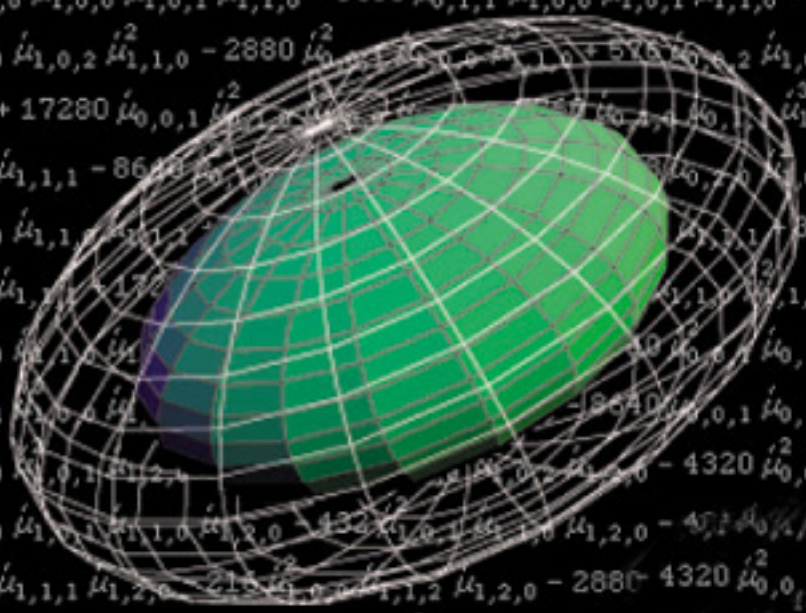
SPRINGER TEXTS IN STATISTICS

MATHEMATICAL STATISTICS

with *Mathematica*®



COLIN ROSE

MURRAY D. SMITH

Mathematical Statistics with *Mathematica*

Chapter 10 – Unbiased Parameter Estimation

Please reference this 2002 edition as:

Rose, C. and Smith, M.D. (2002)
Mathematical Statistics with Mathematica, Springer-Verlag, New York.

Latest edition

For the latest up-to-date edition, please visit: www.mathStatica.com

Chapter 10

Unbiased Parameter Estimation

10.1 Introduction

10.1 A Overview

For any given statistical model, there are any number of estimators that can be constructed in order to estimate unknown population parameters. In the previous chapter, we attempted to distinguish between estimators by specifying a loss structure, from which we hoped to identify the least risk estimator. Unfortunately, this process rarely presents a suitable overall winner. However, two important factors emerged from that discussion (especially for risk computed under quadratic loss), namely, the extent of bias, and the extent of variance inflation. Accounting for these factors yields a search for a preferred estimator from amongst classes of estimators, where the class members are forced to have a specific statistical property. This is precisely the approach taken in this chapter. Attention is restricted to the *class of unbiased estimators*, from which we wish to select the estimator that has least variance. We have already encountered the same type of idea in Chapter 7, where concern lay with unbiased estimation of population moments. In this chapter, on the other hand, we focus on unbiased estimation of the parameters of statistical models.

The chapter begins by measuring the statistical information that is present on a parameter in a given statistical model. This is done using Fisher Information and Sample Information (§10.2). This then leads to the so-called Cramer–Rao Lower Bound (a lower bound on the variance of any unbiased estimator), and to Best Unbiased Estimators, which are the rare breed of estimator whose variance achieves the lower bound (§10.3). The remaining two sections (§10.4 and §10.5) provide for the theoretical development of Minimum Variance Unbiased Estimators (MVUE). Vital to this is the notion of a sufficient statistic, its completeness, and its relation to the MVUE via a famous theorem due to Rao and Blackwell.

The statistical literature on MVUE estimation is extensive. The reference list that follows offers a sample of a range of treatments. In rough order of decreasing technical difficulty are Lehmann (1983), Silvey (1975), Cox and Hinkley (1974), Stuart and Ord (1991), Gourieroux and Monfort (1995), Mittelhammer (1996) and Hogg and Craig (1995).

10.1 B SuperD

In this chapter, it is necessary to activate the **mathStatica** function SuperD. This tool enhances *Mathematica*'s differentiator D (or, equivalently, ∂), allowing differentiation with respect to powers of variables. To illustrate, consider the derivative of $\sigma^{3/2}$ with respect to σ^2:

```
D[σ^(3/2), σ^2]
```

```
– General::ivar : σ^2 is not a valid variable.
```

$$\partial_{\sigma^2}\, \sigma^{3/2}$$

Mathematica does not allow this operation because σ^2 is not a Symbol variable; in fact, it is stored as Power (*i.e.* Head[σ^2] = Power). However, by turning On the **mathStatica** function SuperD:

```
SuperD[On]
```

```
– SuperD is now On.
```

derivatives, such as the former, can now be performed:

```
D[σ^(3/2), σ^2]
```

$$\frac{3}{4\sqrt{\sigma}}$$

At any stage, this enhancement to D may be removed by entering SuperD[Off].

10.2 Fisher Information

10.2 A Fisher Information

Let a random variable X have density $f(x;\theta)$, where θ is an unknown parameter which, for the moment, we assume is a scalar. The amount of statistical information about θ that is contributed per observation on X is defined to be

$$i_\theta = E\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2\right] \tag{10.1}$$

and is termed *Fisher's Information* on θ, after R. A. Fisher who first formulated it.

⊕ ***Example 1:*** Fisher's Information on the Lindley Parameter

Let $X \sim$ Lindley(δ), the Lindley distribution with parameter $\delta \in \mathbb{R}_+$, with pdf $f(x;\delta)$. Then, from **mathStatica**'s *Continuous* palette, the pdf of X is:

```
f = δ^2/(δ + 1) (x + 1) e^(-δ x);
domain[f] = {x, 0, ∞} && {δ > 0};
```

Then i_δ, the Fisher Information on δ, is given by (10.1) as:

```
Expect[D[Log[f], δ]^2, f]
```

$$\frac{2}{\delta^2} - \frac{1}{(1+\delta)^2}$$

⊕ ***Example 2:*** An Imprecise Survey: Censoring a Poisson Variable

Over a 1-week period, assume that the number of over-the-counter banking transactions by individuals is described by a discrete random variable $X \sim \text{Poisson}(\lambda)$, where $\lambda \in \mathbb{R}_+$ is an unknown parameter. Suppose, when collecting data from individuals, a market research company adopts the following survey policy: four or fewer transactions are recorded correctly, whereas five or more are recorded simply as five. Study the loss of statistical information on λ that is incurred by this data recording method.

Solution: Let $f(x; \lambda)$ denote the pmf of X:

```
f = e^(-λ) λ^x/x!;
domain[f] = {x, 0, ∞} && {λ > 0} && {Discrete};
```

Now define a discrete random variable Y, related to X as follows:

$$Y = \begin{cases} X & \text{if } X \le 4 \\ 5 & \text{if } X \ge 5. \end{cases}$$

Notice that the survey method samples Y, not X. Random variable X is said to be *right-censored* at 5. The pmf of Y is given by

$$P(Y = y) = \begin{cases} P(X = y) & \text{if } y \le 4 \\ P(X \ge 5) & \text{if } y = 5. \end{cases}$$

Let $g(y; \lambda)$ denote the pmf of Y in List Form, as shown in Table 1.

