on James-Stein estimator.

Are Bayes estimators unique?

- Bayes estimators may not necessarily be unique in general.
- However, under weak conditions, Bayes estimators are unique, e.g., if the loss function $l(\theta, \delta)$ is squared error loss, or more generally, if it is strictly convex in δ.

  The proof follows easily from that of Corollary 13.1, and hence omitted here.

Can unbiased estimators be Bayes estimators w.r.t. a proper prior?

The answer is NO.

THEOREM **13.2** *No unbiased estimator $\delta(X)$ can be a Bayes solution under L_2-loss unless*

$$r(\pi, \delta) \equiv E[\delta(X) - g(\theta)]^2 = 0. \qquad (E \text{ w.r.t. } X \text{ and } \theta.)$$

Proof. Suppose $\delta(X)$ is a Bayes estimator and is unbiased for estimating $g(\theta)$. Then, we have: (i) $\delta(X) = E[g(\theta)|X]$, a.s. and (ii) $E[\delta(X)|\theta] = g(\theta)$ for all θ. Hence,

$$\begin{aligned} E[g(\theta)\delta(X)] &= E\{\delta(X)E[g(\theta)|X]\} = E[\delta^2(X)] \\ E[g(\theta)\delta(X)] &= E\{g(\theta)E[\delta(X)|\theta]\} = E[g^2(\theta)]. \end{aligned}$$

It follows that $E[\delta(X) - g(\theta)]^2 = E[\delta^2(X)] + E[g^2(\theta)] - 2E[\delta(X)g(\theta)] = 0$.

Example **13.7 Sample means are NOT Bayes estimators**

Let X_i's be i.i.d. $(i = 1, ..., n)$ with $E(X_i) = \theta$ and $var(X_i) = \sigma^2$ (independent of θ), then

$$R(\theta, \bar{X}) = E(\bar{X} - \theta)^2 = \sigma^2/n.$$

For any proper prior,

$$r(\pi, \delta) \equiv ER(\theta, \bar{X}) = \sigma^2/n \neq 0.$$

From the above theorem, $\bar{X}$ is not a Bayes estimator.

Improper prior

In Example 13.1, $\bar{X}$ is the limit of the Bayes estimators as $b \to \infty$. As $b \to \infty$, the prior $\pi(\cdot)$ tends to Lebesgue measure, which is improper prior and non-informative.

Since the Fisher information $I(\theta) = constant$, this is actually the Jeffreys prior. It is easy to check that the posterior distribution calculated from this improper prior is a proper distribution as soon as an observation has been taken. This is not surprising; since X is normally distributed about θ with variance 1, even a single observation provides a good idea of the position of θ.

Connection between Bayes estimation, sufficiency, and likelihood

Let $\mathbf{X} = (X_1, ..., X_n) \sim f(\mathbf{x}|\theta)$ with sufficient statistic T. Its likelihood function is $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$. Given $T = t$, we have

$$f(\mathbf{x}|\theta) = L(\theta|\mathbf{x}) = g(\mathbf{t}|\theta)h(\mathbf{x})$$

For any prior distribution $\pi(\theta)$, the posterior distribution is then

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)d\theta} = \frac{g(\mathbf{t}|\theta)\pi(\theta)}{\int g(\mathbf{t}|\theta)\pi(\theta)d\theta} = \pi(\theta|\mathbf{t}).$$

That is, $\pi(\theta|\mathbf{x})$ depends on $\mathbf{x}$ only through $\mathbf{t}$, and the posterior distribution of θ is the same whether we compute it on the basis of $\mathbf{x}$ or of $\mathbf{t}$. Since Bayesian measures are computed from posterior distributions, we need only focus on functions of a minimal sufficient statistic.

Limits of Bayes estimators

13.4 Empirical Bayes

Procedure

Assume

$$\begin{aligned} X_i|\theta &\sim f(x|\theta), \qquad i=1,..,n \\ \theta|\gamma &\sim \pi(\theta|\gamma). \end{aligned}$$

The hyper-parameter γ can be assigned based on one's own judgements, which could be difficult in practice.

Alternatively, it can be done by empirical Bayes (EB) procedure. In other words, we hope that the data can give us some hints on these hyper-parameters.

1. Compute the marginal pdf of X:
$$m(\mathbf{x}|\gamma) = \int f(\mathbf{x},\theta|\gamma)d\theta = \int f(\mathbf{x}|\theta)\pi(\theta|\gamma)d\theta = \int \prod_{i=1}^{n} f(x_i|\theta)\pi(\theta|\gamma)d\theta.$$

2. Get an MLE $\hat{\gamma}$ for γ
$$\hat{\gamma} = \arg\max_{\gamma} m(\mathbf{x}|\gamma)$$
Other estimation methods include Method of Moment, SURE, etc.

3. Empirical Bayes (EB) estimator minimizes the empirical posterior loss
$$\hat{\delta}^{EB} = \arg\min_{\delta} \int L(\theta,\delta(\mathbf{x}))\pi(\theta|\hat{\gamma}(\mathbf{x}))d\theta. \tag{4.1}$$

Remarks.

