# Chapter 7

# Point Estimation

## 7.1 Introduction

**Definition 7.1.1** *A point estimator is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic is a point estimator.*

Note that an *estimator* is a function of the sample, while an estimate is the realized value of an estimator that is obtained when a sample is actually taken.

## 7.2 Methods of Finding Estimators

### 7.2.1 Methods of Moments

Let $X_1, \ldots, X_n$ be a sample from a population with pdf or pmf $f(x|\theta_1, \ldots, \theta_k)$. Methods of moments estimators are found by equating the first $k$ sample moments to the corresponding $k$ population moments, and solving the resulting system of simultaneous equations. More precisely, define

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad \mu_1' = EX^1$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \mu_2' = EX^2$$

$$\vdots$$

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \mu_k' = EX^k$$

The population moment $\mu_j'$ will typically be a function of $\theta_1, \ldots, \theta_k$, say $\mu_j'(\theta_1, \ldots, \theta_k)$. The method of moments estimator $(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ of $(\theta_1, \ldots, \theta_k)$ is obtained by solving the following system of equations for $(\theta_1, \ldots, \theta_k)$ in terms of $(m_1, \ldots, m_k)$:

$$\begin{aligned} m_1 &= \mu_1'(\theta_1, \ldots, \theta_k), \\ m_2 &= \mu_2'(\theta_1, \ldots, \theta_k), \\ &\vdots \\ m_k &= \mu_k'(\theta_1, \ldots, \theta_k). \end{aligned}$$

**Example 7.2.1** *(***Normal method of moments***) Suppose* $X_1, \ldots, X_n$ *are iid* $N(\mu, \sigma^2)$. *Let* $\theta_1 = \mu$ *and* $\theta_2 = \sigma^2$. *By the method of moments, we have*

$$\bar{X} = \mu,$$
$$\frac{1}{n}\sum X_i^2 = \mu^2 + \sigma^2.$$

*Solving the systems yields the estimators*

$$\tilde{\mu} = \bar{X} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2.$$

**Example 7.2.2** *(***Binomial method of moments***) Let* $X_1, \ldots, X_k$ *be iid Binomial*$(k, p)$*, that is,*

$$P(X_i = x|k, p) = \binom{k}{x} p^x (1-p)^{k-x}, \quad x = 0, 1, \ldots, k.$$

*Here we assume that both* $k$ *and* $p$ *are unknown and we desire point estimators for both parameters. This model has been used to estimate crime rate for crimes that are known to have many unreported occurrences. By the method of moments, we have*

$$\bar{X} = kp$$
$$\frac{1}{n}\sum X_i^2 = kp(1-p) + k^2p^2$$

*Solving for* $k$ *and* $p$ *yields the estimators*

$$\tilde{k} = \frac{\bar{X}^2}{\bar{X} - (1/n)\sum(X_i - \bar{X})^2} \quad \text{and} \quad \tilde{P} = \frac{\bar{X}}{\tilde{k}}.$$

### 7.2.2 Maximum Likelihood Estimators

Recall that if $X_1, \ldots, X_n$ are an iid sample from a population with pdf or pmf $f(x|\theta_1, \ldots, \theta_k)$, the likelihood function is defined by

$$L(\theta|\boldsymbol{x}) = L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k).$$

**Definition 7.2.1** *For each sample point $\boldsymbol{x}$, let $\hat{\theta}(\boldsymbol{x})$ be a parameter value at which $L(\theta|\boldsymbol{x})$ attains its maximum as a function of $\theta$, with $\boldsymbol{x}$ held fixed. A maximum likelihood estimator (MLE) of the parameter $\theta$ based on a sample $\boldsymbol{X}$ is $\hat{\theta}(\boldsymbol{X})$.*

Intuitively, the MLE is a reasonable choice for an estimator. The MLE is the parameter point for which the observed sample is most likely. However, there are two inherent drawbacks associated with the maximum likelihood estimation. The first problem is that of actually finding the global maximum and verifying that, indeed, a global maximum has been found. The second problem is that of numerical sensitivity. That is, how sensitive is the estimate to small changes in the data?

**Example 7.2.3** *(***Normal likelihood***) Let $X_1, \ldots, X_n$ be iid $N(\theta, 1)$, then*

$$L(\theta|\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} \exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\},$$

*and*

$$\log L(\theta|\boldsymbol{x}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2.$$

*The equation $\frac{d}{d\theta}\log L(\theta|\boldsymbol{x}) = 0$ reduces to*

$$\sum_{i=1}^{n}(x_i - \theta) = 0,$$

*which has the solution $\hat{\theta} = \bar{x}$. To verify that $\bar{x}$ is a global maximum of the likelihood function, we can use the following arguments. First, $\bar{x}$ is the only zero of the first derivative. Second, verify that*

$$\frac{d^2}{d\theta^2}\log L(\theta|\boldsymbol{x})|_{\theta=\bar{x}} < 0.$$

*Thus, $\bar{x}$ is the only extreme point in the interior and it is a maximum. To finally verify that $\bar{x}$ is a global maximum, we must the boundaries, $\pm\infty$. By taking limits it is easy to establish that the likelihood is 0 at $\pm\infty$. So $\hat{\theta} = \bar{x}$ is a global maximum and hence $\bar{X}$ is the MLE.*

**Example 7.2.4** *(***Binomial MLE, unknown number of trials***) Let $X_1, \ldots, X_n$ be a random sample from a binomial$(k, p)$ population, where $p$ is known and $k$ is unknown. The likelihood function is*

$$L(k|\boldsymbol{x}, p) = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i}.$$

*Of course, the MLE of $k$ is an integer $k \geq \max_i x_i$ that satisfies $L(k|\boldsymbol{x}, p) \geq L(k-1|\boldsymbol{x}, p)$ and $L(k+1|\boldsymbol{x}, p) < L(k|\boldsymbol{x}, p)$. The ratio of likelihoods is*

$$\frac{L(k|\boldsymbol{x}, p)}{L(k-1|\boldsymbol{x}, p)} = \frac{[k(1-p)]^n}{\prod_{i=1}^{n}(k-x_i)}.$$

*Thus, the condition for a maximum is*

$$[k(1-p)]^n \geq \prod_{i=1}^{n}(k-x_i) \quad \textit{and} \quad [(k+1)(1-p)]^n < \prod_{i=1}^{n}(k+1-x_i).$$

*Dividing by $k^n$ and letting $z = 1/k$, we want to solve*

$$(1-p)^n = \prod_{i=1}^{n}(1-x_i z)$$

*for $0 \leq z \leq 1/max_i x_i$. The right-hand side is clearly a strictly decreasing function of $z$ for $z$ in this range. Thus there is a unique $z$ that solves the equation, and the solution can be found by numerically solving an $n$-th degree polynomial.*

**Theorem 7.2.1** *(***Invariance property of MLEs***) If $\hat{\theta}$ is the MLE of $\theta$, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta}$.*

Proof: We first define for $\tau(\theta)$ the it induced likelihood function $L^*$, given by

$$L^*(\eta|\boldsymbol{x}) = \sup_{\{\theta:\tau(\theta)=\eta\}} L(\theta|\boldsymbol{x}).$$