$P(Y = y)$:	$f(0;\lambda)$	$f(1;\lambda)$	$f(2;\lambda)$	$f(3;\lambda)$	$f(4;\lambda)$	$P(X \ge 5)$
y:	0	1	2	3	4	5

Table 1: List Form pmf of Y

We enter this into *Mathematica* as follows:

```
        g = Append[Table[f, {x, 0, 4}], 1 - Prob[4, f]];
domain[g] = {y, {0, 1, 2, 3, 4, 5}} && {λ > 0} && {Discrete};
```

where $P(Y = 5) = P(X \geq 5) = 1 - P(X \leq 4)$ is used. If an observation on X is recorded correctly, the Fisher Information on λ per observation, denoted by $i_{\lambda,X}$, is equal to:

```
i_{λ,x} = Expect[D[Log[f], λ]^2, f]
```

$$\frac{1}{\lambda}$$

On the other hand, the Fisher Information on λ per observation collected in the actual survey, denoted by $i_{\lambda,Y}$, is:

```
i_{λ,y} = Expect[D[Log[g], λ]^2, g]
```

$$-\left(e^{-\lambda}\left(-144 - 288\,\lambda - 288\,\lambda^2 - 192\,\lambda^3 - 66\,\lambda^4 - 12\,\lambda^5 - \lambda^6 + 6\,e^{\lambda}\left(24 + 24\,\lambda + 12\,\lambda^2 + 4\,\lambda^3 - 4\,\lambda^4 + \lambda^5\right)\right)\right) \Big/ \left(6\,\lambda\left(24 - 24\,e^{\lambda} + 24\,\lambda + 12\,\lambda^2 + 4\,\lambda^3 + \lambda^4\right)\right)$$

Figure 1 plots relative information $i_{\lambda,Y} / i_{\lambda,X}$ against values of λ.

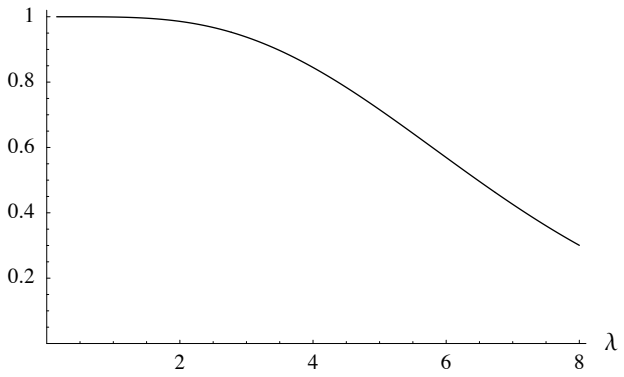


Fig. 1: Relative Fisher Information on λ

The figure shows that as λ increases, relative information declines. When, say, $\lambda = 5$, the relative information is:

```
i_{λ,y} / i_{λ,x} /. λ → 5 // N
```

```
0.715636
```

which means that about 28.5% of relative information on λ per observation has been lost by using this survey methodology. This would mean that to obtain the same amount of statistical information on λ as would be observed in a correctly recorded sample of say 100 individuals, the market research company would need to record data from about 140 (= 100 / 0.716) individuals. ■

10.2 B Alternate Form

Subject to some regularity conditions (*e.g.* Silvey (1975, p. 37) or Gourieroux and Monfort (1995, pp. 81–82)), an alternative expression for Fisher's Information to that given in (10.1) is

$$i_\theta \;=\; -E\Big[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\Big]. \tag{10.2}$$

For a proof of (10.2), see Silvey (1975, p. 40). When it is valid, this form of Fisher's Information can often be more convenient to compute, especially if the second derivative is not stochastic.

⊕ ***Example 3:*** First Derivative Form versus Second Derivative Form

Suppose the discrete random variable $X \sim \text{RiemannZeta}(\rho)$. Then, from **mathStatica**'s *Discrete* palette, the pmf $f(x;\rho)$ of X is given by:

```
f = x^-(ρ+1) / Zeta[1 + ρ];
domain[f] = {x, 1, ∞} && {ρ > 0} && {Discrete};
```

Following (10.1), $\left(\frac{\partial \log f(x;\rho)}{\partial \rho}\right)^2$ is given by:

```
d = D[Log[f], ρ]^2 // Simplify
```

$$\frac{(\text{Log}[x]\ \text{Zeta}[1+\rho] + \text{Zeta}'[1+\rho])^2}{\text{Zeta}[1+\rho]^2}$$

This is a stochastic expression for it depends on x, the values of X. Applying `Expect` yields the Fisher Information on ρ:

```
Expect[d, f]
```

$$\frac{-\text{Zeta}'[1+\rho]^2 + \text{Zeta}[1+\rho]\ \text{Zeta}''[1+\rho]}{\text{Zeta}[1+\rho]^2}$$

Alternately, following (10.2), we find:

```
-D[Log[f], {ρ, 2}] // Simplify
```

$$\frac{-\text{Zeta}'[1+\rho]^2 + \text{Zeta}[1+\rho]\ \text{Zeta}''[1+\rho]}{\text{Zeta}[1+\rho]^2}$$

This output is non-stochastic, and is clearly equivalent to the previous output. In this case, (10.2) yields Fisher's Information on ρ, without the need to even apply `Expect`. ■

⊕ ***Example 4:*** Regularity Conditions

Suppose $X \sim \text{Uniform}(\theta)$, where parameter $\theta \in \mathbb{R}_+$. The pdf of X is:

```
f = 1/θ;   domain[f] = {x, 0, θ} && {θ > 0};
```

According to the definition (10.1), the Fisher Information on θ is:

```
Expect[D[Log[f], θ]^2, f]
```

$$\frac{1}{\theta^2}$$

Next, consider the following output calculated according to (10.2):

```
-Expect[D[Log[f], {θ, 2}], f]
```

$$-\frac{1}{\theta^2}$$

Clearly, this expression cannot be correct, because Fisher Information cannot be negative. The reason why our second computation is incorrect is because a regularity condition is violated—the condition that permits interchangeability between the differential and integral operators. In general, it can be shown (see Silvey (1975, p. 40)) that (10.2) is equivalent to (10.1) if

$$\frac{\partial^2}{\partial\theta^2}\int_0^\theta f\,dx = \int_0^\theta \frac{\partial^2 f}{\partial\theta^2}\,dx \tag{10.3}$$

where $f = 1/\theta$ is the pdf of X. In this case, (10.3) is not true as the value of the pdf at $x = \theta$ is strictly positive. Indeed, as a general rule, the regularity conditions permitting computation of Fisher Information according to (10.2) are violated whenever the domain of support of a random variable depends on unknown parameters, when the density at those points is strictly positive. ■

10.2 C Automating Computation: `FisherInformation`

In light of (10.1) and (10.2), **mathStatica**'s `FisherInformation` function automates the computation of Fisher Information. In an obvious notation, the function's syntax is `FisherInformation[θ, f]`, with options `Method → 1` (default) for computation according to (10.1), or `Method → 2` for computation according to (10.2).

