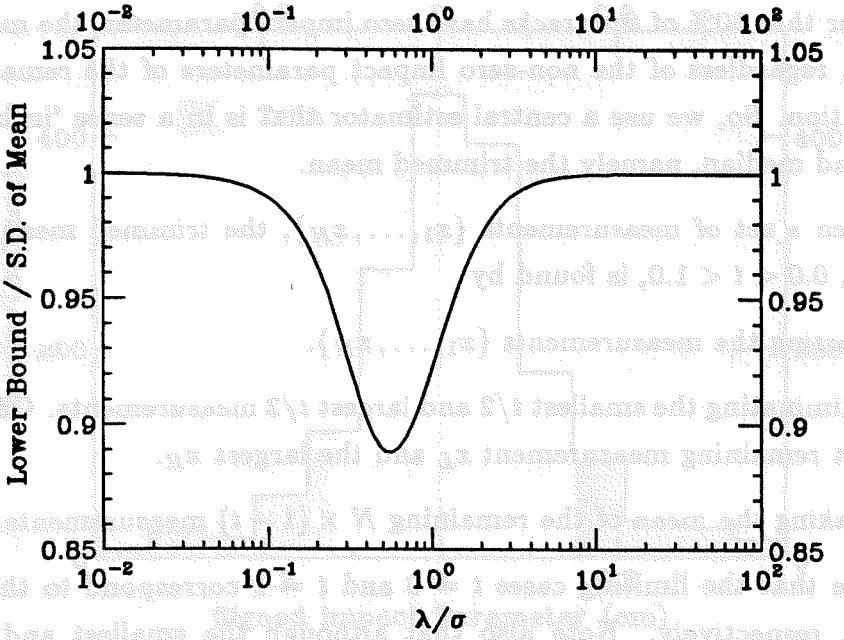
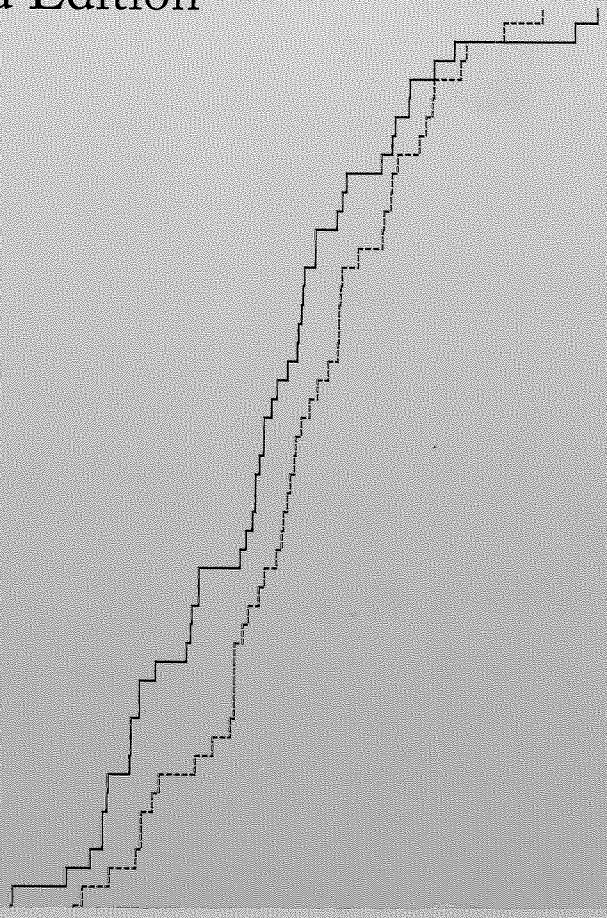


Figure 5.9 **Efficiency of Mean for the Convolution of a Gaussian with an Exponential. The lower bound is computed from $I_\lambda^{-1/2}$, where the information $I_\lambda = \int (\partial \ln f(x,\lambda,\sigma)/\partial\lambda)^2 f(x,\lambda,\sigma)dx$, where $f(x,\lambda,\sigma)$ is the convolution of the exponential characterized by λ and the gaussian by σ. A maximum likelihood fit asymptotically approaches the lower bound for large statistics, but only improves by 11% at best on the mean.**

Frederick James

Statistical Methods in Experimental Physics

2nd Edition



Thus one verifies that the linear least squares estimator (7.23) is unbiased to all orders of N [the terms grouped into $O(N^{-2})$ in Eq. (7.32) are proportional to higher-order derivatives, which all vanish].

