Chapter 3

Point Estimation

3.1 Maximum Likelihood Estimator

3.1.1 Point estimator and point estimate

1• Difference between point estimator and point estimate

- Let the pdf of an r.v. X be $f(x;\boldsymbol{\theta})$ with an unknown parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, where $\boldsymbol{\Theta}$ denotes the corresponding parameter space.
- Thus, we have a family of densities $\{f(x;\boldsymbol{\theta})\colon\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$.
- We need to select one member from the family as the pdf of X.
- This is equivalent to estimating the parameter vector $\boldsymbol{\theta}$.
- To this end, we take a random sample $X_1, \dots, X_n$ from a population with the pdf $f(x;\boldsymbol{\theta})$.

1.1• Remarks

— In Chapters 1–2, we denote the pdf of an r.v. X by $f(x)$, while starting from Chapter 3 we denote it by $f(x;\boldsymbol{\theta})$ to emphasize its dependence on the parameter vector $\boldsymbol{\theta}$.

— For example, if $X \sim N(\mu, \sigma^2)$, we have

$$f(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad \text{where} \quad \boldsymbol{\theta} = (\mu, \sigma^2)^{\top}.$$

— Let $2.18, 2.76, 1.80, 1.73, 1.13, 1.85, 2.02, 2.69, 1.66, 2.59$ be a random sample of size 10 from $N(\mu, \sigma^2)$, how to estimate μ and σ^2? This is the main topic of Chapter 3.

— An advanced reference book is: Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation* (2-nd ed.). Springer, New York.

Definition 3.1 (A statistic). A function of one or more r.v.'s that does not depend on the unknown parameter vector is called a *statistic*. ‖

1.2• Comparison of Definition 3.1 with Definition 2.1 in §2.2

— In Definition 2.1, it is assumed that $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} F(x)$; i.e., $\{X_i\}_{i=1}^n$ is a random sample from $F(x)$.

— In fact, the assumption of independence is not necessary.

— In other words, $\{X_i\}_{i=1}^n$ could be correlated or dependent.

1.3• Definitions of a point estimator and a point estimate

— If a statistic $Y = \varphi(\mathbf{x})$ is used to estimate the parameter θ, where $\mathbf{x} = (X_1, \ldots, X_n)^{\top}$, then the statistic is called a *point estimator* of θ, where Y is a *random variable*.

— If the observations of $X_1, \ldots, X_n$ are $x_1, \ldots, x_n$, then $y = \varphi(\boldsymbol{x})$ is called a *point estimate* of θ, where y is a *real number* and $\boldsymbol{x} = (x_1, \ldots, x_n)^{\top}$.

1.4• Illustration examples

— For example, $\bar{X} = (1/n)\sum_{i=1}^n X_i$ is a point estimator of $\mu = E(X)$ and $\bar{x} = (1/n)\sum_{i=1}^n x_i$ is a point estimate of μ.

— Similarly, $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ is a point estimator of $\sigma^2 = \text{Var}(X)$ and $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$ is a point estimate of σ^2.

1.5• How to understand a point estimator?

— A point estimator is a random variable.

— A point estimator is always related with the estimation of θ. For instance, $\bar{X}$ is a point estimator of $\mu = E(X)$ but $\sum_{i=1}^n X_i$ is not.

— Point estimator is not unique. For example, S^2 is an unbiased estimator of $\sigma^2 = \mathrm{Var}(X)$ while $\hat{\sigma}^2 = (1/n)\sum_{i=1}^n (x_i - \bar{x})^2$ is the moment estimator of σ^2 for any population. In particular, $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2 for the normal population.

3.1.2 Joint density and likelihood function

2• Difference between joint pdf and likelihood function

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the unknown parameter vector and $\boldsymbol{\Theta}$ is the parameter space.

- Let $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ be observations of $\mathbf{x} = (X_1, \ldots, X_n)^\top$, then the *joint density* of $\mathbf{x}$ is $f(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^n f(x_i; \boldsymbol{\theta})$.

- Since $\boldsymbol{x}$ has been observed and its components are therefore fixed real numbers, we regard $f(\boldsymbol{x}; \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, and define

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^n f(x_i; \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

  as the *likelihood function* of the random sample $\mathbf{x}$.

- Alternatively, $L(\boldsymbol{\theta})$ is also called the likelihood function of $\boldsymbol{\theta}$.

2.1• How to understand the likelihood function $L(\boldsymbol{\theta})$?

— The joint density $f(\boldsymbol{x}; \boldsymbol{\theta})$ is a term of Probability while the likelihood function $L(\boldsymbol{\theta})$ is a term of Statistics.

— $f(\boldsymbol{x}; \boldsymbol{\theta})$ emphasizes $\boldsymbol{x}$ while $L(\boldsymbol{\theta})$ emphasizes $\boldsymbol{\theta}$.

— In statistics, in general, $L(\boldsymbol{\theta})$ is concave. That is $\nabla^2 L(\boldsymbol{\theta}) \leqslant 0$. In particular, when $\boldsymbol{\theta}$ is one-dimensional, $L(\theta)$ is concave iff $L''(\theta) \leqslant 0$.

3• The log-likelihood function

- In practice, the natural logarithm of $L(\boldsymbol{\theta})$, called the *log-likelihood*, is mathematically much convenient to work with.

- We define $\ell(\boldsymbol{\theta}) \ \hat{=}\ \log\{L(\boldsymbol{\theta})\} = \sum_{i=1}^{n} \log\{f(x_i;\boldsymbol{\theta})\}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
- Note that there is no loss of information in using $\ell(\boldsymbol{\theta})$ instead of $L(\boldsymbol{\theta})$ because $\log(\cdot)$ is a monotonic increasing function.

3.1.3 Maximum likelihood estimator and maximum likelihood estimate

4• Definition

- Suppose that a statistic

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_p \end{pmatrix} = \begin{pmatrix} u_1(\mathbf{x}) \\ \vdots \\ u_p(\mathbf{x}) \end{pmatrix} \ \hat{=}\ \boldsymbol{u}(\mathbf{x})$$

satisfies

$$L(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}).$$

- Statistically, we can equivalently write above equation as

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}),$$

where "arg" is the abbreviation of "argument".

- Then $\hat{\boldsymbol{\theta}} = \boldsymbol{u}(\mathbf{x})$ is called the *maximum likelihood estimator* (MLE) of $\boldsymbol{\theta}$ and $\boldsymbol{u}(\boldsymbol{x})$ is called a *maximum likelihood estimate* (mle) of $\boldsymbol{\theta}$.

4.1• Remarks

— Note that $L(\boldsymbol{\theta})$ and $\ell(\boldsymbol{\theta})$ share their maxima at the same value of $\boldsymbol{\theta}$, and it is usually easier to find the maximum of $\ell(\boldsymbol{\theta})$.

— In general, the MLE $\hat{\boldsymbol{\theta}}$ is the solution to the score equation

$$\nabla \ell(\boldsymbol{\theta}) \ \hat{=}\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p} \end{pmatrix} = \mathbf{0}_p. \tag{3.1}$$

— There is no guarantee that the MLE exists or if it does whether it is unique.

— Consider the special case of $p = 1$. Then (3.1) becomes

$$\ell'(\theta) = \frac{\mathrm{d}\ell(\theta)}{\mathrm{d}\theta} = 0.$$

If $L(\theta)$ is a monotonic function of θ, then the MLE $\hat{\theta}$ locates at the boundary of Θ or does not exist.

4.2• Stationary point, saddle point and critical point

— A point c satisfying $\varphi'(c) \;\hat{=}\; \varphi'(x)|_{x=c} = 0$ is called a *stationary point* of $\varphi(x)$.

— For instance, $L(\theta)$ and $\ell(\theta)$ have the same stationary points since

$$\ell'(\theta^*) = \ell'(\theta)|_{\theta=\theta^*} = \left.\frac{L'(\theta)}{L(\theta)}\right|_{\theta=\theta^*} = 0;$$

i.e., $\ell'(\theta^*) = 0$ iff $L'(\theta^*) = 0$.

— It is possible for c to be a local rather than a global minimum or maximum or even to be a *saddle point*. For example, $\varphi(x) = x^3$ has a saddle point at 0.

— Let $\varphi(x)$ be defined on the closed interval $[a, b]$. Two endpoints a, b and any stationary points c are known as *critical points* of $\varphi(x)$.

5• Unrestricted MLE

Example 3.1 (Bernoulli distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. Find the MLE of θ.

Solution. The parameter space $\Theta = \{\theta\colon\ 0 < \theta < 1\} = (0, 1)$. Note that the pmf of X_i is given by

X_i	0	1
$p(x_i;\theta) = \Pr(X_i = x_i)$	$1-\theta$	θ

.

Thus, we have $p(x_i;\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$, $x_i = 0, 1$. The joint pmf is

$$p(\boldsymbol{x};\theta) = \prod_{i=1}^{n} p(x_i;\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i},$$

so that the likelihood function is given by

$$L(\theta) = \theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}, \quad 0 < \theta < 1,$$

where $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$. Now

$$\ell(\theta) = n\bar{x}\log(\theta) + (n - n\bar{x})\log(1-\theta)$$

and

$$\ell'(\theta) = \frac{n\bar{x}}{\theta} - \frac{n-n\bar{x}}{1-\theta}. \tag{3.2}$$

Solving $\ell'(\theta) = 0$ for θ, we obtain the solution $\theta = \bar{x}$. To verify that it maximizes $\ell(\theta)$ or $L(\theta)$, we have two alternative methods.

Method I: To check that the second derivative of $\ell(\theta)$ evaluated at $\bar{x}$ is strictly negative; i.e., $\ell''(\bar{x}) < 0$. Now, for any $\theta \in (0, 1)$, uniformly we have

$$\frac{\mathrm{d}^2\ell(\theta)}{\mathrm{d}\theta^2} = -\left\{\frac{n\bar{x}}{\theta^2} + \frac{n-n\bar{x}}{(1-\theta)^2}\right\} < 0.$$

Therefore,

$$\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the MLE of θ and $\bar{x}$ is the mle of θ.

Method II: To check that $\ell'(\theta) > 0$ when $\theta < \bar{x}$ and $\ell'(\theta) < 0$ when $\theta > \bar{x}$. From (3.2), it is easy to check them.

In general, Method II is more convenient than Method I. However, in statistical practice, neither Method I nor Method II is necessary. ‖

Example 3.2 (Normal distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Find the MLEs of μ and σ^2.

Solution. Let $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$. The parameter space is

$$\begin{aligned} \boldsymbol{\Theta} &= \{(\mu, \sigma^2)^\top \colon -\infty < \mu < \infty,\ \sigma^2 > 0\} \\ &= (-\infty, \infty) \times (0, \infty) = \mathbb{R} \times \mathbb{R}_+, \end{aligned}$$

and the likelihood function is

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

Then

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

By differentiating $\ell(\mu, \sigma^2)$ with respect to μ and σ^2 and letting them equal zeros, we have

$$\begin{aligned} \frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} &= \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2} = 0, \\ \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^4} = 0. \end{aligned}$$

The solutions are $\mu = \bar{x}$ and $\sigma^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$. Therefore,

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

are the MLEs of μ and σ^2, respectively. ‖

Example 3.3 (Uniform distribution with one unknown endpoint). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta]$, where $\theta > 0$. Find the MLE of θ.

Solution. The parameter space is $\Theta = (0, \infty) = \mathbb{R}_+$. The joint density of $\mathbf{x} = (X_1, \ldots, X_n)^\top$ is

$$f(\boldsymbol{x}; \theta) = \begin{cases} \dfrac{1}{\theta^n}, & \text{if } \ 0 < x_i \leqslant \theta, \ i = 1, \ldots, n, \\ 0, & \text{elsewhere.} \end{cases}$$

Then, the likelihood function is given by

$$L(\theta) = \begin{cases} \dfrac{1}{\theta^n}, & \text{if } \ \theta \geqslant x_{(n)} \ \hat{=} \ \max(x_1, \ldots, x_n), \\ 0, & \text{elsewhere.} \end{cases} \tag{3.3}$$

Note that $L(\theta)$ is a monotone and decreasing function of θ when $\theta \in [x_{(n)}, \infty)$ as shown in Figure 3.1, and arrives its maximum at $\theta = x_{(n)}$, thus $\hat{\theta} = X_{(n)}$ is the MLE of θ.

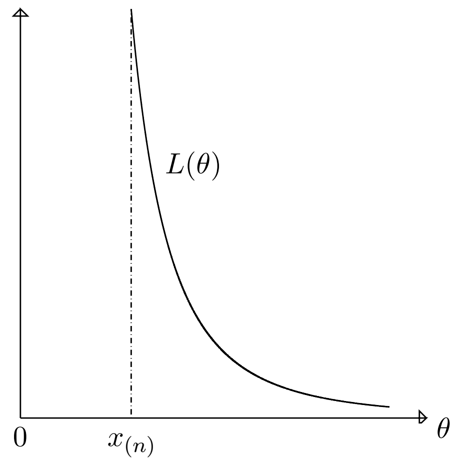

Figure 3.1 The likelihood function $L(\theta)$ defined by (3.3) is a monotone and decreasing function of θ when $\theta \in [x_{(n)}, \infty)$. ||

5.1• Difference between maximum and supremum

— In Example 3.3, if we assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$, then the likelihood function (3.3) becomes

$$L(\theta) = \begin{cases} \dfrac{1}{\theta^n}, & \text{if } \theta > x_{(n)}, \\ 0, & \text{elsewhere.} \end{cases}$$

— Then, the MLE of θ does not exist.

— However, we can obtain

$$\sup L(\theta) = 1/x_{(n)}^n,$$

where "sup" is the abbreviation of "supremum".

— We should realize the difference between "max/min" and "sup/inf", where "inf" is the abbreviation of "infimum".

Example 3.4 (Uniform distribution with two unknown endpoints). Let X_1, $\ldots, X_n \overset{\text{iid}}{\sim} U[\theta - 0.5, \theta + 0.5]$, where $-\infty < \theta < \infty$. Find the MLE of θ.

Solution. The parameter space $\Theta = \mathbb{R}$. The joint density of $\mathbf{x}$ is

$$f(\boldsymbol{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} I_{[\theta-0.5,\ \theta+0.5]}(x_i)$$

so that the likelihood is given by

$$\begin{aligned} L(\theta) &= I_{[x_{(n)}-0.5,\ x_{(1)}+0.5]}(\theta) \qquad (3.4) \\ &= \begin{cases} 1, & \text{if } \ x_{(n)} - 0.5 \leqslant \theta \leqslant x_{(1)} + 0.5, \\ 0, & \text{elsewhere.} \end{cases} \end{aligned}$$

In fact, (3.4) follows since $\prod_{i=1}^{n} I_{[\theta-0.5,\ \theta+0.5]}(x_i)$ is unity iff all $x_1, \ldots, x_n$ are in the interval $[\theta - 0.5,\ \theta + 0.5]$, which is true iff $\theta - 0.5 \leqslant x_{(1)}$ and $x_{(n)} \leqslant \theta + 0.5$ or $x_{(n)} - 0.5 \leqslant \theta \leqslant x_{(1)} + 0.5$. Therefore, any statistic $\hat{\theta}$ satisfying

$$X_{(n)} - 0.5 \leqslant \hat{\theta} \leqslant X_{(1)} + 0.5$$

is an MLE of θ. ‖

Example 3.5 (Laplace distribution). Let $X_1, \ldots, X_n$ be i.i.d. random variables with Laplace density (or double exponential density)

$$f(x;\theta) = \frac{1}{2}\,\mathrm{e}^{-|x-\theta|}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

Find the MLE of θ.