⊕ ***Example 5:*** `FisherInformation`

Suppose that $X \sim N(\mu, 1)$. Then, its pdf is given by:

```
f = 1/Sqrt[2 π] e^(-1/2 (x-μ)^2);  domain[f] = {x, -∞, ∞} && {μ ∈ Reals};
```

The Fisher Information on μ may be derived using the one-line command:

```
FisherInformation[μ, f]
```

```
1
```

It is well worth contrasting the computational efficiency of the two methods of calculation, (10.1) and (10.2):

```
FisherInformation[μ, f, Method → 1] // Timing
```

```
{0.72 Second, 1}
```

```
FisherInformation[μ, f, Method → 2] // Timing
```

```
{0.11 Second, 1}
```

Generally, the second method is more efficient; however, the second method is only valid under regularity conditions. In this example, the regularity conditions are satisfied. ■

10.2 D Multiple Parameters

The discussion so far has been concerned with statistical information on a single parameter. Of course, many statistical models have multiple parameters. Accordingly, we now broaden the definition of Fisher Information (10.1) to the case when θ is a $(k \times 1)$ vector of unknown parameters. Fisher's Information on θ is now a square, symmetric matrix of dimension $(k \times k)$. The $(i, j)^{\text{th}}$ element of the Fisher Information matrix i_θ is

$$E\left[\left(\frac{\partial \log f(X;\theta)}{\partial \theta_i}\right)\left(\frac{\partial \log f(X;\theta)}{\partial \theta_j}\right)\right] \tag{10.4}$$

for $i, j \in \{1, \ldots, k\}$. Notice that when $i = j$, (10.4) becomes (10.1), and is equivalent to the Fisher Information on θ_i. The multi-parameter analogue of (10.2) is given by

$$-E\left[\frac{\partial^2 \log f(X;\theta)}{\partial \theta_i \, \partial \theta_j}\right] \tag{10.5}$$

which corresponds to the $(i, j)^{\text{th}}$ element of i_θ, provided the regularity conditions hold. **mathStatica**'s `FisherInformation` function extends to the multi-parameter setting.

⊕ ***Example 6:*** Fisher Information Matrix for Gamma Parameters

Suppose that $X \sim \text{Gamma}(a, b)$, where $\theta = \begin{pmatrix} a \\ b \end{pmatrix}$ is a (2×1) vector of unknown parameters. Let $f(x; \theta)$ denote the pdf of X:

```
f = x^(a-1) e^(-x/b) / (Γ[a] b^a);    domain[f] = {x, 0, ∞} && {a > 0, b > 0};
```

The elements of Fisher's Information on θ, a (2×2) matrix, are:

```
FisherInformation[{a, b}, f]
```

$$\begin{pmatrix} \text{PolyGamma}[1, a] & \frac{1}{b} \\ \frac{1}{b} & \frac{a}{b^2} \end{pmatrix}$$

where the placement of the elements in the matrix is important; for example, the top-left element corresponds to Fisher's Information on a. ■

10.2 E Sample Information

As estimation of parameters is typically based on a sample of data drawn from a population, it is important to contemplate the amount of information that is contained by a sample about any parameters. Once again, Fisher's formulation may be used to measure statistical information. However, this time we focus upon the joint distribution of the random sample, as opposed to the distribution of the population from which the sample is drawn. We use the symbol I_θ to denote the statistical information contained by a sample, terming this *Sample Information*, as distinct from i_θ for Fisher Information.[1]

Let $\vec{X} = (X_1, \ldots, X_n)$ denote a random sample of size n drawn on a random variable X. Denote the joint density of $\vec{X}$ by $f(\vec{x}; \theta)$, where scalar θ is an unknown parameter. The Sample Information on θ is defined as

$$I_\theta = E\left[\left(\frac{\partial \log f(\vec{X};\theta)}{\partial\theta}\right)^2\right]. \tag{10.6}$$

If $\vec{X}$ is a collection of n independent and identically distributed (iid) random variables, each with density $f(x_i;\theta)$ $(i = 1, \ldots, n)$, equivalent in functional form, then the joint density of the collection $\vec{X}$ is given by $f(\vec{x};\theta) = \prod_{i=1}^n f(x_i;\theta)$. Furthermore, if the regularity condition $E\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)\right] = 0$ is satisfied, then

$$\begin{aligned} I_\theta &= E\left[\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i;\theta)\right)^2\right] \\ &= \sum_{i=1}^n E\left[\left(\frac{\partial \log f(X_i;\theta)}{\partial\theta}\right)^2\right] \\ &= n\, i_\theta\,. \end{aligned} \tag{10.7}$$

If it is valid to do so, it is well worth exploiting (10.7), as the derivation of I_θ through the multivariate expectation (10.6) can be difficult. For example, for n observations collected on $X \sim \text{Lindley}(\delta)$, the Sample Information is simply $n\, i_\delta$, where i_δ was derived in *Example 1*. On the other hand, for models that generate observations according to underlying regimes (*e.g.* the censoring model discussed in *Example 2* is of this type), the relationship between Fisher Information and Sample Information is generally more complicated than that described by (10.7), even if the random sample consists of a collection of iid random variables.

10.3 Best Unbiased Estimators

10.3 A The Cramér–Rao Lower Bound

Let θ denote the parameter of a statistical model, and let $g(\theta)$ be some differentiable function of θ that we are interested in estimating. The *Cramér–Rao Lower Bound* (CRLB) establishes a lower bound below which the variance of an *unbiased estimator* of $g(\theta)$ cannot go. Often the CRLB is written in the form of an inequality — the *Cramér–Rao Inequality*. Let $\hat{g}$ denote an unbiased estimator of $g(\theta)$ constructed from a random sample of n observations. Then, subject to some regularity conditions, the Cramér–Rao Inequality is given by

$$\mathrm{Var}(\hat{g}) \geq \left(\frac{\partial g(\theta)}{\partial \theta}\right)^2 \Big/ I_\theta \tag{10.8}$$

where I_θ denotes Sample Information (§10.2 E). If we are interested in estimating θ, then set $g(\theta) = \theta$, in which case (10.8) simplifies to

$$\mathrm{Var}(\hat{\theta}) \geq 1 / I_\theta \tag{10.9}$$

where $\hat{\theta}$ is an unbiased estimator of θ. When estimating $g(\theta)$, the CRLB is the quantity on the right-hand side of (10.8); similarly, when estimating θ, the CRLB is the right-hand side of (10.9). The inverse relationship between the CRLB and Sample Information is intuitive. After all, the more statistical information that a sample contains on θ, the better should an (unbiased) estimator of θ (or $g(\theta)$) perform. In our present context, 'better' refers to smaller variance.