1. Denote the Bayes estimator by $\delta^{Bayes} = \delta^{Bayes}(\gamma)$. It can be shown that
$$\hat{\delta}^{EB} = \delta^{Bayes}(\hat{\gamma}).$$

2. EB is effective in constructing estimators performing well under Bayesian and frequentist criteria.

3. EB estimators is more robust against mis-specification of the prior distribution.

13.4.1 Example: EB to univariate Normal mean with known variance

- First we point out that the power of EB really lies in high-dimensional problems (e.g. $p \geq 3$ in Guassian case).
- However, we first give a simple univariate ($p = 1$) normal example on how to find EB estimator.
- This example will be used in higher dimensional Guassian problems later.

Example **13.8 (Continuing from Example 13.1)** *We base our inference on the sufficient statistics $Z = \bar{X}$:*

$$\begin{aligned} Z|\mu &\sim N(\mu, \sigma_0^2), \\ \mu &\sim N(M, A), \qquad (\textit{conjugate prior}) \end{aligned}$$

First assume $M = 0$, but A is unknown.

First the Bayes estimator has been shown to be

$$\hat{\mu}^{Bayes} = \left(1 - \frac{\sigma_0^2}{A + \sigma_0^2}\right) z. \tag{4.2}$$

1. *The EB estimator via MLE is*

$$\hat{\mu}^{EB}_{MLE} = \left(1 - \frac{\sigma_0^2}{z^2}\right)_+ z,$$

where $z_+ = z \vee 0$, the positive part of z.

2. *The EB estimator via MoM is*

$$\hat{\mu}^{EB}_{MLE} = \left(1 - \frac{\sigma_0^2}{z^2}\right) z.$$

Questions:

- The Bayes and EB estimators are both biased, but their variances are smaller (shrunk). What about their MSEs in comparisons with the frequentist mean $\bar{X}$?
- Plot of EB is similar to SCAD, MCP, or weighted LASSO.

Proof. First we show that the marginal pdf of Z is

$$Z|A \sim N(0, \sigma_0^2 + A). \tag{4.3}$$

Proof.

$$\begin{aligned}
f(z|A) &= \int \frac{1}{(2\pi\sigma_0^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma_0^2}(z-\mu)^2\right\} \frac{1}{(2\pi A)^{1/2}} \exp\left\{-\frac{\mu^2}{2A}\right\} d\mu \\
&= \frac{1}{(4\pi^2\sigma_0^2 A)^{1/2}} \int \exp\left\{-\frac{(z-\mu)^2}{2\sigma_0^2}\right\} \exp\left\{-\frac{\mu^2}{2A}\right\} d\mu \\
&= \frac{1}{(4\pi^2\sigma_0^2 A)^{1/2}} \int \exp\left\{-\frac{1}{2}\left(\mu^2\left(\frac{1}{A}+\frac{1}{\sigma_0^2}\right) - 2\mu\frac{z}{\sigma_0^2} + \frac{z^2}{\sigma_0^2}\right)\right\} d\mu \\
&= \frac{1}{(4\pi^2\sigma_0^2 A)^{1/2}} \exp\left\{-\frac{z^2}{2\sigma_0^2}\right\} \int \exp\left\{-\frac{1}{2}\left(\frac{1}{A}+\frac{1}{\sigma_0^2}\right)\left(\mu^2 - 2\mu\frac{Az}{\sigma_0^2+A}\right)\right\} d\mu \\
&= \frac{1}{(4\pi^2\sigma_0^2 A)^{1/2}} \exp\left\{-\frac{z^2}{2\sigma_0^2}\right\} \exp\left\{\frac{1}{2}\left(\frac{\sigma_0^2+A}{A\sigma_0^2}\right)\left(\frac{A}{\sigma_0^2+A}\right)^2 z^2\right\} \\
&\quad \times \int \exp\left\{-\frac{1}{2}\left(\frac{\sigma_0^2+A}{A\sigma_0^2}\right)\left(\mu - \frac{Az}{\sigma_0^2+A}\right)^2\right\} d\mu \\
&= \frac{1}{(4\pi^2\sigma_0^2 A)^{1/2}} \exp\left\{-\frac{z^2}{2\sigma_0^2}\right\} \exp\left\{\left(\frac{z^2}{2\sigma_0^2}\right)\left(\frac{A}{\sigma_0^2+A}\right)\right\} \times \left(\frac{2\pi A\sigma_0^2}{\sigma_0^2+A}\right)^{1/2} \\
&= \frac{1}{(2\pi(\sigma_0^2+A))^{1/2}} \exp\left\{-\frac{z^2}{2(\sigma_0^2+A)}\right\}.
\end{aligned}$$

