Chapter 7

Point Estimation

7.1 Introduction

Definition 7.1.1: A point estimator is any function $W(\mathbf{X}) = W(X_1, W_2, \ldots, W_n)$ of a sample; that is, any statistic is a point estimator.

Note:

1. Estimator: function of a sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$
2. Estimate: realized value of an estimator $\mathbf{x} = (x_1, x_2, \ldots, x_n)$

7.2 Methods of Finding Estimators

7.2.1 Method of Moments (MME)

Note:

1. Perhaps, the oldest method of finding point estimators, dating back at least to Kal Pearson in the late 1800s;
2. Idea is simple. In many cases, unfortunately, this method yields estimators that may be improved upon. However, it is a good place to start when other methods prove intractable.

Let $X_1, \ldots, X_n$ be iid from pmf or pdf $f(x|\theta_1, \ldots, \theta_k)$, we have:

1^{st} sample moment: $m_1 = \frac{1}{n}\sum_{i=1}^n X_i$

1^{st} population moment: $\mu_1' = \mathrm{E}X = \mu_1'(\theta_1, \ldots, \theta_k)$

$\ldots$

k^{th} sample moment: $m_k = \frac{1}{n}\sum_{i=1}^n X_i^k$

k^{th} population moment: $\mu_k' = \mathrm{E}X^k = \mu_k'(\theta_1, \ldots, \theta_k)$

To get MME: "Equate" the first k sample moments to the corresponding k population moments and solve the k equations for $(\theta_1, \ldots, \theta_k)$:

$$\begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_1 \\ \frac{1}{n}\sum_{i=1}^n X_i^2 \\ \vdots \\ \frac{1}{n}\sum_{i=1}^2 X_i^k \end{pmatrix} \doteq \begin{pmatrix} \mu_1' \\ \mu_2' \\ \vdots \\ \mu_k' \end{pmatrix} = \begin{pmatrix} \mu_1'(\theta_1, \ldots, \theta_k) \\ \mu_2'(\theta_1, \ldots, \theta_k) \\ \vdots \\ \mu_k'(\theta_1, \ldots, \theta_k) \end{pmatrix}$$

Example 7.2.1: (Normal Method of Moments)

Suppose $X_1, \ldots, X_n$ are iid $\mathrm{n}(\theta, \sigma^2)$. In this case, $k = 2$, $\theta_1 = \theta$ and $\theta_2 = \sigma^2$. We have

$$m_1 = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X} \qquad m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$$

$$\mu_1' = \theta \qquad \mu_2' = \theta^2 + \sigma^2$$

which implies

$$\bar{X} \doteq \theta, \quad \frac{1}{n}\sum_{i=1}^n X_i^2 \doteq \theta^2 + \sigma^2.$$

Then we can get the estimations of θ and σ^2 by solving the two equations above

$$\tilde{\theta} = \bar{X}, \quad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n}S^2.$$

Example 7.2.2: (Binomial Method of Moments)
Suppose $X_1, \ldots, X_n$ are iid binomial(m, p), where both m and p are unknown. In this case, $k = 2$, $\theta_1 = m$ and $\theta_2 = p$. We have

$$\mu_1' = mp \qquad \mu_2' = mp(1-p) + m^2p^2$$

which implies

$$\bar{X} \doteq mp, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \doteq mp(1-p) + m^2p^2.$$

Then we can get the estimations of m and p by solving the two equations above

$$\tilde{m} = \frac{\bar{X}^2}{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \tilde{p} = \frac{\bar{X}}{\tilde{m}}$$

Remark: Method of moments may give estimations that are outside the range of the parameters.

Example 7.2.3: (Satterthwaite Approximation)
The example of Satterthwaite approximation (1946) illustrates one of the most famous uses of the technique called "moment matching", which gives an approximation based on matching moments of distributions.

Let $Y_1, \ldots, Y_k$ be independent $\chi^2_{r_i}$ random variables, i.e., $Y_i \sim \chi^2_{r_i}$. The distribution of $\sum a_i Y_i$ is generally difficult to obtain, where a_i are known constants. However, it seems reasonable to assume that χ^2_ν will provide a good approximation for some value of ν. For given $a_1, \ldots, a_k$, Satterthwaite wanted to find a value of ν so that

$$\sum_{i=1}^{k} a_i Y_i \sim \frac{\chi^2_\nu}{\nu} \qquad \text{(approximately)}.$$

7.2.2 Maximum Likelihood (MLE)

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

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

Definition 7.2.4: For each sample point $\mathbf{x}$, let $\hat{\boldsymbol{\theta}}(\mathbf{x})$ be a parameter value at which $L(\boldsymbol{\theta}|\mathbf{x})$ attains its maximum as a function of $\boldsymbol{\theta}$, with $\mathbf{x}$ held fixed. A *maximum likelihood estimator* (**MLE**) of the parameter $\boldsymbol{\theta}$ based on a sample $\mathbf{X}$ is $\hat{\boldsymbol{\theta}}(\mathbf{X})$.