If θ, or $g(\theta)$, represent vectors of parameters, say θ is $(k \times 1)$ and $g(\theta)$ is $(m \times 1)$ with $m \leq k$, then the CRLB expresses a lower bound on the variance-covariance matrix of unbiased estimators. In this instance, (10.8) becomes

$$\mathrm{Varcov}(\hat{g}) \geqslant G \times I_\theta^{-1} \times G^{\mathrm{T}} \tag{10.10}$$

where the $(m \times k)$ matrix of derivatives

$$G = \frac{\partial g(\theta)}{\partial \theta^{\mathrm{T}}} .$$

Equation (10.9) becomes

$$\mathrm{Varcov}(\hat{\theta}) \geqslant I_\theta^{-1} \tag{10.11}$$

where the notation $A \geqslant B$ indicates that $A - B$ is a positive semi-definite matrix, and I_θ^{-1} denotes the inverse of the Sample Information matrix. For proofs of the Cramér–Rao Inequality for both scalar and vector cases, plus discussion on the regularity conditions, see Silvey (1975), Mittelhammer (1996), or Gourieroux and Monfort (1995).

⊕ ***Example 7:*** The CRLB for the Poisson Parameter

Suppose that $X \sim \text{Poisson}(\lambda)$. Derive the CRLB for all unbiased estimators of λ.

Solution: Let $f(x; \lambda)$ denote the pmf of X:

```
f = e^-λ λ^x / x!;
domain[f] = {x, 0, ∞} && {λ > 0} && {Discrete};
```

The right-hand side of (10.9) gives the general formula for the CRLB for unbiased estimators. Thus, for random samples of size n drawn on X, the CRLB for the Poisson parameter λ is:

```
1
-------------------------------
n FisherInformation[λ, f]
```

$$\frac{\lambda}{n}$$

where we have exploited the relationship between Sample Information and Fisher Information given in (10.7). ■

⊕ ***Example 8:*** The CRLB for the Inverse Gaussian Mean and Variance

Let $X \sim \text{InverseGaussian}(\mu, \lambda)$, and let $\theta = \binom{\mu}{\lambda}$. Derive the CRLB for unbiased estimators of $g(\theta)$, where

$$g(\theta) = g(\mu, \lambda) = \begin{pmatrix} \mu \\ \mu^3/\lambda \end{pmatrix}.$$

Solution: Enter the pdf of X:

```
f = Sqrt[λ / (2 π x^3)] Exp[-λ (x - μ)^2 / (2 μ^2 x)];
domain[f] = {x, 0, ∞} && {μ > 0, λ > 0};
```

The CRLB for θ is equal to the (2×2) matrix:

```
CRLB = Inverse[n FisherInformation[{μ, λ}, f]]
```

$$\begin{pmatrix} \frac{\mu^3}{n\lambda} & 0 \\ 0 & \frac{2\lambda^2}{n} \end{pmatrix}$$

To find the CRLB for $g(\mu, \lambda) = (\mu, \mu^3/\lambda)^{\mathrm{T}}$, a (2×1) vector, the right-hand side of (10.10) must be evaluated. First, we derive the (2×2) matrix of derivatives $G = \partial g(\theta)/\partial \theta^{\mathrm{T}}$ using the **mathStatica** function `Grad`:

```
G = Grad[{μ, μ^3/λ}, {μ, λ}]
```

$$\begin{pmatrix} 1 & 0 \\ \frac{3\,\mu^2}{\lambda} & -\frac{\mu^3}{\lambda^2} \end{pmatrix}$$

Then, the CRLB is given by the (2×2) matrix:

```
G.CRLB.Transpose[G] // Simplify
```

$$\begin{pmatrix} \frac{\mu^3}{n\,\lambda} & \frac{3\,\mu^5}{n\,\lambda^2} \\ \frac{3\,\mu^5}{n\,\lambda^2} & \frac{\mu^6\,(2\,\lambda+9\,\mu)}{n\,\lambda^3} \end{pmatrix}$$

10.3 B Best Unbiased Estimators

Suppose that $\hat{g}$ is an unbiased estimator of $g(\theta)$ that satisfies all regularity conditions, and that $\mathrm{Var}(\hat{g})$ attains the CRLB. In this event, we can do no better (in terms of variance minimisation) by adopting another unbiased estimator of $g(\theta)$; consequently, $\hat{g}$ is preferred over all other unbiased estimators. Because $\mathrm{Var}(\hat{g})$ is equivalent to the CRLB, $\hat{g}$ is referred to as the *Best Unbiased Estimator* (BUE) of $g(\theta)$.

⊕ ***Example 9:*** The BUE of the Poisson Parameter

Suppose that $X \sim \mathrm{Poisson}(\lambda)$, with pmf:

```
f = ℯ^-λ λ^x / x!;   domain[f] = {x, 0, ∞} && {λ > 0} && {Discrete};
```

Let $(X_1, \ldots, X_n)$ denote a random sample of size n drawn on X. We have already seen that the CRLB for unbiased estimators of λ is given by λ/n (see *Example* 7). Consider then the estimator $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$, the sample mean. Whatever the value of index i, X_i is a copy of X, so the mean of $\hat{\lambda}$ is given by:

```
1/n Sum[Expect[x, f], {i, 1, n}]
```

λ

In addition, because X_i is independent of X_j for all $i \neq j$, the variance of $\hat{\lambda}$ is given by:

```
1/n^2 Sum[Var[x, f], {i, 1, n}]
```

$$\frac{\lambda}{n}$$

From these results, we see that $\hat{\lambda}$ is an unbiased estimator of λ, and its variance corresponds to the CRLB. Thus, $\hat{\lambda}$ is the BUE of λ. ■

⊕ ***Example 10:*** Estimation of the Extreme Value Scale Parameter

Let the continuous random variable X have the following pdf:

```
f = 1/σ Exp[-x/σ - e^(-x/σ)];
domain[f] = {x, -∞, ∞} && {σ > 0};
```

Thus, $X \sim$ ExtremeValue, with unknown scale parameter $\sigma \in \mathbb{R}_+$. The CRLB for unbiased estimators of σ is given by:

```
CRLB = 1/(n FisherInformation[σ, f])
```

$$\frac{6\,\sigma^2}{n\,(6\,(-1+\text{EulerGamma})^2+\pi^2)}$$

where n denotes the size of the random sample drawn on X. In numeric terms:

```
CRLB // N
```

$$\frac{0.548342\,\sigma^2}{n}$$

Now consider the expectation $E[|X|]$:

```
Expect[If[x < 0, -x, x], f]
```

```
σ (EulerGamma - 2 ExpIntegralEi[-1])
```

Let γ denote `EulerGamma`, and let Ei(−1) denote `ExpIntegralEi[-1]`. Knowing $E[|X|]$, it is easy to construct an unbiased estimator of the scale parameter σ, namely

$$\hat{\sigma} = \frac{1}{n\,(\gamma - 2\,\text{Ei}(-1))} \sum_{i=1}^{n} |X_i| = \frac{0.984268}{n} \sum_{i=1}^{n} |X_i|$$

where γ and Ei(−1) have been assigned their respective numeric value. Following the method of *Example 9*, the variance of $\hat{\sigma}$ is:

```
Σ_{i=1}^{n} Var[If[x < 0, -x, x], f] / (n (EulerGamma - 2 ExpIntegralEi[-1]))^2 // N
```

$$\frac{0.916362\,\sigma^2}{n}$$

Clearly, Var($\hat{\sigma}$) > CRLB, in which case $\hat{\sigma}$ is *not* the BUE of σ. ■

10.4 Sufficient Statistics

10.4 A Introduction

Unfortunately, there are many statistical models for which the BUE of a given parameter does not exist.[2] In this case, even if it is straightforward to construct unbiased estimators, how can we be sure that the particular estimator we select has least variance? After all, unless we inspect the variance of every unbiased estimator — keep in mind that this class of estimator may well have an infinite number of members — the least variance unbiased estimator may simply not happen to be amongst those we examined. Nevertheless, if our proposed estimator has *used all available statistical information on the parameter of interest*, then intuition suggests that our selection may have least variance. A statistic that retains all information about a parameter is said to be *sufficient* for that parameter.

