Chapter 1

The Fisher Efficiency

1.1. Statistical Experiment

A classical *statistical experiment* $\big(X_1, \ldots, X_n;\, p(x,\theta);\, \theta \in \Theta\big)$ is composed of the following three elements: (i) a set of independent observations $X_1, \ldots, X_n$ where n is the sample size, (ii) a family of probability densities $p(x,\theta)$ defined by a parameter θ, and (iii) a *parameter set* Θ of all possible values of θ.

Unless otherwise stated, we always assume that θ is one-dimensional, that is, $\Theta \subseteq \mathbb{R}$. For discrete distributions, $p\,(x,\theta)$ is the probability mass function. In this chapter we formulate results only for continuous distributions. Analogous results hold for discrete distributions if integration is replaced by summation. Some discrete distributions are used in examples.

Example 1.1. (a) If n independent observations $X_1, \ldots, X_n$ have a normal distribution with an unknown mean θ and a known variance σ^2, that is, $X_i \sim \mathcal{N}(\theta, \sigma^2)$, then the density is

$$p\,(x,\theta) = (2\,\pi\,\sigma^2)^{-1/2}\,\exp\big\{-\,(x-\theta)^2/(2\sigma^2)\big\}\,,\; -\infty < x,\, \theta < \infty,$$

and the parameter set is the whole real line $\Theta = \mathbb{R}$.

(b) If n independent observations have a normal distribution with a known mean μ and an unknown variance θ, that is, $X_i \sim \mathcal{N}(\mu, \theta)$, then the density is

$$p(x,\theta) \;=\; (2\,\pi\,\theta)^{-1/2}\,\exp\big\{-\,(x-\mu)^2/(2\theta)\big\}\,,\; -\infty < x < \infty\,,\; \theta > 0,$$

and the parameter set is the positive half-axis $\Theta = \{\,\theta \in \mathbb{R} :\, \theta > 0\,\}$. $\square$

Example 1.2. Suppose n independent observations $X_1, \ldots, X_n$ come from a distribution with density

$$p(x, \theta) \,=\, p_0(x - \theta),\ -\infty \,<\, x,\, \theta \,<\, \infty,$$

where p_0 is a fixed probability density function. Here θ determines the shift of the distribution, and therefore is termed the *location parameter*. The location parameter model can be written as $X_i = \theta + \varepsilon_i$, $i = 1, \ldots, n$, where $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables with a given density p_0, and $\theta \in \Theta = \mathbb{R}$. □

The independence of observations implies that the joint density of X_i's equals

$$p\,(x_1, \ldots, x_n,\, \theta) \,=\, \prod_{i\,=\,1}^{n} p\,(x_i,\, \theta).$$

We denote the respective expectation by $\mathbb{E}_\theta[\,\cdot\,]$ and variance by $\mathbb{V}ar_\theta[\,\cdot\,]$.

In a statistical experiment, all observations are obtained under the same value of an unknown parameter θ. The goal of the parametric statistical estimation is to assess the true value of θ from the observations $X_1, \ldots, X_n$. An arbitrary function of observations, denoted by $\hat{\theta} = \hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$, is called an *estimator* (or a *point estimator*) of θ.

A random variable

$$l(X_i\,, \theta) \,=\, \ln p(X_i\,, \theta)$$

is referred to as a *log-likelihood function* related to the observation X_i. The *joint log-likelihood function of a sample of size n* (or, simply, the *log-likelihood function*) is the sum

$$L_n(\theta) \,=\, L_n(\theta \,|\, X_1\,, \ldots\,, X_n) \,=\, \sum_{i\,=\,1}^{n} l(X_i\,, \theta) \,=\, \sum_{i\,=\,1}^{n} \ln p(X_i\,, \theta).$$

In the above notation, we emphasize the dependence of the log-likelihood function on the parameter θ, keeping in mind that it is actually a random function that depends on the entire set of observations $X_1, \ldots, X_n$.

The parameter θ may be evaluated by the method of maximum likelihood estimation. An estimator θ_n^* is called the *maximum likelihood estimator* (MLE), if for any $\theta \in \Theta$ the following inequality holds:

$$L_n(\theta_n^*) \,\geq\, L_n(\theta).$$

If the log-likelihood function attains its unique maximum, then the MLE reduces to

$$\theta_n^* \,=\, \operatorname*{argmax}_{\theta \in \Theta} L_n(\theta).$$

If the function L is differentiable at its attainable maximum, then θ_n^* is a solution of the equation

$$\frac{\partial L_n(\theta)}{\partial \theta} \,=\, 0.$$

Note that if the maximum is not unique, this equation has multiple solutions.

The function

$$b_n(\theta) \,=\, b_n(\theta\,,\,\hat{\theta}_n) \,=\, \mathbb{E}_\theta\big[\,\hat{\theta}_n\,\big] \,-\, \theta \,=\, \mathbb{E}_\theta\big[\,\hat{\theta}_n(X_1,\ldots,X_n)\big] \,-\, \theta$$

is called the *bias* of $\hat{\theta}_n$. An estimator $\hat{\theta}_n(X_1,\ldots,X_n)$ is called an *unbiased* estimator of θ if its bias equals zero, or equivalently, $\mathbb{E}_\theta\big[\,\hat{\theta}_n\,\big] \,=\, \theta$ for all $\theta \in \Theta$.

Example 1.3. Assume that the underlying distribution of the random sample $X_1,\ldots,X_n$ is Poisson with mean θ. The probability mass function is given by

$$p_n(x,\theta) \,=\, \frac{\theta^x}{x!}\,e^{-\theta}\,,\;\; \theta \,>\, 0,\; x \in \{0,1,2,\ldots\}.$$

Then the log-likelihood function has the form

$$L_n(\theta) \,=\, \sum_{i\,=\,1}^{n} X_i\,\ln\theta \,-\, n\theta \,-\, \sum_{i\,=\,1}^{n} \ln\,(X_i!).$$

Setting the derivative equal to zero yields the solution $\theta_n^* = \bar{X}_n$, where

$$\bar{X}_n = (X_1 + \cdots + X_n)/n$$

denotes the *sample mean*. In this example, the MLE is unbiased since $\mathbb{E}_\theta\big[\,\theta_n^*\big] \,=\, \mathbb{E}_\theta\big[\,\bar{X}_n\big] \,=\, \mathbb{E}_\theta\big[\,X_1\big] = \theta$. □

Nonetheless, we should not take the unbiased MLE for granted. Even for common densities, its expected value may not exist. Consider the next example.