(iii) The *maximum likelihood method*

Here g is given by Eq. (7.14), and we have already seen in Eq. (7.16) that $E[\xi(\theta_0)] = 0$. Using the relation (7.27) between the second derivatives of g and the information I (per event, on θ_0), and the relation (7.17), the bias and the variance take the form

$$b_N(\hat{\theta}) = E(\hat{\theta} - \theta_0) = \frac{K + 2J}{2NI^2}$$

$$V(\hat{\theta} - \theta_0) = \frac{1}{NI}\,, \tag{7.33}$$

where

$$K \equiv E\left(\frac{\partial^3 \ln f}{\partial\theta^3}\right)_{\theta=\theta_0}$$

$$J \equiv E\left(\frac{\partial^2 \ln f}{\partial\theta^2}\frac{\partial \ln f}{\partial\theta}\right)_{\theta=\theta_0}.$$

In this notation Eq. (7.31) becomes

$$\hat{\theta} - \theta_0 = \frac{X}{I} + \frac{XY}{NI^2} + \frac{1}{2}\frac{X^2K}{NI^3} + O\left(\frac{1}{N^2}\right).$$

When I, J, and K are estimated using the derivatives of the experimental likelihood function at its maximum, random terms of order $N^{-\frac{1}{2}}$ get introduced. Therefore, the expressions (7.33) for the bias and the variance are correct excluding terms of order $N^{-3/2}$.

We shall come back to the question of reduction of bias in Section 8.6. Note that all these expressions are not correct as they stand for the many-dimensional case. However, the same method may be used.

7.4. Information and the Precision of an Estimator

Recall that an estimate (the numerical result of applying an estimator) is a random variable, and it therefore has a probability distribution called the *sampling distribution* of the estimator. In Section 7.1 we defined consistency and unbiasedness in such a way that the sampling distribution of a *consistent estimator* is centred arbitrarily close to the true value θ_0, for a sufficiently large number of observations, and the sampling distribution of an unbiased estimator is always centred on θ_0.

In this section we shall describe the relation between the information of an estimator and the variance of its sampling distribution.

7.4.1. *Lower bounds for the variance — Cramér–Rao inequality*

Let $\mathbf{X}$ be observations from a distribution with density function $f(\mathbf{X}|\theta)$, and let the estimator $\hat{\theta}$ have the sampling distribution $q(\hat{\theta}|\theta)$. Denote the likelihood function of the observations $L_{\mathbf{X}} = L(\mathbf{X}|\theta)$, the likelihood of the estimator $L_{\hat{\theta}} = L(\mathbf{X}|\hat{\theta})$, and the corresponding informations $I_{\mathbf{X}}$ and $I_{\hat{\theta}}$.

From Eq. (7.1), the bias is a function of the true value θ_0

$$b = E(\hat{\theta}) - \theta_0 = \int \hat{\theta}(\mathbf{X}) f(\mathbf{X}|\theta_0)\, d\mathbf{X} - \theta_0\,.$$

Let us, for the time being, drop the subscript on θ. The variance of the sampling distribution,

$$V(\hat{\theta}) = \int [\hat{\theta} - E(\hat{\theta})]^2 q(\hat{\theta}|\theta)\, d\hat{\theta}\,, \tag{7.34}$$

is related to the information by the *Cramér–Rao inequality.*

$$V(\hat{\theta}) \geq \frac{[1 + (db/d\theta)]^2}{I_{\hat{\theta}}} \geq \frac{[1 + (db/d\theta)]^2}{I_{\mathbf{X}}}\,. \tag{7.35}$$

The first part of this important inequality states that the variance of an unbiased estimate is bounded below by the inverse of the information it carries.

The second part of the inequality follows directly from Eq. (5.10), changing t to $\hat{\theta}$, and it states that the variance of any unbiased estimate is bounded below by the inverse of the information contained in the observations. Replacing $I_{\mathbf{X}}$ by its definition (5.2), we have for any estimate

$$V(\hat{\theta}) \geq \frac{\left(\dfrac{d\tau(\theta)}{d\theta}\right)^2}{E\left[\left(\dfrac{\partial \ln L_{\mathbf{X}}}{\partial\theta}\right)^2\right]}\,, \tag{7.36}$$

where

$$\tau(\theta) \equiv E(\hat{\theta}) = \theta + b(\theta)\,. \tag{7.37}$$

The conditions under which the inequalities (7.35) and (7.36) hold are that

(i) the range of observables $\mathbf{X}$ should not depend on θ;
(ii) L_X must be sufficiently regular that differentiation with respect to θ and integration over $\mathbf{X}$ commute.