1. **EB via MLE.**

   The MLE of $\eta = \sigma_0^2 + A$ is

   $$\hat{\eta} = \max\{\sigma_0^2, z^2\}. \tag{4.4}$$

   Thus, an empirical Bayes estimator of μ by MLE is (see the proof below)

   $$\hat{\mu}_{MLE}^{EB} = \left(1 - \frac{\sigma_0^2}{\hat{\eta}}\right) z = \left(1 - \frac{\sigma_0^2}{\max\{\sigma_0^2, z^2\}}\right) z = \left(1 - \frac{\sigma_0^2}{z^2}\right)_+ z.$$

   Proof of (4.4). Let $\eta = \sigma_0^2 + A$, so $\eta \in [\sigma_0^2, \infty)$. We have $L(\eta) \propto \eta^{-1/2}\exp\{-z^2/(2\eta)\}$, and $l(\eta) \propto -\log\eta - z^2/\eta$. So

   $$l'(\eta) \quad \propto \quad -\frac{1}{\eta} + \frac{z^2}{\eta^2} = \frac{1}{\eta^2}(z^2 - \eta)$$

   The solution to $l'(\eta) = 0$ is $\eta_0 = z^2$.

   (a) If $z^2 \in [\sigma_0^2, \infty)$, i.e., $z^2 \geq \sigma_0^2$, the MLE is $\hat{\eta} = z^2$.

   (b) If $z^2 \notin [\sigma_0^2, \infty)$, i.e., $\bar{x}^2 < \sigma_0^2$, the MLE is $\hat{\eta} = \sigma_0^2$. [$l(\mu)$ decreases as $l'(\eta) \leq 0$].

2. **EB via Method of Moment (MoM)**

   From (4.3), $EZ^2 = Var(Z) = \sigma_0^2 + A$. So the moment estimator of $\sigma_0^2 + A$ is z^2, giving

   $$\hat{\mu}_{MM}^{EB} = \left(1 - \frac{\sigma_0^2}{\hat{\eta}}\right) z = \left(1 - \frac{\sigma_0^2}{z^2}\right) z. \quad \blacksquare$$

EB is best suited for large-scale (i.e. high-dim) inference

Empirical Bayes estimation is best suited to large-scale inference, where there are many problems that can be modeled simultaneously in a common way.

Example **13.9 (Empirical Bayes binomial)**

There are K different groups of patients with equal group size n. Each group is given a different treatment for the **same illness**. We have

$$\begin{array}{rl} \text{Group 1}: & X_{11}, \cdots, X_{1n} \sim Ber(p_1) \\ \ldots\ldots & \ldots\ldots\ldots\ldots\ldots\ldots \\ \text{Group K}: & X_{K1}, \cdots, X_{Kn} \sim Ber(p_K) \end{array}$$

where

$$\begin{aligned} X_{kj} &= I\{\text{the } j\text{th patient is successfully treated in } k\text{-th group}\} \sim Bernoulli(p_k) \\ X_k &= \sum_j X_{kj} = \{\text{number of successful treatments in } k\text{-th group}\} \sim Binomial(n, p_k). \end{aligned}$$

Since the groups receive different treatments, we expect different success rates p_k's; however, since we are treating the same illness, these rates should be somewhat related to each other. These considerations suggest the hierarchy

$$\begin{aligned} X_k | p_k &\sim Bin(n, p_k), \\ p_k &\sim Beta(a, b), \qquad k = 1, ..., K, \end{aligned}$$

where the K groups are tied together by the common prior distribution.

As in Example 13.4, the Bayes estimator of p_k under L_2-loss is

$$\hat{p}_k^{Bayes} = E[p_k | x_k] = \frac{a + x_k}{b + n - x_k} = \left(\frac{a+b}{a+b+n}\right)\frac{a}{a+b} + \left(\frac{n}{a+b+n}\right)\frac{x_k}{n}.$$

If a and b are unknown and one can apply EB. First we calculate the marginal distribution

$$\begin{aligned} f(\mathbf{x}|a,b) &= \int_0^1 \cdots \int_0^1 \prod_{i=1}^K \left\{ \binom{n}{x_k} p_k^{x_k}(1-p_k)^{n-x_k} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p_k^{a-1}(1-p_k)^{b-1} dp_k \right\} \\ &= \prod_{k=1}^K \left\{ \binom{n}{x_k} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p_k^{x_k+a-1}(1-p_k)^{n-x_k+b-1} dp_k \right\} \\ &= \prod_{k=1}^K \binom{n}{x_k} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+x_k)\Gamma(n-x_k+b)}{\Gamma(a+b+n)}. \end{aligned}$$

Let $\hat{a}$ and $\hat{b}$ be the MLEs of a and b, which have to be calculated numerically. Therefore,

$$\hat{p}_k^{EB} = E[p_k | x_k] = \frac{\hat{a} + x_k}{\hat{b} + n - x_k} = \left(\frac{\hat{a}+\hat{b}}{\hat{a}+\hat{b}+n}\right)\frac{\hat{a}}{\hat{a}+\hat{b}} + \left(\frac{n}{\hat{a}+\hat{b}+n}\right)\frac{x_k}{n}. \quad \blacksquare$$

Remarks.