Solution. The parameter space $\Theta = \mathbb{R}$. The joint density of $\mathbf{x}$ is

$$f(\boldsymbol{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \frac{1}{2}\,\mathrm{e}^{-|x_i-\theta|}$$

so that the log-likelihood is given by $\ell(\theta) = -n\log(2) - \sum_{i=1}^{n} |x_i - \theta|$. The first derivative is

$$\ell'(\theta) = \sum_{i=1}^{n} \operatorname{sgn}(x_i - \theta), \qquad (3.5)$$

where $\text{sgn}(t) = 1$, 0, or -1 depending on whether $t > 0$, $t = 0$, or $t < 0$. Note that the absolute function

$$h(t) = |t| = \begin{cases} t, & \text{if } \ t > 0, \\ -t, & \text{if } \ t \leqslant 0. \end{cases}$$

When $t = 0$, $h(t)$ is not differentiable. When $t \neq 0$,

$$\begin{aligned} h'(t) &= \begin{cases} 1, & \text{if } \ t > 0, \\ -1, & \text{if } \ t < 0 \end{cases} \\ &= \text{sgn}(t). \end{aligned}$$

To get the solution to the score equation $\ell'(\theta) = 0$, we consider two cases.

- If n is even, then any point in the interval $(x_{(n/2)}, x_{(n/2+1)})$ is an mle of θ;
- If n is odd, then $\text{median}(x_1, \ldots, x_n)$ is the unique mle of θ because the median will make half the terms of the sum in expression (3.5) non-positive and half non-negative.

Therefore, the median($\mathbf{x}$) or any point in $(X_{(n/2)}, X_{(n/2+1)})$ is the MLE $\hat{\theta}$ of θ. $\|$

5.2• Remarks on Example 3.5

— Let $n = 4$ and $x_1 = 0.1$, $x_2 = 0.2$, $x_3 = 0.3$, $x_4 = 0.8$. If let

$$\theta = \text{median}(x_1, \ldots, x_4) = \frac{0.2 + 0.3}{2} = 0.25,$$

then $\ell'(\theta) = \sum_{i=1}^{4} \text{sgn}(x_i - \theta) = -1 - 1 + 1 + 1 = 0$. If fact, *any point* in the open interval $(0.2, 0.3)$ is an mle of θ.

— Let $n = 3$ and $x_1 = -1$, $x_2 = 5$, $x_3 = 100$. If let $\theta = \text{median}(x_1, x_2, x_3) = 5$, then

$$\ell'(\theta) = \text{sgn}(-1 - 5) + \text{sgn}(5 - 5) + \text{sgn}(100 - 5) = -1 + 0 + 1 = 0.$$

Hence, $\text{median}(x_1, x_2, x_3)$ is the unique mle of θ.

6• Restricted MLE

- *Case 1: Equality constraints.* $\mathbf{A}\boldsymbol{\theta} = \boldsymbol{b}$, where $\mathbf{A}_{m\times p}$ and $\boldsymbol{b}_{m\times 1}$ are known, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ is an unknown parameter vector.
- *Case 2: Inequality constraints.* $\boldsymbol{a} \leqslant \mathbf{A}\boldsymbol{\theta} \leqslant \boldsymbol{b}$, where $\boldsymbol{a}_{m\times 1}$ is known.
- *Case 3: Convex constraint.* $\boldsymbol{\theta} \in \mathbb{S}$, where $\mathbb{S}$ is a convex set.

6.1• Definition of a convex set

— Let two points $C \in \mathbb{S}$ and $D \in \mathbb{S}$. If the segment of connecting the point C with the point D still belongs to $\mathbb{S}$, then $\mathbb{S}$ is called a convext set.

Example 3.6 (Multinomial distribution). Consider a multinomial experiment with n trials and p categories. The observed counts are $n_1, \ldots, n_p$ for the p categories. Let θ_j denote the cell probability of category j for $j = 1, \ldots, p$. We have $0 \leqslant \theta_j \leqslant 1$ and $\sum_{j=1}^p \theta_j = 1$. Find the MLE of θ_j subject to the equality constraint $\sum_{j=1}^p \theta_j = 1$.

Solution. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$. The parameter vector space is

$$\mathbb{T}_p = \left\{\boldsymbol{\theta}\colon\ \theta_j \geqslant 0,\ j = 1, \ldots, p,\ \textstyle\sum_{j=1}^p \theta_j = 1\right\}, \tag{3.6}$$

which is the p-dimensional hyperplane. The joint pmf of $n_1, \ldots, n_p$ is

$$f(n_1, \ldots, n_p; \boldsymbol{\theta}) = \binom{n}{n_1, \ldots, n_p} \prod_{j=1}^p \theta_j^{n_j}, \quad n_j \geqslant 0, \quad \sum_{j=1}^p n_j = n.$$

The likelihood function of $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) \propto \prod_{j=1}^p \theta_j^{n_j} = \left(\prod_{j=1}^{p-1} \theta_j^{n_j}\right)\left(1 - \sum_{j=1}^{p-1} \theta_j\right)^{n_p},$$

where

$$\theta_j \geqslant 0 \quad \text{and} \quad \sum_{j=1}^{p-1} \theta_j \leqslant 1. \tag{3.7}$$

Then

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{p-1} n_j \log(\theta_j) + n_p \log\left(1 - \sum_{j=1}^{p-1} \theta_j\right).$$

By differentiating $\ell(\boldsymbol{\theta})$ with θ_j for $j = 1, \ldots, p-1$ and letting them equal zeros, we obtain

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j} = \frac{n_j}{\theta_j} - \frac{n_p}{1 - \sum_{j=1}^{p-1} \theta_j} = \frac{n_j}{\theta_j} - \frac{n_p}{\theta_p} = 0, \quad j = 1, \ldots, p-1.$$

The solutions are given by

$$\hat{\theta}_j = \frac{n_j}{n}, \quad j = 1, \ldots, p-1,$$

which satisfy the constraints specified by (3.7). In addition, $\hat{\theta}_p = n_p/n$. ‖

6.2• Comments on Example 3.6

— Example 3.6 is a case of one equality constraint, in which we transfer the restricted case into an unrestricted case by substituting $\theta_p = 1 - \sum_{j=1}^{p-1} \theta_j$.

Example 3.7 (Normal mean with inequality constraints). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$ subject to $a \leqslant \mu \leqslant b$, where a and b are two fixed constants. Find the MLE of μ.

Solution. The parameter space is $\Theta = [a, b]$. The likelihood function of μ is given by

$$L(\mu) = \prod_{i=1}^{n} \mathrm{e}^{-\frac{(x_i - \mu)^2}{2}}, \quad a \leqslant \mu \leqslant b$$

so that

$$\begin{aligned}
\ell(\mu) &= -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2} \\
&= -\frac{1}{2} \sum_{i=1}^{n} (x_i^2 - 2x_i\mu + \mu^2) \\
&= -\frac{1}{2} \left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 \right) \\
&= -\frac{n}{2} \left\{ (\mu^2 - 2\mu\bar{x} + \bar{x}^2) - \bar{x}^2 + \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right\} \\
&\propto -(\mu - \bar{x})^2, \quad a \leqslant \mu \leqslant b. \qquad (3.8)
\end{aligned}$$

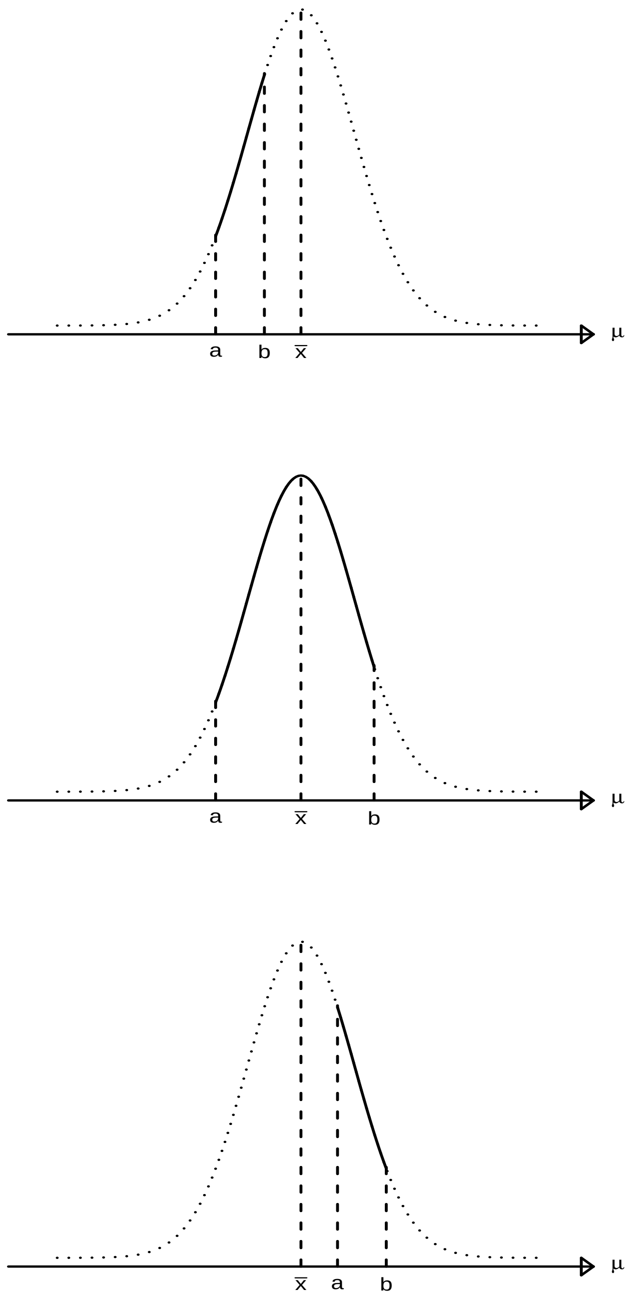

Figure 3.2 Plots of the log-likelihood function $\ell(\mu)$ defined by (3.8) for three cases. Top: $\bar{x} > b$; Middle: $a \leqslant \bar{x} \leqslant b$; Bottom: $\bar{x} < a$.

Figure 3.2 shows that $\ell(\mu)$ is a truncated quadratic function of μ. Hence

$$\begin{aligned}\mu &= \begin{cases} b, & \text{if } \bar{x} > b, \\ \bar{x}, & \text{if } a \leqslant \bar{x} \leqslant b, \\ a, & \text{if } \bar{x} < a \end{cases} \\ &= \text{median}(a, \bar{x}, b)\end{aligned}$$

is the restricted mle of μ and $\hat{\mu} = \text{median}(a, \bar{X}, b)$ is the restricted MLE of μ. As an exercise, to calculate $E(\hat{\mu})$ and $\text{Var}(\hat{\mu})$. $\|$

3.1.4 The invariance property of MLE

7• Reparametrization via a one-to-one map

Theorem 3.1 (Invariance of MLE). Let $\hat{\boldsymbol{\theta}} = \boldsymbol{u}(X_1, \dots, X_n)$ be the MLE of $\boldsymbol{\theta}_{p\times 1} \in \boldsymbol{\Theta}$. If $\boldsymbol{\eta}_{p\times 1} = \boldsymbol{h}(\boldsymbol{\theta}) = (h_1(\boldsymbol{\theta}), \dots, h_p(\boldsymbol{\theta}))^\top$ is a one-to-one transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, then $\hat{\boldsymbol{\eta}} = \boldsymbol{h}(\hat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{\eta}$. $\|$

Proof. Since $\boldsymbol{\eta} = \boldsymbol{h}(\boldsymbol{\theta})$ is a one-to-one map, we have $\boldsymbol{\theta} = \boldsymbol{h}^{-1}(\boldsymbol{\eta})$. The likelihood function is given by

$$L(\boldsymbol{\theta}) = L(\boldsymbol{h}^{-1}(\boldsymbol{\eta})) \ \hat{=}\ L^*(\boldsymbol{\eta}).$$

We want to prove $L^*(\hat{\boldsymbol{\eta}}) \geqslant L^*(\boldsymbol{\eta})$ for all $\boldsymbol{\eta}$. In fact, we have

$$\begin{aligned}L^*(\hat{\boldsymbol{\eta}}) &= L^*(\boldsymbol{h}(\hat{\boldsymbol{\theta}})) = L\boldsymbol{h}^{-1}(\boldsymbol{h}(\hat{\boldsymbol{\theta}})) = L(\hat{\boldsymbol{\theta}}) \\ &\geqslant L(\boldsymbol{\theta}) = L^*(\boldsymbol{\eta}).\end{aligned}$$

Therefore, $\hat{\boldsymbol{\eta}} = \boldsymbol{h}(\hat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{\eta}$. $\square$

7.1• Understanding Theorem 3.1 through Figure 3.3

$$\begin{tikzcd} \boldsymbol{\theta} \arrow[r, "L(\cdot)"] \arrow[d, "\boldsymbol{h}(\cdot)"'] & \hat{\boldsymbol{\theta}} \arrow[d, "\boldsymbol{h}(\cdot)"] \\ \boldsymbol{\eta} \arrow[r, "L^*(\cdot)"] & \hat{\boldsymbol{\eta}} \end{tikzcd}$$

Figure 3.3 An illustration of Theorem 3.1.

7.2• Comments on Figure 3.3

— Figure 3.3 shows that Theorem 3.1 gives two ways to reach $\hat{\boldsymbol{\eta}}$.

— The first way is to first find the $\hat{\boldsymbol{\theta}}$ by maximizing the likelihood function $L(\boldsymbol{\theta})$, then to utilize the map $\boldsymbol{h}(\cdot)$ to obtain $\hat{\boldsymbol{\eta}} = \boldsymbol{h}(\hat{\boldsymbol{\theta}})$.

— The second way is to first utilize the map $\boldsymbol{h}(\cdot)$ to obtain a new parameter vector $\boldsymbol{\eta}$, then to find the $\hat{\boldsymbol{\eta}}$ by maximizing the likelihood function $L^*(\boldsymbol{\eta})$.

7.3• Two illustration examples

— Since $h(\sigma) = \sigma = \sqrt{\sigma^2}$ with $\sigma > 0$ is a one-to-one map between σ^2 and σ, it follows from Example 3.2 that $\hat{\sigma} = \{(1/n)\sum_{i=1}^n (X_i - \bar{X})^2\}^{1/2}$ different from $S = \{\sum_{i=1}^n (X_i - \bar{X})^2/(n-1)\}^{1/2}$, is an MLE of σ.

— Similarly, the MLE of, say $\log(\sigma^2)$, is $\log\left\{(1/n)\sum_{i=1}^n (X_i - \bar{X})^2\right\}$.

8• Can we extend Theorem 3.1?

- It is very natural to ask whether Theorem 3.1 still holds if the assumption that $\boldsymbol{\eta} = \boldsymbol{h}(\boldsymbol{\theta})$ is a one-to-one transformation is removed.

8.1• The MLE of variance in a Bernoulli distribution

— As a first example, assume an estimate of the variance; i.e., $\theta(1-\theta)$, of the Bernoulli(θ) distribution is desired.

— Example 3.1 gives the MLE of θ to be $\bar{X}$, but since $\theta(1-\theta)$ is not a one-to-one function of θ, Theorem 3.1 does not give the MLE of $\theta(1-\theta)$.

— Theorem 3.2 below will give such an estimator and it will be $\bar{X}(1-\bar{X})$.

8.2• The MLE of $\mu^2 + \sigma^2$ in normal distribution

— As a second example, consider the MLE of $\mu^2 + \sigma^2$ in Example 3.2.

— Since $\mu^2 + \sigma^2$ is not a one-to-one function of μ and σ^2, Theorem 3.1 does not give the MLE of $\mu^2 + \sigma^2$.

— Such an estimator will be obtainable from Theorem 3.2 below and it will be $\bar{X}^2 + (1/n)\sum_{i=1}^n (X_i - \bar{X})^2$.