Example 1.4. For the exponential distribution with the density

$$p(x,\theta) \,=\, \theta\,\exp\big\{\,-\,\theta\,x\big\}\,,\;\; x \,>\, 0,\;\; \theta \,>\, 0,$$

the MLE $\theta_n^* \,=\, 1/\bar{X}_n$ has the expected value $\mathbb{E}_\theta\big[\,\theta_n^*\big] \,=\, n\,\theta/(n-1)$ (see Exercise 1.6). In particular, for $n = 1$, the expectation does not exist since $\int_0^\infty\, x^{-1}\,\theta\,\exp\big\{\,-\,\theta\,x\big\}\,dx = \infty$. □

In this example, however, an unbiased estimator may be found for $n > 1$. Indeed, the estimator $(n-1)\theta_n^*/n$ is unbiased. As the next example shows, an unbiased estimator may not exist at all.

Example 1.5. Let X be a Binomial$(n\,,\,\theta^{\,2})$ observation, that is, a random number of successes in n independent Bernoulli trials with the probability of a success $p = \theta^{\,2}$, $0 < \theta < 1$. An unbiased estimator of the parameter θ does not exist. In fact, if $\hat{\theta} = \hat{\theta}(X)$ were such an estimator, then its expectation would be an even polynomial of θ,

$$\mathbb{E}_\theta\big[\,\hat{\theta}(X)\,\big] \,=\, \sum_{k\,=\,0}^{n} \hat{\theta}(k)\,\binom{n}{k}\,\theta^{\,2k}\,(1-\theta^{\,2})^{n-k},$$

which cannot be identically equal to θ. □

1.2. The Fisher Information

Introduce the *Fisher score function* as the derivative of the log-likelihood function with respect to θ,

$$l'(X_i\,,\theta) \,=\, \frac{\partial \ln p(X_i\,,\theta)}{\partial\theta} \,=\, \frac{\partial p(X_i\,,\theta)/\partial\theta}{p(X_i\,,\theta)}.$$

Note that the expected value of the Fisher score function is zero. Indeed,

$$\mathbb{E}_\theta\big[\,l'(X_i\,,\theta)\,\big] \,=\, \int_{\mathbb{R}} \frac{\partial p(x\,,\theta)}{\partial\theta}\,dx \,=\, \frac{\partial \int_{\mathbb{R}}\, p(x\,,\theta)\,dx}{\partial\theta} \,=\, 0.$$

The *total Fisher score function* for a sample $X_1\,,\ldots\,,X_n$ is defined as the sum of the score functions for each individual observation,

$$L_n'(\theta) \,=\, \sum_{i\,=\,1}^{n} l'(X_i\,,\theta).$$

The *Fisher information* of one observation X_i is the variance of the Fisher score function $l'(X_i\,,\theta)$,

$$I(\theta) \,=\, \mathbb{V}ar_\theta\big[\,l'(X_i\,,\theta)\,\big] = \mathbb{E}_\theta\Big[\,\big(l'(X_i\,,\theta)\big)^2\Big]$$

$$= \mathbb{E}_\theta\,\Big[\Big(\frac{\partial \ln\, p\,(X,\,\theta)}{\partial\,\theta}\Big)^2\Big] \,=\, \int_{\mathbb{R}} \Big(\,\frac{\partial \ln p(x\,,\theta)}{\partial\theta}\,\Big)^{\,2}\,p(x\,,\theta)\,dx$$

$$= \int_{\mathbb{R}} \frac{\big(\,\partial p(x\,,\theta)/\partial\theta\,\big)^2}{p(x\,,\theta)}\,dx.$$

Remark 1.6. In the above definition of the Fisher information, the density appears in the denominator. Thus, it is problematic to calculate the Fisher information for distributions with densities that may be equal to zero for some values of x; even more so, if the density turns into zero as a function of x on sets that vary depending on the value of θ. A more general approach to the concept of information that overcomes this difficulty will be suggested in Section 4.2. □

The *Fisher information for a statistical experiment of size n* is the variance of the total Fisher score function,

$$\begin{aligned} I_n(\theta) \,&=\, \mathbb{V}ar_\theta\big[\, L_n'(\theta)\,\big] \,=\, \mathbb{E}_\theta\,\big[\,\big(\,L_n'(\theta)\,\big)^{\,2}\,\big] \\ &= \mathbb{E}_\theta\Big[\Big(\frac{\partial \ln\, p\,(X_1,\ldots,X_n,\,\theta)}{\partial\,\theta}\Big)^2\,\Big] \\ = \int_{\mathbb{R}^n} &\frac{(\partial\, p\,(x_1,\ldots,x_n,\theta)/\partial\theta)^2}{p\,(x_1,\ldots,x_n,\theta)}\; dx_1\,\ldots\,dx_n\,. \end{aligned}$$

Lemma 1.7. *For independent observations, the Fisher information is additive. In particular, for any* $\theta \in \Theta$, *the equation holds* $I_n(\theta) = n\,I(\theta)$.

Proof. As the variance of the sum of n independent random variables,

$$\begin{aligned} I_n(\theta) \,&=\, \mathbb{V}ar_\theta\big[\, L_n'(\theta)\,\big] \,=\, \mathbb{V}ar_\theta\big[\, l'(X_1\,,\theta) + \,\ldots\, + l'(X_n\,,\theta)\,\big] \\ &=\, n\,\mathbb{V}ar_\theta\big[\, l'(X_1\,,\theta)\,\big] \,=\, n\,I(\theta). \end{aligned}$$

□

In view of this lemma, we use the following definition of the Fisher information for a random sample of size n:

$$I_n(\theta) \,=\, n\,\mathbb{E}_\theta\,\Big[\Big(\frac{\partial \ln\, p\,(X,\,\theta)}{\partial\,\theta}\Big)^2\,\Big].$$

Another way of computing the Fisher information is presented in Exercise 1.1.