1. $\hat{p}_k^{EB}$ is $\hat{p}_k^{Bayes}(\hat{a}, \hat{b})$.

2. Baseball example is a special case of this.

Example **13.10 (Empirical Bayes Multivariate Normal)**

Write $\mu = (\mu_1, ..., \mu_p)'$, $z = (z_1, ..., z_p)'$, and I as the $p \times p$ identify matrix.

Suppose

$$\begin{aligned} z_i|\mu_i &\sim_{ind} N(\mu_i, \sigma_0^2), \qquad i = 1, ..., p \\ \mu_i &\sim_{ind} N(0, A), \qquad \text{(conjugate prior)}. \end{aligned}$$

Equivalently,

$$\begin{aligned} z|\mu &\sim N(\mu, \sigma_0^2 I), \\ \mu &\sim N(M, AI). \end{aligned}$$

Again, we assume $M = 0$, but A is unknown. The posterior distribution of z is

$$\mu|z \sim N(Bz, B\sigma_0^2 I), \qquad \text{where } B = \frac{A}{A + \sigma_0^2}.$$

The Bayes estimator has been shown to be

$$\hat{\mu}^{Bayes} = \left(1 - \frac{\sigma_0^2}{A + \sigma_0^2}\right) z. \tag{4.5}$$

Note A is unknown. We now proceed to find EB estimators. The marginal distribution of z is

$$z \sim N(0, (A + \sigma_0^2) I),$$

i.e.

$$f(z|A) = \frac{1}{(2\pi(\sigma_0^2 + A))^{p/2}} \exp\left\{-\frac{\|z\|^2}{2(\sigma_0^2 + A)}\right\}.$$

1. An MLE of A is ?????????.

   Hence, The EB estimator via MLE is

   $$\hat{\mu}^{EB}_{MLE} = ...$$

2. An MoM estimator of A is ....

   Hence, the EB estimator via MoM is

   $$\hat{\mu}^{EB}_{MLE} = ...$$

13.5 Empirical Bayes and the James-Stein Estimator

1. MLEs were in common use for more than a century, and are still in use.
2. Charles Stein (1955) showed that MLEs for Gaussian models were inadmissible beyond simple 1- or 2-dimensional situations.
3. Stein-type estimators have pointed the way toward a radically different empirical Bayes approach to high-dimensional statistical inference.
4. Empirical Bayes ideas are powerful for estimation, testing, and prediction.
5. Empirical Bayes estimation is best suited to large-scale inference, where many problems can be modelled simultaneously in a common way.
6. Many contemporary techniques in large-dimensional inference, such as LASSO, can trace its roots back to James-Stein estimations.

A bit of history

Although the connection was not immediately recognized, Stein's work was half of an energetic post-war empirical Bayes initiative. The other half, explicitly named "empirical Bayes" by its principal developer Herbert Robbins, was less shocking but more general in scope, aiming to show how frequentists could achieve full Bayesian efficiency in large-scale parallel studies. Large-scale parallel studies were rare in the 1950s, however, and Robbins's theory did not have the applied impact of Stein's shrinkage estimators, which are useful in much smaller data sets. All of this has changed in the 21st century. New scientific technologies, epitomized by the microarray, routinely produce studies of thousands of parallel cases - well-suited for the Robbins point of view.

Stein's theory concerns estimation, whereas the Robbins branch of empirical Bayes allows for hypothesis testing, that is, for situations where many or most of the true effects pile up at a specific point, usually called 0. Empirical Bayes theory blurs the distinction between estimation and testing as well as between frequentist and Bayesian methods.

Model

Suppose

$$\begin{aligned} z_i|\mu_i &\sim_{ind} N(\mu_i, \sigma_0^2), \qquad i = 1, ..., p \\ \mu_i &\sim_{ind} N(0, A), \qquad \text{(conjugate prior)}. \end{aligned}$$

Write $\mu = (\mu_1, ..., \mu_p)'$, $z = (z_1, ..., z_p)'$, and I as the $p \times p$ identify matrix. Then,

$$\begin{aligned} z|\mu &\sim N(\mu, \sigma_0^2 I), \\ \mu &\sim N(0, AI). \end{aligned}$$

The posterior distribution and the marginal distribution of z are

$$\mu|z \sim N(Bz, B\sigma_0^2 I), \qquad \text{where } B = \frac{A}{A + \sigma_0^2}.$$

and

$$z \sim N(0, (A + \sigma_0^2)I),$$

which have been verified component-wise before.