Theorem 3.2 (Extension of Theorem 3.1). Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \boldsymbol{\Theta}$. If $\boldsymbol{\eta}_{r\times 1} = \boldsymbol{h}(\boldsymbol{\theta}) = (h_1(\boldsymbol{\theta}), \ldots, h_r(\boldsymbol{\theta}))^\top$ for $1 \leqslant r \leqslant p$ is a many-to-few transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, then $\hat{\boldsymbol{\eta}} = \boldsymbol{h}(\hat{\boldsymbol{\theta}}) = (h_1(\hat{\boldsymbol{\theta}}), \ldots, h_r(\hat{\boldsymbol{\theta}}))^\top$ is the MLE of $\boldsymbol{\eta}$. ‖

Proof. Let $\mathbb{H}$ denote the range space of the map $\boldsymbol{h}(\cdot) = (h_1(\cdot), \ldots, h_r(\cdot))^\top$. $\mathbb{H}$ is an r-dimensional space. Define

$$M(\boldsymbol{h}) = \max_{\{\boldsymbol{\theta}:\, \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{h}\}} L(\boldsymbol{\theta}),$$

which is called the likelihood function induced by $\boldsymbol{h}(\cdot)$. It suffices to show

$$M(\boldsymbol{h}) \leqslant M(\boldsymbol{h}(\hat{\boldsymbol{\theta}})) \quad \text{for any } \boldsymbol{h} \in \mathbb{H},$$

which follows immediately from the inequality

$$\begin{aligned} M(\boldsymbol{h}) &= \max_{\{\boldsymbol{\theta}:\, \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{h}\}} L(\boldsymbol{\theta}) \\ &\leqslant \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}) = L(\hat{\boldsymbol{\theta}}) \\ &= \max_{\{\boldsymbol{\theta}:\, \boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{h}(\hat{\boldsymbol{\theta}})\}} L(\boldsymbol{\theta}) \\ &= M(\boldsymbol{h}(\hat{\boldsymbol{\theta}})), \end{aligned}$$

for any $\boldsymbol{h} \in \mathbb{H}$. □

8.3• Understanding Theorem 3.2

— This property of invariance of MLEs allows us in our discussion of maximum likelihood estimation to consider estimating $(\theta_1, \ldots, \theta_p)^\top$ rather than the more general $h_1(\theta_1, \ldots, \theta_p), \ldots, h_r(\theta_1, \ldots, \theta_p)$.

3.2 Moment Estimator

9• Three basic methods of estimation

- The first procedure for estimating parameters is the method of *maximum likelihood estimation.*

- The second procedure for estimating parameters is the *method of moments* proposed by the great British statistician Karl Pearson near the turn of the twentieth century.

- The third procedure is called *Bayesian estimation*.

10• Background for the maximum likelihood estimation

- Let $x_1 = 0.099$, $x_2 = -1.146$, $x_3 = -1.172$, $x_4 = -0.290$, $x_5 = 1.435$ and $x_6 = -0.657$ be corresponding observations of a random sample of size six from the population r.v. X.
- We guess that $X_1, \ldots, X_6 \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ or $X \sim N(\mu, \sigma^2)$ and we want to find the mles of μ and σ^2.
- We wonder if or not our guess is correct, which can be tested by statistical methods (e.g., the goodness-of-fit test, see §5.5); i.e.,

$$H_0\text{: The distribution of } X \text{ is normal}$$

against

$$H_1\text{: The distribution of } X \text{ is not normal.}$$

10.1• If H_0 is accepted, what can we do next step?

— Based on the observed data $\{x_i\}_{i=1}^6$, if H_0 is accepted, then by using the method of ML estimation as shown in Example 3.2, the mles of μ and σ^2 are given by

$$\bar{x} = -0.2885 \quad \text{and} \quad \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = 0.7954,$$

respectively.

— We could claim that $\{x_i\}_{i=1}^6$ are observation of a random sample of size six from the most possible population $N(-0.2885, 0.7954)$.

10.2• If H_0 is rejected, what can we do next step?

— One way is to guess another population distribution. If the new H_0 was accepted, we could repeat the above process.

— Alternatively, we can estimate the first and second moments of the unknown population distribution $F(\cdot)$ by using the *method of moments*.

— Of course, when the family of distribution is known but the parameters are unknown, the method of moments can also be applied.

11• Moment estimators

- By first equating the *sample moments*
$$\frac{1}{n}\sum_{i=1}^n X_i, \quad \frac{1}{n}\sum_{i=1}^n X_i^2, \quad \ldots, \quad \frac{1}{n}\sum_{i=1}^n X_i^r$$
to the corresponding *population moments*
$$E(X), \quad E(X^2), \quad \ldots, \quad E(X^r),$$
then solving the system of equations, we can obtain *moment estimators* of parameters.

- Specifically, if there are a total of r parameters, the moment estimators can be obtained from solving the system of equations:
$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n X_i &= E(X),\\ \frac{1}{n}\sum_{i=1}^n X_i^2 &= E(X^2),\\ &\vdots\\ \frac{1}{n}\sum_{i=1}^n X_i^r &= E(X^r).\end{aligned}$$

Example 3.8 (Gamma distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$. Find the moment estimators of α and β.

Solution. Let $X \sim \text{Gamma}(\alpha, \beta)$, from Appendix A.2.4, we have $E(X) = \alpha/\beta$ and $\text{Var}(X) = \alpha/\beta^2$. Thus
$$E(X^2) = \text{Var}(X) + \{E(X)\}^2 = \frac{\alpha(\alpha+1)}{\beta^2}.$$
The moment estimators of α and β must satisfy
$$\begin{aligned}\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i &= E(X) = \frac{\alpha}{\beta}, \quad \text{and}\\ \frac{1}{n}\sum_{i=1}^n X_i^2 &= E(X^2) = \frac{\alpha(\alpha+1)}{\beta^2}.\end{aligned}$$

Thus,

$$\hat{\beta}^{\mathrm{M}} = \frac{n\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{and} \quad \hat{\alpha}^{\mathrm{M}} = \frac{n\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

are the corresponding moment estimators of α and β. ‖

Example 3.9 (Beta distribution). Let $x_1 = 0.42$, $x_2 = 0.10$, $x_3 = 0.65$ and $x_4 = 0.23$ be observations of random variables of size $n = 4$ from the pdf

$$f(x;\theta) = \theta x^{\theta-1}, \quad 0 \leqslant x \leqslant 1.$$

Find the moment estimate of θ.

Solution. Let $X \sim f(x;\theta)$, we have

$$E(X) = \int_0^1 x \cdot \theta x^{\theta-1}\,\mathrm{d}x = \frac{\theta}{\theta+1}.$$

Let $E(X)$ equal to the first sample moment

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{0.42 + 0.10 + 0.65 + 0.23}{4} = 0.35,$$

we obtain $\theta/(\theta+1) = \bar{x}$. Thus the moment estimate for θ is

$$\hat{\theta}^{\mathrm{M}} = \frac{\bar{x}}{1-\bar{x}} = \frac{0.35}{1-0.35} = 0.54.$$ ‖

Example 3.10 (Normal distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Find the moment estimators of μ and σ^2.

Solution. Let $X \sim N(\mu, \sigma^2)$, we have $E(X) = \mu$ and $E(X^2) = \sigma^2 + \mu^2$. The moment estimators of μ and σ^2 must satisfy

$$\bar{X} = \mu \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2.$$

Hence,

$$\hat{\mu}^{\mathrm{M}} = \bar{X} \quad \text{and} \quad \hat{\sigma}^{2\mathrm{M}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

are the corresponding moment estimators of μ and σ^2. ‖

11.1• The application range of the method of moments

— The method of moments can be applied to both *parametric* and *nonparametric* statistics.

11.2• What is the parametric statistics?

— Make inferences (i.e., estimation and testing hypothesis) on parameters in a known/specified family of distributions $\{f(x;\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ based on an i.i.d. sample $\{x_i\}_{i=1}^n$ or more than one i.i.d. sample.

11.3• What is the nonparametric (or distribution-free) statistics?

— Make inferences (i.e., estimation and test) on an unknown distribution itself $F(\cdot)$ based on an i.i.d. sample $\{x_i\}_{i=1}^n$ or on two unknown distributions $F(\cdot)$ and $G(\cdot)$ based on two i.i.d. samples $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$.

3.3 Bayesian Estimator

12• Three reference books for Bayesian statistics

- Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2-nd ed.). Springer, New York, USA.

- Carlin, B.P. and Louis, T.A. (2008). *Bayesian Methods for Data Analysis* (3-rd ed.). Chapman & Hall/CRC (Texts in Statistical Science), Boca Raton, USA.

- Gelman, A., Carlin, J.P., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis* (3-rd ed.). Chapman & Hall/CRC (Texts in Statistical Science), Boca Raton, USA.

13• Main features of Bayesian method

- In the ML estimation method and the method of moments, we have assumed that the parameters are *fixed* but *unknown* constants.

- In the Bayesian method, we assume that $\boldsymbol{\theta}$ is a random vector with a density $\pi(\boldsymbol{\theta})$, which is called the *prior density* of $\boldsymbol{\theta}$.

- Then the joint density or likelihood function (in the ML estimation method) of $\mathbf{x} = (X_1, \ldots, X_n)^\top$ becomes the conditional density (in the Bayesian method) of $\mathbf{x}$ given $\boldsymbol{\theta}$, denoted by $f(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.

13.1• The basic idea of Bayesian estimation

— The basic idea of Bayesian estimation is to utilize both the information from the prior density of $\boldsymbol{\theta}$ and the likelihood function of the observed data $\boldsymbol{x}$.

14• Three steps for determining Bayesian estimators

- Given a random sample $\mathbf{x} = (X_1, \ldots, X_n)^\top$, determine the joint density of $\mathbf{x}$ and $\boldsymbol{\theta}$:

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{\theta}) &= \text{Likelihood} \times \text{Prior} \\ &= f(\boldsymbol{x}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \\ &= \left\{ \prod_{i=1}^n f(x_i|\boldsymbol{\theta}) \right\} \times \pi(\boldsymbol{\theta}), \end{aligned} \tag{3.9}$$

  where $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.

- Determine the *posterior density* (i.e., the conditional density of $\boldsymbol{\theta}$ given $\mathbf{x} = \boldsymbol{x}$) of $\boldsymbol{\theta}$,

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &= \frac{f(\boldsymbol{x}, \boldsymbol{\theta})}{f(\boldsymbol{x})} = c^{-1} f(\boldsymbol{x}, \boldsymbol{\theta}) \qquad (3.10) \\ &\propto f(\boldsymbol{x}, \boldsymbol{\theta}) = \text{Likelihood} \times \text{Prior}, \end{aligned}$$

  where $f(\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{x}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \mathrel{\hat{=}} c$ is the normalizing constant of $p(\boldsymbol{\theta}|\boldsymbol{x})$ because $\mathbf{x} = \boldsymbol{x}$ is given.

- The Bayesian estimate of $\boldsymbol{\theta}$ (i.e., the conditional expectation of $\boldsymbol{\theta}$) is defined by

$$E(\boldsymbol{\theta}|\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta} \cdot p(\boldsymbol{\theta}|\boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta}. \tag{3.11}$$

Example 3.11 (Bernoulli–beta distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ and the prior distribution of θ be $\text{Beta}(\alpha, \beta)$. Find the Bayesian estimate of θ.

Solution. Note that $f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$, then the joint density of $\mathbf{x}$ and θ is

$$\begin{aligned} f(\boldsymbol{x}, \theta) &= \left\{ \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \right\} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha+x_+-1}(1-\theta)^{\beta+n-x_+-1}, \quad 0 < \theta < 1, \end{aligned}$$

where $x_+ \hat{=} \sum_{i=1}^n x_i$. The posterior density of θ is given by

$$p(\theta|\boldsymbol{x}) \propto \theta^{\alpha+x_+-1}(1-\theta)^{\beta+n-x_+-1}, \quad 0 < \theta < 1;$$

i.e., $\theta|\boldsymbol{x} \sim \text{Beta}(\alpha + x_+, \beta + n - x_+)$. Therefore,

$$E(\theta|\boldsymbol{x}) = \frac{\alpha + x_+}{\alpha + \beta + n}$$

is the Bayesian estimate of θ, and $(\alpha + n\bar{X})/(\alpha + \beta + n)$ is the Bayesian estimator of θ. ‖

Example 3.12 (Poisson–gamma distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$ and the prior distribution of θ be $\text{Gamma}(a, b)$. Find the Bayesian estimate of θ.

Solution. Note that $f(x_i|\theta) = \mathrm{e}^{-\theta}\theta^{x_i}/x_i!$, $x_i = 0, 1, 2, \ldots$, then the joint density of $\mathbf{x}$ and θ is

$$\begin{aligned} f(\boldsymbol{x}, \theta) &= \left\{ \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} \mathrm{e}^{-\theta} \right\} \times \frac{b^a}{\Gamma(a)} \theta^{a-1} \mathrm{e}^{-b\theta} \\ &= \frac{b^a}{\Gamma(a) \prod_{i=1}^n x_i!} \theta^{a+x_+-1} \mathrm{e}^{-(b+n)\theta}, \quad \theta > 0, \end{aligned}$$

where $x_+ \hat{=} \sum_{i=1}^n x_i$. The posterior density of θ is given by

$$p(\theta|\boldsymbol{x}) \propto \theta^{a+x_+-1} \mathrm{e}^{-(b+n)\theta}, \quad \theta > 0;$$

i.e., $\theta|\boldsymbol{x} \sim \text{Gamma}(a + x_+, b + n)$. Therefore,

$$E(\theta|\boldsymbol{x}) = \frac{a + x_+}{b + n}$$

is the Bayesian estimate of θ, and $(a+n\bar{X})/(b+n)$ is the Bayesian estimator of θ. ‖

15• Differences between MLE and Bayesian estimator

Table 3.1 A comparison of MLE with Bayesian estimator

	MLE	Bayesian estimator
1	$\boldsymbol{\theta}$: A fixed and unknown parameter vector	$\boldsymbol{\theta}$: A random vector with a prior density $\pi(\boldsymbol{\theta})$
2	$f(\boldsymbol{x};\boldsymbol{\theta})$: The joint density of $\mathbf{x} = (X_1, \ldots, X_n)^{\top}$	$f(\boldsymbol{x}\vert\boldsymbol{\theta})$: The conditional density of $\mathbf{x}$ given $\boldsymbol{\theta}$
3	$L(\boldsymbol{\theta})$: Likelihood function	$p(\boldsymbol{\theta}\vert\boldsymbol{x}) \propto L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$: Posterior density
4	$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta})$: MLE	$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(\boldsymbol{\theta}\vert\boldsymbol{x})$: Posterior mode

15.1• The non-informative prior

— When the non-informative prior (i.e., $\pi(\boldsymbol{\theta}) \propto 1$) is taken as the prior of $\boldsymbol{\theta}$, or when $\pi(\boldsymbol{\theta})$ is flat, we have $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$.