1.3. The Cramér-Rao Lower Bound

A statistical experiment is called *regular* if its Fisher information is continuous, strictly positive, and bounded for all $\theta \in \Theta$. Next we present an inequality for the variance of any estimator of θ in a regular experiment. This inequality is termed the *Cramér-Rao inequality*, and the lower bound is known as the *Cramér-Rao lower bound*.

Theorem 1.8. *Consider an estimator* $\hat\theta_n = \hat\theta_n(X_1,\ldots,X_n)$ *of the parameter* θ *in a regular experiment. Suppose its bias* $b_n(\theta) \,=\, \mathbb{E}_\theta\big[\,\hat\theta_n\big] - \theta$ *is continuously differentiable. Let* $b_n'\,(\theta)$ *denote the derivative of the bias. Then the variance of* $\hat\theta_n$ *satisfies the inequality*

$$\mathbb{V}ar_\theta\big[\,\hat\theta_n\big] \,\geq\, \frac{\big(\,1\,+\,b_n'(\theta)\,\big)^2}{I_n(\theta)}\,, \quad \theta \in \Theta. \tag{1.1}$$

Proof. By the definition of the bias, we have that

$$\theta + b_n(\theta) = \mathbb{E}_\theta\big[\,\hat\theta_n\big] \,=\, \int_{\mathbb{R}^n} \hat\theta_n(x_1,\ldots,x_n)\,p\,(x_1,\ldots,x_n,\theta)\,dx_1\ldots dx_n.$$

In the regular case, the differentiation and integration are interchangeable, hence differentiating in θ, we get the equation,

$$1 + b_n'(\theta) = \int_{\mathbb{R}^n} \hat{\theta}_n(x_1, \dots, x_n) \left[\partial p(x_1, \dots, x_n, \theta)/\partial\theta\right] dx_1 \dots dx_n$$

$$= \int_{\mathbb{R}^n} \hat{\theta}_n(x_1, \dots, x_n) \Big(\frac{\partial p(x_1, \dots, x_n, \theta)/\partial\theta}{p(x_1, \dots, x_n, \theta)}\Big) p(x_1, \dots, x_n, \theta)\, dx_1 \dots dx_n$$

$$= \mathbb{E}_\theta\big[\hat{\theta}_n\, L_n'(\theta)\big] = \mathbb{C}ov_\theta\big[\hat{\theta}_n\,,\, L_n'(\theta)\big]$$

where we use the fact that $\mathbb{E}_\theta\big[L_n'(\theta)\big] = 0$. The correlation coefficient ρ_n of $\hat{\theta}_n$ and $L_n'(\theta)$ does not exceed 1 in its absolute value, so that

$$1 \geq \rho_n^2 = \frac{\big(\mathbb{C}ov_\theta\big[\hat{\theta}_n\,,\, L_n'(\theta)\big]\big)^2}{\mathbb{V}ar_\theta[\hat{\theta}_n]\ \mathbb{V}ar_\theta[L_n'(\theta)]} = \frac{(1 + b_n'(\theta))^2}{\mathbb{V}ar_\theta[\hat{\theta}_n]\ I_n(\theta)}. \qquad \square$$

1.4. Efficiency of Estimators

An immediate consequence of Theorem 1.8 is the formula for unbiased estimators.

Corollary 1.9. *For an unbiased estimator $\hat{\theta}_n$, the Cramér-Rao inequality* (1.1) *takes the form*

$$\mathbb{V}ar_\theta\big[\hat{\theta}_n\big] \geq \frac{1}{I_n(\theta)}, \quad \theta \in \Theta. \qquad (1.2) \qquad \square$$

An unbiased estimator $\theta_n^* = \theta_n^*(X_1, \dots, X_n)$ in a regular statistical experiment is called *Fisher efficient* (or, simply, *efficient*) if, for any $\theta \in \Theta$, the variance of θ_n^* reaches the Cramér-Rao lower bound, that is, the equality in (1.2) holds:

$$\mathbb{V}ar_\theta\big[\theta_n^*\big] = \frac{1}{I_n(\theta)}, \quad \theta \in \Theta.$$

Example 1.10. Suppose, as in Example 1.1(a), the observations $X_1, \dots, X_n$ are independent $\mathcal{N}(\theta, \sigma^2)$ where σ^2 is assumed known. We show that the sample mean $\bar{X}_n = (X_1 + \dots + X_n)/n$ is an efficient estimator of θ. Indeed, $\bar{X}_n$ is unbiased and $\mathbb{V}ar_\theta\big[\bar{X}_n\big] = \sigma^2/n$. On the other hand,

$$\ln p(X, \theta) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(X-\theta)^2}{2\sigma^2}$$

and

$$l'(X, \theta) = \frac{\partial \ln p(X, \theta)}{\partial\theta} = \frac{X - \theta}{\sigma^2}.$$

Thus, the Fisher information for the statistical experiment is

$$I_n(\theta) = n\,\mathbb{E}_\theta\big[\big(l'(X, \theta)\big)^2\big] = \frac{n}{\sigma^4}\,\mathbb{E}_\theta\big[(X-\theta)^2\big] = \frac{n\sigma^2}{\sigma^4} = \frac{n}{\sigma^2}.$$

Therefore, for any value of θ, the variance of $\bar{X}_n$ achieves the Cramér-Rao lower bound $1/I_n(\theta) = \sigma^2/n$. □

The concept of the Fisher efficiency seems to be nice and powerful. Indeed, besides being unbiased, an efficient estimator has the minimum possible variance uniformly in $\theta \in \Theta$. Another feature is that it applies to any sample size n. Unfortunately, this concept is extremely restrictive. It works only in a limited number of models. The main pitfalls of the Fisher efficiency are discussed in the next chapter.