Remark:

1. Finding the MLE can be difficult in some cases
2. MLE may not be obtained through differentiation but in some cases differentiation will not work.
3. When differentiation will be used to find the MLE, it will be easier to deal with the natural log of the likelihood.
4. Maximization should be only over the range of the parameter.
5. If MLE cannot be obtained analytically, it can be obtained numerically.

Example 7.2.5: (Normal Likelihood)
Let $X_1, \ldots, X_n$ be iid n$(\mu, 1)$. Show that $\bar{X}$ is the MLE of μ using derivatives.

Step 1: Find the solutions from the following equation:

$$\frac{\mathrm{d}}{\mathrm{d}\mu} L(\mu|\mathbf{x}) = 0.$$

Step 2: Verify whether the solution achieves the global maximum,

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu^2} L(\mu|\mathbf{x})) < 0, \quad \text{in this case.}$$

Step 3: Check the boundaries ($\mu = \pm\infty$ in this case; it is not necessary in this case).

Example 7.2.6: Recall Theorem 5.2.4 (page 212) Part (a): If $x_1, \ldots, x_n$ are any numbers and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, then for any real numbers, we have

$$\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2$$

with equality if and only if $a = \bar{x}$. It implies that for any μ,

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \geq \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

with equality if and only if $\mu = \bar{x}$. So $\bar{X}$ is the MLE.

Example 7.2.7: (Bernoulli MLE)
Let $X_1, \ldots, X_n$ be iid Bernoulli(p) . Find the MLE of p where $0 \geq p \geq 1$. Note that we include the possibility that $p = 0$ or $p = 1$.

Solution. User the natural log of the likelihood function. □

Example 7.2.8: (Restricted Range MLE)
Let $X_1, \ldots, X_n$ be iid n(θ, 1), where $\theta \geq 0$. Find the MLE of p.

Solution. Without any restriction, $\bar{X}$ is the MLE. So when $\bar{x} \geq 0$, $\hat{\theta} = \bar{x}$. When $\bar{x} < 0$, $L(\theta|\mathbf{x})$ achieves its maximum at $\hat{\theta} = 0$ for $\theta \geq 0$, so $\hat{\theta} = 0$ in this situation, i.e.,

$$\hat{\theta} = \bar{X} \cdot I_{[0,\infty)}(\bar{X}) = \begin{cases} \bar{X} & \text{if } \bar{X} \geq 0; \\ 0 & \text{if } \bar{X} < 0. \end{cases}$$

□

Example 7.2.9: (Binomial MLE, Unknown Number of Trails)
Let $X_1, \ldots, X_n$ be iid binomial(k, p). Find the MLE of k where p is known and k is unknown. (Example where differentiation cannot be used to obtain the MLE.)

Solution. The likelihood function is

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

Then consider the ratio:

$$\frac{L(k|p, \mathbf{x})}{L(k-1|p, \mathbf{x})}.$$

□

Invariance Property of Maximum Likelihood Estimators

Definition: Consider a function $\tau(\theta)$ which may not necessarily be one-to-one function so that for a given value η, there may be more than one θ value such that $\tau(\theta) = \eta$. Then *induced likelihood function*, L^*, of $\tau(\theta)$ is given by:

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

The value $\hat{\eta}$ that maximizes $L^*(\eta|\mathbf{x})$ will be called the MLE of $\eta = \tau(\theta)$.

Theorem 7.2.10: (Invariance Property of MLEs)
If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Example: Let $X_1, \ldots, X_n$ be iid n$(\theta, 1)$, the MLE of θ^2 is $\bar{X}^2$.

Example: Let $X_1, \ldots, X_n$ be iid binomial(k, p) where k is known and p is unknown. Find the MLE of the variance and standard deviation of X_1.

Solution. First to verify that the MLE of p is

$$\hat{p} = \frac{1}{nk} \sum_{i=1}^{n} X_i = \frac{1}{k} \bar{X}.$$

Then

$$\eta = kp(1-p), \text{so } \hat{\eta} = k\hat{p}(1-\hat{p}).$$
$$\eta = \sqrt{kp(1-p)}, \text{so } \hat{\eta} = \sqrt{k\hat{p}(1-\hat{p})}.$$

□

Example: Let $X_1, \ldots, X_n$ be iid Poisson(λ). Find the MLE of $P(X = 0)$.