16• Statistical interpretation of Bayesian estimator

- *Loss function*: $l(\boldsymbol{\theta}, \boldsymbol{a})$ computes the loss incurred when $\boldsymbol{\theta}$ is the true state of nature and the action $\boldsymbol{a} \in \mathcal{A}$ is taken.
- *Squared error loss*: $l(\boldsymbol{\theta}, \boldsymbol{a}) = \|\boldsymbol{\theta} - \boldsymbol{a}\|^2 = (\boldsymbol{\theta} - \boldsymbol{a})^{\top}(\boldsymbol{\theta} - \boldsymbol{a})$.
- *Posterior risk*: $\rho(p, \boldsymbol{a}) \ \hat{=}\ \int_{\boldsymbol{\Theta}} \|\boldsymbol{\theta} - \boldsymbol{a}\|^2 p(\boldsymbol{\theta}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{\theta}$.
- *Bayesian estimator*: $E(\boldsymbol{\theta}|\boldsymbol{x})$ is the action $\boldsymbol{a}^*$ such that the posterior risk reaches its minimum. That is,

$$E(\boldsymbol{\theta}|\boldsymbol{x}) = \boldsymbol{a}^* = \arg\min_{\boldsymbol{a}\in\mathcal{A}} \rho(p, \boldsymbol{a})$$

or

$$\rho(p, E(\boldsymbol{\theta}|\boldsymbol{x})) \leqslant \rho(p, \boldsymbol{a}), \quad \forall \boldsymbol{a} \in \mathcal{A}.$$

3.4 Properties of Estimators

3.4.1 Unbiasedness

17• Measures for comparing two point estimators

Definition 3.2 (Unbiased estimator and bias). An estimator $\varphi(\mathbf{x})$ is an *unbiased estimator* of the parameter θ if $E\{\varphi(\mathbf{x})\} = \theta$ for $\theta \in \Theta$. Otherwise, the estimator is biased and the bias is defined by

$$b(\theta) = E\{\varphi(\mathbf{x})\} - \theta, \tag{3.12}$$

where $\mathbf{x} = (X_1, \dots, X_n)^\top$. ‖

Example 3.13 (Distribution with a finite second-order moment). Let X_1, $\dots, X_n$ be a random sample from a population (which is not necessary to be a normal population) with mean μ and variance $\sigma^2 < \infty$. According to Eq. (2.9), we can see that

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{3.13}$$

are unbiased estimators of μ and σ^2, respectively. ‖

Example 3.14 (Uniform distribution). Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$, then

1) the n-th order statistic $X_{(n)}$ is a biased estimator of θ;

2) $\frac{n+1}{n} X_{(n)}$ is an unbiased estimator of θ; and

3) $2\bar{X}$ is also an unbiased estimator of θ.

Solution. 1) From Example 2.16, we know that the pdf of $X_{(n)}$ is

$$f_n(x) = nx^{n-1}/\theta^n, \quad 0 < x < \theta.$$

Hence,

$$E\{X_{(n)}\} = \int_0^\theta x f_n(x)\,\mathrm{d}x = \frac{n}{n+1} \cdot \theta \neq \theta, \tag{3.14}$$

indicating that $X_{(n)}$ is a biased estimator of θ.

2) Clearly, $\frac{n+1}{n} X_{(n)}$ is an unbiased estimator of θ.

3) Since

$$E(X_1) = \int_0^\theta x_1 \cdot \frac{1}{\theta}\,\mathrm{d}x_1 = \frac{\theta}{2},$$

we have $E(2\bar{X}) = 2E(\bar{X}) = 2E(X_1) = \theta$. ‖

Definition 3.3 (MSE). Given an estimator $Y = \varphi(\mathbf{x})$ of θ, the *mean square error* (MSE) of the estimator is defined by

$$\text{MSE} = E\big\{\varphi(\mathbf{x}) - \theta\big\}^2. \qquad \|$$

17.1• Remarks on Definition 3.3

— It is easy to verify that

$$\begin{aligned}
& \text{MSE} \\
= \;& E\big\{Y - E(Y) + E(Y) - \theta\big\}^2 \\
= \;& E\{Y - E(Y)\}^2 + \{E(Y) - \theta\}^2 + E\big[2\{Y - E(Y)\}\underbrace{\{E(Y) - \theta\}}_{\text{constant}}\big] \\
= \;& \text{Var}\{\varphi(\mathbf{x})\} + b^2(\theta).
\end{aligned}$$

— Clearly, if an estimator $\varphi(\mathbf{x})$ is unbiased, then

$$\text{MSE} = \text{Var}\{\varphi(\mathbf{x})\}.$$

— Smaller MSE means greater precision.

3.4.2 Efficiency

18• Why need we the notion of efficiency?

- It is possible that there are several unbiased estimators for the same unknown parameter of interest.
- For instance, in Example 3.14, both $\frac{n+1}{n}X_{(n)}$ and $2\bar{X}$ are unbiased estimators of θ.
- Which one should we choose?
- Answer: The unbiased estimator with the *smaller* variance is the desired.
- Comparing two variances is equivalent to comparing two efficiencies.

18.1• Efficiency of an estimator

— Efficiency of an estimator $\hat{\theta}$ is proportional to the reciprocal of its variance:

$$\mathrm{Eff}_{\hat{\theta}}(\theta) \propto \frac{1}{\mathrm{Var}(\hat{\theta})}.$$

18.2• Relative efficiency of two estimators

Definition 3.4 (Relative efficiency). Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators for a parameter θ. If

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2), \tag{3.15}$$

we say that $\hat{\theta}_1$ is *more efficient* than $\hat{\theta}_2$. The *relative efficiency* of $\hat{\theta}_1$ to $\hat{\theta}_2$ is defined by the ratio

$$\frac{\mathrm{Eff}_{\hat{\theta}_1}(\theta)}{\mathrm{Eff}_{\hat{\theta}_2}(\theta)} = \frac{\mathrm{Var}(\hat{\theta}_2)}{\mathrm{Var}(\hat{\theta}_1)}.$$ ‖

Example 3.15 (Example 3.14 revisited). In Example 3.14, we have shown that $\hat{\theta}_1 = \frac{n+1}{n}X_{(n)}$ and $\hat{\theta}_2 = 2\bar{X}$ are two unbiased estimators of θ. Which estimator is more efficient?

Solution. From Appendix A.2.1, since $X_1 \sim U(0,\theta)$, we have $\mathrm{Var}(X_1) = \theta^2/12$. Hence,

$$\mathrm{Var}(\hat{\theta}_2) = \mathrm{Var}(2\bar{X}) = \frac{4}{n^2}\cdot n\mathrm{Var}(X_1) = \frac{\theta^2}{3n}.$$

On the other hand, similar to (3.14), we have

$$E(X_{(n)}^2) = \int_0^\theta x^2 \cdot \frac{nx^{n-1}}{\theta^n}\,\mathrm{d}x = \frac{n}{n+2}\theta^2. \tag{3.16}$$

Thus, based on (3.14) and (3.16), we obtain

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_1) &= \frac{(n+1)^2}{n^2}\mathrm{Var}\{X_{(n)}\} \\
&= \frac{(n+1)^2}{n^2}\Big[E(X_{(n)}^2) - \{E(X_{(n)})\}^2\Big] \\
&= \frac{(n+1)^2}{n^2}\left\{\frac{n\theta^2}{n+2} - \frac{n^2}{(n+1)^2}\theta^2\right\} = \frac{\theta^2}{n(n+2)}.
\end{aligned}$$

When $n > 1$, we have $\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$, indicating that $\hat{\theta}_1$ has a smaller variance (and hence is more efficient) than $\hat{\theta}_2$. ‖

19• Why need we the Cramér–Rao inequality

- Let $\mathcal{U} = \{\hat{\theta}\colon\ E(\hat{\theta}) = \theta\}$ denote the family of unbiased estimators of θ. The goal is to find the $\hat{\theta}^* \in \mathcal{U}$ with the smallest variance.
- Let $m = \#\mathcal{U}$ denote the number of elements in $\mathcal{U}$. If m is finite, we write $\mathcal{U} = \{\hat{\theta}_1, \ldots, \hat{\theta}_m\}$. Hence, we can choose the $\hat{\theta}_{k_0}$ such that

$$\mathrm{Var}(\hat{\theta}_{k_0}) \leqslant \mathrm{Var}(\hat{\theta}_j), \quad j \neq k_0, \quad j = 1, \ldots, m.$$

- If m is infinite, how to find the $\hat{\theta}^*$ with the smallest variance?

19.1• A motivation

— If we could find a constant c_0 satisfying

$$\mathrm{Var}(\hat{\theta}) \geqslant c_0, \quad \forall\, \hat{\theta} \in \mathcal{U},$$

then, this inequality can guide us to choose the $\hat{\theta}^*$ with variance being c_0.

— Thus, finding the $\hat{\theta}^*$ is equivalent to finding the lower bound c_0, which was found by Cramér and Rao.

— The c_0 is closely related to two new concepts: Score function and Fisher information.

19.2• Score function

— Let $X_1, \ldots, X_n$ be a random sample from the population r.v. X with density $f(x;\theta)$. Define $\mathbf{x} = (X_1, \ldots, X_n)^\top$ and $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ are their realizations. In the previous sections, we denote the likelihood function by

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta).$$

— If we replace x_i in $L(\theta; x_1, \ldots, x_n)$ by X_i, then the resultant $L(\theta; \mathbf{x})$ is also a random variable and depends on the parameter θ.

— When the expectation and variance of a specific function of $L(\theta; \mathbf{x})$ are calculated, we also denote the likelihood function by

$$L(\theta) = L(\theta; X_1, \ldots, X_n) = L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(X_i; \theta)$$

to emphasize its dependence on $\mathbf{x}$.

— Let $\ell(\theta) = \log\{L(\theta)\}$ denote the log-likelihood function of θ, we call

$$S(\theta) = S(\theta; \mathbf{x}) \mathrel{\hat{=}} \frac{\mathrm{d}\ell(\theta)}{\mathrm{d}\theta} = \ell'(\theta) = \frac{L'(\theta)}{L(\theta)} \tag{3.17}$$

the *score function*.

19.3• Understanding the score function

— $S(\theta)$ is a function of θ.

— $S(\theta) = S(\theta; \mathbf{x})$ is also a function of $\mathbf{x}$ so that

$$E\{S(\theta)\} = E_{\mathbf{x}}\{S(\theta; \mathbf{x})\} = \int S(\theta; \boldsymbol{x}) \prod_{i=1}^{n} f(x_i; \theta)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

— $S(\theta)$ is not a statistic because it depends on the unknown parameter θ.

19.4• Fisher information

— We call

$$I_n(\theta) = \mathrm{Var}\{S(\theta)\} = \mathrm{Var}_{\mathbf{x}}\{S(\theta; \mathbf{x})\} \tag{3.18}$$

the *Fisher information*, which is a way of measuring the amount of information that $\mathbf{x}$ carries about the unknown parameter θ.

— In many statistical problems, we have $E\{S(\theta)\} = 0$ so that (3.18) becomes

$$I_n(\theta) = E\{S^2(\theta; \mathbf{x})\} = E\left\{\left(\frac{\mathrm{d}\log L(\theta; \mathbf{x})}{\mathrm{d}\theta}\right)^2\right\}. \tag{3.19}$$

— However, it is possible in practice that $E\{S(\theta)\} \neq 0$ as shown in the following example.

Example 3.16 (Example 3.14 revisited). Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} U(0, \theta)$, where $\theta > 0$. Find the score function $S(\theta)$, $E\{S(\theta)\}$ and the Fisher information.

Solution. The population density is $f(x;\theta) = 1/\theta$, $x \in (0, \theta)$ depending on θ. We can rewrite $f(x;\theta) = (1/\theta) I_{(0,\theta)}(x)$ so that the likelihood function is

$$L(\theta) = \prod_{i=1}^n f(X_i;\theta) = \theta^{-n} \prod_{i=1}^n \xi_i,$$

where $\xi_i \;\hat{=}\; I_{(0,\theta)}(X_i) \sim \text{Bernoulli}(p_i)$ with $p_i = \Pr(0 < X_i < \theta)$. From (3.17), we have

$$S(\theta; \mathbf{x}) = \frac{L'(\theta)}{L(\theta)} = \frac{-n\theta^{-n-1}}{\theta^{-n}} \prod_{i=1}^n \xi_i = -\frac{n}{\theta} \prod_{i=1}^n \xi_i.$$

Thus,

$$\begin{aligned} E\{S(\theta;\mathbf{x})\} &= -\frac{n}{\theta} \prod_{i=1}^n E(\xi_i) = -\frac{n}{\theta} \prod_{i=1}^n \Pr(0 < X_i < \theta) \\ &= -\frac{n}{\theta} \left(\int_0^\theta \frac{1}{\theta} \, \mathrm{d}x \right)^n = -\frac{n}{\theta} \neq 0. \end{aligned}$$

Similarly, we have $E\{S^2(\theta;\mathbf{x})\} = n^2/\theta^2$ and $I_n(\theta) = \text{Var}\{S(\theta;\mathbf{x})\} = 0$. ‖

19.5• A basic result on Bernoulli r.v. used in Example 3.16

— Let $\xi \sim \text{Bernoulli}(p)$, then $\xi \stackrel{\text{d}}{=} \xi^r$ for any positive integer r.

— Clearly, we have $E(\xi) = E(\xi^r)$.

20• The Cramér–Rao inequality

Theorem 3.3 (The general CR inequality). Let $\tau(\theta)$ be an arbitrary function of the unknown parameter θ. If (i) $\hat{\theta} = T(\mathbf{x})$ is an unbiased estimator of $\tau(\theta)$, and (ii) the support of the population density $f(x;\theta)$ does not depend on the parameter θ, then

$$\text{Var}(\hat{\theta}) \geqslant \frac{\{\tau'(\theta)\}^2}{I_n(\theta)}, \tag{3.20}$$

where $I_n(\theta)$ is the Fisher information. ‖

Proof. On the one hand,

$$1 = \int \prod_{i=1}^{n} f(x_i;\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n.$$

By differentiating both sides of this identity with respect to θ, we have

$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\int\cdots\int L(\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n.$$

Since the supports of x_i's do not depend on the parameter θ, we can interchange differentiation and integration with respect to θ, yielding

$$\begin{aligned} 0 &= \int L'(\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\ &\overset{(3.17)}{=} \int S(\theta)\prod_{i=1}^{n} f(x_i;\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\ &= E\{S(\theta)\}. \end{aligned} \tag{3.21}$$

On the other hand, $\hat{\theta}$ is unbiased, then

$$\begin{aligned} \tau(\theta) = E(\hat{\theta}) &= \int T(\boldsymbol{x})\prod_{i=1}^{n} f(x_i;\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\ &= \int T(\boldsymbol{x})L(\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n. \end{aligned}$$

By differentiating both sides of the this equality with respect to θ, we obtain

$$\begin{aligned} \tau'(\theta) &= \int T(\boldsymbol{x})L'(\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\ &= \int T(\boldsymbol{x})\frac{L'(\theta)}{L(\theta)}\cdot L(\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\ &\overset{(3.17)}{=} \int T(\boldsymbol{x})S(\theta)\cdot\prod_{i=1}^{n} f(x_i;\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\ &= E\{\hat{\theta}\times S(\theta)\} \\ &\overset{(3.21)}{=} \mathrm{Cov}\{\hat{\theta}, S(\theta)\}. \end{aligned}$$

By the Cauchy–Schwarz inequality,

$$\{\tau'(\theta)\}^2 = [\mathrm{Cov}\{\hat{\theta}, S(\theta)\}]^2 \leqslant \mathrm{Var}(\hat{\theta})\times\mathrm{Var}\{S(\theta)\} = \mathrm{Var}(\hat{\theta})\times I_n(\theta),$$

which indicates (3.20). □

20.1• Comments on Theorem 3.3

— The result in (3.20) is not valid if the support of $f(x;\theta)$ depends on θ, see Example 3.16.

— The Cauchy–Schwarz inequality states that $\{E(XY)\}^2 \leqslant E(X^2)E(Y^2)$ or equivalently $\{\mathrm{Cov}(X,Y)\}^2 \leqslant \mathrm{Var}(X)\cdot\mathrm{Var}(Y)$, see Theorem 1.5.

— The right hand side of (3.20) is called the *Cramér–Rao lower bound.*

— In particular, if $\tau(\theta)=\theta$, then (3.20) becomes

$$\mathrm{Var}(\hat{\theta}) \geqslant \frac{1}{I_n(\theta)}. \tag{3.22}$$

20.2• Two identities related to the Fisher information

— Theorem 3.4 below provides another way to calculate $I_n(\theta)$.

