AMA563

Chapter 3

Theory of Point Estimation

Contents

Recall that if we have a random sample $X_1, \ldots, X_n$ drawn independently from a distribution with density (or mass function) $f(x;\theta)$, a statistic, or an estimator $T = T(X_1, \ldots, X_n)$ is a function of the random sample. Note that although we always focus on the case that each random variable X_i in the sample takes only real number, it is also possible that these X_i's are all vectors, or all complex numbers, etc. Also, note that the parameter θ may also be complex, or a vector, although in this subject we usually work on the case $\theta \in \mathbb{R}$.

We have just learnt two important methods of point estimation: method of moments, and maximum likelihood estimation. Also, by definition there could well be a lot of possible estimators for a specific parameter. The following questions come up naturally.

1. Is there any way to measure or judge whether an estimator is good or not?
2. Is there any "best" estimator in some sense?

3. As the sample size goes larger and larger, would the statistic really "converge" to the parameter we are looking for?

4. We all know that a statistic is a function of the random sample, and therefore it is also a random number. Now, with a sufficiently large sample size, what is the distribution of the statistic?

We try to give answers to the questions in this chapter.

3.1 Bias and Relative Efficiency

Definition 3.1.1. *Suppose we have a random sample $X_1, \ldots, X_n$ and an estimator $T = T(X_1, \ldots, X_n)$ for the parameter θ.*

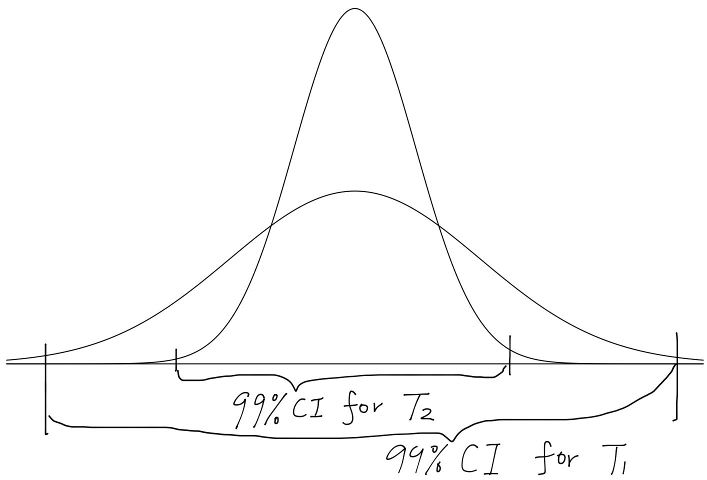
*a. The **bias** of T is defined by $\mathrm{bias}_\theta(T) = E(T - \theta) = E(T) - \theta$. If $\mathrm{bias}(T) = 0$ (or equivalently, $E(T) = \theta$), we say that the estimator T is **unbiased** for θ.*

b. The standard error of T is its standard deviation $\sigma_T = \sqrt{\mathrm{Var}(T)}$.

Definition 3.1.2. *Let T_1 and T_2 be two **unbiased** estimators for some parameter. The **relative efficiency** of T_1 to T_2 is defined by*

$$\mathrm{RE}(T1, T2) = \frac{\mathrm{Var}(T_2)}{\mathrm{Var}(T_1)}.$$

Remark 3.1.3. *For two unbiased estimators T_1 and T_2, if $\mathrm{RE}(T_1, T_2) < 1$, then we say that T_2 is more efficient than T_1. We use the following figure to illustrate the intuition. Recall that T_1 and T_2 are both random variables and now $\mathrm{Var}(T_2) < \mathrm{Var}(T_1)$. We plot the density function with the assumption that they both have the normal distribution.*



We see that the, say, the 99% confidence interval (CI) for T_2 is shorter, meaning that T_2 is often more close to the unknown parameter than T_1, and therefore T_2 is more precise. In another way to tell the story, we will soon learn that the length of CI will decrease as the sample size increases, and usually people maintain a reasonable sample size, not too large to save money, and not too small to make the CI satisfactorily short. With a shorter CI, people would need only a smaller sample size to achieve the same CI, therefore the estimator T_2 is more "efficient".

A useful tool to find the MLE is the indicator function, defined below.

Definition 3.1.4 (indicator function)**.** *Let A be a subset of $\mathbb{R}$. The indicator function $\chi_A(x)$ is defined on $\mathbb{R}$ by*

$$\chi_A(x) = \begin{cases} 1, & x \in A, \\ 0, & \textit{otherwise.} \end{cases}$$

Example 3.1.5. *For the density function $f(x|\theta) = \frac{1}{\theta}$ on $0 \leq x \leq \theta$ and zero elsewhere, with an unknown parameter $\theta > 0$ and a random sample $X_1, \ldots, X_n$, find the MME and MLE for θ. Find their bias and variance. Compare the relative efficiency between MME and the estimator $\widehat{\theta}_* = \frac{n+1}{n} X_{(n)}$.*

Solution. First, we find the MME. Since $\mu_1 = \theta/2$, we match μ_1 with m_1 to give $\widehat{\theta}_{\mathsf{MME}} = 2\overline{X}$.