Solution. The MLE of λ is $\hat{\lambda} = \bar{X}$. Since $P(X = 0) = \exp(-\lambda)$, the MLE of $P(X = 0)$ is $\exp\big(-\bar{X}\big)$. □

Remark: Theorem 7.2.10 includes the multivariate case. 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 parameter vector, then by the invariance property of the MLE, the MLE of $\tau(\theta_1, \ldots, \theta_k)$ is $\tau(\hat{\theta}_1, \ldots, \hat{\theta}_k)$.

Example 7.2.11: (Normal MLEs, μ and σ^2 Unknown)
Let $X_1, \ldots, X_n$ be iid n(μ, σ^2) where both μ and σ^2 are unknown. Then the MLE of μ is $\hat{\mu} = \bar{X}$ and the MLE of σ^2 is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2$.

Solution. Verify these estimators using (a) univariate calculus (Example 7.2.11) and (b) multivariate calculus (Example 7.2.12). □

Note:

1. MLE is susceptible to problems associated with numerical instability if the MLEs cannot be solved explicitly.
2. How sensitive is the MLE to measurement error in the data? (See Example 7.2.13)

7.2.3 Bayes Estimators

Bayesian Approach to Statistics

1. The parameter θ is a random quantity described a probability distribution known as the prior distribution.
2. A sample is then taken from the population indexed by θ.
3. The prior distribution is update with sample information to get what is known as the posterior distribution using Bayes' rule (Theorem 1.3.5 page 23). Let $\pi(\theta)$ denote the prior distribution of θ and let $f(\mathbf{x}|\theta)$ denote the sampling distribution. The posterior distribution of θ given the sample $\mathbf{x}$ is given by

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

where $m(\mathbf{x})$ is the marginal distribution of $\mathbf{x}$, i.e.,

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

4. The posterior distribution is then used to make statements about θ, which is still considered a random quantity. For instance, the mean of the posterior distribution can be used as a point estimate of θ.

Example 7.2.14: (Binomial Bayes Estimation)
Let $X_1, \ldots, X_n$ be iid Bernoulli(p), where p is unknown. Then $Y = \sum_{i=1}^{n} X_i$ is binomial(n, p). We assume the prior distribution of p is beta(α, β). Then posterior distribution of p given $Y = y$, $\pi(p|y)$, is beta($y + \alpha, n - y + \beta$). Naturally, the Bayes estimator of p is constructed by the mean of the posterior distribution, i.e.,

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

Note that the mean of the prior distribution is $\dfrac{\alpha}{\alpha + \beta}$ and $\hat{p}_B$ can be written as

$$\hat{p}_B = \left(\frac{n}{\alpha + \beta + n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)\left(\frac{\alpha}{\alpha + \beta}\right)$$

Thus $\hat{p}_B$ is a linear combination of the sample mean and the prior mean, with the weights determined by α, β and n.

Remark: Both the prior and posterior distributions are beta distribution.

Definition 7.2.15: Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x|\theta)$ (indexed by θ). A class Π of prior distributions is a *conjugate family* for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π, and all $x \in \mathcal{X}$.

Example 7.2.16: (Normal Bayes Estimation) Let $X \sim \text{n}(\theta, \sigma^2)$, where σ^2 is known. We assume the prior distribution on θ is $\text{n}(\mu, \tau^2)$. The posterior distribution of θ given $X = x$, is also normal with mean and variance:

$$\text{E}(\theta|x) = \frac{\tau^2}{\tau^2 + \sigma^2}x + \frac{\sigma^2}{\tau^2 + \sigma^2}\mu \quad \text{and} \quad \text{Var}(\theta|x) = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

Remark:

1. The normal family is its own conjugate family.
2. If the prior information is vague (i.e., τ^2 is very large), then more weight is given to the sample data.
3. If the prior information is good (i.e., $\sigma^2 > \tau^2$), then more weight is given to the prior mean.

7.3 Methods of Evaluating Estimators

7.3.1 Mean Squared Error

Definition 7.3.1: The mean squared error (MSE) of an estimator W of a parameter θ is defined as:

$$\text{MSE} = \text{E}_\theta(W-\theta)^2 = \text{Var}_\theta W + (\text{Bias}_\theta W)^2,$$

where $\text{Bias}_\theta W = \text{E}_\theta W - \theta$.