— That is, using (3.23) to calculate $I_n(\theta)$ is much easier than using (3.18).

Theorem 3.4 (Alternative expression). Let $I_n(\theta)$ denote the Fisher information. If $E\{S(\theta)\}=0$, then

$$I_n(\theta) = E\left\{-\frac{\mathrm{d}^2\log L(\theta;\mathbf{x})}{\mathrm{d}\theta^2}\right\} = nI(\theta), \tag{3.23}$$

where

$$I(\theta) = E\left[\left\{\frac{\mathrm{d}\log f(X;\theta)}{\mathrm{d}\theta}\right\}^2\right] = E\left\{-\frac{\mathrm{d}^2\log f(X;\theta)}{\mathrm{d}\theta^2}\right\} \tag{3.24}$$

denotes the Fisher information for a single sample. ‖

Proof. From (3.21), we have

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int S(\theta)L(\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\
&= \int\left\{\frac{\mathrm{d}S(\theta)}{\mathrm{d}\theta}L(\theta)+S(\theta)L'(\theta)\right\}\mathrm{d}x_1\cdots\mathrm{d}x_n \\
&= E\left\{\frac{\mathrm{d}S(\theta)}{\mathrm{d}\theta}\right\}+\int S(\theta)S(\theta)L(\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\
&= E\left\{\frac{\mathrm{d}^2\log L(\theta)}{\mathrm{d}\theta^2}\right\}+E\{S^2(\theta)\} \\
&= E\left\{\frac{\mathrm{d}^2\log L(\theta;\mathbf{x})}{\mathrm{d}\theta^2}\right\}+I_n(\theta).
\end{aligned}$$

Therefore, the first equation in (3.23) follows.

Since $L(\theta;\mathbf{x}) = \prod_{i=1}^n f(X_i;\theta)$, we have

$$\log L(\theta;\mathbf{x}) = \sum_{i=1}^n \log f(X_i;\theta),$$

and

$$\frac{\mathrm{d}^2 \log L(\theta;\mathbf{x})}{\mathrm{d}\theta^2} = \sum_{i=1}^n \frac{\mathrm{d}^2 \log f(X_i;\theta)}{\mathrm{d}\theta^2}.$$

Therefore,

$$\begin{aligned}
I_n(\theta) &= E\left\{-\frac{\mathrm{d}^2 \log L(\theta;\mathbf{x})}{\mathrm{d}\theta^2}\right\} \\
&= \sum_{i=1}^n E\left\{-\frac{\mathrm{d}^2 \log f(X_i;\theta)}{\mathrm{d}\theta^2}\right\} \\
&= nE\left\{-\frac{\mathrm{d}^2 \log f(X;\theta)}{\mathrm{d}\theta^2}\right\} \\
&= nI(\theta).
\end{aligned}$$

This means the second equation in (3.23). □

20.3• How to check the condition $E\{S(\theta)\} = 0$ in Theorem 3.4?

— If the support of the population density $f(x;\theta)$ does not depend on θ, $\Longrightarrow E\{S(\theta)\} = 0$.

— That is, that the support of $f(x;\theta)$ is free from θ is a sufficient condition for $E\{S(\theta)\} = 0$. We only need to check the support of $f(x;\theta)$.

20.4• How to understand (3.24)?

— In (3.17), we consider the case of $n = 1$ and we have

$$S(\theta) = \ell'(\theta) = \frac{\mathrm{d}\log f(X;\theta)}{\mathrm{d}\theta} = \frac{f'(X;\theta)}{f(X;\theta)},$$

where $f'(X;\theta)$ is the derivative of $f(X;\theta)$ with respect to θ.

— On the one hand, if $\ell'(\theta)$ is close to zero, then the r.v. X does not provide much information about θ.

— On the other hand, if $|\ell'(\theta)|$ or $\{\ell'(\theta)\}^2$ is large, then the r.v. X provides much information about θ.

— Thus, we can use $\{\ell'(\theta)\}^2$ to measure the amount of information provided by X.

— However, since X is a random variable, we should consider the average case. Thus, the Fisher information (for θ) contained in the r.v. X should be defined by

$$I(\theta) = E\{\ell'(\theta)\}^2,$$

which is (3.24).

21• UMVUE AND EFFICIENT ESTIMATOR

- The inequality (3.22) told us that for any $\hat{\theta} \in \mathcal{U} = \{\hat{\theta}\colon\ E(\hat{\theta}) = \theta\}$ with infinite elements, we have $\text{Var}(\hat{\theta}) \geqslant 1/I_n(\theta)$, which can guide us to find a $\hat{\theta}^*$ such that

$$\hat{\theta}^* \in \mathcal{U} \quad \text{and} \quad \text{Var}(\hat{\theta}^*) = \min_{\hat{\theta}\in\mathcal{U}} \text{Var}(\hat{\theta}), \tag{3.25}$$

- This is a mathematical definition of a *uniformly minimum variance unbiased estimator* (UMVUE), which is to be shown in Definition 3.5.

- If $\hat{\theta}^*$ satisfies $\text{Var}(\hat{\theta}^*) = 1/I_n(\theta)$, then $\hat{\theta}^*$ is called efficient estimator of θ. For the general case, see Definition 3.6.

Definition 3.5 (UMVUE). An estimator $\hat{\theta}^*$ is called a UMVUE of θ if it is unbiased and has the smallest variance among all unbiased estimators. ‖

Definition 3.6 (Efficient estimator). If an unbiased estimator $\hat{\theta} = T(\mathbf{x})$ for $\tau(\theta)$ has variance equal to the Cramér–Rao lower bound, then $\hat{\theta}$ is called an *efficient estimator* for $\tau(\theta)$. ‖

21.1• Efficient estimator versus UMVUE

— Obviously, an efficient estimator for $\tau(\theta)$ is a UMVUE for $\tau(\theta)$; i.e.,

$$\text{efficient estimator} \implies \text{UMVUE}.$$

— However, the converse is not always true; i.e.,

$$\text{efficient estimator} \;\not\Leftarrow\; \text{UMVUE.}$$

— In other words, it is possible that a UMVUE whose variance does not attain the CR lower bound. See Example 3.20 .

Example 3.17 (Bernoulli distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. Then $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ is a UMVUE of θ.

Solution. Let $X \sim \text{Bernoulli}(\theta)$, then the pmf of X is $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, $x = 0, 1$. Then, from (3.24), we have

$$I(\theta) = E\left\{\frac{\mathrm{d}\log f(X;\theta)}{\mathrm{d}\theta}\right\}^2 = E\left(\frac{X}{\theta} - \frac{1-X}{1-\theta}\right)^2$$

$$= E\left\{\frac{X-\theta}{\theta(1-\theta)}\right\}^2 = \frac{\text{Var}(X)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$$

and

$$I_n(\theta) = nI(\theta) = \frac{n}{\theta(1-\theta)}.$$

Now, $\bar{X}$ is unbiased and

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{\theta(1-\theta)}{n} = \frac{1}{I_n(\theta)};$$

i.e., the variance attains the CR lower bound. Then $\bar{X}$ is a UMVUE of θ. $\|$

Example 3.18 (Normal distribution with known variance). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$ with known σ_0^2 and unknown μ. Show that $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ is a UMVUE for μ.

Solution. Let $X \sim N(\mu, \sigma_0^2)$, then the pdf of X is

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}\,\sigma_0}\exp\left\{-\frac{(x-\mu)^2}{2\sigma_0^2}\right\}$$

and

$$\log f(x;\mu) = -\log(\sqrt{2\pi}\,\sigma_0) - \frac{(x-\mu)^2}{2\sigma_0^2}.$$

From (3.24), we have

$$I(\mu) = E\left\{\frac{\mathrm{d}\log f(X;\mu)}{\mathrm{d}\mu}\right\}^2 = E\left(\frac{X-\mu}{\sigma_0^2}\right)^2 = \frac{1}{\sigma_0^2}.$$

and

$$I_n(\mu) = nI(\mu) = \frac{n}{\sigma_0^2}.$$

Now, $\bar{X}$ is unbiased and

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\sigma_0^2}{n} = \frac{1}{I_n(\mu)},$$

reaching the CR lower bound. Then $\bar{X}$ is a UMVUE of μ. ‖

21.2• Efficiency of an unbiased estimator

— In general, the *efficiency* of an unbiased estimator $\hat{\theta}$ for θ is defined by

$$\mathrm{eff}_{\hat{\theta}}(\theta) = \frac{\text{Cramér–Rao lower bound}}{\text{Actual variance}} = \frac{1/I_n(\theta)}{\mathrm{Var}(\hat{\theta})}. \tag{3.26}$$

Example 3.19 (Normal distribution with known mean). Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu_0, \theta)$ with known μ_0 and unknown θ. Calcualte the efficiency of $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$.

Solution. Let $X \sim N(\mu_0, \theta)$, then the pdf of X is

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(x-\mu_0)^2}{2\theta}\right\}$$

and

$$\log f(x;\theta) = -\frac{1}{2}\log(2\pi\theta) - \frac{(x-\mu_0)^2}{2\theta}.$$

From (3.24), we have

$$I(\theta) = E\left\{-\frac{\mathrm{d}^2\log f(X;\theta)}{\mathrm{d}\theta^2}\right\} = E\left\{-\frac{1}{2\theta^2} + \frac{(X-\mu_0)^2}{\theta^3}\right\} = \frac{1}{2\theta^2}.$$

and

$$I_n(\theta) = nI(\theta) = \frac{n}{2\theta^2}.$$

Since $(n-1)S^2/\theta \sim \chi^2(n-1)$, we have $E(S^2) = \theta$ and

$$\mathrm{Var}(S^2) = \frac{2\theta^2}{n-1} > \frac{2\theta^2}{n} = \frac{1}{I_n(\theta)}.$$

Therefore, S^2 is unbiased and its efficiency is

$$\mathrm{eff}_{S^2}(\theta) = \frac{1/I_n(\theta)}{\mathrm{Var}(S^2)} = \frac{n-1}{n} \to 1 \quad \text{as } n \to \infty;$$

i.e., S^2 is asymptotically efficient. ‖

Example 3.20 (Poisson distribution). Let $X \sim \mathrm{Poisson}(\theta)$ and $\tau(\theta) = \mathrm{e}^{-\theta}$. Define

$$\hat{\theta} = g(X) = \begin{cases} 1, & \text{if } X = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Use Theorem 3.7 to show that $\hat{\theta}$ is the unique UMVUE of $\tau(\theta)$, but $\mathrm{Var}(\hat{\theta})$ is larger than the CR lower bound.

Solution. In Example 3.21 let $n = 1$, we know that $T(X) = X$ is sufficient for θ. Next, we need to prove that $T(X) = X$ is also complete. If

$$E\{h(X)\} = \sum_{x=0}^{\infty} h(x) \frac{\theta^x}{x!} \mathrm{e}^{-\theta} = 0,$$

for $\theta > 0$, we have

$$\sum_{x=0}^{\infty} h(x) \frac{\theta^x}{x!} = 0.$$

Since $\theta^x/x! > 0$ for any $\theta > 0$ and $x \geqslant 0$, we obtain $h(X) \equiv 0$. Then $T = X$ is also complete. Since $\hat{\theta} = g(X) = g(T)$ is unbiased for $\tau(\theta)$, it is the unique UMVUE for $\tau(\theta)$ according to Theorem 3.7.

Since $I(\theta) = 1/\theta$, and the CR lower bound is

$$\frac{\{\tau'(\theta)\}^2}{I(\theta)} = \theta\,\mathrm{e}^{-2\theta},$$

we have

$$\mathrm{Var}(\hat{\theta}) = \mathrm{e}^{-\theta}(1 - \mathrm{e}^{-\theta}) > \theta\,\mathrm{e}^{-2\theta},$$

which completes the proof. ‖

21.3• Is the UMVUE unique?

— If $\mathcal{U} = \{\hat{\theta}\colon\ E(\hat{\theta}) = \theta\}$ is not an empty set, then there exists *at most* one UMVUE of θ. In other words, the number of UMVUEs is zero or one.

21.4• How to find the unique UMVUE?

— In this subsection, we provide a sufficient condition; i.e.,

$$\text{if } \hat{\theta} \text{ is an efficient estimator } \Longrightarrow \hat{\theta} \text{ is the unique UMVUE.}$$

— §3.4.4 will provide a sufficient and necessary condition, which involves two important notions: Sufficiency (§3.4.3) and completeness (§3.4.3).

3.4.3 Sufficiency

22• Motivation from data reduction

- In many of the estimation problems, we need to summarize the information contained in the sample $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.
- That is, we need to find some function of the sample that tells us just as much about θ as the sample itself.
- Such a function would be sufficient for estimation purposes and accordingly is called a *sufficient statistic*.

22.1• Raw data and reduced data

— Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \mu, \sigma^2)$, then $X_1, \ldots, X_n$ are called *raw data*.

— The quantities such as the sample mean $\bar{X}$, the sample variance S^2, the smallest order statistic $X_{(1)}$ and the largest order statistic $X_{(n)}$ are called *reduced data*.

— Given raw data, any reduce data can be determined uniquely; while the converse may not be true:

$$\text{raw data} \;\overset{\rightarrow}{\not\leftarrow}\; \text{reduced data.}$$

22.2• Intuitive interpretation on a sufficient statistic

— To estimate the population mean μ, we only need to use the reduced datum $\bar{X}$, which contains all information about μ.

— In other words, using $\bar{X}$ to estimate μ will not lose any information.

— However, using $\sum_{i=1}^{n-1} X_i/(n-1)$ to estimate μ will lose information from X_n.

— Hence, $\bar{X}$ is a sufficient estimator of μ while $\sum_{i=1}^{n-1} X_i/(n-1)$ is not a sufficient estimator of μ.

23• Single sufficient statistic

Definition 3.7 (Sufficient statistic). A statistic $T(\mathbf{x})$ is said to be a *sufficient statistic* of θ if the conditional distribution of $\mathbf{x}$, given $T(\mathbf{x}) = t$, does not depend on θ for any value of t. In discrete case, this means that

$$\Pr\{X_1 = x_1, \ldots, X_n = x_n; \theta | T(\mathbf{x}) = t\} = h(\boldsymbol{x})$$

does not depend on θ. ‖

23.1• Deeply understanding Definition 3.7

— The definition says that if you know the value of the sufficient statistic, then the sample values themselves are not needed and can tell you nothing more about θ.

— This is true since the distribution of the sample given the sufficient statistic does not depend on θ.

— The joint distribution of $\mathbf{x}$ and $T(\mathbf{x})$ is

$$\begin{aligned} & \Pr\{X_1 = x_1, \ldots, X_n = x_n, T(\mathbf{x}) = t; \theta\} \\ = \; & \Pr\{X_1 = x_1, \ldots, X_n = x_n; \theta | T(\mathbf{x}) = t\} \times \Pr\{T(\mathbf{x}) = t; \theta\} \\ = \; & h(\boldsymbol{x}) \times \Pr\{T(\mathbf{x}) = t; \theta\}, \end{aligned}$$

where the left-hand side is, in general, the joint distribution of $\mathbf{x}$ subject to the constraint $T(\mathbf{x}) = t$.

— Thus, the MLE $\hat{\theta}$ can be obtained by maximizing $\log[\Pr\{T(\mathbf{x}) = t; \theta\}]$.

— Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ and $T(\mathbf{x}) = \sum_{i=1}^n X_i$. We have

$$\begin{aligned}
& \Pr\{X_1 = x_1, \ldots, X_n = x_n, T(\mathbf{x}) = t; \theta\} \\
= \; & \Pr\{X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n = t - \textstyle\sum_{j=1}^{n-1} x_j; \theta\} \\
= \; & \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \\
= \; & \theta^t(1-\theta)^{n-t}.
\end{aligned}$$

On the other hand, since $T(\mathbf{x}) = \sum_{i=1}^n X_i \sim \text{Binomial}(n, \theta)$, we obtain

$$\Pr\{T(\mathbf{x}) = t; \theta\} = \binom{n}{t}\theta^t(1-\theta)^{n-t}.$$

The MLE $\hat{\theta} = T(\mathbf{x})/n = \bar{X}$.