Bayes estimator

So the Bayes estimator of μ is

$$\hat{\mu}^{Bayes} = Bz = \left(1 - \frac{\sigma_0^2}{A + \sigma_0^2}\right) z. \tag{5.6}$$

Theorem **13.3**

$$\begin{aligned} R(\mu, \hat{\mu}^{Bayes}) &= p\sigma_0^2 B^2 + (1-B)^2 \|\mu\|^2, \\ r(\mu, \hat{\mu}^{Bayes}) &= p\sigma_0^2 B. \quad \blacksquare \end{aligned}$$

Proof.

$$\begin{aligned} R(\mu, \hat{\mu}^{Bayes}) &= E_\mu \|\hat{\mu}^{Bayes} - \mu\|^2 = Var(\hat{\mu}^{Bayes}) + Bias^2 = pB^2\sigma_0^2 + (1-B)^2\|\mu\|^2, \\ r(\mu, \hat{\mu}^{Bayes}) &= ER(\mu, \hat{\mu}^{Bayes}) = B^2 p\sigma_0^2 + (1-B)^2 E\|\mu\|^2 = B^2 p\sigma_0^2 + (1-B)^2 pA \\ &= p\left(\frac{A^2\sigma_0^2}{(A+\sigma_0^2)^2} + \frac{A\sigma_0^2}{(A+\sigma_0^2)^2}\right) = p\sigma_0^2\left(\frac{A}{A+\sigma_0^2}\right) = p\sigma_0^2 B. \quad \blacksquare \end{aligned}$$

MLE

The obvious estimator of μ is the MLE $\hat{\mu}^{MLE} = z$. Its risk and Bayes risks are

$$\begin{aligned} R^{(MLE)}(\mu) &= \sum_{i=1}^{p} E_\mu (z_i - \mu_i)^2 = \sum_{i=1}^{p} V_\mu(z_i) = p\sigma_0^2, \\ r^{(MLE)}(\mu) &= ER^{(MLE)}(\mu) = p\sigma_0^2. \end{aligned}$$

James-Stein Estimation = Empirical Bayes Estimation

Since A is assumed unknown, we can use empirical Bayes rules below. Note that

$$\|z\|^2 = \sum_{i=1}^{p} z_i^2 \sim (A+\sigma_0^2)\chi_p^2.$$

It can be shown that

$$E\left(\frac{A+\sigma_0^2}{\|z\|^2}\right) = \frac{1}{p-2}. \tag{5.7}$$

Thus, if we replace $\sigma_0^2/(A+\sigma_0^2)$ in (5.6) by an unbiased estimator $(p-2)\sigma_0^2/\|z\|^2$, we have

$$\hat{\mu}^{JS} = \left(1 - \frac{(p-2)\sigma_0^2}{\|z\|^2}\right) z,$$

the James-Stein estimator. It was discovered by Stein (1956) and later shown by James and Stein (1961) to have a smaller mean squared error than the MLE z for all μ. Its empirical Bayes derivation can be found in Efron and Morris (1972).

Proof of (5.7). If $Y \sim Gamma(a,b)$, then

$$EY^{-1} = \int_0^\infty y^{-1}\frac{b^a}{\Gamma(a)}y^{a-1}e^{-by}dy = \frac{b\Gamma(a-1)}{\Gamma(a)}\int_0^\infty \frac{b^{a-1}}{\Gamma(a-1)}y^{(a-1)-1}e^{-by}dy = \frac{b}{a-1}.$$

Since $\|z\|^2/(A+\sigma_0^2) \sim \chi_p^2 = Gamma(a,b)$, with $(a,b) = (p/2, 1/2)$, we have

$$E\left(\frac{A+\sigma_0^2}{\|z\|^2}\right) = \frac{1/2}{p/2-1} = \frac{1}{p-2}, \qquad \Longrightarrow \qquad E\left(\frac{(p-2)\sigma_0^2}{\|z\|^2}\right) = \frac{\sigma_0^2}{A+\sigma_0^2}. \quad \blacksquare$$

Since the James-Stein estimator (or any EB estimator) cannot attain as small a Bayes risk as the Bayes estimator, it is of interest to compare their risks, which, in effect, tell us the penalty we are paying for estimating A.

Stein Unbiased Risk Estimation (SURE)

The Bayes risk $r(\theta, \delta)$ gives a criterion for selecting optimal estimators, but it is unknown.

Stein provided an unbiased risk estimation to $r(\theta, \delta)$, referred to as SURE.

Theorem **13.4 (SURE)**

Let $X \sim N_p(\mu, \sigma^2 I)$, and let $\delta(x) = x - g(x)$ be an estimator of μ. Then

$$\begin{aligned} E_\mu\|\delta(x) - \mu\|^2 &= p\sigma^2 + E_\mu\|g(x)\|^2 - 2\sigma^2 \sum_{i=1}^{p} E_\mu \frac{dg_i}{dx_i}. \\ (MSE &= Var + Bias^2 - 2Cov.) \end{aligned}$$

Hence, Stein's unbiased risk estimate is

$$SURE(\delta(x)) = p\sigma^2 + \|g(x)\|^2 - 2\sigma^2 \sum_{i=1}^{p} \frac{dg_i}{dx_i}. \quad \blacksquare$$

Remarks.