Definition 7.3.2: The bias of a point estimator W of a parameterθ is the difference between the expected value of W and θ. An estimator whose bias is identically (in θ) equal to 0 is called *unbiased* and satisfies $E_\theta W = \theta$ for all θ.

If W is unbiased then,

$$\text{MSE} = \text{E}_\theta(W-\theta)^2 = \text{Var}_\theta W.$$

Example 7.3.3: (Normal MSE) Let $X_1, \ldots, X_n$ be iid $\text{n}(\mu, \sigma^2)$. We know that $\bar{X}$ and S^2 are unbiased estimators of μ and σ^2 respectively,

$$\text{E}\bar{X} = \mu \quad \text{and} \quad \text{E}S^2 = \sigma^2, \quad \text{for all } \mu \text{ and } \sigma^2,$$

which is true even without normality see Theorem 5.2.6. Thus,

$$\text{MSE}(\bar{X}) = \text{E}(\bar{X}-\mu)^2 = \frac{\sigma^2}{n},$$
$$\text{MSE}(S^2) = \text{E}(S^2-\sigma^2)^2 = \text{Var}S^2 = \frac{2\sigma^4}{n-1}.$$

Recall that $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, a chi-squared distribution with $n-1$ degrees of freedom. Since χ^2_p is gamma$(p/2, 2)$ and the variance of gamma$(\alpha, \beta) = \alpha\beta^2$, we have $\text{Var}\left(\dfrac{(n-1)S^2}{\sigma^2}\right) = 2(n-1)$.

Example 7.3.4: Let $X_1, \ldots, X_n$ be iid n(μ, σ^2) . Recall that both the MLE (Maximum Likelihood Estimator) and MME (Method of Moments Estimator) of σ^2 are

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}S^2.$$

Note that

$$\mathrm{E}(\hat{\sigma}^2) = \mathrm{E}\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}\sigma^2 \quad \text{and}$$
$$\mathrm{Var}(\hat{\sigma}^2) = \mathrm{Var}\left(\frac{n-1}{n}S^2\right) = \frac{(n-1)^2}{n^2}\mathrm{Var}(S^2) = \frac{2(n-1)}{n^2}\sigma^4,$$

so that

$$\mathrm{MSE}(\hat{\sigma}^2) = \mathrm{Var}(\hat{\sigma}^2) + \big(\mathrm{Bias}(\hat{\sigma}^2)\big)^2 = \frac{2(n-1)}{n^2}\sigma^4 + \frac{1}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4.$$

It can be verified that $\hat{\sigma}^2$ has smaller MSE than S^2.

Example 7.3.5: (MSE of Binomial Bayes Estimator)
Let $X_1, \ldots, X_n$ be iid Bernoulli(p). We compare the MLE of p and the Bayes estimator of p below:
(1) MLE: $\hat{p} = \bar{X}$ is unbiased estimator for p and

$$\mathrm{MSE}(\hat{p}) = \mathrm{E}_p(\hat{p} - p)^2 = \mathrm{Var}_p(\bar{X}) = \frac{p(1-p)}{n}.$$

(2) Bayes Estimator: $\hat{p}_B = \dfrac{Y + \alpha}{\alpha + \beta + n}$ is a biased estimator, where $Y = \sum_{i=1}^{n} X$, because $\mathrm{E}_p(\hat{p}_B) = \dfrac{np + \alpha}{\alpha + \beta + n} = p + \dfrac{\alpha - (\alpha + \beta)p}{\alpha + \beta + n}$. Then

$$\begin{aligned}\mathrm{MSE}(\hat{p}_B) &= \mathrm{Var}(\hat{p}_B) + \big(\mathrm{E}_p(\hat{p}_B - p)\big)^2 \\ &= \frac{np(1-p)}{(\alpha + \beta + n)^2} + \left(\frac{\alpha - \alpha p - \beta p}{\alpha + \beta + n}\right)^2\end{aligned}$$

If we choose $\alpha = \beta = \sqrt{n/4}$, it yields $\hat{p}_B = \dfrac{Y + \sqrt{n/4}}{n + \sqrt{n}}$ and $\mathrm{MSE}(\hat{p}_B) = \dfrac{n}{4(n + \sqrt{n})^2}$ as a constant for all p. In this situation, we can determine which of these two estimators is better in terms of the MSE.