Let X denote the population of interest, dependent on some unknown parameter θ (which may be a vector). Then, the 'information' referred to above is that which is derived from a size n random sample drawn on X, the latter denoted by $\vec{X} = (X_1, \ldots, X_n)$. A sufficient statistic S is a function of the random sample; that is, $S = S(\vec{X})$. Obviously $S(\vec{X})$ is a random variable, but for a particular set of observed data, $\vec{x} = (x_1, \ldots, x_n)$, $S(\vec{x})$ must be numeric.

A statistic S, whose values we shall denote by s, is sufficient for a parameter θ if the conditional distribution of $\vec{X}$ given $S = s$ does not depend on θ. Immediately, then, the identity statistic $S = \vec{X}$ must be sufficient; however, it is of no use as it has dimension n. This is because the key idea behind sufficiency is to reduce the dimensionality of $\vec{X}$, without losing information. Finally, if another statistic $T = T(\vec{X})$ is such that it *loses* all information about a parameter, then it is termed *ancillary* for that parameter. It is also possible that a statistic $U = U(\vec{X})$ can be neither sufficient nor ancillary for a parameter.

⊕ ***Example 11:*** Sufficiency in Bernoulli Trials

Let $X \sim \text{Bernoulli}(p)$, where $p = P(X = 1)$ denotes the success probability. Given a random sample $\vec{X}$, we would expect the number of successes $S = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$ to be influential when estimating the success probability p. In fact, for values $x_i \in \{0, 1\}$, and value $s \in \{0, 1, \ldots, n\}$ such that $s = \sum_{i=1}^{n} x_i$, the conditional distribution of $\vec{X}$ given $S = s$ is

$$P\left(\vec{X} \,\middle|\, S = s\right) = \frac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(S = s)} = \frac{p^n (1-p)^{n-s}}{\binom{n}{s} p^n (1-p)^{n-s}} = \frac{1}{\binom{n}{s}}.$$

As the conditional distribution does not depend on p, the one-dimensional statistic $S = \sum_{i=1}^{n} X_i$ is sufficient for p. On the other hand, the statistic T, defined here as the chronological order in which observations occur, contributes nothing to our knowledge of the success probability: T is ancillary for p. A third statistic, the sample median M, is neither sufficient for p, nor is it ancillary for p.

It is interesting to examine the loss in Sample Information incurred as a result of using M to estimate p. For simplicity, set $n = 4$. Then, the sample sum $S \sim$ Binomial(4, p), with pmf $f(s; p)$:

```
f = Binomial[4, s] p^s (1 - p)^(4-s);

domain[f] = {s, 0, 4} && {0 < p < 1} && {Discrete};
```

From *Example 3* of Chapter 9, when $n = 4$, the sample median M has pmf $g(m, p)$, as given in Table 2.

$P(M = m)$:	$P(S \le 1)$	$P(S = 2)$	$P(S \ge 3)$
m:	0	$\frac{1}{2}$	1

Table 2: The pmf of M when $n = 4$

We enter the pmf of M in List Form:

```
g = {Prob[1, f],   f /. s → 2,  1 - Prob[2, f]}
```

$$\{-(-1+p)^3\,(1+3p),\ 6\,(1-p)^2\,p^2,\ 4\,p^3 - 3\,p^4\}$$

with domain of support:

```
domain[g] = {m, {0, 1/2, 1}} && {Discrete};
```

To compute the Sample Information on p, we use the fact that it is equivalent to the Fisher Information on p per observation on the sufficient statistic S:

```
FisherInformation[p, f]
```

$$\frac{4}{p - p^2}$$

Similarly, the amount of Sample Information on p that is captured by statistic M is equivalent to the Fisher Information on p per observation on M:

```
FisherInformation[p, g]
```

$$-\frac{24\,(4 - p + p^2)}{(-4 + 3p)\,(1 + 3p)}$$

Figure 2 plots the amount of Sample Information captured by each statistic against values of p. Evidently, the farther the true value of p lies from $\frac{1}{2}$, the greater is the loss of information about p incurred by the sample median M.

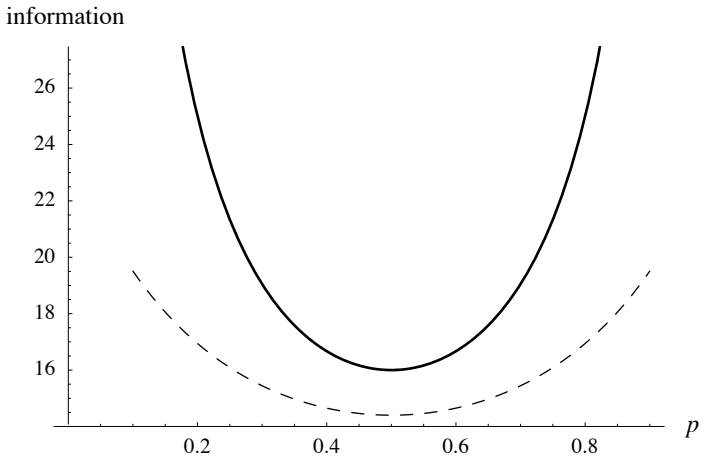


Fig. 2: Information on p due to statistics S (——) and M (– – –) when $n = 4$

10.4 B The Factorisation Criterion

The *Factorisation Criterion* provides a way to identify sufficient statistics. Once again, let X denote the population of interest, dependent on some unknown parameter θ, and let $\vec{X}$ denote a size n random sample drawn on X with joint density $f_*(\vec{x}; \theta)$. A necessary and sufficient condition for a statistic $S = S(\vec{X})$ to be sufficient for θ is that the density of $\vec{X}$ can be factored into the product,

$$f_*(\vec{x}; \theta) \;=\; g_*(s; \theta)\; h_*(\vec{x}) \tag{10.12}$$

where $g_*(s; \theta)$ denotes the density of S, and $h_*(\vec{x})$ is a non-negative function that does not involve θ; for discussion of the proof of this result, see Stuart and Ord (1991, Chapter 17). The factorisation (10.12) requires knowledge of the density of S which can, on occasion, add unnecessary difficulties. Fortunately, (10.12) can be weakened to

$$f_*(\vec{x}; \theta) \;=\; g(s; \theta)\; h(\vec{x}) \tag{10.13}$$

where $g(s; \theta)$ is a non-negative function (not necessarily a density function), and $h(\vec{x})$ is a non-negative function that does not involve θ. From now on, we shall adopt (10.13) to identify sufficient statistics.[3]

The **mathStatica** function `Sufficient[f]` constructs the joint density $f_*(\vec{x}; \theta)$ of a size n random sample $\vec{X} = (X_1, \ldots, X_n)$ drawn on a random variable X, and then simplifies it. The output from `Sufficient` can be useful when attempting to identify sufficient statistics for a parameter.