The proof of the inequality is given below.

Let us note first that, by definition,

$$\int L_{\hat{\theta}} d\hat{\theta} = \int q(\hat{\theta}|\theta) d\hat{\theta} = 1 .$$

It is, therefore, immaterial whether we call the sampling distribution $q(\hat{\theta}|\theta)$ or $L_{\hat{\theta}}$. Differentiating with respect to θ, gives

$$\int \frac{\partial q}{\partial \theta} d\hat{\theta} = \int \frac{\partial (\ln q)}{\partial \theta} q d\hat{\theta} = 0 . \tag{7.38}$$

Differentiating Eq. (7.37) with respect to θ gives

$$\frac{\partial}{\partial \theta} \int \hat{\theta} q(\hat{\theta}|\theta) d\hat{\theta} = \int \hat{\theta} \frac{\partial (\ln q)}{\partial \theta} q d\hat{\theta} = 1 + \frac{db}{d\theta} . \tag{7.39}$$

Noting that $\theta + b(\theta)$ is a constant, we can multiply it into Eq. (7.38), to obtain zero, and subtract it from Eq. (7.39):

$$\int [\hat{\theta} - \theta - b(\theta)] \frac{(\ln q)}{\partial \theta} q d\hat{\theta} = 1 + \frac{db}{d\theta} .$$

Application of Schwarz' inequality now yields

$$\int [\hat{\theta} - \theta - b(\theta)]^2 q(\hat{\theta}|\theta) d\hat{\theta} \int \left[\frac{\partial (\ln q)}{\partial \theta}\right]^2 q(\hat{\theta}|\theta) d\hat{\theta} \geq \left(1 + \frac{db}{d\theta}\right)^2 . \tag{7.40}$$

Since the first integral is an expression for $V(\hat{\theta})$, as defined by Eqs. (7.34) and (7.37), the inequality (7.40) becomes

$$V(\hat{\theta}) \geq \frac{\left(1 + \frac{db}{d\theta}\right)^2}{E\left[\left(\frac{\partial \ln q}{\partial \theta}\right)^2\right]} = \frac{\left(1 + \frac{db}{d\theta}\right)^2}{E\left[\left(\frac{\partial \ln L_{\hat{\theta}}}{\partial \theta}\right)^2\right]} = \frac{\left(1 + \frac{db}{d\theta}\right)^2}{I_{\hat{\theta}}} ,$$

which proves the Cramér–Rao inequality (7.35).

7.4.2. *Efficiency and minimum variance*

Consider first the conditions under which the inequality (7.40), and the first inequality in (7.35), become equalities. This occurs when

$$\frac{\partial \ln L_{\hat{\theta}}}{\partial \theta} = A(\theta)[\hat{\theta} - \theta - b(\theta)] . \tag{7.41}$$

It follows from Eq. (7.41) that *minimum variance* for the statistic $\hat{\theta}$,

$$V(\hat{\theta}) = (I_{\hat{\theta}})^{-1} \left(1 + \frac{db}{d\theta}\right)^2 , \tag{7.42}$$

is attained when the sampling distribution of $\hat{\theta}$ is of the exponential form (5.8)

$$L_{\hat{\theta}} = q(\hat{\theta}|\theta) = \exp[a(\theta)\hat{\theta} + \beta(\hat{\theta}) + c(\theta)] . \tag{7.43}$$

Here $a(\theta)$ is the integral of the arbitrary function $A(\theta)$ in Eq. (7.41), $\beta(\hat{\theta})$ is an integration constant and $c(\theta)$ is an arbitrary function.

When also the second inequality in Eq. (7.35) holds as equality,

$$I_{\hat{\theta}} = I_{\mathbf{X}} , \tag{7.44}$$

the variance $V(\hat{\theta})$ attains its *minimum variance bound*, and the statistic $\hat{\theta}$ is said to be an *efficient estimator*. It is important to note the distinction between

(i) *minimum variance* when the particular estimator $\hat{\theta}$ has the lowest variance of the family of estimators considered.
(ii) *minimum variance bound* for the set of data $\mathbf{X}$, attained when both Eqs. (7.42) and (7.44) hold. (The estimator $\hat{\theta}$ is *efficient*.)