7.3.2 Best Unbiased Estimator

Consider the class of estimators

$$\mathcal{C}_\tau = \big\{W : \mathrm{E}_\theta W = \tau(\theta)\big\}.$$

For any $W_1, W_2 \in \mathcal{C}_\tau$, we have $\mathrm{Bias}_\theta W_1 = \mathrm{Bias}_\theta W_2$, so

$$\begin{aligned}\mathrm{MSE}(W_1) - \mathrm{MSE}(W_2) &= \mathrm{E}_\theta(W_1 - \theta)^2 - \mathrm{E}_\theta(W_2 - \theta)^2 \\ &= \mathrm{Var}_\theta W_1 - \mathrm{Var}_\theta W_2\end{aligned}$$

and MSE comparisons, within the class $\mathcal{C}_\tau$, can be base on variance alone.

Definition 7.3.7: An estimator W^* is a *best unbiased estimator* of $\tau(\theta)$ if it satisfies $\mathrm{E}_\theta W^* = \tau(\theta)$ for all θ and, for any other estimator W with $\mathrm{E}_\theta W = \tau(\theta)$, we have $\mathrm{Var}_\theta W^* \leq \mathrm{Var}_\theta W$ for all θ. W^* is also called a *uniform minimum variance unbiased estimator* (**UMVUE**) of $\tau(\theta)$.

Remark:

1. UMVUE may not necessarily exist.
2. If UMVUE exists, it is unique (from Theorem 7.3.19).

Example 7.3.8: (Poisson Unbiased Estimation)
Let $X_1, \ldots, X_n$ be iid Poisson(λ). Note that $\mathrm{E}_\lambda(\bar{X}) = \lambda$ and $\mathrm{E}_\lambda(S^2) = \lambda$ for all λ. Thus, both $\bar{X}$ and S^2 are unbiased estimators of λ. Also, note that the class of estimators given by

$$W_a(\bar{X}, S^2) = a\bar{X} + (1-a)S^2$$

is a class of unbiased estimators for $0 \leq a \leq 1$.

To determine which estimator has the smallest MSE, we need to calculate $\mathrm{Var}_\lambda(\bar{X})$, $\mathrm{Var}_\lambda(S^2)$ and $\mathrm{Var}_\lambda(a\bar{X} + (1-a)S^2)$. The calculation can be lengthy.

Question: how can we find the best, i.e., smallest variance, of these unbiased estimators?

Theorem 7.3.9: (Cramér-Rao Inequality)
Let $X_1, \ldots, X_n$ be a sample with pdf $f(\mathbf{x}|\theta)$ and let $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ be any estimator satisfying

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta W(\mathbf{X}) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\Big[W(\mathbf{x})f(\mathbf{x}|\theta)\Big]d\mathbf{x}$$

$$\text{and} \quad \mathrm{Var}_\theta\big(W(\mathbf{X})\big) < \infty.$$

Then

$$\mathrm{Var}_\theta\big(W(\mathbf{X})\big) \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta W(\mathbf{X})\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right)}.$$

Corollary 7.3.10: (Cramér-Rao Inequality, iid case)
If the assumption of Theorem 7.3.9 are satisfied and, additionally, $X_1, \ldots, X_n$ are iid with pdf $f(x|\theta)$, then

$$\mathrm{Var}_\theta\big(W(\mathbf{X})\big) \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta W(\mathbf{X})\right)^2}{n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)}.$$

Remark:

1. The quantity $\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right)$ is called information number, or Fisher information, of the sample.
2. The information number gives a bound on the variance of the best unbiased estimator of θ.
3. As the information number increases, we have more information about θ, and we have a smaller bound.

The following lemma helps in the computation of the **CRLB** (Cramér-Rao Lower Bounds).

Lemma 7.3.11: If $f(x|\theta)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right) = \int \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right) f(x|\theta)\right] dx$$

(true for an exponential family), then

$$\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right) = -\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right).$$

Example 7.3.12: (Conclusion of Example 7.3.8)
Recall the Poisson problem. We will show that $\bar{X}$ is the UMVUE of λ.

Note: Key assumption of the Cramér-Rao Theorem is that one can differentiate under the integral sign. Below is an example where this assumption is not satisfied.

Example 7.3.13: (Unbiased Estimator for the Scale Uniform)
Let $X_1, \ldots, X_n$ be iid with pdf $f(x|\theta) = 1/\theta, 0 < x < \theta$.

Note: Cramér-Rao Lower Bound (CRLB) is not guaranteed to be sharp, i.e., there is no guarantee that the CRLB can be attained.

Example 7.3.14: (Normal Variance Bound)
Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$. We have

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

Hence S^2 does not attain the Cramér-Rao Lower Bound.

Question: How do we know if there exists an unbiased estimator that achieves the CRLB?