Exercises

Exercise 1.1. Show that the Fisher information can be computed by the formula

$$I_n(\theta) = -n\,\mathbb{E}_\theta\Big[\frac{\partial^2 \ln p\,(X,\theta)}{\partial\,\theta^2}\Big].$$

Hint: Make use of the representation (show!)

$$\Big(\frac{\partial \ln p\,(x,\theta)}{\partial\,\theta}\Big)^2 p\,(x,\,\theta) = \frac{\partial^2 p\,(x,\,\theta)}{\partial\theta^2} - \Big(\frac{\partial^2 \ln p\,(x,\,\theta)}{\partial\theta^2}\Big)\,p\,(x,\,\theta).$$

Exercise 1.2. Let $X_1, \ldots, X_n$ be independent observations with the $\mathcal{N}(\mu,\,\theta)$ distribution, where μ has a known value (refer to Example 1.1(b)). Prove that

$$\theta_n^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

is an efficient estimator of θ. Hint: Use Exercise 1.1 to show that $I_n(\theta) = n/(2\,\theta^2)$. When computing the variance of θ_n^*, first notice that the variable $\sum_{i=1}^n (X_i-\mu)^2/\theta$ has a chi-squared distribution with n degrees of freedom, and, thus, its variance equals $2n$.

Exercise 1.3. Suppose that independent observations $X_1, \ldots, X_n$ have a Bernoulli distribution with the probability mass function

$$p\,(x,\,\theta) = \theta^{\,x}\,(1-\theta)^{1-x}, \quad x \in \{\,0,1\,\},\ \ 0 < \theta < 1.$$

Show that the Fisher information is of the form

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

and verify that the estimator $\theta_n^* = \bar{X}_n$ is efficient.

Exercise 1.4. Assume that $X_1, \ldots, X_n$ are independent observations from a Poisson distribution with the probability mass function

$$p\,(x,\,\theta) = \frac{\theta^{\,x}}{x!}\,e^{-\theta}, \quad x \in \{\,0, 1, \ldots\,\}, \quad \theta > 0.$$

Prove that the Fisher information in this case is $I_n(\theta) = n/\theta$, and show that $\bar{X}_n$ is an efficient estimator of θ.

Exercise 1.5. Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with the density

$$p\,(x,\,\theta) = \frac{1}{\theta}\,e^{-\,x/\theta}, \quad x > 0, \quad \theta > 0.$$

Verify that $I_n(\theta) = n/\theta^2$, and prove that $\bar{X}_n$ is efficient.

Exercise 1.6. Show that in the exponential model with the density $p(x\,,\theta) = \theta\,\exp\{-\theta\,x\}\,,\ x\,,\theta > 0$, the MLE $\theta_n^* = 1/\bar{X}_n$ has the expected value $\mathbb{E}_\theta[\,\theta_n^*\,] = n\,\theta/(n-1)$. What is the variance of this estimator?

Exercise 1.7. Show that for the location parameter model with the density $p(x\,,\theta) = p_0(x-\theta)$, introduced in Example 1.2, the Fisher information is a constant if it exists.

Exercise 1.8. In the Exercise 1.7, find the values of α for which the Fisher information exists if $p_0(x) = C\,\cos^\alpha x\,, \quad -\pi/2 < x < \pi/2\,,$ and $p_0(x) = 0$ otherwise, where $C = C(\alpha)$ is the normalizing constant. Note that p_0 is a probability density if $\alpha > -1\,.$

Chapter 2

The Bayes and Minimax Estimators

2.1. Pitfalls of the Fisher Efficiency

Fisher efficient estimators defined in the previous chapter possess two major unattractive properties, which prevent the Fisher efficiency from being widely used in statistical theory. First, the Fisher efficient estimators rarely exist, and second, they need to be unbiased. In effect, the Fisher efficiency does not provide an answer to how to compare biased estimators with different bias functions. A lesser issue is that the comparison of estimators is based on their variances alone.

Before we proceed to an illustrative example, we need several notions defined below. A function $w(u),\ u \in \mathbb{R}$, is called a *loss function* if: (i) $w(0) = 0$, (ii) it is symmetric, $w(u) = w(-u)$, (iii) it is non-decreasing for $u > 0$, and (iv) it is not identically equal to zero. Besides, we require that w is bounded from above by a power function, that is, (v) $w(u) \le k(1 + |u|^a)$ for all u with some constants $k > 0$ and $a > 0$.

The loss function $w(\hat{\theta}_n - \theta)$ measures the deviation of the estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ from the true parameter θ. In this book, we do not go far beyond: (i) *quadratic loss function*, $w(u) = u^2$, (ii) *absolute loss function*, $w(u) = |u|$, or (iii) *bounded loss function*, $w(u) = \mathbb{I}(\,|u| > c\,)$ with a given positive c, where $\mathbb{I}(\cdot)$ denotes the indicator function.

The *normalized risk function* (or simply, the *normalized risk*) $R_n(\theta, \hat{\theta}_n, w)$ is the expected value of the loss function w evaluated at $\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta)$,

that is,

$$R_n(\theta, \hat{\theta}_n, w) \;=\; \mathbb{E}_\theta\Big[\, w\big(\sqrt{I_n(\theta)}(\hat{\theta}_n - \theta)\,\big)\,\Big]$$

$$= \int_{\mathbb{R}^n} w\Big(\sqrt{I_n(\theta)}(\hat{\theta}_n(x_1,\ldots,x_n) - \theta)\Big)\, p(x_1,\ldots,x_n\,,\theta)\, dx_1 \ldots dx_n.$$

Example 2.1. For the quadratic loss function $w(u) = u^2$, the normalized risk (commonly termed the *normalized quadratic risk*) of an estimator $\hat{\theta}_n$ can be found as

$$R_n(\theta, \hat{\theta}_n, u^2) = \mathbb{E}_\theta\Big[\, I_n(\theta)\big(\hat{\theta}_n - \theta\,\big)^2\,\Big] \;=\; I_n(\theta)\,\mathbb{E}_\theta\Big[\, \big(\hat{\theta}_n - \mathbb{E}_\theta\,[\,\hat{\theta}_n\,] + \mathbb{E}_\theta\,[\,\hat{\theta}_n\,] - \theta\big)^2\,\Big]$$

$$= I_n(\theta)\,\Big[\, \mathbb{V}ar_\theta[\,\hat{\theta}_n\,] \;+\; b_n^2(\theta\,,\hat{\theta}_n)\,\Big] \tag{2.1}$$

where $b_n(\theta\,,\hat{\theta}_n) = \mathbb{E}_\theta\,[\,\hat{\theta}_n\,] - \theta$ is the bias of $\hat{\theta}_n$. □