An important case is when the sampling distribution $q(\hat{\theta}|\theta)$ is Normal. It is then completely equivalent to speak of the estimate $\hat{\theta}$ reaching the minimum variance bound, and maximum information.

Equation (7.44) is a necessary and sufficient condition for $\hat{\theta}$ to be a *sufficient statistic* for θ.

When a sufficient statistic t exists, the probability density $f(\mathbf{X}|\theta)$ is of the exponential form (5.8), by Darmois' theorem (Section 5.3.4). Inversely, when $f(\mathbf{X}, \theta)$ is of the exponential form (5.8), then

$$t = N^{-1} \sum_{i=1}^{N} \alpha(X_i) \tag{7.45}$$

is a sufficient statistic, from a sampling distribution of exponential form,

$$q(t|\theta) = \exp[a(\theta)t + \beta_1(t) + c(\theta)]\,,$$

where β_1 is deduced from β by Eq. (7.45). It can be shown that

$$E(t) = -(dc/d\theta)/(da/d\theta) \equiv r(\theta)\,.$$

The statistic (7.45) is therefore an unbiased estimator[e] of $r(\theta)$ and since $q(t|\theta)$ is of the exponential form, it follows that

$$V(t) = I_t^{-1}[r(\theta)] = I_{\mathbf{X}}^{-1}[r(\theta)]\,.$$

In conclusion, a minimum variance bound unbiased estimator t exists if, and only if the distribution $f(\mathbf{X},\theta)$ admits sufficient statistics. In this case, the function $r(\theta)$ estimated by t is unique, to within a linear transformation [Fourgeaud, p. 206].

Examples

To illustrate these results, consider the examples of Darmois theorem in Section 5.3.4.

In case (i), with μ unknown and σ^2 known,

$$r(\mu) = -\frac{\mu}{\sigma^2}\,,$$

and μ can be estimated efficiently.

In case (ii), with μ known and σ^2 unknown

$$r(\sigma) = \frac{\mu^2}{2} - \frac{\sigma^2}{2\sqrt{2\pi}}\,,$$

and only the function σ^2 can be estimated efficiently without bias.

Finally, in case (iii), when both μ and σ^2 are unknown, we have

$$r_1(\mu) = \mu$$

and

$$r_2(\sigma) = \frac{\mu^2}{2} - \frac{\sigma^2}{2\sqrt{2\pi}}\,.$$

Thus μ and σ^2 can be estimated efficiently without bias.

[e]It is also the maximum likelihood estimate.

This example shows why one usually estimates σ^2 rather than σ: σ^2 is the unique function of σ which can be estimated efficiently without bias.

On the other hand, it is possible to construct an unbiased estimate of σ, which then does not attain the minimum variance bound. In other words, it does not contain the maximum information about σ, it is less efficient. Such an estimator is

$$T = \sqrt{\frac{N}{2}}\,\frac{\Gamma\left(\frac{N}{2}\right)}{\Gamma\left(\frac{N+1}{2}\right)}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2}\,, \tag{7.46}$$

[Fourgeaud, p. 204]. Among all unbiased estimators, T is of minimum variance. If one insisted upon an estimator of even smaller variance, one would have to give up unbiasedness. This brings us to an important result, of greater generality than the above example.

If a sufficient statistic t exists, a necessary and sufficient condition, for an unbiased estimate T to have minimum variance, is that it is a function of t only

$$T = T(t)\,. \tag{7.47}$$

In the example of Eq. (7.46), T is a function of the sufficient statistic

$$t = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2}\,.$$

The statistic T is not, in general, efficient, that is, it does not necessarily attain the minimum variance bound, since there is only one function of the parameter which can be estimated in an unbiased and efficient way.

7.4.3. *Cramér–Rao inequality for several parameters*

The above results generalize in a straightforward way to the case of several parameters,

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)\,.$$

As shown in Section 5.2.1, the information becomes a matrix $\mathcal{I}_{\mathbf{X}}$, with elements given by Eq. (5.3). The *Cramér–Rao inequality* becomes, for unbiased estimators $\widehat{\boldsymbol{\theta}}$,

$$V(\hat{\theta}_i) \geq \left[\,\mathcal{I}_{\hat{\theta}}^{-1}\,(\boldsymbol{\theta})\right]_{ii} \quad \geq \left[\,\mathcal{I}_{\mathbf{X}}^{-1}\,(\boldsymbol{\theta})\right]_{ii}\,, \quad i = 1, \ldots, k\,, \tag{7.48}$$