Example 3.21 (Poisson distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$, where $\theta > 0$. Show that $T(\mathbf{x}) = \sum_{i=1}^n X_i$ is a sufficient statistic of θ.

Solution. From Example 2.12, we have $T(\mathbf{x}) = \sum_{i=1}^n X_i \sim \text{Poisson}(n\theta)$. Since the conditional distribution

$$\begin{aligned}
& \Pr\{X_1 = x_1, \ldots, X_n = x_n;\ \theta | T(\mathbf{x}) = t\} \\
= \; & \frac{\Pr(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = t - \sum_{i=1}^{n-1} x_i;\ \theta)}{\Pr(\sum_{i=1}^n X_i = t)} \\
= \; & \left(\prod_{i=1}^{n-1} \frac{\theta^{x_i}\,\mathrm{e}^{-\theta}}{x_i!}\right) \cdot \frac{\theta^{t - \sum_{i=1}^{n-1} x_i}\,\mathrm{e}^{-\theta}}{(t - \sum_{i=1}^{n-1} x_i)!} \Bigg/ \frac{(n\theta)^t\,\mathrm{e}^{-n\theta}}{t!} \\
= \; & \frac{t!}{x_1! \cdots x_{n-1}!(t - \sum_{i=1}^{n-1} x_i)!} \cdot \frac{1}{n^t}
\end{aligned}$$

does not depend on θ for any value of t, $T(\mathbf{x})$ is a sufficient statistic of θ. $\|$

Example 3.22 (Bernoulli distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, where $\theta > 0$. Show that $T(\mathbf{x}) = \sum_{i=1}^n X_i$ is a sufficient statistic of θ.

Solution. Note that $T(\mathbf{x}) = \sum_{i=1}^n X_i \sim \text{Binomial}(n, \theta)$, we have

$$\begin{aligned}
& \Pr\{X_1 = x_1, \dots, X_n = x_n;\ \theta | T(\mathbf{x}) = t\} \\
= \ & \frac{\Pr(X_1 = x_1, \dots, X_n = x_n;\ \theta)}{\Pr\{T(\mathbf{x}) = t\}} \\
= \ & \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
= \ & \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \qquad [\because \textstyle\sum_{i=1}^n x_i = t] \\
= \ & \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
= \ & \frac{1}{\binom{n}{t}},
\end{aligned}$$

which does not depend on θ for any value of t. Therefore, $T(\mathbf{x})$ is a sufficient statistic of θ. ‖

23.2• How to find a sufficient statistic?

Theorem 3.5 (Factorization theorem). A statistic $T(\mathbf{x})$ is a sufficient statistic of the unknown parameter θ iff the joint pdf (or pmf) can be written in the form

$$f(x_1, \dots, x_n; \theta) = f(\boldsymbol{x}; \theta) = g(T(\boldsymbol{x}); \theta) \times h(\boldsymbol{x}), \tag{3.27}$$

where $h(\boldsymbol{x})$ does not depend on θ, $g(T; \theta)$ is a function of both T and θ, and it depends on $x_1, \dots, x_n$ only through T. ‖

Proof. We give a proof for the discrete case.

" $\Longleftarrow$ " (Sufficiency). Assume that $\Pr(\mathbf{x} = \boldsymbol{x}; \theta) = g(T(\boldsymbol{x}); \theta) \times h(\boldsymbol{x})$. Note that

$$\begin{aligned}
\Pr\{T(\mathbf{x}) = t; \theta\} &= \sum_{T(\boldsymbol{x})=t} \Pr(\mathbf{x} = \boldsymbol{x}; \theta) \\
&= \sum_{T(\boldsymbol{x})=t} g(T(\boldsymbol{x}); \theta) \times h(\boldsymbol{x}) \\
&= g(t; \theta) \sum_{T(\boldsymbol{x})=t} h(\boldsymbol{x}),
\end{aligned}$$

we obtain

$$
\begin{aligned}
\Pr\{\mathbf{x}=\boldsymbol{x};\theta|T(\mathbf{x})=t\} &= \begin{cases} 0, & \text{if } T(\boldsymbol{x})\neq t,\\ \dfrac{\Pr\{\mathbf{x}=\boldsymbol{x},T(\mathbf{x})=t;\theta\}}{\Pr\{T(\mathbf{x})=t;\theta\}}, & \text{if } T(\boldsymbol{x})=t,\end{cases}\\
&= \begin{cases} 0, & \text{if } T(\boldsymbol{x})\neq t,\\ \dfrac{\Pr(\mathbf{x}=\boldsymbol{x};\theta)}{\Pr\{T(\mathbf{x})=t;\theta\}}, & \text{if } T(\boldsymbol{x})=t,\end{cases}\\
&= \begin{cases} 0, & \text{if } T(\boldsymbol{x})\neq t,\\ \dfrac{h(\boldsymbol{x})}{\sum_{T(\boldsymbol{x})=t} h(\boldsymbol{x})}, & \text{if } T(\boldsymbol{x})=t.\end{cases}
\end{aligned}
$$

It does not depend on θ, then $T(\mathbf{x})$ is sufficient for θ.

" $\Longrightarrow$ " (Necessity). Assume that $T(\mathbf{x})$ is sufficient, then

$$\Pr(\mathbf{x}=\boldsymbol{x};\theta)=\Pr\{T(\mathbf{x})=t\}\times\Pr\{\mathbf{x}=\boldsymbol{x};\theta|T(\mathbf{x})=t\} \tag{3.28}$$

with $T(\boldsymbol{x})=t$. Let

$$\Pr\{T(\mathbf{x})=t\}=g(t;\theta) \quad \text{and} \quad \Pr\{\mathbf{x}=\boldsymbol{x};\theta|T(\mathbf{x})=t\}=h(\boldsymbol{x}),$$

then (3.28) becomes

$$\Pr(\mathbf{x}=\boldsymbol{x};\theta)=g(t;\theta)\times h(\boldsymbol{x})$$

and (3.27) follows. □

Example 3.23 (Normal distribution with known variance). Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} N(\theta,\sigma_0^2)$ with known σ_0^2. Then $\bar{X}$ is a sufficient statistic for θ.

Solution. The joint pdf of $X_1,\ldots,X_n$ is

$$
\begin{aligned}
f(x_1,\ldots,x_n;\theta) &= \frac{1}{(\sqrt{2\pi}\,\sigma_0)^n}\exp\left\{-\frac{\sum_{i=1}^n(x_i-\theta)^2}{2\sigma_0^2}\right\}\\
&= \frac{1}{(\sqrt{2\pi}\,\sigma_0)^n}\exp\left\{-\frac{\sum_{i=1}^n(x_i-\bar{x}+\bar{x}-\theta)^2}{2\sigma_0^2}\right\}\\
&= \frac{1}{(\sqrt{2\pi}\,\sigma_0)^n}\exp\left\{-\frac{n(\bar{x}-\theta)^2}{2\sigma_0^2}\right\}\\
&\quad \times\ \exp\left\{-\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{2\sigma_0^2}\right\}
\end{aligned}
$$

Then $T=\bar{X}$ is sufficient for θ. ‖

Example 3.24 (Shift exponential distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \theta)$, where

$$f(x;\theta) = \begin{cases} \exp\{-(x-\theta)\}, & \text{if } \ x > \theta, \\ 0, & \text{elsewhere.} \end{cases}$$

Then $X_{(1)} = \min(X_1, \ldots, X_n)$ is a sufficient statistic for θ.

Solution. The joint pdf of $X_1, \ldots, X_n$ is

$$\begin{aligned} f(x_1, \ldots, x_n; \theta) &= \prod_{i=1}^{n} \exp\{-(x_i - \theta)\} \cdot I_{(\theta,\infty)}(x_i) \\ &= \mathrm{e}^{-\sum_{i=1}^{n} x_i + n\theta} \prod_{i=1}^{n} I_{(\theta,\infty)}(x_i) \\ &= \mathrm{e}^{n\theta} I_{(\theta,\infty)}(x_{(1)}) \times \mathrm{e}^{-\sum_{i=1}^{n} x_i}. \end{aligned}$$

Then $X_{(1)}$ is sufficient for θ. $\|$

Example 3.25 (A special beta distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \theta)$, where

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1}, & \text{if } \ 0 < x < 1, \\ 0, & \text{elsewhere,} \end{cases}$$

where $\theta > 0$. Then $\prod_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Solution. The joint pdf of $X_1, \ldots, X_n$ is

$$f(x_1, \ldots, x_n; \theta) = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1} \times 1.$$

Then $\prod_{i=1}^{n} X_i$ is sufficient for θ. The function $h(\boldsymbol{x})$ in (3.27) may be a constant as shown in this example. $\|$

23.3• Why do we need a sufficient statistic?

— Given a sufficient statistic T and an unbiased estimator Y of θ, we can immediately find another unbiased estimator Z with a smaller variance, see the beginning of §3.4.4.

— Usefulness in finding a unique UMVUE of θ: If a sufficient statistic $T(\mathbf{x})$ is also a complete statistic simultaneously, then we can immediately identify a unique UMVUE of θ, see Theorem 3.7.

23.4• Why does it take the name of sufficient statistic?

— For the normal population with known variance, we know that the sample mean $\bar{X}$ is an unbiased estimator of the population mean θ.

— From Example 3.23, we can see that $\bar{X}$ is also a sufficient statistic of θ.

— $\bar{X}$ contains *all/sufficient* information from the random sample $X_1, \ldots, X_n$ to estimate θ.

— $\sum_{i=1}^{n-1} X_i/(n-1)$ is also unbiased estimator of θ but it is not a sufficient statistic.

23.5• Is a sufficient statistic unique?

— First, we note that a sufficient statistic is not unique.

— If $Y_1 = T(\mathbf{x})$ is a sufficient statistic for θ and $Y_2 = g(Y_1)$, where $g(\cdot)$ is a *one-to-one* function, then Y_2 is also sufficient.

— For instance, in Example 3.23, $\sum_{i=1}^{n} X_i = n\bar{X}$ is another sufficient statistic of θ but $\bar{X}^2$ is not sufficient.

23.6• Sufficient statistic versus sufficient estimator

— An estimator is a meaningful statistic.

— In Example 3.23, $\bar{X}$ is a sufficient statistic of θ, and it is also a sufficient estimator of θ.

— Note that $\sum_{i=1}^{n} X_i$ is just a sufficient statistic of θ, not a sufficient estimator of θ.

23.7• Sufficient statistic versus unbiased estimator

— For the normal population with known variance, both $\bar{X}$ and $n\bar{X}$ are sufficient statistics for θ. The former is unbiased while the latter is biased.

— Both $\bar{X}$ and $\sum_{i=1}^{n-1} X_i/(n-1)$ are unbiased estimators for θ. The former is sufficient while the latter is not sufficient.

23.8• Statistic versus estimator

— An estimator $\Longrightarrow$ a statistic.

— Based on different criteria, estimators could be classified into:

$$\begin{cases} \text{biased estimator, unbiased estimator;} \\ \text{MLE, moment estimator, Bayesian estimator;} \\ \text{efficient estimator, UMVUE;} \\ \text{sufficient estimator, complete estimator.} \end{cases}$$

— For example,

$$\frac{1}{a}\sum_{i=1}^{n} X_i: \qquad \text{statistic (for any non-zero constant } a\text{),}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i: \qquad \begin{cases} \text{MLE,} \\ \text{moment estimator,} \\ \text{unbiased estimator,} \\ \text{UMVUE,} \\ \text{sufficient estimator.} \end{cases}$$

24• Joint sufficient statistics

- For some problems, no single sufficient statistic exists.
- However, there will always exist joint sufficient statistics.

Definition 3.8 (Joint sufficient statistics). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \boldsymbol{\theta})$. The statistics $T_1(\mathbf{x}), \ldots T_r(\mathbf{x})$ are said to be *jointly sufficient* if the conditional distribution of $\mathbf{x}$, given $T_1 = t_1, \ldots, T_r = t_r$, does not depend on $\boldsymbol{\theta}$. ‖

Theorem 3.6 (Factorization theorem with joint sufficient statistics). A set of statistics $T_1(\mathbf{x}), \ldots, T_r(\mathbf{x})$ is jointly sufficient for the parameter vector $\boldsymbol{\theta}$ iff the joint pdf (or pmf) can be written in the form

$$f(x_1, \ldots, x_n; \boldsymbol{\theta}) = f(\boldsymbol{x}; \boldsymbol{\theta}) = g(T_1(\boldsymbol{x}), \ldots, T_r(\boldsymbol{x}); \boldsymbol{\theta}) \times h(\boldsymbol{x}), \tag{3.29}$$

where $h(\boldsymbol{x})$ does not depend on $\boldsymbol{\theta}$, $g(T_1, \ldots, T_r; \boldsymbol{\theta})$ depends on $x_1, \ldots, x_n$ only through $T_1, \ldots, T_r$. ‖

24.1• Comments on Theorem 3.6

— If $T_1(\mathbf{x}), \ldots, T_r(\mathbf{x})$ is a set of jointly sufficient statistics, then any set of one-to-one functions/transformations of $T_1(\mathbf{x}), \ldots, T_r(\mathbf{x})$ is also jointly sufficient.

— In addition, the function $h(\boldsymbol{x})$ in (3.29) may be a constant.

Example 3.26 (Normal distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Find jointly sufficient statistics for $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Solution. The joint pdf of $X_1, \ldots, X_n$ is

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

Hence, $\sum X_i$ and $\sum X_i^2$ are jointly sufficient. It can be shown that $\bar{X}$ and $S^2 = \{1/(n-1)\} \sum (X_i - \bar{X})^2$ are one-to-one functions of $\sum X_i$ and $\sum X_i^2$; so $\bar{X}$ and S^2 are also jointly sufficient. ‖

3.4.4 Completeness

25• Why need we the notion of the complete statistic?

- Assume that $T(\mathbf{x})$ is *sufficient* for θ, and $Y(\mathbf{x})$ is an *unbiased estimator* of $\tau(\theta)$, a function of θ.

- Let $Z \hat{=} E(Y|T) = \varphi(T)$, then we have

$$E(Z) \overset{(1.27)}{=} E(Y) = \tau(\theta), \quad \text{and} \tag{3.30}$$

$$\begin{aligned} \text{Var}(Z) &= 0 + \text{Var}(Z) \\ &\leqslant E\Big\{\underbrace{\text{Var}(Y|T)}_{\text{non-negative}}\Big\} + \text{Var}(Z) \\ &= E\{\text{Var}(Y|T)\} + \text{Var}\{E(Y|T)\} \overset{(1.28)}{=} \text{Var}(Y). \end{aligned}$$

- Thus, from a sufficient statistic T and an unbiased estimator Y, we can find a set of unbiased estimators $\{Z_j\}_{j=1}^{m}$ satisfying

$$\begin{aligned} Z_1 &= E(Y|T), \\ Z_2 &= E(Z_1|T), \\ Z_3 &= E(Z_2|T), \\ &\vdots \\ Z_m &= E(Z_{m-1}|T), \end{aligned}$$

  and $\mathrm{Var}(Z_m) \leqslant \mathrm{Var}(Z_{m-1}) \leqslant \cdots \leqslant \mathrm{Var}(Z_1) \leqslant \mathrm{Var}(Y)$.

- Let $\mathcal{U} = \{Z\colon\ E(Z) = \tau(\theta)\}$ and $\#\mathcal{U}$ is infinite.