By (2.1), for any unbiased estimator $\hat{\theta}_n$, the normalized quadratic risk function has the representation $R_n(\theta, \hat{\theta}_n, u^2) = I_n(\theta)\mathbb{V}ar_\theta[\,\hat{\theta}_n\,]$. The Cramér-Rao inequality (1.2) can thus be written as

$$R_n(\theta, \hat{\theta}_n, u^2) \;=\; \mathbb{E}_\theta\,\Big[\, I_n(\theta)\big(\hat{\theta}_n - \theta\,\big)^2\,\Big] \;\geq\; 1, \quad \theta \in \Theta, \tag{2.2}$$

with the equality attained for the Fisher efficient estimators θ_n^*,

$$R_n(\theta, \theta_n^*, u^2) \;=\; \mathbb{E}_\theta\,\Big[\, I_n(\theta)\big(\theta_n^* \,-\, \theta\,\big)^2\,\Big] \;=\; 1, \quad \theta \in \Theta. \tag{2.3}$$

Next, we present an example of a biased estimator that in a certain interval performs more efficiently than the Fisher efficient unbiased estimator, if we define a *more efficient estimator* as the one with a smaller normalized quadratic risk.

Example 2.2. Let $X_1, \ldots, X_n$ be independent observations from the $\mathcal{N}(\theta,\, \sigma^2)$ distribution, where σ^2 is known. Consider two estimators: (i) $\theta_n^* = \bar{X}_n$, which is efficient by Example 1.10, and (ii) a constant-value estimator $\hat{\theta} = \theta_0$, where θ_0 is a fixed point. The normalized quadratic risk of θ_n^* equals the unity by (2.3), while that of $\hat{\theta}$ is

$$R_n(\theta, \hat{\theta}, u^2) = \mathbb{E}_\theta\big[\, I_n(\theta)(\hat{\theta} - \theta)^2\,\big] \;=\; \frac{n}{\sigma^2}\,(\theta_0 - \theta)^2.$$

Note that $\hat{\theta}$ is a biased estimator with the bias $b_n(\theta) = \theta_0 - \theta$.

It is impossible to determine which of the two normalized quadratic risks is smaller (refer to Figure 1). If θ is within $\theta_0 \pm \sigma/\sqrt{n}$, then $\hat{\theta}$ is more efficient, whereas for all other values of θ, θ^* is a more efficient estimator. □

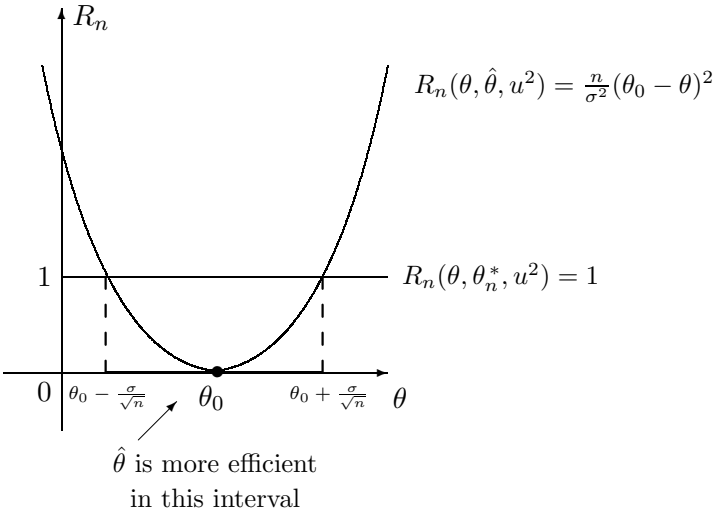

Figure 1. The normalized quadratic risk functions in Example 2.2.

This example illustrates the difficulty in comparing normalized risks of two estimators as functions of $\theta \in \Theta$. To overcome it, we could try to represent each risk function by a positive number. In statistics, there are two major ways to implement this idea. One approach is to integrate the normalized risk over the parameter set Θ, whereas the other one is to take the maximum value of the normalized risk function over Θ. These are called the Bayes and the minimax approaches, respectively. They are explored in the next three sections.

2.2. The Bayes Estimator

In what follows, we study only regular statistical models, which by definition have a strictly positive, continuous Fisher information.

Assume that there is a probability density $\pi(\theta)$ defined on the parameter set Θ. The density $\pi(\theta)$ is called a *prior density* of θ. It reflects the judgement of how likely values of θ are before the data are obtained. The *Bayes risk* of $\hat{\theta}_n$ is the integrated value of the normalized risk function,

$$\beta_n(\hat{\theta}_n,\, w,\, \pi) \,=\, \int_\Theta R_n(\theta, \hat{\theta}_n, w)\, \pi(\theta)\, d\theta\,. \tag{2.4}$$

An estimator $t_n = t_n(X_1, \ldots, X_n)$ is called the *Bayes estimator* of θ, if for any other estimator $\hat{\theta}_n$, the following inequality holds:

$$\beta_n(t_n,\, w, \pi) \,\le\, \beta_n(\hat{\theta}_n,\, w, \pi).$$

In other words, the Bayes estimator minimizes the Bayes risk. Loosely speaking, we can understand the Bayes estimator as a solution of the minimization problem,

$$t_n \,=\, \mathrm{argmin}_{\hat{\theta}_n}\ \beta(\hat{\theta}_n,\, w, \pi),$$

though we should keep in mind that the minimum value may not exist or may be non-unique.

In the case of the quadratic loss $w(u) = u^2$, the Bayes estimator can be computed explicitly. Define the *posterior density* of θ as the conditional density, given the observations $X_1, \ldots, X_n$; that is,

$$f(\theta \,|\, X_1, \ldots, X_n) \,=\, C_n\, p(X_1, \ldots, X_n\,,\, \theta)\, \pi(\theta), \quad \theta \in \Theta,$$

where $C_n = C_n\,(X_1, \ldots, X_n)$ is the normalizing constant. Assuming that

$$\int_\Theta I_n(\theta)\, f(\theta \,|\, X_1\,, \ldots\,, X_n)\, d\theta \,<\, \infty,$$

we can introduce the *weighted posterior density* as

$$\tilde{f}(\theta \,|\, X_1\,, \ldots\,, X_n) \,=\, \tilde{C}_n\, I_n(\theta)\, f(\theta \,|\, X_1\,, \ldots\,, X_n), \quad \theta \in \Theta,$$

with the normalizing constant $\tilde{C}_n = \big[\, \int_\Theta\, I_n(\theta)\, f(\theta \,|\, X_1\,, \ldots\,, X_n)\, d\theta\big]^{-1}$, which is finite under our assumption.