Finally, sufficient statistics are not unique; indeed, if a statistic S is sufficient for a parameter θ, then so too is a one-to-one function of S. To illustrate, suppose that statistic $S = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient for a parameter θ. Then, $T = \left(\overline{X}, \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2\right)$ is also sufficient for θ, as T and S are related by a one-to-one transformation.

⊕ ***Example 12:*** A Sufficient Statistic for the Poisson Parameter

Let $X \sim$ Poisson(λ) with pmf $f(x; \lambda)$:

```
f = e^-λ λ^x / x!;   domain[f] = {x, 0, ∞} && {λ > 0} && {Discrete};
```

The joint density of $\vec{X}$, a random sample of size n drawn on X, is given by $f_*(\vec{x}; \lambda) = \prod_{i=1}^{n} f(x_i; \lambda)$. This is derived by `Sufficient` as follows:

```
Sufficient[f]
```

$$e^{-n\lambda}\ \lambda^{\sum_{i=1}^{n} x_i}\ \prod_{i=1}^{n} \frac{1}{x_i!}$$

If we define $S = \sum_{i=1}^{n} X_i$, and let $g(s; \lambda) = e^{-n\lambda}\,\lambda^s$ and $h(\vec{x}) = \prod_{i=1}^{n} \frac{1}{x_i!}$, then, in view of (10.13), it follows that S is sufficient for λ. ■

⊕ ***Example 13:*** Sufficient Statistics for the Normal Parameters

Let $X \sim N(\mu, \sigma^2)$ with pdf $f(x; \mu, \sigma^2)$:

```
f = 1/(σ √(2 π)) Exp[-(x - μ)^2/(2 σ^2)];
domain[f] = {x, -∞, ∞} && {μ ∈ Reals, σ > 0};
```

Let $\vec{X}$ denote a random sample of size n drawn on X. Identify sufficient statistics when: (i) μ is unknown and σ^2 is known, (ii) μ is known and σ^2 unknown, (iii) both μ and σ^2 are unknown, and (iv) $\mu = \sigma = \theta$ is unknown.

Solution: In each case we must inspect the joint density of $\vec{X}$ produced by:

```
Sufficient[f]
```

$$e^{-\frac{n\mu^2 - 2\mu\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2}{2\sigma^2}}\ (2\pi)^{-n/2}\ \sigma^{-n}$$

(i) Define $S_1 = \sum_{i=1}^{n} X_i$. Because the value of σ^2 is known, let

$$g(s_1; \mu) = \exp\left(-\frac{n\mu^2 - 2\mu s_1}{2\sigma^2}\right)$$

$$h(\vec{x}) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right)(2\pi)^{-n/2}\,\sigma^{-n}.$$

Then, by (10.13), it follows that S_1 is sufficient for μ.

(ii) Define $S_2 = n\,\mu^2 - 2\,\mu \sum_{i=1}^n X_i + \sum_{i=1}^n X_i^2 = \sum_{i=1}^n (X_i - \mu)^2$. As μ is known, let

$$g(s_2; \sigma^2) = \exp\Big(-\frac{s_2}{2\,\sigma^2}\Big)\,\sigma^{-n}$$

$$h(\vec{x}) = (2\,\pi)^{-n/2}.$$

Since $g(s_2; \sigma^2)\,h(\vec{x})$ is equivalent to the joint density of $\vec{X}$, it follows that S_2 is sufficient for σ^2.

(iii) Define $S_3 = (S_{31}, S_{32}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$. Setting

$$g(s_3; \mu, \sigma^2) = \exp\Big(-\frac{n\,\mu^2 - 2\,\mu\,s_{31} + s_{32}}{2\,\sigma^2}\Big)\,\sigma^{-n}$$

$$h(\vec{x}) = (2\,\pi)^{-n/2}$$

it follows that the two-dimensional statistic S_3 is sufficient for (μ, σ^2).

(iv) For $\mu = \sigma = \theta$, and S_3 as defined in part (iii), set

$$g(s_3; \theta) = \exp\Big(\frac{2\,\theta\,s_{31} - s_{32}}{2\,\theta^2}\Big)\,\theta^{-n}$$

$$h(\vec{x}) = e^{-n/2}\,(2\,\pi)^{-n/2}.$$

Then, the two-dimensional statistic S_3 is sufficient for the scalar parameter θ. This last example serves to illustrate a more general point: the number of sufficient statistics need not match the number of unknown parameters. ■

10.5 Minimum Variance Unbiased Estimation

10.5 A Introduction

So far, we have armed ourselves with a sufficient statistic that captures all the statistical information that exists about a parameter. The next question is then how to use that statistic to construct an unbiased estimator of the unknown parameter. Intuition suggests that such an estimator should distinguish itself by having least variance. In other words, the estimator should be a *minimum variance unbiased estimator* (MVUE). This section focuses on the search for the MVUE of a parameter. Important to this development are theorems due to Rao and Blackwell (§10.5 B) and Lehmann and Scheffé (§10.5 D), and the notion of a complete sufficient statistic (§10.5 C).

10.5 B The Rao–Blackwell Theorem

The following theorem, due to Rao and Blackwell, is critical in the search for a MVUE:

Theorem (Rao–Blackwell): Let $S = S(\vec{X})$ be a sufficient statistic for a parameter θ, and let another statistic $T = T(\vec{X})$ be an unbiased estimator of $g(\theta)$ with finite variance. Define the function $\hat{g}(s) = E[T \mid S = s]$. Then:

(i) $E[\hat{g}(S)] = g(\theta)$; that is, $\hat{g}(S)$ is an unbiased estimator of $g(\theta)$.

(ii) $\text{Var}(\hat{g}(S)) \le \text{Var}(T)$.

Proof: See, for example, Silvey (1975, pp. 28–29). For discussion, see Hogg and Craig (1995, p. 326).

⊕ ***Example 14:*** A Conditional Expectation

Let $X \sim N(\mu, 1)$, and let $\overline{X}$ denote the sample mean from a random sample of size $n = 2r + 1$ drawn on X (for integer $r \ge 1$). Derive $E[T \mid \overline{X} = \overline{x}]$, where $T = T(\vec{X})$ denotes the sample median.

Solution (partial): We know from *Example 13*(i) that $S = \sum_{i=1}^{n} X_i$ is sufficient for μ. Thus, $\overline{X}$ will also be sufficient for μ as it is a one-to-one function of S. It follows that $E[T \mid \overline{X} = \overline{x}]$ can only be some function of $\overline{x}$, say $\hat{g}(\overline{x})$; that is, $E[T \mid \overline{X} = \overline{x}] = \hat{g}(\overline{x})$. The next step is to try and narrow down the possibilities for $\hat{g}(\overline{x})$. This is where part (i) of the Rao–Blackwell Theorem is used, for after deriving $E[T] = g(\mu)$, we may then be able to deduce those functions $\hat{g}(\overline{x})$ satisfying $E[\hat{g}(\overline{X})] = g(\mu)$, as we know $\overline{X} \sim N(\mu, \frac{1}{n})$.