Let $\hat{\eta}$ denote the value that maximize $L^*(\eta|\boldsymbol{x})$. We must show that $L^*(\hat{\eta}|\boldsymbol{x}) = L^*(\tau(\hat{\theta})|\boldsymbol{x})$.

$$\begin{aligned} L^*(\hat{\eta}|\boldsymbol{x}) &= \sup_{\eta} L^*(\eta|\boldsymbol{x}) = \sup_{\eta} \sup_{\{\theta:\tau(\theta)=\eta\}} L(\theta|\boldsymbol{x}) \\ &= \sup_{\theta} L(\theta|\boldsymbol{x}) = L(\hat{\theta}|\boldsymbol{x}). \end{aligned}$$

The first equality of the second line follows because the iterated maximization is equal to the unconditional maximization over $\theta$, which is attained at $\hat{\theta}$. Furthermore

$$L(\hat{\theta}|\boldsymbol{x}) = \sup_{\{\theta:\tau(\theta)=\tau(\hat{\theta})\}} L(\theta|\boldsymbol{x}) = L^*(\tau(\hat{\theta})|\boldsymbol{x}).$$

Hence, $L^*(\hat{\eta}|\boldsymbol{x}) = L^*(\tau(\hat{\theta})|\boldsymbol{x})$ and $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$. □

Using this theorem, we see that the MLE of $\mu^2$, the square of a normal mean, is $\bar{X}^2$; the MLE of $\sqrt{p(1-p)}$, where $p$ is a binomial probability, is given by $\sqrt{\hat{p}(1-\hat{p})}$. Note that the invariance property of MLEs also holds in the multivariate case. There is nothing in the above proof that precludes $\theta$ from being a vector. If the MLE of $(\theta_1, \ldots, \theta_k)$ is $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$, and if $\tau(\theta_1, \ldots, \theta_k)$ is any function of the parameters, the MLE of $\tau(\theta_1, \ldots, \theta_k)$ is $\tau(\hat{\theta}_1, \ldots, \hat{\theta}_k)$.

**Example 7.2.5** *(***Bivariate Normal***) To use two-variate calculus to verify that a function $H(\theta_1, \theta_2)$ has a local maximum at $(\hat{\theta}_1, \hat{\theta}_2)$, it must be shown that the following three conditions hold.*

*a. The first-order partial derivatives are 0,*

$$\frac{\partial}{\partial\theta_1} H(\theta_1, \theta_2)|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} = 0 \quad \textit{and} \quad \frac{\partial}{\partial\theta_2} H(\theta_1, \theta_2)|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} = 0$$

*b. At least one second-order partial derivative is negative,*

$$\frac{\partial^2}{\partial\theta_1^2} H(\theta_1, \theta_2)|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} < 0 \quad \textit{and} \quad \frac{\partial^2}{\partial\theta_2^2} H(\theta_1, \theta_2)|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} < 0.$$

*c. The Jacobian of the second-order partial derivatives is positive,*

$$\left| \begin{matrix} \frac{\partial^2}{\partial\theta_1^2} H(\theta_1, \theta_2) & \frac{\partial^2}{\partial\theta_1\partial\theta_2} H(\theta_1, \theta_2) \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1} H(\theta_1, \theta_2) & \frac{\partial^2}{\partial\theta_2^2} H(\theta_1, \theta_2) \end{matrix} \right|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} > 0.$$

*For the normal log likelihood, we have*

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2} \log L(\theta, \sigma^2|\boldsymbol{x}) &= -\frac{n}{\sigma^2}, \\ \frac{\partial^2}{\partial(\sigma^2)^2} \log L(\theta, \sigma^2|\boldsymbol{x}) &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \theta)^2, \\ \frac{\partial^2}{\partial\theta\partial\sigma^2} \log L(\theta, \sigma^2|\boldsymbol{x}) &= -\frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \theta). \end{aligned}$$

*Properties (a) and (b) are easily seen to hold, and the Jacobian is*

$$\left|\begin{matrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^n(x_i-\theta) \\ -\frac{1}{\sigma^4}\sum_{i=1}^n(x_i-\theta) & \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^n(x_i-\theta)^2 \end{matrix}\right|_{\theta=\bar{x},\sigma^2=\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^6}\frac{n^2}{2} > 0.$$

*Thus, the calculus conditions are satisfied and we have indeed found a maximum.*

Recall that the likelihood function is a function of the parameter, $\theta$, with the data, $\boldsymbol{x}$, held constant. However, since the data are measured with error, we might ask how small changes in the data might affect the MLE.

**Example 7.2.6** *(***Continuation of Example 7.2.2***) Olkin, Petkau, and Zidek (1981) demonstrate that the MLSs of* $k$ *and* $p$ *in binomial sampling can be highly unstable. They illustrate their case with the following example. Five realizations of a binomial*$(k, p)$ *experiment are observed, where both* $k$ *and* $p$ *are unknown. The first data set is* $(16, 18, 22, 25, 27)$. *For this data set, the MLE of* $k$ *is* $\hat{k} = 99$. *If a second data set is* $(16, 18, 22, 25, 28)$, *then the MLE of* $k$ *is* $\hat{k} = 190$, *demonstrating a large amount of variability.*

Such occurrences happen when the likelihood function is very flat in the neighborhood of its maximum or when there is no finite maximum. When the MLEs can be found explicitly, as will often be the case in our examples, this is usually not a problem. However, in many instances, such as in the above example, the MLE cannot be solved for explicitly and must be found by numerical methods. When faced with such a problem, it is often wise to spend a little extra time investigating the stability of the solution.

### 7.2.3 Bayes Estimators

In the classical approach the parameter, $\theta$, is thought to be an unknown, but fixed, quantity. A random sample, $X_1, \ldots, X_n$, is drawn from a population indexed by $\theta$ and, based on the observed values in the sample, knowledge about the value of $\theta$ is obtained. In the Bayesian approach $\theta$ is considered to be a quantity whose variation can be described by a probability distribution (called the prior distribution). This is a subjective distribution, based on the experimenter's belief, and is formulated before the data are seen (hence the name prior distribution). A sample is then taken from a population indexed by $\theta$ and the prior distribution is updated with this sample information. The updated prior is called the posterior distribution. This updating is done with the use of Bayes' Rule, hence the name Bayesian statistics.