Theorem 2.3. *If $w(u) = u^2$, then the Bayes estimator t_n is the weighted posterior mean*

$$t_n \,=\, t_n(X_1\,, \ldots\,, X_n) \,=\, \int_\Theta\ \theta\ \tilde{f}(\theta \,|\, X_1, \ldots, X_n)\, d\theta.$$

In particular, if the Fisher information is a constant independent of θ, then the Bayes estimator is the non-weighted posterior mean,

$$t_n \,=\, t_n(X_1\,, \ldots\,, X_n) \,=\, \int_\Theta\ \theta\, f(\theta \,|\, X_1, \ldots, X_n)\, d\theta.$$

Proof. The Bayes risk of an estimator $\hat{\theta}_n$ with respect to the quadratic loss can be written in the form

$$\beta_n(\hat{\theta}_n,\, \pi) \,=\, \int_\Theta \int_{\mathbb{R}^n} I_n(\theta)\, (\hat{\theta}_n - \theta)^2\, p(x_1, \ldots, x_n\,,\, \theta)\, \pi(\theta)\ dx_1 \ldots dx_n\, d\theta$$

$$= \int_{\mathbb{R}^n} \Big[\int_\Theta (\hat{\theta}_n - \theta)^2\, \tilde{f}(\,\theta \,|\, x_1, \ldots, x_n)\, d\theta \Big]\, \tilde{C}_n^{-1}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n.$$

Thus, the minimization problem of the Bayes risk is tantamount to minimization of the integral

$$\int_\Theta (\hat{\theta}_n - \theta)^2\, \tilde{f}(\,\theta \,|\, x_1, \ldots, x_n)\, d\theta$$

with respect to $\hat{\theta}_n$ for any fixed values $x_1, \dots, x_n$. Equating to zero the derivative of this integral with respect to $\hat{\theta}_n$ produces a linear equation, satisfied by the Bayes estimator t_n,

$$\int_\Theta (t_n - \theta)\, \tilde{f}(\theta \,|\, x_1, \dots, x_n)\, d\theta = 0.$$

Recalling that $\int_\Theta \tilde{f}(\theta \,|\, x_1, \dots, x_n)\, d\theta = 1$, we obtain the result,

$$t_n = \int_\Theta \theta\, \tilde{f}(\theta \,|\, x_1, \dots, x_n)\, d\theta. \qquad \square$$

In many examples, the weighted posterior mean t_n is easily computable if we choose a prior density $\pi(\theta)$ from a *conjugate family* of distributions. A *conjugate prior distribution* $\pi(\theta)$ is such that the posterior distribution belongs to the same family of distributions for any sample $X_1, \dots, X_n$. If the posterior distribution allows a closed-form expression of expectations, then t_n can be found without integration. The following example illustrates the idea.

Example 2.4. Consider independent Bernoulli observations $X_1, \dots, X_n$ with the probability mass function

$$p(x, \theta) = \theta^x\, (1-\theta)^{1-x}, \quad x \in \{0, 1\}, \quad 0 < \theta < 1,$$

where θ is assumed to be a random variable. The joint distribution function of the sample is

$$p(X_1, \dots, X_n, \theta) = \theta^{\sum X_i}\, (1-\theta)^{n - \sum X_i}.$$

As a function of θ, it has an algebraic form or a beta distribution. Thus, we select a beta density as a prior density,

$$\pi(\theta) = C(\alpha, \beta)\, \theta^{\alpha-1}\, (1-\theta)^{\beta-1}, \quad 0 < \theta < 1,$$

where α and β are positive parameters, and $C(\alpha, \beta)$ is the normalizing constant. The posterior density is then also a beta density,

$$f\big(\theta \,|\, X_1, \dots, X_n\big) = C(\alpha, \beta)\, \theta^{\sum X_i + \alpha - 1}\, (1-\theta)^{n - \sum X_i + \beta - 1}, \quad 0 < \theta < 1.$$

By Exercise 1.3, the Fisher information is equal to $I_n(\theta) = n/[\,\theta(1-\theta)]$. Thus, the weighted posterior density is a beta density as well,

$$\tilde{f}\big(\theta \,|\, X_1, \dots, X_n\big) = \tilde{C}_n \theta^{\sum X_i + \alpha - 2}\, (1-\theta)^{n - \sum X_i + \beta - 2}, \quad 0 < \theta < 1,$$

where $\alpha > 1$ and $\beta > 1$. The weighted posterior mean therefore is equal to

$$t_n = \frac{\sum X_i + \alpha - 1}{\sum X_i + \alpha - 1 + n - \sum X_i + \beta - 1} = \frac{\sum X_i + \alpha - 1}{n + \alpha + \beta - 2}.$$

More examples of the conjugate families are in the exercises.

2.3. Minimax Estimator. Connection Between Estimators

Define a *maximum normalized risk* of an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ with respect to a loss function w by

$$r_n(\hat{\theta}_n,\, w) \,=\, \sup_{\theta\in\Theta} R_n(\theta, \hat{\theta}_n\,, w) \,=\, \sup_{\theta\in\Theta} \mathbb{E}_\theta\Big[w\big(\,\sqrt{I_n(\theta)}\,(\hat{\theta}_n \,-\, \theta)\,\big)\Big]\,.$$

An estimator $\theta_n^* = \theta_n^*(X_1, \dots, X_n)$ is called *minimax* if its maximum normalized risk does not exceed that of any other estimator $\hat{\theta}_n$. That is, for any estimator $\hat{\theta}_n$,

$$r_n(\theta_n^*,\, w) \le r_n(\hat{\theta}_n,\, w).$$

The maximum normalized risk of a minimax estimator, $r_n(\theta_n^*,\, w)$, is called the *minimax risk*.

In contrast with the Bayes estimator, the minimax estimator represents a different concept of the statistical optimality. The Bayes estimator is optimal in the averaged (integrated) sense, whereas the minimax one takes into account the "worst-case scenario".