Our strategy requires that we determine $E[T]$. Enter f, the pdf of X:

$$\mathbf{f = \frac{1}{\sqrt{2\,\pi}}\ Exp\left[-\frac{(x-\mu)^2}{2}\right];}$$

domain[f] = {x, -∞, ∞} && {μ ∈ Reals};

In a sample of size $n = 2r + 1$, the sample median T corresponds to the $(r+1)^{\text{th}}$ order statistic. We can use `OrderStat` to determine the pdf of T:

g = OrderStat[r + 1, f, 2 r + 1]

$$\frac{2^{-\frac{1}{2}-2r}\ e^{-\frac{1}{2}(x-\mu)^2}\left(1-\text{Erf}\left[\frac{x-\mu}{\sqrt{2}}\right]^2\right)^r\ (1+2r)!}{\sqrt{\pi}\ r!^2}$$

domain[g] = {x, -∞, ∞} && {μ ∈ Reals};

Transforming $T \to Q$, such that $Q = T - \mu$, yields the pdf of Q:

```
h = Transform[q == x - μ, g]
```

$$\frac{2^{-\frac{1}{2}-2\,r}\, e^{-\frac{q^2}{2}} \left(1 - \text{Erf}\left[\frac{q}{\sqrt{2}}\right]^2\right)^r (1 + 2\,r)!}{\sqrt{\pi}\; r!^2}$$

```
domain[h] = {q, -∞, ∞};
```

From this, we find $E[Q]$:

```
Expect[q, h]
```

```
0
```

Thus, $E[Q] = E[T - \mu] = 0$; that is, $E[T] = g(\mu) = \mu$. Substituting into part (i) of Rao–Blackwell's Theorem finds $E[\hat{g}(\overline{X})] = \mu$.

Now it is also true that $E[\overline{X}] = \mu$, as $\overline{X} \sim N(\mu, \frac{1}{n})$. Therefore, one solution for $\hat{g}(\overline{x})$ is the identity function $\hat{g}(\overline{x}) = \overline{x}$; that is,

$$E[T \mid \overline{X} = \overline{x}] = \overline{x}.$$

However, we cannot at this stage eliminate the possibility of other solutions to the conditional expectation (at least not under the Rao–Blackwell Theorem). In fact, for our solution to be unique, the concept of a *complete sufficient statistic* is required. We turn to this next. ■

10.5 C Completeness and MVUE

Suppose that a statistic S is sufficient for a parameter θ. Let $h(S)$ denote any function of S such that $E[h(S)] = 0$; note that the expectation is taken with respect to distributions of S. If this expectation only holds in the degenerate case when $h(S) = 0$, for all θ, then *the family of distributions of S is complete*.[4] A slightly different nomenclature is to refer to S as a *complete sufficient statistic*. We will not concern ourselves with establishing the completeness of a sufficient statistic; in fact, with the exception of the sufficient statistic derived in *Example 13*(iv), every other sufficient statistic we have encountered has been complete.

Completeness is important because of the uniqueness it confers on expectations of a sufficient statistic. In particular, if S is a complete sufficient statistic such that $E[S] = g(\theta)$, then there can be no other function of S that is unbiased for $g(\theta)$. In other words, completeness ensures that S is the *unique unbiased estimator* of $g(\theta)$. We may now finish *Example 14*. Since the sufficient statistic $S = \sum_{i=1}^{n} X_i$ is complete, our tentative solution is, in fact, the only solution. Thus, $E[T \mid \overline{X} = \overline{x}] = \overline{x}$.

The presence of a complete sufficient statistic in the Rao–Blackwell Theorem yields a MVUE. To see this, let S be a complete sufficient statistic for θ. Now, for any other statistic T that is unbiased for $g(\theta)$, the Rao–Blackwell Theorem yields, without exception, the function $\hat{g}(S)$, which is *unbiased* for $g(\theta)$; that is, $E[\hat{g}(S)] = g(\theta)$. By completeness, $\hat{g}(S)$

is the *unique* unbiased estimator of $g(\theta)$ amongst all functions of S. Furthermore, by the Rao–Blackwell Theorem, $\hat{g}(S)$ has variance *no larger* than that of any other unbiased estimator of $g(\theta)$. In combination, these facts ensure that $\hat{g}(S)$ is the MVUE of $g(\theta)$.

⊕ ***Example 15:*** Estimation of Probabilities

Let random variable $X \sim$ Exponential(λ), with pdf $f(x; \lambda)$:

```
f = 1/λ e^(-x/λ) ;        domain[f] = {x, 0, ∞} && {λ > 0};
```

and let $\vec{X} = (X_1, \ldots, X_n)$ denote a random sample of size n drawn on X. In this example, we shall derive the MVUE of the survival function $g(\lambda) = P(X > k)$, namely:

```
g = 1 - Prob[k, f]
```

$$e^{-\frac{k}{\lambda}}$$

where k is a known positive constant. Estimation of probabilistic quantities, such as $g(\lambda)$, play a prominent role in many continuous time statistical models, especially duration models (*e.g.* see Lancaster (1992)).

The first thing we must do is to identify a complete sufficient statistic for λ. This is quite straightforward after we apply `Sufficient`:

```
Sufficient[f]
```

$$e^{-\frac{\sum_{i=1}^{n} x_i}{\lambda}}\ \lambda^{-n}$$

Here, $S = \sum_{i=1}^{n} X_i$ fills our requirements (we state completeness of S without proof). Next, consider statistics $T = T(\vec{X})$ that are unbiased for $g(\lambda)$. One such statistic is the Bernoulli random variable defined as[5]

$$T = \begin{cases} 0 & \text{if } X_n \le k \\ 1 & \text{if } X_n > k. \end{cases}$$

Then, let

$$\hat{g}(s) = E[T \,|\, S = s] = P(T = 1 \,|\, S = s) = P(X_n > k \,|\, S = s). \tag{10.14}$$

By the Rao–Blackwell Theorem, $\hat{g}(S)$ is the MVUE of $g(\lambda)$. The next step is therefore clear. We must find $P(X_n > k \,|\, S = s)$.