Corollary 7.3.15: (Attainment) Let $X_1, \ldots, X_n$ be iid with pdf $f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the Cramér-Rao Theorem. Let $L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta)$ denote the likelihood function. If $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramér-Rao Lower Bound if and only if

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

for some function $a(\theta)$.

Example 7.3.16: (Continuation of Example 7.3.14)

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

so that

$$\frac{\partial^2}{\partial\theta^2} \log L(\mu, \sigma^2|\mathbf{x}) = \frac{n}{2\sigma^4}\left(\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} - \sigma^2\right)$$

If μ is known, CRLB can be achieved and the UMVUE is $W(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$. Otherwise, no unbiased estimator of σ^2 will achieve the CRLB.

Question:

1. What can we do to find the "best" estimator if $f(x|\theta)$ does not satisfy the assumption of the Cramér-Rao Theorem?
2. What if the CRLB is not attainable, how do we know whether our estimator is the "best"?

7.3.3 Sufficiency and Unbiasedness

Recall two important results:

$$\mathrm{E}(X) = \mathrm{E}\left[\mathrm{E}\left(X|Y\right)\right]$$
$$\mathrm{Var}(X) = \mathrm{Var}\left[\mathrm{E}\left(X|Y\right)\right] + \mathrm{E}\left[\mathrm{Var}\left(X|Y\right)\right]$$

Theorem 7.3.17: (Rao-Blackwell)
Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient statistic for θ. Define $\phi(T) = \mathrm{E}(W|T)$. Then $\mathrm{E}_\theta\phi(T) = \tau(\theta)$ and $\mathrm{Var}_\theta\phi(T) \leq \mathrm{Var}_\theta W$; that is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.

Remark:

1. Conditioning any unbiased estimator on a sufficient statistic will result in an improved estimator.
2. To find the UMVUE, only need to consider functions of he sufficient statistic.
3. Sufficiency is needed so that the resulting quantity (estimator) after conditioning on the sufficient statistic will not depend on θ.

Example 7.3.18: (Conditioning on an Insufficient Statistic)
Let X_1 and X_2 be iid $\mathrm{n}(\theta, 1)$. Then $\bar{X}$ is an unbiased estimator (and a sufficient statistic) of θ. Suppose we condition $\bar{X}$ on X_1 which is not a sufficient statistic. Let $\phi(X_1) = \mathrm{E}_\theta(X_1) = \mathrm{E}(\bar{X}|X_1)$. Then $\phi(X_1)$ is unbiased for θ and has a smaller variance than $\bar{X}$ but it is not a valid estimator.

Theorem 7.3.19: If W is a best unbiased estimator of $\tau(\theta)$, then W is unique.

Let W be such that $\mathrm{E}_\theta(W) = \tau(\theta)$ and let U be such that $\mathrm{E}_\theta(U) = 0$ for all θ. Then

$$\phi_a = W + aU,$$

where a is a constant forming a class of unbiased estimators of θ with

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

Question: Which is a better estimator, W or ϕ_a?

Theorem 7.3.20: If $\mathrm{E}_\theta(W) = \tau(\theta)$, W is the best unbiased estimator of $\tau(\theta)$ if and only if W is uncorrelated with all unbiased estimators of 0.

Example 7.3.21: (Unbiased Estimators of Zero)
Let X be an observation from uniform$(\theta, \theta + 1)$ distribution. Then

$$\mathrm{E}X = \int_\theta^{\theta+1} x dx = \theta + \frac{1}{2} \quad \text{and} \quad \mathrm{Var}_\theta X = \frac{1}{12}.$$

Therefore, $X - \frac{1}{2}$ is an unbiased estimator of θ. We will show that $X - \frac{1}{2}$ is correlated with an unbiased estimator of 0, and hence cannot be a best unbiased estimator of θ.

Remark: If a family of pdfs $f(x|\theta)$ has the property that there are no unbiased estimators of 0 other than 0 itself, then our search would be ended since $\mathrm{Cov}(W, 0) = 0$. What is the property called?

Example 7.3.22: (Continuation of Example 7.3.13)
Le $X_1, \ldots, X_n$ be iid uniform$(0, \theta)$. Then $\frac{n+1}{n}Y$ where $Y = X_{(n)}$ is an unbiased estimator of θ.

Solution.

1. Conditions of Cramér-Rao Theorem were not satisfied.
2. By Rao-Blackwell Theorem, we only need to consider unbiased estimator of θ based on Y.
3. Y is a complete sufficient statistic, therefore Y is correlated with all unbiased estimators of 0 since this would just be 0 itself.
4. $\frac{n+1}{n}Y$ is the best unbiased estimator of θ.