If we denote the prior distribution by $\pi(\theta)$ and the sampling distribution by $f(\boldsymbol{x}|theta)$, then the posterior distribution is

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})},$$

where $m(\boldsymbol{x})$ is the marginal distribution of $\boldsymbol{X}$, that is,

$$m(\boldsymbol{x}) = \int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta.$$

**Example 7.2.7** *(***Binomial Bayes estimation***) Let $X_1, \dots, X_n$ be iid Bernoulli(p). Then $Y = \sum X_i$ is binomial $(n, p)$. We assume the prior distribution on $p$ is beta$(\alpha, \beta)$. The joint distribution of $Y$ and $p$ is*

$$\begin{aligned} f(y,p) &= \Big[\binom{n}{y} p^y (1-p)^{n-y}\Big]\Big[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\Big] \\ &= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}. \end{aligned}$$

*The marginal of $Y$ is*

$$f(y) = \int_0^1 f(y,p)dp = \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}.$$

*The posterior is*

$$f(p|y) = \frac{f(y,p)}{f(y)} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1},$$

*which is beta$(y+\alpha, n-y+\beta)$. A natural estimate for $p$ is the mean of the posterior distribution, which would give us the Bayes estimator of $p$,*

$$\hat{p}_B = \frac{y+\alpha}{\alpha+\beta+n}.$$

**Definition 7.2.2** *let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x|\theta)$ (indexed by $\theta$). A class $\prod$ of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the class $\prod$ for all $f \in \mathcal{F}$, all priors in $\prod$, and all $x \in \mathcal{X}$.*

**Example 7.2.8** *(***Normal Bayes estimators***) Let $X \sim N(\theta, \sigma^2)$, and suppose that the prior on $\theta$ is $N(\mu, \tau^2)$. The posterior distribution of $\theta$ is also normal, with mean variance by*

$$E(\theta|x) = \frac{\tau^2}{\tau^2 + \sigma^2}x + \frac{\sigma^2}{\sigma^2 + \tau^2}\mu,$$

*and*

$$Var(\theta|x) = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}$$

*Note that the normal family is its own conjugate family.*

### 7.2.4 The EM Algorithm

The EM (Expectation-maximization) algorithm was originally introduced by Dempster et al (1977) to overcome the difficulties in maximizing likelihoods. It is based on the idea of replacing one difficult likelihood maximization with a sequence of easier maximizations whose limit is the answer ti the original problem. It is particularly suited to "missing data" problems.

Suppose that we observe $X_1, \dots, X_n$, iid from $g(x|\theta)$ and want to compute $\hat{\theta} = \arg\max L(\theta|\boldsymbol{x}) = \prod_{i=1}^n g(x_i|\theta)$. We augment the data with $\boldsymbol{z}$, where $(\boldsymbol{X}, \boldsymbol{Z}) \sim f(\boldsymbol{x}, \boldsymbol{z}|\theta)$ and note the identity

$$k(\boldsymbol{z}|\theta, \boldsymbol{x}) = \frac{f(\boldsymbol{x}, \boldsymbol{z}|\theta)}{g(\boldsymbol{x}|\theta)},$$

where $k(\boldsymbol{z}|\theta, \boldsymbol{x})$ is the conditional distribution of the missing data $Z$ given the observed data $\boldsymbol{x}$. This identity leads to the following relationship between complete-data likelihood $L(\theta|\boldsymbol{x}, \boldsymbol{z})$ and the observed data likelihood $L(\theta|\boldsymbol{x})$. For any value $\theta_0$,

$$\log L(\theta|\boldsymbol{x}) = E_{\theta_0}[\log L(\theta|\boldsymbol{x}, \boldsymbol{z})|\theta_0, \boldsymbol{x}] - E_{\theta_0}[\log k(\boldsymbol{z}|\theta, \boldsymbol{x})|\theta_0, \boldsymbol{x}], \tag{7.1}$$

where the expectation is with respect to $k(\boldsymbol{z}|\theta_0, \boldsymbol{x})$. To maximize $\log L(\theta|\boldsymbol{x})$, the later theorem shows that we only need to deal with the first term on the right side of (7.1), as the other term can be ignored.

Common EM notation is to denote the expected log-likelihood by

$$Q(\theta|\theta_0, \boldsymbol{x}) == E_{\theta_0}[\log L(\theta|\boldsymbol{x}, \boldsymbol{z})|\theta_0, \boldsymbol{x}].$$

We then maximize $Q(\theta|\theta_0, \boldsymbol{x})$, and if $\hat{\theta}_{(1)}$ is the value of $\theta$ maximizing $Q(\theta|\theta_0, \boldsymbol{x})$, the process can then be repeated with $\theta_0$ replaced by the updated value $\hat{\theta}_{(1)}$. In this manner, a sequence of estimators $\hat{\theta}_{(j)}, j = 1, 2, \ldots$, is obtained iteratively, i.e.,

$$Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j-1)}, \boldsymbol{x}) = \max_{\theta} Q(\theta|\hat{\theta}_{(j-1)}, \boldsymbol{x}). \tag{7.2}$$

**The EM algorithm**

1. (E-step) Compute

$$Q(\theta|\hat{\theta}_{(m)}, \boldsymbol{x}) = E_{\hat{\theta}_{(m)}}[\log L(\theta|\boldsymbol{x}, \boldsymbol{z})],$$

   where the expectation is with respect to $k(\boldsymbol{z}|\hat{\theta}_{(m)}, \boldsymbol{x})$.

2. (M-step) Maximize $Q(\theta|\hat{\theta}_{(m)}, \boldsymbol{x})$ in $\theta$ and take

$$\hat{\theta}_{(m+1)} = \arg\max_{\theta} Q(\theta|\hat{\theta}_{(m)}, \boldsymbol{x}).$$

   The iterations are conducted until a fixed point of $Q$ is obtained.

**Theorem 7.2.2** *(***Monotonic EM sequence***) The sequence $(\hat{\theta}_{(j)})$ defined by (7.2) satisfies*

$$L(\hat{\theta}_{(j+1)}|\boldsymbol{x}) \geq L(\hat{\theta}_{(j)}|\boldsymbol{x}),$$

*with equality holding if and only if $Q(\hat{\theta}_{(j+1)}|\hat{\theta}_{(j)},\boldsymbol{x}) = Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j)},\boldsymbol{x})$.*

Proof: On successive iterations, it follows from the definition of $\hat{\theta}_{(j+1)}$ that

$$Q(\hat{\theta}_{(j+1)}|\hat{\theta}_{(j)},\boldsymbol{x}) \geq Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j)},\boldsymbol{x}).$$

Thus, if we can show that

$$E_{\hat{\theta}_{(j)}}[\log k(\boldsymbol{Z}|\hat{\theta}_{(j+1)},\boldsymbol{x})|\hat{\theta}_{(j)},\boldsymbol{x}] \leq E_{\hat{\theta}_{(j)}}[\log k(\boldsymbol{Z}|\hat{\theta}_{(j)},\boldsymbol{x})|\hat{\theta}_{(j)},\boldsymbol{x}], \tag{7.3}$$

it will follow from (7.1) that the value of the likelihood is increased at each iteration. Since

$$E_{\hat{\theta}_{(j)}}\big[\log\big(\frac{k(\boldsymbol{Z}|\hat{\theta}_{(j+1)},\boldsymbol{x})}{k(\boldsymbol{Z}|\hat{\theta}_{(j)},\boldsymbol{x})}\big)|\hat{\theta}_{(j)},\boldsymbol{x}\big] \leq \log E_{\hat{\theta}_{(j)}}\big[\frac{k(\boldsymbol{Z}|\hat{\theta}_{(j+1)},\boldsymbol{x})}{k(\boldsymbol{Z}|\hat{\theta}_{(j)},\boldsymbol{x})}|\hat{\theta}_{(j)},\boldsymbol{x}\big] = 0,$$

where the inequality follows from Jensen's inequality. The theorem is therefore proved. □