1. One can use SURE (instead of Bayes risk) as a criterion in selecting estimators.
2. Suppose that an EB estimator is indexed by λ:

   $$\delta_\lambda = (1 + \lambda)^{-1} z.$$

   One could use SURE to find the optimal $\hat{\lambda}^{SURE}$.
3. SURE can be used to find the tuning parameters (soft-thresholding level).
   See Donoho and Johnston's work in 1990's.
4. One can obtain general expression for the Bayes risk for exponential family; see Theorem 3.5 of Lehmann and Casella.

   In the normal case, we can get an unbiased estimator of the risk for a fairly wide class of estimators.

Proof.

$$\begin{aligned}
E_\mu\|\delta(x)-\mu\|^2 &= E_\mu\|(x-\mu)-g(x)\|^2 \\
&= E_\mu\|x-\mu\|^2 + E_\mu\|g(x)\|^2 - 2E_\mu g(x)^T(x-\mu) \\
&= p\sigma^2 + E_\mu\|g(x)\|^2 - 2E_\mu g(x)^T(x-\mu) \\
&= p\sigma^2 + E_\mu\|g(x)\|^2 - 2Cov_\mu(g(x),x) \qquad (5.8) \\
&= p\sigma^2 + E_\mu\|g(x)\|^2 - 2\sigma^2\sum_{i=1}^{p} E_\mu \frac{dg_i}{dx_i}. \qquad (5.9) \\
(MSE &= Var + Bias^2 - 2Cov.)
\end{aligned}$$

It suffices to show the step from (5.8) to (5.9). For simplicity, we only prove it for $p = 2$. For $g(x) = (g_1(x), g_2(x))^T$, we have

$$E_\mu g(x)^T(x-\mu) = E_\mu g_1(x_1,x_2)(x_1-\mu_1) + E_\mu g_2(x_1,x_2)(x_2-\mu_2).$$

Using integration by parts and $\phi'(t) = -t\phi(t)$, we have

$$\begin{aligned}
& E_\mu g(x_1,x_2)(x_1-\mu_1) \\
= & \int\int \frac{1}{\sigma}\phi\left(\frac{x_1-\mu_1}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x_2-\mu_2}{\sigma}\right) g_1(x_1,x_2)(x_1-\mu_1)dx_1dx_2 \\
= & \int\int (-\sigma)\phi'\left(\frac{x_1-\mu_1}{\sigma}\right)\phi\left(\frac{x_2-\mu_2}{\sigma}\right) g_1(x_1,x_2)dx_1dx_2 \\
= & -\sigma\int_{x_2\in R}\phi\left(\frac{x_2-\mu_2}{\sigma}\right)\int_{x_1\in R}\phi'\left(\frac{x_1-\mu_1}{\sigma}\right) g_1(x_1,x_2)dx_1dx_2 \\
= & -\sigma^2\int_{x_2\in R}\phi\left(\frac{x_2-\mu_2}{\sigma}\right)\int_{x_1\in R} g_1(x_1,x_2)d\phi\left(\frac{x_1-\mu_1}{\sigma}\right)dx_2 \\
= & \sigma^2\int_{x_2\in R}\phi\left(\frac{x_2-\mu_2}{\sigma}\right)\int_{x_1\in R}\frac{\partial g_1(x_1,x_2)}{\partial x_1}\phi\left(\frac{x_1-\mu_1}{\sigma}\right)dx_1dx_2 \\
= & \sigma^2\int_{x_1\in R}\int_{x_2\in R}\frac{\partial g_1(x_1,x_2)}{\partial x_1}\phi\left(\frac{x_1-\mu_1}{\sigma}\right)\phi\left(\frac{x_2-\mu_2}{\sigma}\right)dx_1dx_2 \\
= & \sigma^2 E_\mu\frac{dg_1}{dx_1}.
\end{aligned}$$

Similarly, $E_\mu g(x_1,x_2)(x_2-\mu_2) = \sigma^2 E_\mu \frac{dg_2}{dx_2}$. ∎

Theorem **13.5** *For JS estimator* $\hat{\mu}^{JS} = \left(1 - \frac{(p-2)\sigma_0^2}{\|z\|^2}\right) z$, *we have*

$$\begin{aligned}
R(\mu, \hat{\mu}^{JS}) &= p\sigma^2 - (p-2)^2\sigma^4 E_\mu\left(\frac{1}{\|z\|^2}\right) \\
SURE(\hat{\mu}^{JS}) &= p\sigma^2 - \frac{(p-2)^2\sigma^4}{\|z\|^2} \\
r(\mu, \hat{\mu}^{JS}) &= p\sigma^2 - (p-2)^2\sigma^4 E\left(\frac{1}{\|z\|^2}\right) \\
&= p\sigma^2 - \frac{(p-2)\sigma^4}{\sigma^2 + A} \qquad \textit{(by (5.7))} \\
&= p\frac{\sigma^2 A}{\sigma^2 + A} + 2\frac{\sigma^4}{\sigma^2 + A}
\end{aligned}$$

Note: $E_\mu\left(\|z\|^{-2}\right) \neq E\left(\|z\|^{-2}\right)$.