- We wonder if we could find a $Z^* \in \mathcal{U}$ such that

$$\mathrm{Var}(Z^*) \leqslant \mathrm{Var}(Z), \quad \forall\, Z \in \mathcal{U}.$$

  In other words, Z^* is the UMVUE of $\tau(\theta)$.

- We wish that $Z^* = Z_m = \cdots = Z_1$. The notion of "complete statistic" facilitates this purpose.

26• Definition of a complete statistic

Definition 3.9 (Completeness). Let $X_1, \ldots, X_n$ denote a random sample from the pdf (or pmf) $f(x;\theta)$ with parameter space Θ and let $T(\mathbf{x})$ be a statistic, where $\mathbf{x} = (X_1, \ldots, X_n)^\top$. The statistic T is said to be *complete* if

$$E\{h(T)\} = 0 \quad \text{for all } \theta \in \Theta$$

implies that $h(T) = 0$ with probability 1; i.e.,

$$\Pr\{h(T) = 0\} = 1 \quad \text{for all } \theta \in \Theta,$$

where the function $h(T)$ is a statistic. ‖

26.1• Alternative statement

— Alternatively, we can say: T is complete iff the *only* unbiased estimator of 0 that is a function of T is the statistic that is identically 0 with probability 1.

27• How to understand the completeness?

- We need two "bridges" to reach the *uniqueness* of UMVUE.

27.1• Case I: $\mathcal{U} = \{\hat{\theta}\colon E(\hat{\theta}) = \theta\}$ with finite elements

— Let $\hat{\theta}_i \in \mathcal{U}$ for $i = 1, 2$.

— The first bridge is the *sufficiency*; i.e., suppose that we have found a sufficient statistic T for θ.

— From (3.30), we know that $Z_i = E(\hat{\theta}_i|T) = h_i(T)$, $i = 1, 2$, are two unbiased estimators of θ so that $E(Z_1 - Z_2) = \theta - \theta = 0$,

$$\mathrm{Var}(Z_1) \leqslant \mathrm{Var}(\hat{\theta}_1) \quad \text{and} \quad \mathrm{Var}(Z_2) \leqslant \mathrm{Var}(\hat{\theta}_2).$$

— Which one should we choose? Z_1 or Z_2?

— Of course, we choose Z_1 if $\mathrm{Var}(Z_1) \leqslant \mathrm{Var}(Z_2)$. Otherwise, we choose Z_2.

— In other words, we do not need the "second bridge" (i.e., the completeness) for Case I.

27.2• Case II: $\mathcal{U} = \{\hat{\theta}\colon E(\hat{\theta}) = \theta\}$ with infinite elements

— Let $\hat{\theta}_i \in \mathcal{U}$ for $i = 1, 2, \ldots$.

— Define $Z_i = E(\hat{\theta}_i|T) = h_i(T)$, $i = 1, 2, \ldots$, we have $E(Z_i - Z_j) = E\{h_i(T) - h_j(T)\} = \theta - \theta = 0$, and

$$\mathrm{Var}(Z_i) \leqslant \mathrm{Var}(\hat{\theta}_i), \quad i = 1, 2, \ldots.$$

— Which Z_i should we choose?

— Then, we wish to find a second "bridge" such that $Z_i = Z_j$ with probability 1. The second bridge is the completeness.

Example 3.27 (Uniform distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$, where $\Theta = \{\theta\colon\ \theta > 0\}$. Show that $X_{(n)} = \max(X_1, \ldots, X_n)$ is complete.

Solution. We must show that if $E\{h(X_{(n)})\} = 0$ for all $\theta > 0$, then $\Pr\{h(X_{(n)}) = 0\} = 1$ for all $\theta > 0$. From Example 2.16, the density of $X_{(n)}$ is

$$f_n(x) = nx^{n-1}/\theta^n, \quad 0 < x < \theta.$$

Note that

$$E\{h(X_{(n)})\} = \int h(x) f_n(x)\,\mathrm{d}x = \int_0^\theta h(x)\theta^{-n} n x^{n-1}\,\mathrm{d}x,$$

and $E\{h(X_{(n)})\} = 0$ for all $\theta > 0$ when and only when

$$\int_0^\theta h(x) x^{n-1}\,\mathrm{d}x = 0 \quad \text{for all } \theta > 0.$$

Differentiating both sides of this identity with respect to θ produces

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

which in turn implies $h(\theta) = 0$ for all $\theta > 0$. ‖

28• How to find the unique UMVUE?

Theorem 3.7 (Lehmann–Scheffé theorem). Let $T(\mathbf{x})$ is a complete sufficient statistic for θ. If $g(T)$ is an unbiased estimator of $\tau(\theta)$, then $g(T)$ is the unique UMVUE for $\tau(\theta)$. ‖

Proof. Let Y be any unbiased estimator of $\tau(\theta)$ and let $\varphi(T) = E(Y|T)$, then

$$E\{\varphi(T)\} = \tau(\theta) \quad \text{and} \quad \mathrm{Var}\{\varphi(T)\} \leqslant \mathrm{Var}(Y).$$

Therefore,

$$E\{g(T) - \varphi(T)\} = \tau(\theta) - \tau(\theta) = 0 \quad \text{for all } \theta.$$

As T is complete, this implies that $g(T) = \varphi(T)$ with probability 1 and

$$\mathrm{Var}\{g(T)\} = \mathrm{Var}\{\varphi(T)\} \leqslant \mathrm{Var}(Y).$$

Consequently, $g(T)$ is the unique function of T which is unbiased and has a smaller variance than any other unbiased estimator has. Then $g(T)$ is the unique UMVUE of $\tau(\theta)$. □

Example 3.28 (Bernoulli distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Bernoulli(θ), where $\Theta = \{\theta\colon 0 < \theta < 1\}$. Show that the statistic $T = \sum_{i=1}^n X_i$ is a complete sufficient statistic for θ. Find the UMVUE for θ.

Solution. The joint pdf of $X_1, \ldots, X_n$ is

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^t(1-\theta)^{n-t},$$

where $t = \sum_{i=1}^n x_i$. By Theorem 3.5, $T = \sum_{i=1}^n X_i$ is sufficient, and $T \sim$ Binomial(n, θ). Now assume that a function $h(T)$ satisfies

$$E\{h(T)\} = \sum_{t=0}^{n} h(t)\Pr(T = t) = \sum_{t=0}^{n} h(t)\binom{n}{t}\theta^t(1-\theta)^{n-t} = 0, \tag{3.31}$$

for $0 < \theta < 1$. Let $y = \theta/(1-\theta)$, then (3.31) becomes

$$\sum_{t=0}^{n} h(t)\binom{n}{t} y^t = 0, \quad y > 0.$$

A polynomial is identical to zero, then all coefficients are zero. Thus

$$h(t)\binom{n}{t} = 0 \qquad \text{for } t = 0, 1, \ldots, n.$$

Hence $h(T) \equiv 0$. Then T is also complete. Since $\bar{X} = T/n$ is unbiased for θ, it is the unique UMVUE for θ according to Theorem 3.7. ‖

28.1• Remarks on Example 3.28

— Note that $T = \sum_{i=1}^n X_i$ is sufficient for θ and $Y = T/n$ is an unbiased estimator of θ.

— We have

$$\begin{aligned} Z_1 &= E(Y|T) = E\left(\frac{T}{n}\middle|T\right) = \frac{T}{n} = Y \quad \text{and} \\ Z_2 &= E(Z_1|T) = \frac{T}{n} = Y \end{aligned}$$

so that $\mathcal{U} = \{\hat{\theta}\colon E(\hat{\theta}) = \theta\} = \{Y\}$ and $\#\mathcal{U} = 1$.

28.2• Remarks on Example 3.23

— From Example 3.23, we know that $T = \bar{X} = \sum_{i=1}^{n} X_i/n$ is sufficient for θ and $Y = T$ is an unbiased estimator of θ.

— We have $Z = E(Y|T) = E(T|T) = T$ so that $\mathcal{U} = \{\hat{\theta}\colon\ E(\hat{\theta}) = \theta\} = \{\bar{X}\}$ and $\#\mathcal{U} = 1$.

3.5 Limiting Properties of MLE

29• MLE weakly converges in probability to its true value

- In §3.1.4, we have stated the invariance property of MLE. In this section, we introduce limiting properties of MLE.

- We rewrite Definition 2.3 as follows: A sequence of r.v.'s $\{X_n\}_{n=1}^{\infty}$ is said to weakly converge in probability to an r.v. X, denoted by $X_n \xrightarrow{\text{P}} X$, if for any $\varepsilon > 0$, $\lim_{n\to\infty} \Pr(|X_n - X| \geqslant \varepsilon) = 0$.

- Let $\{X_n\}_{n=1}^{\infty}$ be i.i.d. from a population with pdf $f(x;\theta)$. Let $\hat{\theta}_n$ be the MLE of θ based on $X_1, \dots, X_n$. Then under certain regularity conditions, we have
$$\hat{\theta}_n \xrightarrow{\text{P}} \theta \quad \text{as} \quad n \to \infty. \tag{3.32}$$

29.1• MLE also converges in distribution to its true value

— The conclusion in (3.32) states that when $n \to \infty$, the MLE $\hat{\theta}_n$ weakly converges in probability to the true value of the parameter.

— From Property 2.1 in §2.5.3, we obtain $\hat{\theta}_n \xrightarrow{\text{L}} \theta$; i.e., the MLE $\hat{\theta}_n$ converges in distribution to the true value of the parameter.

30• MLE is asymptotically normally distributed

- Let $\{X_n\}_{n=1}^{\infty} \stackrel{\text{iid}}{\sim} f(x;\theta)$ and $\hat{\theta}_n$ be the MLE of θ based on $X_1, \dots, X_n$.

- Let $S(\theta; \mathbf{x})$ with $\mathbf{x} = (X_1, \dots, X_n)^\top$ and $I_n(\theta) = nI(\theta)$ denote the score function and the Fisher information, respectively.

- If $E\{S(\theta;\mathbf{x})\}=0$ and $\mathrm{Var}\{S(\theta;\mathbf{x})\}=nI(\theta)$, then

$$\frac{S(\theta;\mathbf{x})}{\sqrt{nI(\theta)}} \xrightarrow{\mathrm{L}} N(0,1) \quad \text{as} \quad n\to\infty \tag{3.33}$$

  and

$$\{nI(\theta)\}^{1/2}(\hat\theta_n-\theta) \xrightarrow{\mathrm{L}} N(0,1) \quad \text{as} \quad n\to\infty. \tag{3.34}$$

- The corresponding proofs of (3.33) and (3.34) are given in §3.6.

30.1• Remarks on (3.34)

— The MLE $\hat\theta_n$ is an asymptotically unbiased estimator of θ.

— The MLE $\hat\theta_n$ is an asymptotically UMVUE because it reaches the CR lower bound in the sense that

$$\lim_{n\to\infty}\mathrm{eff}_{\hat\theta_n}(\theta)=\lim_{n\to\infty}\frac{1/I_n(\theta)}{\mathrm{Var}(\hat\theta_n)}=1.$$

— The MLE $\hat\theta_n$ is asymptotically normally distributed.

Example 3.29 (A special beta distribution). Let $X_1,\dots,X_n \overset{\mathrm{iid}}{\sim} f(x;\theta)$, where

$$f(x;\theta)=\begin{cases}\theta x^{\theta-1}, & \text{if } 0<x<1,\\ 0, & \text{elsewhere,}\end{cases}$$

and $\Theta=\{\theta\colon\ \theta>0\}$. Find the MLE of θ and study its limiting properties.

Solution. The likelihood function is

$$L(\theta)=\prod_{i=1}^n \theta x_i^{\theta-1}=\theta^n\left(\prod_{i=1}^n x_i\right)^{\theta-1}.$$

The MLE of θ is $\hat\theta_n=-n/\sum_{i=1}^n\log X_i$. Since

$$I(\theta)=E\left\{-\frac{\mathrm{d}^2\log f(X;\theta)}{\mathrm{d}\theta^2}\right\}=\frac{1}{\theta^2},$$

we have

$$(n/\theta^2)^{1/2}(\hat\theta_n-\theta)\xrightarrow{\mathrm{L}} N(0,1).$$

For large n, we have approximately

$$(n/\hat\theta_n^2)^{1/2}(\hat\theta_n-\theta)\ \dot\sim\ N(0,1). \qquad \|$$

3.6 Some Challenging Questions

Example 3.30 (Grouped Dirichlet distribution). Let $(x_1, \ldots, x_4, x_{12}, x_{34})$ be observed values of random variables $(X_1, \ldots, X_4, X_{12}, X_{34})$, respectively. Assume that the likelihood function of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)^\top$ is

$$L(\boldsymbol{\theta}) = \left(\prod_{i=1}^{4} \theta_i^{x_i}\right) \cdot (\theta_1 + \theta_2)^{x_{12}} (\theta_3 + \theta_4)^{x_{34}}, \quad \boldsymbol{\theta} \in \mathbb{T}_4.$$

Find the MLE of $\boldsymbol{\theta}$ subject to the constraints $\theta_i \geqslant 0$ and $\sum_{i=1}^4 \theta_i = 1$.

Solution. The log-likelihood function is given by

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{4} x_i \log \theta_i + x_{12} \log(\theta_1 + \theta_2) + x_{34} \log(\theta_3 + \theta_4) \\ &= \sum_{i=1}^{3} x_i \log \theta_i + x_4 \log(1 - \theta_1 - \theta_2 - \theta_3) \\ &\quad + x_{12} \log(\theta_1 + \theta_2) + x_{34} \log(1 - \theta_1 - \theta_2). \end{aligned}$$

Solving the following system of equations

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} &= \frac{x_1}{\theta_1} - \frac{x_4}{1 - \theta_1 - \theta_2 - \theta_3} \\ &\quad + \frac{x_{12}}{\theta_1 + \theta_2} - \frac{x_{34}}{1 - \theta_1 - \theta_2} = 0, \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_2} &= \frac{x_2}{\theta_2} - \frac{x_4}{1 - \theta_1 - \theta_2 - \theta_3} \\ &\quad + \frac{x_{12}}{\theta_1 + \theta_2} - \frac{x_{34}}{1 - \theta_1 - \theta_2} = 0, \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_3} &= \frac{x_3}{\theta_3} - \frac{x_4}{1 - \theta_1 - \theta_2 - \theta_3} = 0, \end{aligned} \tag{3.35}$$

we obtain

$$\begin{aligned} \frac{x_1}{\theta_1} &= \frac{x_2}{\theta_2} = \frac{x_1 + x_2}{\theta_1 + \theta_2}, \\ \frac{x_3}{\theta_3} &= \frac{x_4}{\theta_4} = \frac{x_3 + x_4}{\theta_3 + \theta_4}. \end{aligned}$$

Hence, from (3.35), we have

$$\frac{x_1+x_2}{\theta_1+\theta_2}-\frac{x_3+x_4}{\theta_3+\theta_4}+\frac{x_{12}}{\theta_1+\theta_2}-\frac{x_{34}}{\theta_3+\theta_4}=0,$$

or

$$\frac{x_1+x_2+x_{12}}{\theta_1+\theta_2}=\frac{x_3+x_4+x_{34}}{\theta_3+\theta_4}=\frac{N}{1},$$

where $N \hat{=} \sum_{i=1}^4 x_i + x_{12} + x_{34}$, resulting in

$$\theta_1+\theta_2=\frac{x_1+x_2+x_{12}}{N}.$$

Therefore, the MLE of θ_i is

$$\hat{\theta}_i=\frac{X_i}{N}\left\{\frac{X_1+X_2+X_{12}}{X_1+X_2}\cdot I_{(1\leqslant i\leqslant 2)}+\frac{X_3+X_4+X_{34}}{X_3+X_4}\cdot I_{(3\leqslant i\leqslant 4)}\right\}. \qquad \|$$

31• Proof of (3.33) and (3.34)

31.1• Recall the central limit theorem

— Let $\{Y_n\}_{n=1}^\infty$ be i.i.d. random variables with the common mean μ and common variance $\sigma^2 > 0$.