**Example 7.2.9** *(***Censored data likelihood***) Suppose that we observe $X_1, \ldots, X_n$, iid, from $N(\theta, 1)$ and we ordered the observations so that $(x_1, \ldots, x_m)$ are uncensored and $(x_{m+1}, \ldots, x_n)$ are censored (and equal to $a$). The likelihood function is*

$$L(\theta|\boldsymbol{x}) = \frac{1}{(2\pi)^{m/2}} \exp\{-\frac{1}{2}\sum_{i=1}^{m}(x_i - \theta)^2\}[1 - \Phi(a - \theta)]^{n-m}$$

*and the complete-data log-likelihood is*

$$\log L(\theta|\boldsymbol{x}, \boldsymbol{z}) = -\frac{n}{2}\log\{2\pi\} - \frac{1}{2}\sum_{i=1}^{m}(x_i - \theta)^2 - \frac{1}{2}\sum_{i=m+1}^{n}(z_i - \theta)^2,$$

*and $k(\boldsymbol{z}|\theta, \boldsymbol{x})$ is a truncated normal distribution,*

$$k(\boldsymbol{z}|\theta, \boldsymbol{x}) = \frac{\exp\{-\frac{1}{2}(z - \theta)^2\}}{\sqrt{2\pi}[1 - \Phi(a - \theta)]}, \quad z > a.$$

*At the $j$-th step of the EM sequence, we have*

$$Q(\theta|\hat{\theta}_{(j)}, \boldsymbol{x}) = -\frac{n}{2}\log\{2\pi\} - \frac{1}{2}\sum_{i=1}^{m}(x_i - \theta)^2 - \frac{1}{2}\sum_{i=m+1}^{n}\int_a^\infty (z_i - \theta)^2 k(z|\hat{\theta}_{(j)}, \boldsymbol{x}$$

*and differentiating with respect to $\theta$ yields*

$$m(\bar{x} - \theta) + (n - m)(E[Z|\hat{\theta}_{(j)}, \boldsymbol{x}) - \theta) = 0$$

*that is,*

$$\hat{\theta}_{(j+1)} = \frac{m\bar{x} + (n - m)E(Z|\hat{\theta}_{(j)}, \boldsymbol{x})}{n},$$

*where*

$$E(Z|\hat{\theta}_{(j)}, \boldsymbol{x}) = \int_a^\infty z k(\boldsymbol{z}|\theta, \boldsymbol{x})dz = \hat{\theta}_{(j)} + \frac{\phi(a - \hat{\theta}_{(j)})}{1 - \Phi(a - \hat{\theta}_{(j+1)})}.$$

*Thus, the EM sequence is defined by*

$$\hat{\theta}_{(j+1)} = \frac{m}{n}\bar{x} + \frac{n-m}{n}\big[\hat{\theta}_{(j)} + \frac{\phi(a-\hat{\theta}_{(j)})}{1-\Phi(a-\hat{\theta}_{(j+1)})}\big],$$

*which converges to the MLE* $\hat{\theta}$*.*

## 7.3 Methods of Evaluating Estimators

### 7.3.1 Mean Squared Error

**Definition 7.3.1** *The bias of a point estimator $W$ of a parameter $\theta$ is the difference between the expected value of $W$ and $\theta$; that is, $Bias_\theta W = E_\theta W - \theta$. An estimator whose bias is identically (in $\theta$) equal to 0 is called unbiased and satisfies $E_\theta W = \theta$ for all $\theta$.*

**Definition 7.3.2** *The mean squared error (MSE) of an estimator $W$ of a parameter $\theta$ is the function of $\theta$ defined by $E_\theta(W-\theta)^2$.*

Note that

$$E_\theta(W - \theta 0^2 = \mathrm{Var}_\theta W + (E_\theta W - \theta)^2 = \mathrm{Var}_\theta W + (\mathrm{Bias}_\theta W)^2.$$

Thus, MSE incorporates two components, one measuring the variability of the estimator (precision) and the other measuring its bias (accuracy). It is sometimes the case that a trade-off occurs between variance and bias in such a way that a small increase in bias can be traded for a larger decrease in variance, resulting in an improvement in MSE.

**Example 7.3.1** *(***Normal MSE***) Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. The statistics $\bar{X}$ and $S^2$ are both unbiased estimators since*

$$E\bar{X} = \mu, \quad ES^2 = \sigma^2, \quad \textit{for all } \mu \textit{ and } \sigma^2.$$

*The MSEs of these estimators are given by*

$$E(\bar{X} - \mu)^2 = Var\bar{X} = \frac{\sigma^2}{n},$$

$$E(S^2 - \sigma^2)^2 = VarS^2 = \frac{2\sigma^4}{n-1}.$$

*An alternative estimator for $\sigma^2$ is the MLE $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$. It is straightforward to calculate*

$$E(\hat{\sigma}^2 - \sigma^2) = E(\hat{\sigma}^2 - \frac{n-1}{n}\sigma^2 + \frac{n-1}{n}\sigma^2 - \sigma^2)^2 = \frac{2n-1}{n^2}\sigma^4.$$

*We thus have*

$$E(\hat{\sigma}^2 - \sigma^2) = \frac{2n-1}{n^2}\sigma^4 < \frac{2}{n-1}\sigma^4 = E(S^2 - \sigma^2)^2,$$

*showing that $\hat{\sigma}^2$ has smaller MSE than $S^2$. Thus, by trading off variance for bias, the MSE is improved.*

### 7.3.2 Best Unbiased Estimators

**Definition 7.3.3** *An estimator* $W^*$ *is a best unbiased estimator of* $\tau(\theta)$ *if it satisfies* $E_\theta W^* = \tau(\theta)$ *for all* $\theta$ *and, for any other estimator* $W$ *with* $E_\theta W = \tau(\theta)$*, we have* $Var_\theta W^* \leq Var_\theta W$ *for all* $\theta$*.* $W^*$ *is also called a uniform minimum variance unbiased estimator (UMVUE) of* $\tau(\theta)$*.*

**Theorem 7.3.1** *(***Cramér-Rao Inequality***) Let $X_1, \ldots, X_n$ be a sample with pdf $f(X|\theta)$, and let $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$ be any estimator satisfying*

$$\frac{d}{d\theta}E_\theta W(\boldsymbol{X}) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}[W(\boldsymbol{x})f(\boldsymbol{x}|\theta)]d\boldsymbol{x}$$

*and*

$$Var_\theta W(\boldsymbol{X}) < \infty.$$

*Then*

$$Var_\theta(W(\boldsymbol{X})) \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\boldsymbol{X})\right)^2}{E_\theta\left((\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta))^2\right)}.$$

Proof: The proof of this theorem is an elegant application of the Cauchy-Schwarz Inequality, or, stated statistically, the fact that for any two random variables $X$ and $Y$,

$$[\mathrm{Cov}(X, Y)]^2 \leq (\mathrm{Var}X)(\mathrm{Var}Y).$$

First note that

$$\begin{aligned}\frac{d}{d\theta}E_\theta W(\boldsymbol{X}) &= \int_{\mathcal{X}} W(\boldsymbol{x})[\frac{\partial}{\partial\theta}f(\boldsymbol{x}|\theta)]dx = E_\theta\big[W(\boldsymbol{X})\frac{\frac{\partial}{\partial\theta}f(\boldsymbol{X}|\theta)}{f(\boldsymbol{X}|\theta)}\big] \\ &= E_\theta\big[W(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\big].\end{aligned}$$