Proof. Here $\hat{\mu}^{JS} = z - g(z)$ with $g(z) = (p-2)\sigma_0^2 z/\|z\|^2$. Hence,

$$\begin{aligned}
R(\mu, \hat{\mu}^{JS}) &= p\sigma^2 + E_\mu \frac{(p-2)^2\sigma_0^4\|z\|^2}{\|z\|^4} - 2\sigma_0^2 \sum_{i=1}^{p} E_\mu \left(\frac{\partial}{\partial z_i} \frac{(p-2)\sigma_0^2 z_i}{\|z\|^2} \right) \\
&= p\sigma^2 + (p-2)^2\sigma_0^4 E_\mu \frac{1}{\|z\|^2} - 2(p-2)\sigma_0^4 \sum_{i=1}^{p} E_\mu \left(\frac{\|z\|^2 - 2z_i^2}{\|z\|^4} \right) \\
&= p\sigma^2 + (p-2)^2\sigma_0^4 E_\mu \frac{1}{\|z\|^2} - 2(p-2)\sigma_0^4 E_\mu \left(\frac{p\|z\|^2 - 2\|z\|^2}{\|z\|^4} \right) \\
&= p\sigma^2 - (p-2)^2\sigma_0^4 E_\mu \frac{1}{\|z\|^2}.
\end{aligned}$$

And

$$\begin{aligned}
r(\mu, \hat{\mu}^{JS}) &= ER(\mu, \hat{\mu}^{JS}) \\
&= p\sigma^2 - (p-2)^2\sigma^4 E\left(\frac{1}{\|z\|^2}\right) \\
&= p\sigma^2 - \frac{(p-2)\sigma^4}{\sigma^2 + A} \qquad \text{(by (5.7))} \\
&= p\frac{\sigma^2 A}{\sigma^2 + A} + 2\frac{\sigma^4}{\sigma^2 + A} \quad \blacksquare
\end{aligned}$$

THEOREM **13.6** *It is known*

$$r(\mu, \hat{\mu}^{Bayes}) = p\frac{\sigma^2 A}{\sigma^2 + A}$$

Hence the relative risk is

$$\frac{r(\mu, \hat{\mu}^{JS})}{r(\mu, \hat{\mu}^{Bayes})} = 1 + \frac{2\frac{\sigma^4}{\sigma^2+A}}{p\frac{\sigma^2 A}{\sigma^2+A}} = 1 + \frac{2\sigma^2}{pA}$$

e.g. for $p = 10$ and $A = \sigma^2$, $r(\mu, \hat{\mu}^{JS})$ is only 20% greater than the true Bayes risk.

13.6 Comparisons: Bayes, EB=JS, and MLE

13.6.1 James-Stein is everywhere better than MLE for $p \geq 3$ (1961)

We have shown

$$r(\mu, \hat{\mu}^{Bayes}) < r(\mu, \hat{\mu}^{JS}) < r(\mu, \hat{\mu}^{MLE}),$$

The shock the James-Stein estimator provided the statistical world didn't come from the above average risk comparisons. These are based on the zero-centric Bayesian model, where the MLE, which doesn't favor values of μ near 0, might be expected to be bested.

THEOREM **13.7** *For $p \geq 3$, we have*

$$R(\mu, \hat{\mu}^{Bayes}) < R(\mu, \hat{\mu}^{JS}) < R(\mu, \hat{\mu}^{MLE}),$$

for every choice of μ. ∎

Remarks

- This result is frequentist rather that Bayesian:

  $\hat{\mu}^{JS}$ is superior no matter what one's prior beliefs about μ may be.
- The apparent uniform inferiority of $\hat{\mu}^{MLE}$ was a cause for alarm, since versions of $\hat{\mu}^{MLE}$ dominate popular statistical techniques such as linear regression,
- The fact that linear regression applications continue unabated reflects some virtues of $\hat{\mu}^{MLE}$.

13.7 Using SURE to find Estimators

A standard application of SURE is to choose a parametric form for an estimator, and then optimize the values of the parameters to minimize the risk estimate. This technique has been applied in several settings. For example, a variant of the James-Stein estimator can be derived by finding the optimal shrinkage estimator. The technique has also been used by Donoho and Johnstone to determine the optimal shrinkage factor in a wavelet denoising setting.

Recall

$$\hat{\mu}^{Bayes} = \left(1 - \frac{1}{A+1}\right) z = z - g(z)$$

where $g(z) = z/(A+1)$ and $\partial g_i(z)/\partial z_i = 1/(A+1)$. From Theorem 13.4, we have

$$\text{SURE}(A) = p\sigma^2 + \|g(x)\|^2 - 2\sigma^2 \sum_{i=1}^{p} \frac{dg_i}{dx_i} = p\sigma^2 + \frac{\|z\|^2}{(A+1)^2} - \frac{2p\sigma^2}{(A+1)}$$

A SURE estimate of A is

$$\hat{A} = \arg\min_{A \geq 0} \text{SURE}(A),$$

which can be found by setting

$$\frac{\partial \text{SURE}(A)}{\partial A} = -\frac{2\|z\|^2}{(A+1)^3} + \frac{2p\sigma^2}{(A+1)^2} = 0.$$

Thus, a SURE estimate of $\dfrac{1}{(1+A)}$ is $\dfrac{1}{(1+\hat{A})} = \dfrac{p\sigma^2}{\|z\|^2}$. Hence, we obtain a SURE type estimator of μ as

$$\hat{\mu}^{SURE} = \left(1 - \frac{p\sigma^2}{\|z\|^2}\right) z.$$

Note that this is very similar to the James-Stein estimator $\hat{\mu}^{JS}$.