□

Theorem 7.3.23: Let T be a complete sufficient statistic for a parameter θ, and let $\phi(T)$ be any estimator based only on T. Then $\phi(T)$ is the unique best unbiased estimator of its expected value.

Remark:

1. What is critical is the completeness of the family of distributions of the sufficient statistics not the completeness of the original family.
2. If T is complete sufficient statistic for a parameter θ and $h(\mathbf{X})$ is any unbiased estimator of $\tau(\theta)$, then $\phi(T) = \mathrm{E}(h(\mathbf{X})|T)$ is the unique best unbiased estimator of $\tau(\theta)$.

Example 7.3.24: (Binomial Best Unbiased Estimation)
Let $X_1, \ldots, X_n$ be iid binomial(k, θ). We want ot estimate $\tau(\theta) = P_\theta(X = 1) = k\theta(1-\theta)^{k-1}$.

Solution. Recall that $\sum_{i=1}^{n} X_i \sim$ binomial(kn, θ) is a complete sufficient statistic for θ. □

Question: How about an unbiased estimator for $\tau(\theta)$? Once we find an unbiased estimator, how do we get the best unbiased estimator?

7.3.4 Loss Function Optimality

Decision Theory:

1. After the data $\mathbf{X} = \mathbf{x}$ are observed, where $\mathbf{X} \sim f(\mathbf{x}|\theta)$ $\theta \in \Theta$, a decision regarding θ is made.
2. The set of allowable decisions is the *action space*, denoted by $\mathcal{A}$. (Often in point estimation problems, $\mathcal{A} = \Theta$.)
3. The objective function is defined based on *loss function*.

Definition: Loss function is a nonnegative function that generally increases as the distance between an action, a, and θ increases.

Note:

1. $L(\theta, \theta) = 0$ (What does this mean? - the loss is minimum if the action is correct)
2. If θ is real-valued, two commonly used loss function are:
   (a) **Absolute Error Loss** $L(\theta, a) = |a - \theta|$: more penalty on small discrepancies.
   (b) **Squared Error Loss** $L(\theta, a) = (a-\theta)^2$: more penalty on large discrepancies.
   (c) Other examples:
   $$L(\theta, a) = \begin{cases} (a-\theta)^2 & \text{if } a \leq \theta \\ 10(a-\theta)^2 & \text{if } a > \theta \end{cases},$$
   which penalizes overestimation more than underestimation.
   (d) **Relative Squared Error Loss** $L(\theta, a) = \dfrac{(a-\theta)^2}{|\theta| + 1}$, which penalizes errors in estimation more if θ is near 0 than if $|\theta|$ is large.

Definition: In decision theoretic analysis, the quality of an estimator, $\delta(\mathbf{X})$, is quantified by its *risk function* defined by

$$R(\theta, \delta) = \mathrm{E}_\theta L\big(\theta, \delta(\mathbf{X})\big),$$

i.e., at a given θ, the risk function is the average loss that will be incurred if the estimator $\delta(\mathbf{X})$ is used.

Remark:

1. MSE is an example of a risk function with respect to the squared error loss

$$R(\theta, \delta) = \mathrm{E}_\theta L(\theta, \delta(\mathbf{X})) = \mathrm{E}_\theta(\theta - \delta(\mathbf{X}))^2 = \mathrm{Var}_\theta \delta(\mathbf{X}) + (\mathrm{Bias}_\theta \delta(\mathbf{X}))^2$$

2. We want to find an estimator that has a smaller risk function for all θ compared to another estimator. However, most of the time the risk functions of two estimators cross.

Example 7.3.25: (Binomial Risk Functions)
Recall Example 7.3.5 comparing the Bayes estimator and the MLE of the Bernoulli parameter p:

$$\hat{p}_B = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}} \quad \text{and} \quad \hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i.$$

Example 7.3.26: (Risk of Normal Variance)
Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$. We want to estimate σ^2 considering the estimators of the form $\delta_b(\mathbf{X}) = bS^2$.

Solution. Recall that $\mathrm{E}S^2 = \sigma^2$ and $\mathrm{Var}(S^2) = \dfrac{2\sigma^4}{n-1}$. The risk function with respect to the squared error loss is

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

□

Remark:

1. The resulting risk function does not depend on μ.
2. This risk function can be minimized by setting $b = \dfrac{n-1}{n+1}$. Thus, for every value of (μ, σ^2), the estimator with the smallest risk among all estimators fo the form $\delta_b(\mathbf{X}) = bS^2$ is

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

See Figure 7.3.2, page 351.