Let $W(\boldsymbol{X}) = 1$, we have

$$E_\theta\big[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\big] = \frac{d}{d\theta}E_\theta[1] = 0.$$

Thus

$$\mathrm{Cov}_\theta\big(W(\boldsymbol{X}), \frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\big) = E_\theta\big[W(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\big] = \frac{d}{d\theta}E_\theta W(\boldsymbol{X})$$

and

$$\mathrm{Var}_\theta\big(\frac{\partial}{\partial\theta}f(\boldsymbol{X}|\theta)\big) = E_\theta\big((\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta))^2\big).$$

Using the Cauchy-Schwarz Inequality, we obtain

$$\mathrm{Var}_\theta(W(\boldsymbol{X})) \geq \frac{\big(\frac{d}{d\theta}E_\theta W(\boldsymbol{X})\big)^2}{E_\theta\big((\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta))^2\big)},$$

proving the theorem. □

If we add the assumption of independent samples, then the calculation of the lower bound is simplified. The expectation in the denominator becomes a univariate calculation, as the following corollary shows.

**Corollary 7.3.1** *(* **Cramér-Rao Inequality, iid case** *) If the assumptions of Theorem 7.3.1 are satisfied and, additionally, if $X_1, \ldots, X_n$ are iid with pdf $f(x|theta)$, then*

$$Var_\theta(W(\boldsymbol{X})) \geq \frac{\big(\frac{d}{d\theta}E_\theta W(\boldsymbol{X})\big)^2}{nE_\theta\big((\frac{\partial}{\partial\theta}\log f(X|\theta))^2\big)}.$$

Proof: Since $X_1, \ldots, X_n$ are independent,

$$E_\theta\big((\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta))^2\big) = E_\theta\big((\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(X_i|\theta))^2\big)$$

$$= \sum_{i=1}^n E_\theta\big((\frac{\partial}{\partial\theta}\log f(X_i|\theta))^2\big) + \sum_{i\neq j} E_\theta\big(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\frac{\partial}{\partial\theta}\log f(X_j|\theta)\big).$$

For $i \neq j$ we have

$$E_\theta\big(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\frac{\partial}{\partial\theta}\log f(X_j|\theta)\big) = E_\theta\big(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\big)E_\theta\big(\frac{\partial}{\partial\theta}\log f(X_j|\theta)$$

Therefore

$$E_\theta\big((\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta))^2\big) = nE_\theta\big((\frac{\partial}{\partial\theta}\log f(X|\theta))^2\big),$$

which establishes the corollary. □

The quantity $E_\theta\big((\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta))^2\big)$ is called the information number, or Fisher information of the sample. It reflects the fact that the information number gives a bound on the variance of the best unbiased estimator of $\theta$. As the information number gets bigger and we have more information about $\theta$, we have a smaller bound on the variance of the best unbiased estimator. The following lemma presents a computational result that aids in the application of this theorem.

**Lemma 7.3.1** *if $f(x|\theta)$ satisfies*

$$\frac{d}{d\theta}E_\theta\big(\frac{\partial}{\partial\theta}\log f(X|\theta)\big) = \int \frac{\partial}{\partial\theta}\big[\big(\frac{\partial}{\partial\theta}\log f(x|\theta)\big)f(x|\theta)\big]dx$$

*(true for an exponential family), then*

$$E_\theta\big((\frac{\partial}{\partial\theta}\log f(X|\theta))^2\big) = -E_\theta\big(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\big).$$

**Example 7.3.2** *(***Poisson unbiased estimation***) Let* $X_1, \dots, X_n$ *be iid Poisson($\lambda$), and $\tau(\lambda) = \lambda$. Since we have an exponential family, the above lemma gives us*

$$\begin{aligned}
&E_\lambda\big((\frac{\partial}{\partial\lambda}\log\prod_{i=1}^{n} f(X_i|\lambda))^2\big) = -nE_\lambda\big(\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda)\big) \\
&\quad = -nE_\lambda\big(\frac{\partial^2}{\partial\lambda^2}\log\big(\frac{e^{-\lambda}\lambda^X}{X!}\big)\big) \\
&= -nE_\lambda\big(\frac{\partial^2}{\partial\lambda^2}(-\lambda + X\log\lambda - \log X!)\big) \\
&= -nE_\lambda\big(-\frac{X}{\lambda^2}\big) = \frac{n}{\lambda}.
\end{aligned}$$

*Hence for any unbiased estimator, $W$, of $\lambda$, we must have*

$$Var_\lambda W \geq \frac{\lambda}{n}.$$

*Since Var($\bar{X}$) $= \lambda/n$, $\bar{X}$ is the UMVUE of $\lambda$.*

**Example 7.3.3** *(***Unbiased estimator for the scale uniform***)* *Let $X_1, \ldots, X_n$ be iid with pdf $f(x|\theta) = 1/\theta$, $0 < x < \theta$. Since $\frac{\partial}{\partial\theta} \log f(x|\theta) = -1/\theta$, we have*

$$E_\theta\big((\frac{\partial}{\partial\theta} \log f(X|theta))^2\big) = \frac{1}{\theta^2}.$$

*The C-Rao theorem would seem to indicate that if $W$ is any unbiased estimator of $\theta$,*

$$Var_\theta W \geq \frac{\theta^2}{n}.$$

*Consider the sufficient statistic $Y = \max(X_1, \ldots, X_n)$, the largest order statistic. The pdf of $Y$ is $f_Y(y|\theta) = ny^{n-1}/theta^n$, $0 < y < \theta$, so*

$$E_\theta Y = \int_0^\theta \frac{ny^n}{\theta^n} dy = \frac{n}{n+1}\theta,$$

*showing that $\frac{n+1}{n}Y$ is an unbiased estimator of $\theta$. We next calculate*

$$\begin{aligned} Var_\theta\big(\frac{n+1}{n}Y\big) &= \big(\frac{n+1}{n}\big)^2 Var_\theta Y \\ &= \big(\frac{n+1}{n}\big)^2 [E_\theta Y^2 - (\frac{n}{n+1}\theta)^2] = \frac{1}{n(n+2)}\theta^2, \end{aligned}$$

*which is uniformly smaller than $\theta^2/n$. This indicates that the C-Rao Theorem is not applicable to this pdf. According to Leibnitz's*

*Rule, we have*

$$\begin{aligned}\frac{d}{d\theta}\int_0^\theta h(x)f(x|\theta)dx = \frac{d}{d\theta}\int_0^\theta h(x)\frac{1}{\theta}dx &= \frac{h(\theta)}{\theta} + \int_0^\theta h(x)\frac{\partial}{\partial\theta}(\frac{1}{\theta})dx \\ &\neq \int_0^\theta h(x)\frac{\partial}{\partial\theta}f(x|\theta)dx,\end{aligned}$$

*unless* $h(\theta)/\theta = 0$ *for all* $\theta$*. Hence, the C-Rao Theorem does not apply. In general, if the range of the pdf depends on the parameter, the theorem will not be applicable.*