It follows from the above definition that a minimax estimator θ_n^* solves the optimization problem

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta\Big[w\big(\,\sqrt{I_n(\theta)}\,(\hat{\theta}_n \,-\, \theta)\,\big)\Big] \;\to\; \inf_{\hat{\theta}_n}.$$

Finding the infimum over all possible estimators $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$, that is, over all functions of observations $X_1\,, \dots\,, X_n\,$, is not an easily tackled task. Even for the most common distributions, such as normal or binomial, the direct minimization is a hopeless endeavor. This calls for an alternative route in finding minimax estimators.

In this section we establish a connection between the Bayes and minimax estimators that will lead to some advances in computing the latter. The following theorem shows that if the Bayes estimator has a constant risk, then it is also minimax.

Theorem 2.5. *Let $t_n = t_n(X_1, \dots, X_n)$ be a Bayes estimator with respect to a loss function w. Suppose that the normalized risk function of the Bayes estimator is a constant for any $\theta \in \Theta$, that is,*

$$R_n(\theta, t_n\,, w) \,=\, \mathbb{E}_\theta\Big[\, w\big(\sqrt{I_n(\theta)}\,(\,t_n - \theta)\big)\,\Big] \,=\, c$$

for some $c > 0$. Then t_n is also a minimax estimator.

Proof. Notice that since the risk function of t_n is a constant, the Bayes and maximum normalized risks of t_n are the same constants. Indeed, letting

$\pi(\theta)$ denote the corresponding prior density, we write

$$\beta_n(t_n\,,\,w,\pi) = \int_\Theta R_n(\theta, t_n\,, w)\,\pi(\theta)\,d\theta = c\int_\Theta \pi(\theta)\,d\theta = c$$

and

$$r_n(t_n,\,w) = \sup_{\theta\in\,\Theta} R_n(\theta, t_n\,, w) \;=\; \sup_{\theta\in\,\Theta} c = c.$$

Further, for any estimator $\hat\theta_n$,

$$r_n(\hat\theta_n, w) = \sup_{\theta\in\,\Theta} R_n(\theta, \hat\theta_n\,, w) \ge \int_\Theta R_n(\theta, \hat\theta_n\,, w)\,\pi(\theta)\,d\theta$$

$$= \beta_n(\hat\theta_n\,,\,w,\pi) \ge \beta_n(t_n\,,\,w,\pi) = c = r_n(t_n\,, w). \qquad \square$$

Unfortunately, Theorem 2.5 does not provide a recipe for choosing a prior density for which the normalized risk function is a constant on Θ. Moreover, constant-risk priors rarely exist. Below we give two examples where we try to explain why it happens.

Example 2.6. Consider independent Bernoulli observations $X_1, \ldots, X_n$ with parameter θ. As shown in Example 2.4, the weighted posterior mean of θ is

$$t_n \;=\; \frac{\sum X_i + \alpha - 1}{n + \alpha + \beta - 2}.$$

If we now select $\alpha = \beta = 1$, then t_n becomes the sample mean $\bar X_n$. From Exercise 1.3 we know that $\bar X_n$ is an efficient estimator of θ, and therefore its weighted quadratic risk is equal to 1, a constant. However, $\alpha = \beta = 1$ is not a legitimate choice in this instance, because the weighted posterior density

$$\tilde f(\theta\,|\,X_1, \ldots, X_n) \;=\; \tilde C_n \theta^{\sum X_i - 1}\,(1-\theta)^{\,n - \sum X_i - 1}$$

does not exist for $\sum X_i = 0$. Indeed, $\theta^{-1}\,(1-\theta)^{\,n-1}$ is not integrable at zero, and therefore the normalizing constant $\tilde C_n$ does not exist. $\square$

Example 2.7. Let $X_1, \ldots, X_n$ be independent observations from the $\mathcal{N}(\theta,\,1)$ distribution. If we choose the prior density of θ to be $\mathcal{N}(0,\,b^2)$ for some positive real b, then, by Exercise 2.10, the weighted posterior distribution is also normal,

$$\mathcal{N}\Big(\frac{n\,b^2\,\bar X_n}{n\,b^2+1},\;\frac{b^2}{n\,b^2+1}\Big).$$

Here the weighted posterior mean $t_n \;=\; n\,b^2\,\bar X_n/(n\,b^2+1)$ is the Bayes estimator with respect to the quadratic loss function. If we let $b\to\infty$, then t_n equals $\bar X_n$, which is Fisher efficient (see Example 1.10) and thus has a constant normalized quadratic risk. The flaw in this argument is that no normal prior density exists with infinite b. $\square$

2.4. Limit of the Bayes Estimator and Minimaxity

Assume that we can find a family of prior distributions with the densities $\pi_b(\theta)$ indexed by a positive real number b. If the Bayes risks of the respective Bayes estimators have a limit as b goes to infinity, then this limit guarantees a minimax lower bound. A rigorous statement is presented in the following theorem.

Theorem 2.8. *Let $\pi_b(\theta)$ be a family of prior densities on Θ that depend on a positive real parameter b, and let $t_n(b) = t_n(X_1, \ldots, X_n, b)$ be the respective Bayes estimators for a loss function w. Suppose that the Bayes risk $\beta_n\big(t_n(b), w, \pi_b\big)$ has a limit,*

$$\lim_{b\to\infty} \beta_n\big(t_n(b), w, \pi_b\big) = c > 0.$$

Then the minimax lower bound holds for any n,

$$\inf_{\hat\theta_n} r_n(\hat\theta_n, w) = \inf_{\hat\theta_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta\Big[w\big(\sqrt{I_n(\theta)}\,(\hat\theta_n - \theta)\big)\Big] \geq c.$$

Proof. As in the proof of Theorem 2.5, for any estimator $\hat\theta_n$, we can write

$$r_n(\hat\theta_n, w) = \sup_{\theta\in\Theta} R_n(\theta, \hat\theta_n, w) \geq \int_\Theta R_n(\theta, \hat\theta_n, w)\,\pi_b(\theta)\,d\theta$$

$$= \beta_n(\hat\theta_n, w, \pi_b) \geq \beta_n\big(t_n(b), w, \pi_b\big).$$

Now take the limit as $b \to \infty$. Since the left-hand side is independent of b, the theorem follows. □