13.8 James-Stein Estimator, shrinking to non-zero point

The above James-Stein estimator shrinks each observed value z_i toward 0. We don't have to take 0 as the preferred shrinking point. Now suppose

$$\begin{aligned} z_i|\mu_i &\sim N(\mu_i, \sigma_0^2), \qquad i = 1, ..., p \\ \mu_i &\sim N(M, A), \qquad \text{(conjugate prior)}, \end{aligned}$$

the (μ_i, z_i) pairs being independent of each other. We can write this compactly as

$$\begin{aligned} z|\mu &\sim N(\mu, \sigma_0^2 I), \\ \mu &\sim N(M, AI). \end{aligned}$$

The posterior distribution is

$$\mu_i|z_i \sim N((1-B)M + Bz_i, B\sigma_0^2 I), \qquad [B = A/(A+\sigma_0^2)],$$

and the marginal distribution of z is

$$z_i \sim N(M, A + \sigma_0^2).$$

Proof. Let $\eta = \theta - M$. Then, similar to earlier derivation, we have

$$\begin{aligned} f(z|b) &= \int \frac{1}{(2\pi A)^{1/2}} \exp\left\{-\frac{1}{2A}(z-\theta)^2\right\} \frac{1}{(2\pi b^2)^{1/2}} \exp\left\{-\frac{(\theta - M)^2}{2b^2}\right\} d\theta \\ &= \int \frac{1}{(2\pi A)^{1/2}} \exp\left\{-\frac{1}{2A}(z - M - \eta)^2\right\} \frac{1}{(2\pi b^2)^{1/2}} \exp\left\{-\frac{\eta^2}{2b^2}\right\} d\eta \\ &= \frac{1}{(2\pi(A+b^2))^{1/2}} \exp\left\{-\frac{(z-M)^2}{2(A+b^2)}\right\}. \quad \blacksquare \end{aligned}$$

Bayes estimator

The Bayes estimator of μ is

$$\hat{\mu}^{Bayes} = E(\mu|z) = M + B(z - M) = M + \left(1 - \frac{\sigma_0^2}{A + \sigma_0^2}\right)(z - M). \tag{8.10}$$

THEOREM **13.8**

$$\begin{aligned} R(\mu, \hat{\mu}^{Bayes}) &= E_\mu \|\hat{\mu}^{Bayes} - \mu\|^2 = Var(\hat{\mu}^{Bayes}) + Bias^2 \\ &= B^2 p \sigma_0^2 + (1 - B)^2 \|M - \mu\|^2 \\ r(\mu, \hat{\mu}^{JS}) &= ER(\mu, \hat{\mu}^{Bayes}) = B^2 p \sigma_0^2 + (1 - B)^2 E\|M - \mu\|^2 \\ &= B^2 p \sigma_0^2 + (1 - B)^2 pA = p\left(\frac{A^2\sigma^2}{(A + \sigma^2)^2} + \frac{A\sigma^2}{(A + \sigma^2)^2}\right) \\ &= p\sigma^2 \left(\frac{A}{A + \sigma^2}\right). \quad \blacksquare \end{aligned}$$

James-Stein Estimation

Since A is assumed unknown, we can use empirical Bayes rules below. Note that

$$\|z - \bar{z}\|^2 = \sum_{i=1}^{p} (z_i - \bar{z})^2 \sim (A + \sigma_0^2)\chi_{p-1}^2.$$

Similarly to the previous derivations, we have

$$E\left(\frac{(p-3)\sigma_0^2}{\|z - \bar{z}\|^2}\right) = \frac{\sigma_0^2}{A + \sigma_0^2}. \tag{8.11}$$

Thus, if we replace $1/(A+1)$ in (8.10) by an unbiased estimator $(p-3)/\|z - \bar{z}\|^2$, we have

$$\hat{\mu}^{JS} = \left(1 - \frac{(p-3)\sigma_0^2}{\|z - \bar{z}\|^2}\right) z,$$

the James-Stein estimator.

Theorem 13.7 continues to hold, except now $p \geq 4$.

THEOREM **13.9** *For $p \geq 4$, we have*

$$R(\mu, \hat{\mu}^{Bayes}) < R(\mu, \hat{\mu}^{JS}) < R(\mu, \hat{\mu}^{MLE}),$$

for every choice of μ. ∎