**Corollary 7.3.2** *(***Attainment***) Let $X_1, \dots, X_n$ be iid $f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the C-Rao Theorem. Let $L(\theta|\boldsymbol{x}) = \prod_{i=1}^n f(x_i|\theta)$ denote the likelihood function. If $W(X_1, \dots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(\boldsymbol{X})$ attains the C-Rao Lower Bound if and only if*

$$a(\theta)[W(\boldsymbol{x}) - \tau(\theta)] = \frac{\partial}{\partial\theta} \log L(\theta|\boldsymbol{x})$$

*for some function $a(\theta)$.*

Proof: The C-Rao inequality can be written as

$$\left[\mathrm{Cov}_\theta\left(W(\boldsymbol{X}), \frac{\partial}{\partial\theta} \log L(\theta|\boldsymbol{x})\right)\right]^2 \le \mathrm{Var}_\theta W(\boldsymbol{X}) \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta} \log L(\theta|\boldsymbol{x})\right),$$

and, recalling that $E_\theta W = \tau(\theta)$, $E_\theta \frac{\partial}{\partial\theta} \log L(\theta|\boldsymbol{x}) = 0$, we have equality if and only if $W(\boldsymbol{x}) - \tau(\theta)$ is proportional to $\frac{\partial}{\partial\theta} \log L(\theta|\boldsymbol{x})$. That is, $a(\theta)[W(\boldsymbol{x}) - \tau(\theta)] = \frac{\partial}{\partial\theta} \log L(\theta|\boldsymbol{x})$. □

**Example 7.3.4** *(***Normal variance bound***) Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, and consider estimation of $\sigma^2$, where $\mu$ is unknown.*

$$L(\mu, \sigma^2|\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^n (x_i-\mu)^2/\sigma^2},$$

*and hence*

$$\frac{\partial}{\partial\sigma^2} \log L(\mu, \sigma^2|\boldsymbol{x}) = \frac{n}{2\sigma^4}\Big(\sum_{i=1}^n \frac{(x_i-\mu)^2}{n} - \sigma^2\Big).$$

*Thus, taking $a(\sigma^2) = n/(2\sigma^4)$ shows that the best unbiased estimator of $\sigma^2$ is $\sum_{i=1}^n (x_i-\mu)^2/n$, which is calculable only if $\mu$ is known. If $\mu$ is known, the bound cannot be attained. The bound can be calculated as follows.*

$$-E\Big(\frac{\partial^2}{\partial(\sigma^2)^2} \log f(X|\mu, \sigma^2)|\mu, \sigma^2\Big) = -E\Big(\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}|\mu, \sigma^2\Big) = \frac{1}{2\sigma^4}.$$

*Thus, any unbiased estimator, $W$, of $\sigma^2$ must satisfy*

$$Var(W|\mu, \sigma^2) \geq \frac{2\sigma^4}{n}.$$

*In Example 7.3.1 we saw*

$$Var(S^2|\mu, \sigma^2) = \frac{2\sigma^4}{n-1},$$

*so $S^2$ does not attain the C-Rao lower bound.*

### 7.3.3 Sufficiency and Unbiasedness

**Theorem 7.3.2** *(***Rao-Blackwell***) Let $W$ be any unbiased estimator of $\tau(\theta)$, and let $T$ be a sufficient statistic for $\theta$. Define $\phi(T) = E(W|T)$. Then $E_\theta\phi(T) = \tau(\theta)$ and $Var_\theta\phi(T) \leq Var_\theta W$ for all $\theta$; that is, $\phi(T)$ is uniformly better unbiased estimator of $\tau(\theta)$.*

Proof: Since

$$\tau(\theta) = E_\theta W = E_\theta[E(W|T)] = E_\theta\phi(T)$$

so $\phi(T)$ is unbiased for $\tau(\theta)$. Also,

$$\begin{aligned}\text{Var}_\theta W &= \text{Var}_\theta[E(W|T)] + E_\theta[\text{Var}(W|T)] \\ &= \text{Var}_\theta\phi(T) + E_\theta[\text{Var}(W|T)] \geq \text{Var}_\theta\phi(T).\end{aligned}$$

Hence $\phi(T)$ is uniformly better than $W$, and it only remains to show that $\phi(T)$ is indeed an estimator. That is, we must show that $\phi(T) = E(W|T)$ is a function of only the sample and, in particular, is independent of $\theta$. But it follows from the definition of sufficiency, and the fact that $W$ is a function only of the sample, that the distribution of $W|T$ is independent of $\theta$. Hence $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$. □

**Example 7.3.5** *(***Conditioning on an insufficient statistic***)* *Let $X_1$, $X_2$ be iid $N(\theta, 1)$. The statistic $\bar{X} = \frac{1}{2}(X_1 + X_2)$ has*

$$E_\theta \bar{X} = \theta \quad and \quad Var_\theta \bar{X} = \frac{1}{2}.$$

*Consider conditioning on $X_1$, which is not sufficient. Let $\phi(X_1) = E_\theta(\bar{X}|X_1)$. It follows that $E_\theta \phi(X_1) = \theta$ and $Var_\theta \phi(X_1) \leq Var_\theta \bar{X}$, so $\phi(X_1)$ is better than $\bar{X}$. However,*

$$\phi(X_1) = E_\theta([\frac{1}{2}(X_1 + X_2)|X_1] = \frac{1}{2}X_1 + \frac{1][2}{\theta}.$$

*Hence, $\phi(X_1)$ is not an estimator.*

We now know that, in looking for a best unbiased estimator of $\tau(\theta)$, we need consider only estimators based on a sufficient statistic. The question now arises that if we have $E_\theta \phi = \tau(\theta)$ and $\phi$ is based on a sufficient statistic, that is, $E(\phi|T) = \phi$, how do we know that $\phi$ is best unbiased? Of course, if $\phi$ attains the C-Rao lower bound, then it is best unbiased, but if it does not, have we gained anything? For example, if $\phi^*$ is another unbiased estimator of $\tau(\theta)$, how does $E(\phi^*|T)$ compare to $\phi$? The next theorem answers this question in part by showing that a best unbiased estimator is unique.

**Theorem 7.3.3** *If $W$ is a best unbiased estimator of $\tau(\theta)$, then $W$ is unique.*

PROOF: Suppose $W'$ is another best unbiased estimator, and consider the estimator $W^* = \frac{1}{2}(W + W')$. Note that $E_\theta W = \tau(\theta)$ and

$$\begin{aligned}\mathrm{Var}_\theta W^* &= \mathrm{Var}_\theta\big(\frac{1}{2}W + \frac{1}{2}W'\big) = \frac{1}{4}\mathrm{Var}_\theta W + \frac{1}{4}\mathrm{Var}_\theta W' + \frac{1}{2}\mathrm{Cov}_\theta(W, W')\\ &\le= \frac{1}{4}\mathrm{Var}_\theta W + \frac{1}{4}\mathrm{Var}_\theta W' + \frac{1}{2}[(\mathrm{Var}_\theta W)(\mathrm{Var}_\theta W')]^{1/2} = \mathrm{Var}_\theta W.\end{aligned}$$