— The central limit theorem presented in Theorem 2.9 states that

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i-\sqrt{n}\mu}{\sigma}=\frac{\sqrt{n}(\sum_{i=1}^n Y_i/n-\mu)}{\sigma}\stackrel{\text{L}}{\to} N(0,1) \tag{3.36}$$

as $n \to \infty$.

31.2• Proof of (3.33)

— The likelihood function of θ is $L(\theta;\mathbf{x})=\prod_{i=1}^n f(X_i;\theta)$ so that the log-likelihood function is $\ell(\theta;\mathbf{x})=\sum_{i=1}^n \log f(X_i;\theta)$.

— The score function is

$$S(\theta;\mathbf{x})=\frac{\mathrm{d}\ell(\theta;\mathbf{x})}{\mathrm{d}\theta}=\sum_{i=1}^n\frac{\mathrm{d}\log f(X_i;\theta)}{\mathrm{d}\theta} \;\hat{=}\; \sum_{i=1}^n Y_i, \tag{3.37}$$

so that

$$E\{S(\theta;\mathbf{x})\}=nE(Y_1) \quad \text{and} \quad \mathrm{Var}\{S(\theta;\mathbf{x})\}=n\mathrm{Var}(Y_1), \tag{3.38}$$

where $\{Y_i\}_{i=1}^{\infty}$ be i.i.d. random variables with the common mean

$$\mu = E(Y_1) \stackrel{(3.38)}{=} \frac{E\{S(\theta;\mathbf{x})\}}{n} = 0 \tag{3.39}$$

and the common variance

$$\sigma^2 = \mathrm{Var}(Y_1) \stackrel{(3.38)}{=} \frac{\mathrm{Var}\{S(\theta;\mathbf{x})\}}{n} = \frac{I_n(\theta)}{n} = I(\theta). \tag{3.40}$$

— Thus

$$\begin{aligned}
\frac{S(\theta;\mathbf{x}) - E\{S(\theta;\mathbf{x})\}}{\sqrt{\mathrm{Var}\{S(\theta;\mathbf{x})\}}} &\stackrel{(3.39)}{=} \frac{S(\theta;\mathbf{x})}{\sqrt{nI(\theta)}} \\
&\stackrel{(3.37)}{=} \frac{\sum_{i=1}^n Y_i}{\sqrt{nI(\theta)}} \stackrel{(3.40)}{=} \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i}{\sigma} \\
&\stackrel{(3.36)}{=} \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i - \sqrt{n}\times 0}{\sigma} \\
&\stackrel{\mathrm{L}}{\to} N(0,1) \quad \text{as} \quad n \to \infty,
\end{aligned}$$

which completes the proof of (3.33). □

31.3• Proof of (3.34)

— By applying the first-order Taylor expansion to the score function $S(\theta;\mathbf{x})$ around the MLE $\hat{\theta}_n$ and noting $S(\hat{\theta}_n;\mathbf{x}) = 0$, we have

$$S(\theta;\mathbf{x}) = S(\hat{\theta}_n;\mathbf{x}) + (\theta - \hat{\theta}_n)\left.\frac{\mathrm{d}S(\theta;\mathbf{x})}{\mathrm{d}\theta}\right|_{\theta=\theta^*} \hat{=}\ 0 + (\theta - \hat{\theta}_n)H(\theta^*;\mathbf{x}),$$

where θ^* is a point between θ and $\hat{\theta}_n$. Thus

$$\frac{S(\theta;\mathbf{x})}{\sqrt{nI(\theta)}} = \sqrt{nI(\theta)}\,(\hat{\theta}_n - \theta) \times \frac{-H(\theta^*;\mathbf{x})/n}{I(\theta)}.$$

— We only need to prove that

$$-\frac{H(\theta^*;\mathbf{x})}{n} \stackrel{\mathrm{P}}{\to} I(\theta) \quad \text{as} \quad n \to \infty. \tag{3.41}$$

— According to the weak law of large number (see, Theorem 2.7), we have

$$\begin{aligned}
-\frac{H(\theta;\mathbf{x})}{n} &= -\frac{1}{n}\cdot\frac{\mathrm{d}S(\theta;\mathbf{x})}{\mathrm{d}\theta} \overset{(3.37)}{=} \frac{1}{n}\sum_{i=1}^{n} -\frac{\mathrm{d}^2\log f(X_i;\theta)}{\mathrm{d}\theta^2} \\
&\hat{=} \frac{1}{n}\sum_{i=1}^{n} Z_i \overset{\mathrm{P}}{\to} E(Z_1) \quad \text{as} \quad n\to\infty. \qquad (3.42)
\end{aligned}$$

From (3.24), we obtain

$$E(Z_1) = E\left\{-\frac{\mathrm{d}^2\log f(X_1;\theta)}{\mathrm{d}\theta^2}\right\} = I(\theta).$$

From (3.32), since $\hat{\theta}_n \overset{\mathrm{P}}{\to} \theta$ as $n\to\infty$, we have

$$\begin{aligned}
-\frac{H(\theta^*;\mathbf{x})}{n} &= -\frac{H(\theta;\mathbf{x})}{n}\times\frac{H(\theta^*;\mathbf{x})}{H(\theta;\mathbf{x})} \qquad \text{[by using (3.42)]} \\
&\overset{\mathrm{P}}{\to} I(\theta)\times 1 = I(\theta) \quad \text{as} \quad n\to\infty,
\end{aligned}$$

implying (3.41). □

Exercise 3

3.1 Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} U[\theta_1,\theta_2]$. Find the MLEs of θ_1 and θ_2.

3.2 A sample of size n_1 is drawn from $N(\mu_1,\sigma_1^2)$. A second sample of size n_2 is drawn from $N(\mu_2,\sigma_2^2)$. Assume that the two samples are independent.

(a) What is the MLE of $\theta = \mu_1 - \mu_2$?

(b) If we assume that the total sample size $n = n_1 + n_2$ is fixed, how should the n observations be approximately divided between the two populations in order to minimize the variance of the $\hat{\theta}$?

3.3 The joint pmf of N_1, N_2, N_3 and N_4 is assumed to be

$$p(n_1,\ldots,n_4;\boldsymbol{\theta}) = \binom{n}{n_1,\ldots,n_4}\prod_{i=1}^{4}\theta_i^{n_i},$$

where $n_i \geqslant 0$, $\sum_{i=1}^4 n_i = n$, and $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_4)^\top \in \mathbb{T}_4$. Let $\theta_1 = \alpha\beta$, $\theta_2 = \alpha(1-\beta)$, $\theta_3 = (1-\alpha)\beta$, and $\theta_4 = (1-\alpha)(1-\beta)$, where $0<\alpha<1$ and $0<\beta<1$. Find the MLEs of α and β.

3.4 Let $X_{i1}, \ldots, X_{in} \stackrel{\text{iid}}{\sim} N(\mu_i, \sigma^2)$ for $i = 1, \ldots, 4$, where $\mu_1 = a + b + c$, $\mu_2 = a + b - c$, $\mu_3 = a - b + c$, and $\mu_4 = a - b - c$. The four samples are independent. What are the MLEs of a, b, c and σ^2?

3.5 Let $X_1, \ldots, X_n \sim U[\mu - \sqrt{3}\,\sigma,\ \mu + \sqrt{3}\,\sigma]$, where $\mu \in \mathbb{R}$ and $\sigma > 0$.

(a) Find the MLEs of μ and σ.

(b) Find the moment estimators of μ and σ.

3.6 Let $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} f(x; \theta)$ with $f(x; \theta) = \mathrm{e}^{-(x-\theta)}$ for $x \geqslant \theta$ and $\theta \in \mathbb{R}$.

(a) Find the MLE of θ.

(b) Find the moment estimator of θ.

(c) Using the prior density $\pi(\theta) = \mathrm{e}^{-\theta} I_{(0,\infty)}(\theta)$, find the Bayesian estimator of θ.

3.7 Let $X \sim \text{Bernoulli}(\theta)$. Let $t_1(X) = X$ and $t_2(X) = 1/2$.

(a) Are both $t_1(X)$ and $t_2(X)$ unbiased?

(b) Compare the MSE of $t_1(X)$ with that of $t_2(X)$.

3.8 Let $\{Y = 1\}$ denote the class of people who possess a sensitive characteristic (e.g., drug-taking, shoplifting, driving under influence and so on) and $\{Y = 0\}$ denote the complementary class. Let W be a non-sensitive dichotomous variate and be independent of Y. The interviewer should select a suitable W so that the proportion $p = \Pr(W = 1)$ can be estimated easily. Without loss of generality, p is assumed to be known. For example, we may define $W = 1$ if the respondent was born between August and December and $W = 0$ otherwise. Hence, it is reasonable to assume that $p \approx 5/12 = 0.41667$. Our aim is to estimate the proportion $\pi = \Pr(Y = 1)$.

To collect sensitive information, the interviewer may adopt the format at the left-hand side of Table 3.2. The interviewee is then asked to put a tick in either the open circle or in the triangle formed by the three solid dots in Table 3.2 according to his/her truthful answer. In this case, $\{Y = 0,\ W = 0\}$ means that the interviewee was neither a drug user nor born between August and December. That is, $\{Y = 0,\ W = 0\}$ represents a non-sensitive subclass. On the other hand, a tick in the triangle may possibly indicates the interviewee was born between August and December (i.e., $\{W = 1\}$). Therefore, respondents who are drug users are well covered their true identities by those who are between–August–December born non-drug users, and are willing to circle the triangle formed by the three dots. Such a design encourages the respondents to not only participate in the survey but also provide their truthful responses.

Table 3.2 The triangular model and its cell probabilities

Category	$W = 0$	$W = 1$
$Y = 0$	○	●
$Y = 1$	●	●

	$W = 0$	$W = 1$	Total
$Y = 0$	$(1 - \pi)(1 - p)$	$(1 - \pi)p$	$1 - \pi$
$Y = 1$	$\pi(1 - p)$	πp	π
Total	$1 - p$	p	1

Note: Please truthfully put a tick in the circle (i.e., ○) or circle the triangle formed by the three dots (i.e., ●).

Let $Y_{\text{obs}} = \{y_i\colon i = 1, \ldots, n\}$ denote the observed data for n respondents with $y_i = 1$ if the i-th respondent puts a tick in the triangle; $y_i = 0$ otherwise.

(a) Find the MLE $\hat{\pi}$ of π.

(b) Find the expectation of $\hat{\pi}$.

3.9 A discrete random variable Y is said to follow a *zero-truncated binomial* (ZTB) distribution if its pmf is

$$\Pr(Y = y) = \binom{m}{y} \pi^y (1 - \pi)^{m-y} / [1 - (1 - \pi)^m], \quad 1 \leqslant y \leqslant m,$$

where $\pi \in (0, 1)$ is an unknown parameter, and m is a known positive integer. We will write $Y \sim \text{ZTB}(m, \pi)$.

Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{ZTB}(m, \pi)$. Find the MLE of π by using the Fisher scoring algorithm.

3.10 Let $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} N(\mu_0, \theta)$, where μ_0 is known and $\theta > 0$.

(a) Find the MLE $\hat{\theta}$ of θ?

(b) What is the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$?

3.11 Let $X_1, \ldots, X_n$ be a random sample from a distribution with density

$$f(x; \theta) = \frac{g(x)}{h(\theta)}, \quad a(\theta) \leqslant x \leqslant b(\theta),$$

where $g(x)$ is a function of x only and $h(\theta) = \int_{a(\theta)}^{b(\theta)} g(x)dx$ a function of θ only. Let $a^{-1}(\theta)$ and $b^{-1}(\theta)$ be the inverse functions of $a(\theta)$ and $b(\theta)$, respectively. Prove that

(a) If $a(\theta)$ and $b(\theta)$ are monotone-increasing and monotone-decreasing functions of θ, respectively, then the sufficient statistic for θ is $\hat{\theta} = \min\{a^{-1}(X_{(1)}),\ b^{-1}(X_{(n)})\}$, where $X_{(1)}$ and $X_{(n)}$ are the smallest and largest order statistics, respectively.

(b) If $a(\theta)$ and $b(\theta)$ are monotone-decreasing and monotone-increasing functions of θ, respectively, then the sufficient statistic for θ is $\hat{\theta} = \max\{a^{-1}(X_{(1)}),\ b^{-1}(X_{(n)})\}$.

(c) The $\hat{\theta}$ is also the MLE of θ.

3.12 Let $Y = 1$ if a respondent is a drug user and $Y = 0$ otherwise. Let U denote the number of travel out of Hong Kong per year for the same respondent in a population in Hong Kong. Obviously, Y is a sensitive binary r.v. (thus it is not observable if the question is asked directly) and U is a non-sensitive random variable. Define $X = Y + U$. Let $Y \sim \text{Bernoulli}(\theta)$, $U \sim \text{Poisson}(\lambda)$, and $Y \perp\!\!\!\perp U$. The interviewer could ask the i-th respondent to report the sum $X_i = U_i + Y_i$ according to his/her truthful answer, $i = 1, \ldots, n$. Let the observed data be $X_1, \ldots, X_n$.

(a) Find the moment estimators of θ and λ.

(b) Find the MLEs of θ and λ.

3.13 Let $X_1, \ldots, X_n$ be a random sample from $f(x; \theta) = \mathrm{e}^{-(x-\theta)} I_{(\theta,\infty)}(x)$ for $-\infty < \theta < \infty$ and $Y_1 = \min(X_1, \ldots, X_n)$.

(a) Show that Y_1 is a complete sufficient statistic for θ.

(b) Find the function of Y_1 which is the unique UMVUE of θ.

3.14 Let a random sample of size n be taken from a discrete distribution with pmf $f(x;\theta) = 1/\theta$, $x = 1, 2, \ldots, \theta$, where θ is an unknown positive integer.

(a) Show that the largest observation $X_{(n)} \hat{=} Y$ is a complete sufficient statistic for θ.

(b) Prove that
$$\frac{Y^{n+1} - (Y-1)^{n+1}}{Y^n - (Y-1)^n}$$
is the unique UMVUE of θ.

3.15 Let $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. Define $\tau(\theta) = \text{Var}(X) = \theta(1-\theta)$.

(a) Find the Cramér–Rao lower bound for the unbiased estimator of $\tau(\theta)$.

(b) Find the unique UMVUE of $\tau(\theta)$ if such exists.

3.16 Let $X_i \sim \text{Poisson}(\lambda_i)$ for $i = 0, 1, 2$, and X_0, X_1, X_2 are independent. Define $Y_1 = X_0 + X_1$ and $Y_2 = X_0 + X_2$. Then $(Y_1, Y_2)^\top$ is said to follow the two-dimensional Poisson distribution with parameters $(\lambda_0, \lambda_1, \lambda_2)$, denoted by $(Y_1, Y_2)^\top \sim \text{MP}_2(\lambda_0, \lambda_1, \lambda_2)$.

(a) Find the joint probability mass function of $(Y_1, Y_2)^\top$.

(b) Let $\mathbf{y}_1, \ldots, \mathbf{y}_n \stackrel{\text{iid}}{\sim} \text{MP}_2(\lambda_0, \lambda_1, \lambda_2)$, where $\mathbf{y}_j = (Y_{1j}, Y_{2j})^\top$ and $\boldsymbol{y}_j = (y_{1j}, y_{2j})^\top$ denotes the realization of $\mathbf{y}_j$, $j = 1, \ldots, n$. Furthermore, let $\min(\boldsymbol{y}_j) = \min(y_{1j}, y_{2j}) = 0$ for all $j = 1, \ldots, n$. Find the MLEs of $(\lambda_0, \lambda_1, \lambda_2)$.