Example 7.3.27: (Variance Estimation Using Stein's Loss)
Let $X_1, \ldots, X_n$ be iid from a population with positive finite variance, σ^2. We want to estimate σ^2.

Solution. Cosidering estimators of the form $\delta_b(\mathbf{X}) = bS^2$ and the loss function (attributed to Stein)

$$L(\sigma^2, a) = \frac{a}{\sigma^2} - 1 - \log\frac{a}{\sigma^2}.$$

In this case, the risk function is given by

$$R(\sigma^2, \delta_b) = \mathrm{E}\left(\frac{bS^2}{\sigma^2} - 1 - \log\frac{bS^2}{\sigma^2}\right) = b - 1 - \log b - \mathrm{E}\left(\log\frac{S^2}{\sigma^2}\right).$$

Note that $\mathrm{E}\left(\log\dfrac{S^2}{\sigma^2}\right)$ does not depend on b. Thus, $R(\sigma^2, \delta_b)$ is minimized in b, for all σ^2, by the value of b that minimizes $b - \log b$, that is, $b = 1$. Therefore the estimator of the form bS^2 that has the smallest risk for all values of σ^2 is

$$\delta_1(\mathbf{X}) = S^2$$

□

Bayesian Approach to Loss Function Optimality

Definition: Given a prior distribution $\pi(\theta)$, the Bayes risk is

$$\int_\Theta R(\theta,\delta)\pi(\theta)d\theta = \int_\Theta \left(\int_\mathcal{X} L(\theta,\delta(\mathbf{x}))f(\mathbf{x}|\theta)d\mathbf{x}\right)\pi(\theta)d\theta$$

The estimator that results in the smallest value of the Bayes risk is known as the *Bayes rule with respect to a prior* $\pi(\theta)$ and is often denoted δ^π. If we write $f(\mathbf{X}|\theta)\pi(\theta) = \pi(\theta|\mathbf{x})m(\mathbf{x})$, where $\pi(\theta|\mathbf{x})$ is the posterior distribution of θ and $m(\mathbf{x})$ is the marginal distribution of $\mathbf{X}$, the Bayes risk can be expressed as

$$\int_\Theta R(\theta,\delta)\pi(\theta)d\theta = \int_\mathcal{X} \left[\int_\Theta L(\theta,\delta(\mathbf{x}))\pi(\theta|\mathbf{x})d\theta\right] m(\mathbf{x})d\mathbf{x}$$

where the quantity in the square brackets is known as the *posterior expected loss*. The Action $\delta(\mathbf{X})$ that minimizes the posterior expected loss will also minimize the Bayes risk.

Example 7.3.28: (Two Bayes Rules)

Consider a point estimation problem for a real-valued parameter θ.

1. For squared error loss, the posterior expected loss is

   $$\int_\Theta (\theta-a)^2\pi(\theta|\mathbf{x})d\theta = \mathrm{E}\big((\theta-a)^2|\mathbf{X}=\mathbf{x}\big)$$

   where $\theta \sim \pi(\theta|\mathbf{x})$. The expected value is minimized by $\delta^\pi = \mathrm{E}(\theta|\mathbf{x})$. So the Bayes rule is the mean of the posterior distribution.

2. For absolute error loss, the posterior expected loss is

   $$\int_\Theta |\theta-a|\pi(\theta|\mathbf{x})d\theta = \mathrm{E}\big(|\theta-a||\mathbf{X}=\mathbf{x}\big)$$

   This is minimized by choosing $\delta^\pi(\mathbf{x}) =$ median of $\pi(\theta|\mathbf{x})$.

Example 7.3.29: (Normal Bayes Estimates)
Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\theta, \sigma^2)$ and let $\pi(\theta)$ be $\mathrm{n}(\mu, \tau^2)$, where σ^2, τ^2 and μ are known. From Example 7.2.16 and Exercise 7.22, the posterior distribution of θ given $\bar{X} = \bar{x}$ is normal with mean and variance:

$$\mathrm{E}(\theta|\bar{x}) = \frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{x} + \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\mu,$$
$$\mathrm{Var}(\theta|\bar{x}) = \frac{\tau^2\sigma^2/n}{\tau^2 + \sigma^2/n}.$$

1. For squared error loss,

$$\delta^{\pi}(\mathbf{x}) = \mathrm{E}(\theta|\bar{x}).$$

2. For absolute error loss,

$$\delta^{\pi}(\mathbf{x}) = \text{median of the posterior distribution } \pi(\theta|\mathbf{x}),$$

which is equal to $\mathrm{E}(\theta|\bar{x})$ because the posterior distribution is normal and it is symmetric about its mean.