But if the above inequality is strict, then the best unbiasedness of $W$ is contradicted, so we must have equality for all $\theta$. Sine the equality is an application of Cauchy-Schwarz, we can have equality only if $W' = a(\theta)W + b(\theta)$ and $\mathrm{Cov}_\theta(W, W') = \mathrm{Var}_\theta(W)$. Now using properties of covariance, we have

$$\mathrm{Cov}_\theta(W, W') = \mathrm{Cov}_\theta[W, a(\theta)W + b(\theta)] = \mathrm{Cov}_\theta[W, a(\theta)W] = a(\theta)\mathrm{Var}_\theta W,$$

Hence $a(\theta) = 1$ and, since $W'$ is unbiased, we must have $b(\theta) = 0$. This shows that $W$ is unique. □

To see when an unbiased estimator is best unbiased, we might ask how could we improve upon a given unbiased estimator? Suppose that $W$ satisfies $E_\theta W = \tau(\theta)$, and we have another estimator, $U$, that satisfies $E_\theta U = 0$ for all $\theta$, that is, $U$ is an unbiased estimator of 0. The estimator

$$\phi_a = W + aU,$$

where $a$ is a constant, satisfies $E_\theta \phi_a = \tau(\theta)$ and hence is also an unbiased estimator of $\tau(\theta)$. The variance of $\phi_a$ is

$$\mathrm{Var}_\theta \phi_a = \mathrm{Var}_\theta(W + aU) = \mathrm{Var}_\theta W + 2a\mathrm{Cov}_\theta(W, U) + a^2\mathrm{Var}_\theta U.$$

Now, if for some $\theta = \theta_0$, $\mathrm{Cov}_\theta(W, U) < 0$, then we can make the sum of the last two terms be negative bu choosing a suitable value of $a$. Hence, $\phi_a$ will be better than $W$ at $\theta = \theta_0$ and $W$ cannot be best unbiased. If $\mathrm{Cov}_\theta(W, U) > 0$ and we let $\phi_a = W - aU$, we will have the similar arguments. Thus, the relationship of $W$ with unbiased estimators of 0 is crucial in evaluating whether $W$ is best unbiased.

**Theorem 7.3.4** *If $E_\theta W = \tau(\theta)$, $W$ is the best unbiased estimator of $\tau(\theta)$ if and only if $W$ is uncorrelated with any unbiased estimator of 0.*

Proof: If $W$ is best unbiased, the above argument shows that $W$ must satisfy $\text{Cov}_\theta(W, U) = 0$ for all $\theta$, for any $U$ with $E_\theta U = 0$. Hence, the necessity is established.

Suppose now that we have an unbiased estimator $W$ that is uncorrelated with all unbiased estimators of 0. Let $W'$ be any other estimator satisfying $E_\theta W' = E_\theta W = \tau(\theta)$. Write

$$W' = W + (W' - W),$$

and calculate

$$\begin{aligned}\text{Var}_\theta W' &= \text{Var}_\theta W + \text{Var}_\theta(W' - W) + 2\text{Cov}_\theta(W, W' - W)\\ &= \text{Var}_\theta W + \text{Var}_\theta(W' - W) \geq \text{Var}_\theta W.\end{aligned}$$

Since $W'$ is arbitrary, it follows that $W$ is the best unbiased estimator of $\tau(\theta)$. □

Characterizing the unbiased estimators of zero is not an easy task and requires conditions on the pdf (or pmf) with which we are working. However, if a family of pdfs or pmfs $f(x|\theta)$ has the property that there are no unbiased estimators of zero (other than zero itself), then our search would be ended, since any statistic $W$ satisfies $\mathrm{Cov}_\theta(W, 0) = 0$. Recall that the property of completeness, defined in chapter 6, guarantees such a situation.

**Example 7.3.6** *(***Continuation of Example 7.3.3***) For $X_1, \ldots, X_n$ iid uniform$(0, \theta)$, we saw that $\frac{n+1}{n}Y$ is an unbiased estimator of $\theta$, where $T = \max\{X_1, \ldots, X_n\}$. The conditions of the C-Rao Theorem are not satisfied, and we have not yet established whether this estimator is best unbiased. However, we have shown that $Y$ is a complete sufficient statistic. This means that the family of pdfs of $Y$ is complete, and there are no unbiased estimators of zero that are based on $Y$. Therefore, $\frac{n+1}{n}Y$ is uncorrelated with all unbiased estimators of zero (since the only one is zero itself) and thus $\frac{n+1}{n}Y$ is the best unbiased estimator of $\theta$.*

We sum up the relationship between completeness and best unbiasedness in the following theorem.

**Theorem 7.3.5** *Let $T$ be a complete sufficient statistic for a parameter $\theta$, and let $\phi(T)$ be any estimator based only on $T$. Then $\phi(T)$ is the unique best unbiased estimator of its expected value.*

The theory tells us that in the presence of completeness, if we can find any unbiased estimator, we can find the best unbiased estimator. If $T$ is a complete sufficient statistic for a parameter $\theta$ and $h(X_1, \ldots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $\phi(T) = E(h(X_1, \ldots, X_n)|T)$ is the best unbiased estimator of $\tau(\theta)$.

**Example 7.3.7** *(***Binomial best unbiased estimation***) Let $X_1, \ldots, X_n$ be iid binomial$(k, \theta)$. The problem is to estimate the probability of exactly one success from a binomial$(k, \theta)$, that is, estimate*

$$\tau(\theta) = P_\theta(X = 1) = k\theta(1-\theta)^{k-1}.$$

*Note $\sum_{i=1}^n X_i \sim Bi(kn, \theta)$ is a complete sufficient statistic, and*

$$h(X_1) = \begin{cases} 1 & \text{if } X_1 = 1 \\ 0 & \text{otherwise} \end{cases}$$

*is a simple unbiased estimator for $\tau(\theta)$. The theory tells that the estimator*

$$\phi(\sum_{i=1}^n X_i) = E(h(X_1)|\sum_{i=1}^n X_i)$$

*is the best unbiased estimator of* $\tau(\theta)$.

$$\begin{aligned}
\phi(t) &= E(h(X_1)|\sum_{i=1}^{n} X_i = t) = P(X_1 = 1|\sum_{i=1}^{n} X_i = t) \\
&= \frac{P_\theta(X_1 = 1, \sum_{i=1}^{n} X_i = t)}{P_\theta(\sum_{i=1}^{n} X_i = t)} = \frac{P_\theta(X_1 = 1, \sum_{i=2}^{n} X_i = t-1)}{P_\theta(\sum_{i=1}^{n} X_i = t)} \\
&= \frac{P_\theta(X_1 = 1)P_\theta(\sum_{i=1}^{n} X_i = t)}{P_\theta(\sum_{i=1}^{n} X_i = t)} \\
&= \frac{[k\theta(1-\theta)^{k-1}][\binom{k(n-1)}{t-1}\theta^{t-1}(1-\theta)^{k(n-1)-(t-1)}]}{\binom{kn}{t}\theta^t(1-\theta)^{kn-t}} \\
&= k\frac{\binom{k(n-1)}{t-1}}{\binom{kn}{t}}.
\end{aligned}$$

*Hence, the best unbiased estimator of* $\tau(\theta)$ *is*

$$\phi(\sum_{i=1}^{n} X_i) = k\frac{\binom{k(n-1)}{\sum X_i - 1}}{\binom{kn}{\sum X_i}}.$$