Example 2.9. Let $X_1, \ldots, X_n$ be independent $\mathcal{N}(\theta, 1)$ observations. We will show that conditions of Theorem 2.8 are satisfied under the quadratic loss function $w(u) = u^2$, and therefore the lower bound for the corresponding minimax risk holds:

$$\inf_{\hat\theta_n} r_n(\hat\theta_n, w) = \inf_{\hat\theta_n} \sup_{\theta\in\mathbb{R}} E_\theta\Big[\big(\sqrt{n}\,(\hat\theta_n - \theta)\big)^2\Big] \geq 1.$$

As shown in Example 2.7, for a $\mathcal{N}(0, b^2)$ prior density, the weighted posterior mean $t_n(b) = n\,b^2\,\bar X_n/(n\,b^2+1)$ is the Bayes estimator with respect to the quadratic loss function. Now we will compute its Bayes risk. This estimator has the variance

$$\mathbb{V}ar_\theta\big[t_n(b)\big] = \frac{n^2\,b^4\,\mathbb{V}ar_\theta\big[\bar X_n\big]}{(n\,b^2+1)^2} = \frac{n\,b^4}{(n\,b^2+1)^2}$$

and the bias

$$b_n\big(\theta, t_n(b)\big) = \mathbb{E}_\theta\big[t_n(b)\big] - \theta = \frac{n^2\,b^2\,\theta}{n\,b^2+1} - \theta = -\frac{\theta}{n\,b^2+1}.$$

Therefore, its normalized quadratic risk is expressed as

$$R_n\big(\theta\,, t_n(b), w\big) = E_\theta\Big[\big(\sqrt{n}\,(t_n(b)\,-\,\theta)\big)^2\Big] = n\Big[\,\mathbb{V}ar_\theta[\,t_n(b)\,]\,+\,b_n^2\big(\theta\,,\,t_n(b)\big)\Big]$$

$$= \frac{n^2\,b^4}{(n\,b^2+1)^2}\,+\,\frac{n\,\theta^2}{(n\,b^2+1)^2}.$$

With the remark that $\int_{\mathbb{R}}\,\theta^2\,\pi_b\,(\theta)\,d\theta = b^2$, the Bayes risk of $t_n(b)$ equals

$$\beta_n\big(t_n(b), w, \pi_b\big)\;=\;\int_{\mathbb{R}}\,\Big[\,\frac{n^2\,b^4}{(n\,b^2+1)^2}\,+\,\frac{n\,\theta^2}{(n\,b^2+1)^2}\,\Big]\,\pi_b\,(\theta)\,d\theta$$

$$= \frac{n^2\,b^4}{(n\,b^2+1)^2}\,+\,\frac{n\,b^2}{(n\,b^2+1)^2}\,\to\,1\;\text{ as }\;b\to\infty.$$

Applying Theorem 2.8, we obtain the result with $c\,=\,1$. Taking a step further, note that the minimax lower bound is attained for the estimator $\bar{X}_n$, which is thus minimax. Indeed, $\mathbb{E}_\theta\,\Big[\big(\sqrt{n}\,(\bar{X}_n-\theta)\big)^2\Big]\,=\,1$. □

In subsequent chapters we present additional useful applications of Theorem 2.8.

Exercises

Exercise 2.9. Suppose the random observations $X_1,\ldots,X_n$ come from a Poisson distribution with the probability mass function

$$p(x\,,\theta) = \frac{\theta^{\,x}\,e^{-\theta}}{x!}\,,\;\;x\in\{\,0\,,1\,,\ldots\},$$

where θ is a random variable. Show that the conjugate prior density of θ is a gamma density, $\pi(\theta)\,=\,C(\alpha,\,\beta)\,\theta^{\,\alpha-1}\,e^{-\,\beta\,\theta}$, $\theta\,>\,0$, for some positive parameters α and β, and the normalizing constant $C(\alpha,\,\beta)$. Find the weighted posterior mean of θ.

Exercise 2.10. Consider a set of independent observations $X_1,\ldots,X_n\,\sim\,\mathcal{N}(\theta,\sigma^2)$, where θ is assumed random with the prior density $\mathcal{N}(\mu,\sigma_\theta^2)$. Show that the weighted posterior distribution of θ is also normal with the mean $\big(n\,\sigma_\theta^2\,\bar{X}_n+\mu\sigma^2\big)/\big(n\,\sigma_\theta^2+\sigma^2\big)$ and variance $\sigma^2\,\sigma_\theta^2/(n\sigma_\theta^2+\sigma^2)$. Note that the family of normal distributions is self-conjugate.

Exercise 2.11. Find a conjugate distribution and the corresponding Bayes estimator for the parameter θ in the exponential model with $p(x,\theta)\;=\;\theta\exp\{-\,\theta\,x\}$,$\;\;x\,,\theta\,>\,0$.

Exercise 2.12. Consider n independent Bernoulli observations $X_1, \ldots, X_n$ with $p(x,\theta) = \theta^x (1-\theta)^{1-x}$, $x \in \{0, 1\}$, and $\Theta = (0,1)$. Define the estimator

$$\theta_n^* = \frac{\sum X_i + \sqrt{n}/2}{n + \sqrt{n}}.$$

(i) Verify that θ_n^* is the non-weighted posterior mean with respect to the conjugate prior density $\pi(\theta) = C\left[\theta(1-\theta)\right]^{\sqrt{n}/2-1}$, $0 < \theta < 1$.

(ii) Show that the non-normalized quadratic risk of θ_n^* (with the factor $\sqrt{I_n(\theta)}$ omitted) is equal to

$$\mathbb{E}_\theta\left[(\theta_n^* - \theta)^2\right] = \frac{1}{4(1+\sqrt{n})^2}.$$

(iii) Verify that Theorem 2.5 is valid for a non-normalized risk function, and argue that θ_n^* is minimax in the appropriate sense.

Exercise 2.13. Refer to the Bernoulli model in Example 2.4. Show that the prior beta distribution with $\alpha = \beta = 1 + b^{-1}$ defines the weighted posterior mean $t_n(b)$ which is minimax for $b = \infty$.