To derive the distribution of $X_n \,|\, (S = s)$, we first require the bivariate distribution of (S, X_n). Now this bivariate distribution is found from the joint density of the n random variables in the random sample $\vec{X}$. Superficially the problem appears complicated: we must transform $\vec{X}$ to $(S, X_2, \ldots, X_n)$, followed by $n-2$ integrations to remove the unwanted variables $(X_2, \ldots, X_{n-1})$. However, if we define $S_{(n)} = \sum_{i=1}^{n-1} X_i$ (the sum of the first $n-1$ components of $\vec{X}$), with density $f_{(n)}(s_{(n)}; \lambda)$, then, by independence, the joint

density of $(S_{(n)}, X_n)$ is equal to the product $f_{(n)}(s_{(n)}; \lambda)\, f(x_n; \lambda)$. The joint density of (S, X_n) is then found by a simple transformation, because $S = S_{(n)} + X_n$. Determining $f_{(n)}(s_{(n)}; \lambda)$ is the key; fortunately, §4.5 contains a number of useful results concerning the density of sums of random variables. For our particular case, from *Example 22* of Chapter 4, we know that $S_{(n)} \sim \text{Gamma}(n-1, \lambda)$. Thus, the joint density of $(S_{(n)}, X_n)$ is given by:

```
h1 = (sn^(a-1) e^(-sn/b) / Γ[a] b^a /. {a → n - 1, b → λ}) * (f /. x → xn);
domain[h1] =
  {{sn, 0, ∞}, {xn, 0, ∞}} && {λ > 0, n > 1, n ∈ Integers};
```

Transforming $(S_{(n)}, X_n)$ to (S, X_n), where $S = S_{(n)} + X_n$, gives the pdf of (S, X_n):

```
h2 = Transform[{s == sn + xn, y == xn}, h1] /. y → xn
```

$$\frac{e^{-\frac{s}{\lambda}}\, \lambda^{-n}\, (s - x_n)^{-2+n}}{\Gamma[-1+n]}$$

The domain of support for (S, X_n) is all points in $\mathbb{R}^2_+$ such that $0 < x_n < s < \infty$. Thus:

```
domain[h2] =
  {{s, xn, ∞}, {xn, 0, s}} && {λ > 0, n > 1, n ∈ Integers};
```

The conditional distribution $X_n \,\big|\, (S = s)$ is given by:

```
h3 = Conditional[xn, h2]
domain[h3] = {xn, 0, s} && {n > 1, n ∈ Integers};
```

– Here is the conditional pdf h2($x_n \,|\, s$):

$$\frac{s^{1-n}\, \Gamma[n]\, (s - x_n)^{-2+n}}{\Gamma[-1+n]}$$

We now have all the ingredients in place ready to evaluate $\hat{g}(s) = P(X_n > k \,|\, S = s)$ and so determine the functional form of the MVUE:

```
Simplify[1 - Prob[k, h3], s > 0]
```

$$\left(1 - \frac{k}{s}\right)^{-1+n}$$

We conclude that $\hat{g}(S)$, the MVUE of $g(\lambda) = e^{-k/\lambda}$, is given by

$$\hat{g} = \begin{cases} 0 & \text{if } \sum_{i=1}^{n} X_i \le k \\ \left(1 - \dfrac{k}{\sum_{i=1}^{n} X_i}\right)^{n-1} & \text{if } \sum_{i=1}^{n} X_i > k. \end{cases}$$

Notice that $\hat{g}$ is a function of the complete sufficient statistic $S = \sum_{i=1}^{n} X_i$. ■

10.5 D Conclusion

In the previous example, the fact that the sufficient statistic was complete enabled us to construct the MVUE of $g(\lambda)$ by direct use of the Rao–Blackwell Theorem. Now, if in a given problem there exists a complete sufficient statistic, the key feature to notice from the Rao–Blackwell Theorem is that the MVUE will be *a function of the complete sufficient statistic*. We can, therefore, confine ourselves to examining the expectation of functions of complete sufficient statistics in order to derive minimum variance unbiased estimators. The following theorem summarises:

Theorem (Lehmann–Scheffé): Let S be a complete sufficient statistic for a parameter θ. If there is a function of S that has expectation $g(\theta)$, then this function is the MVUE of $g(\theta)$.

Proof: See, for example, Silvey (1995, p. 33). Also, Hogg and Craig (1995, p. 332).

⊕ ***Example 16:*** MVUE of the Normal Parameters

Let $X \sim N(\mu, \sigma^2)$ and define (see *Example 13*(iii)),

$$S = \begin{pmatrix} \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i^2 \end{pmatrix}$$

which is a complete sufficient statistic for (μ, σ^2). Let

$$T = \begin{pmatrix} \overline{X} \\ \hat{\sigma}^2 \end{pmatrix}$$

where $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ denotes the sample mean, and $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ is the sample variance. T is related one-to-one with S, and therefore it too is complete and sufficient for (μ, σ^2). Now we know that

$$E[T] = \begin{pmatrix} E[\overline{X}] \\ E[\hat{\sigma}^2] \end{pmatrix} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}.$$

Therefore, by the Rao–Blackwell and Lehmann–Scheffé theorems, $\overline{X}$ is the MVUE of μ, and $\hat{\sigma}^2$ is the MVUE of σ^2. ■

MVUE estimation relies on the existence of a complete sufficient statistic (whose variance exists). Without such a statistic, the rather elegant theory encapsulated in the Rao–Blackwell and Lehmann–Scheffé theorems cannot be applied. If it so happens that MVUE estimation is ruled out, how then do we proceed to estimate unknown parameters? We can return to considerations based on asymptotically desirable properties (Chapter 8), or choice based on decision loss criteria (Chapter 9), or choice based on maximising the content of statistical information (§10.2). Fortunately, there is another estimation technique — maximum likelihood estimation — which combines together features of each of these methods; the last two chapters of this book address aspects of this topic.

10.6 Exercises

1. Let the random variable $X \sim \text{Rayleigh}(\sigma)$, where parameter $\sigma > 0$. Derive Fisher's Information on σ.

2. Let the random variable $X \sim \text{Laplace}(\mu, \sigma)$. Obtain the CRLB for (μ, σ^2).

3. Let the random variable $X \sim \text{Lindley}(\delta)$. The sample mean $\overline{X}$ is the BUE of

$$g(\delta) = \frac{2+\delta}{\delta+\delta^2}.$$

Using *Mathematica*'s `SolveAlways` function, show that

$$h(\delta) = \frac{(3\,\delta+2)\,(2\,\delta+1)}{2\,\delta(\delta+1)}$$

is a linear function of $g(\delta)$. Hence, obtain the BUE of $h(\delta)$.

4. Let the random variable $X \sim \text{Laplace}(0, \sigma)$, and $(X_1, \ldots, X_n)$ denote a random sample of size n collected on X. Show that $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n} |X_i|$ is the BUE of σ.

5. Referring to *Example 10*, show that the estimator $\tilde{\sigma} = \frac{1}{n\gamma}\sum_{i=1}^{n} X_i$ is unbiased for σ. Give reasons as to why $\hat{\sigma}$, given in *Example 10*, is preferred to $\tilde{\sigma}$ as an estimator of σ.

6. Let $X \sim \text{RiemannZeta}(\rho)$, and let $(X_1, \ldots, X_n)$ denote a random sample of size n drawn on X. Use the Factorisation Criterion to identify a sufficient statistic for ρ.

7. Let the pair (X, Y) be bivariate Normal with $E[X] = E[Y] = 0$, $\text{Var}(X) = \text{Var}(Y) = 1$ and correlation coefficient ρ. Use the Factorisation Criterion to identify a sufficient statistic for ρ.

8. Let $X \sim \text{Gamma}(a, b)$, and let $(X_1, \ldots, X_n)$ denote a random sample of size n drawn on X. Use the Factorisation Criterion to identify a sufficient statistic for (a, b).

9. Using the technique of *Example 15*, obtain the MVUE of $P(X = 0) = e^{-\lambda}$, where $X \sim \text{Poisson}(\lambda)$.