### 7.3.4 Loss Function Optimality

In a loss function or decision theoretic analysis, the quality of an estimator is quantified in its risk function; that is, for an estimator $\delta(\boldsymbol{x})$ of $\theta$, the risk function, a function of $\theta$ is

$$R(\theta, \delta) = E_\theta L(\theta, \delta(\boldsymbol{x})).$$

At a given $\theta$, the risk function is the average loss that will be incurred if the estimator $\delta(\boldsymbol{x})$ is used. Two commonly used loss functions are

$$\text{absolute err loss} \quad L(\theta, \delta(\boldsymbol{x})) = |\delta(\boldsymbol{x}) - \theta|,$$

and

$$\text{squared error loss} \quad L(\theta, \delta(\boldsymbol{x})) = (\delta(\boldsymbol{x}) - \theta)^2.$$

For the squared error loss, the risk function is the mean squared error. For two estimators, $\delta_1$ and $\delta_2$, if $R(\theta, \delta_1) < R(\theta, \delta_2)$ for all $\theta \in \Theta$, then $\delta_1$ is the preferred estimator because $\delta_1$ performs better for all $\theta$. More typically, the two risk functions will cross. Then the judgment as to which estimator is better may not be so clear-cut.

**Example 7.3.8** *(***Risk of normal variance***) Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population. Consider estimating $\sigma^2$ using squared error loss. We will consider estimators of the form $\delta_b(\boldsymbol{X}) = bs^2$, where $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ and $b$ can be any nonnegative constant. Recall that $ES^2 = \sigma^2$ and $VarS^2 = 2\sigma^4/(n-1)$. The risk function for $\delta_b$ is*

$$\begin{aligned} R(\sigma^2, \delta_b) = & \ Var(bS^2) + (E(bS^2) - \sigma^2)^2 = b^2 VarS^2 + (bES^2 - \sigma^2)^2 \\ & = [\frac{2b^2}{n-1} + (b-1)^2]\sigma^4. \end{aligned}$$

*The value of $b$ that gives the overall minimum value of $c_b = \frac{2b^2}{n-1} + (b-1)^2$ yields the best estimator $\delta_b$ in this class. Standard calculus methods show that $b = (n-1)/(n+1)$ is the minimizing value. Thus, at every value of $(\mu, \sigma^2)$, the estimator*

$$\tilde{S}^2 = \frac{n-1}{n+1}S^2 = \frac{1}{n+1}\sum_{i=1}^n (X_i - \bar{X})^2$$

*has the smallest risk among all estimators of the form $bS^2$.*

We can also use a Bayesian approach to the problem of loss function optimality, where we would have a prior distribution, $\pi(\theta)$. The Bayes risk is defined as

$$\int_\Theta R(\theta, \delta)\pi(\theta)d\theta.$$

Averaging the risk function gives us one number for assessing the performance of an estimator with respect to a given loss function. An estimator that yields the smallest value of the Bayes risk is called the *Bayes rule with respect to a prior* $\pi$ and is often denoted by $\delta^\pi$.

For $\boldsymbol{X} \sim f(\boldsymbol{x}|\theta)$ and $\theta \sim \pi$, the Bayes risk of a decision rule $\delta$ can be written as

$$\begin{aligned}&\int_\Theta R(\theta, \delta)\pi(\theta)d\theta = \int_\Theta \big( \int_\mathcal{X} L(\theta, \delta(\boldsymbol{x}))f(\boldsymbol{x}|\theta)d\boldsymbol{x}\big)\pi(\theta)d\theta \\ &= \int_\mathcal{X} \big[ \int_\Theta L(\theta, \delta(\boldsymbol{x}))\pi(\theta|\boldsymbol{x})d\theta\big] m(\boldsymbol{x})d\boldsymbol{x}.\end{aligned}$$

The quantity in square brackets is the expected value of the loss function with respect to the posterior distribution, call the *posterior expected loss*. It is a function only of $\boldsymbol{x}$, and not a function of $\theta$. Thus, for each $\boldsymbol{x}$, if we choose the action $\delta(\boldsymbol{x})$ to minimize the posterior expected loss, we will minimize the Bayes risk.

**Example 7.3.9** *(***Two Bayes rules***) Consider a point estimation problem for a real-valued parameter $\theta$.*

*a. For squared error loss, the posterior expected loss is*

$$\int_{\Theta} (\theta - \delta)^2 \pi(\theta|\boldsymbol{x}) d\theta = E((\theta - \delta)^2 | \boldsymbol{X} = \boldsymbol{x}).$$

*This expected value is minimized by $\delta^{\pi}(\boldsymbol{x}) = E(\theta|\boldsymbol{x})$. So the Bayes rule is the mean of the posterior distribution.*

*b. For absolute error loss, the posterior expected loss is $E(|\theta - \delta||\boldsymbol{X} = \boldsymbol{x})$. It is minimized by choosing $\delta^{\pi}(\boldsymbol{x}) =$ median of $\pi(\theta|\boldsymbol{x})$.*

**Example 7.3.10** *(***Binomial Bayes estimates***) Let $X_1, \dots, X_n$ be iid Bernoulli(p). Suppose the prior on p is beta($\alpha, \beta$). Then posterior of p is beta($y+\alpha, n-y+\beta$), where $y = \sum_{i=1}^n x_i$. Hence, $\delta^\pi(y) = E(p|y) = (y+\alpha)/(n+\alpha+\beta)$ is the Bayes estimator of p for squared error loss.*

*For absolute loss, we need to find the median of $\pi(p|\boldsymbol{x}) =$ beta($y+\alpha, n-y+\beta$) numerically, i.e., solving the equation*

$$\int_0^m \frac{\Gamma(\alpha+\beta+n)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} dp = \frac{1}{2}$$

*for m. The authors (Casella and Berger, 2002) have done this for $n = 10$ and $\alpha = \beta = 1$. The following table compares the Bayes estimator for absolute error loss, the Bayes estimator for squared error loss, and the MLE, $\hat{p} = y/n$. It is typical of Bayes estimators that they would not take on extreme values in the parameter space. No matter how large the sample size, the prior always has some influence on the estimator and tends to draw it away from the extreme values. In the table, you can see that even if $y = 0$ and n is large, the Bayes estimator is a positive number.*

Table 7.1: Three estimators for a binomial $p$.

| $n = 10$ | prior $\pi(p) \sim \text{uniform}(0, 1)$ | | |
|---|---|---|---|
| $y$ | MLE | Bayes absolute error | Bayes squared error |
| 0 | .0000 | .0611 | .0833 |
| 1 | .1000 | .1480 | .1667 |
| 2 | .2000 | .2358 | .2500 |
| 3 | .3000 | .3238 | .3333 |
| 4 | .4000 | .4119 | .4167 |
| 5 | .5000 | .5000 | .5000 |
| 6 | .6000 | .5881 | .5833 |
| 7 | .7000 | .6762 | .6667 |
| 8 | .8000 | .7642 | .7500 |
| 9 | .9000 | .8520 | .8333 |
| 10 | 1.0000 | .9389 | .9137 |