Now we find the MLE. We write $x_{(n)} = \max\{x_1, \ldots, x_n\}$. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \begin{cases} \frac{1}{\theta^n}, & \theta \geq x_{(n)}, \\ 0, & \text{otherwise}, \end{cases}$$

which is maximized at $\theta = x_{(n)}$. Therefore the MLE is $\widehat{\theta}_{\mathsf{MLE}} = X_{(n)}$. Note that since it is trivial to obtain this global maximizer, there is no need to check, e.g., the second order derivative, etc..

Note, that one may also use the language of indicator function to build the likelihood function which in some cases would be very elegant. Since $f(x|\theta) = \frac{1}{\theta}\chi_{[0,\theta]}(x) = \frac{1}{\theta}\chi_{[x,\infty)}(\theta)$ for any $x \geq 0$, we have

$$L(\theta) = \frac{1}{\theta^n}\prod_{i=1}^{n}\chi_{[0,\theta]}(x_i) = \frac{1}{\theta^n}\prod_{i=1}^{n}\chi_{[x_i,\infty)}(\theta) = \frac{1}{\theta^n}\chi_{\cap_{1\leq i\leq n}[x_i,\infty)}(\theta) = \frac{1}{\theta^n}\chi_{[x_{(n)},\infty)}(\theta).$$

Now we find the bias and variance of these estimators.

$$\begin{aligned}
E(\widehat{\theta}_{\mathsf{MME}}) &= 2E(\overline{X}) = \theta, \\
\text{bias}(\widehat{\theta}_{\mathsf{MME}}) &= E(\widehat{\theta}_{\mathsf{MME}} - \theta) = 0, \\
\text{Var}(\widehat{\theta}_{\mathsf{MME}}) &= \text{Var}\left(2 \times \frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{4}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{\theta^2}{3n}. \\
f_{X_{(n)}}(t) &= nF(t)^{n-1}f(t) = n\left(\frac{t}{\theta}\right)^{n-1}\frac{1}{\theta}, \\
E(\widehat{\theta}_{\mathsf{MLE}}) &= E(X_{(n)}) = \frac{n}{\theta^n}\int_0^\theta t^n dt = \frac{n\theta^{n+1}}{(n+1)\theta^n} = \frac{n\theta}{n+1}, \\
\text{bias}(\widehat{\theta}_{\mathsf{MLE}}) &= E(\widehat{\theta}_{\mathsf{MLE}} - \theta) = -\frac{\theta}{n+1}, \\
E(\widehat{\theta}^2_{\mathsf{MLE}}) &= E(X^2_{(n)}) = \frac{n}{\theta^n}\int_0^\theta t^{n+1} dt = \frac{n\theta^2}{n+2}, \\
\text{Var}(\widehat{\theta}_{\mathsf{MLE}}) &= \text{Var}(X_{(n)}) = \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} = \frac{n\theta^2}{(n+1)(n+2)^2}
\end{aligned}$$

Therefore the MLE is biased and the MME is unbiased. Note that therefore we cannot compare their efficiency. On the other hand,

$$E(\widehat{\theta}_*) = \theta,$$

so it is unbiased. Also,

$$\begin{aligned}
\mathrm{Var}(\widehat{\theta}_*) &= \frac{(n+1)^2}{n^2}\mathrm{Var}(X_{(n)}) = \frac{(n+1)\theta^2}{n(n+2)^2},\\
\mathrm{RE}(\widehat{\theta}_{\mathsf{MME}}, \widehat{\theta}_*) &= \frac{\mathrm{Var}(\widehat{\theta}_*)}{\mathrm{Var}(\widehat{\theta}_{\mathsf{MME}})} = \frac{3(n+1)}{(n+2)^2} \leq \frac{n+1}{n+2} < 1,
\end{aligned}$$

therefore the estimator $\widehat{\theta}_*$ is more efficient than $\widehat{\theta}_{\mathsf{MME}}$.

Example 3.1.6 (for the tutorial)**.** *A random sample of size 2, X_1 and X_2, is drawn from the pdf*

$$f(x;\theta) = 2x\theta^2, \quad 0 \leq x \leq 1/\theta, \quad \theta > 0.$$

Find the constant C to make the statistic $C(X_1 + 2X_2)$ unbiased for $\frac{1}{\theta}$.

Example 3.1.7 (for the tutorial)**.** *Let* $X_1, \ldots, X_n$ *be a random sample from the distribution* $N(\mu, \sigma^2)$*. Find the bias of* $\overline{X}^2$ *as an estimator for* μ^2*.*

Example 3.1.8 (for the tutorial)**.** *Let* X_1, X_2, X_3 *be a random sample from a distribution with mean* μ *and variance* σ^2*. Which of the following is a more efficient estimator for* μ*?*

$$\widehat{\mu}_1 = \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3, \quad \widehat{\mu}_2 = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3.$$

Example 3.1.9 (for the tutorial)**.** *Suppose that X_1 and X_2 is a random sample of size 2 drawn from a distribution with mean μ and variance $\sigma^2 > 0$. Consider the estimator $\widehat{\mu} = c_1 X_1 + c_2 X_2$ for μ. For what values of c_1 and c_2 will $\widehat{\mu}$ be unbiased and have the smallest variance?*

Exercise 3.1

1. With a random sample of size n from $\mathsf{Exponential}(\lambda)$, find the bias of $\overline{X}^2$ as an estimator for $1/\lambda^2$.

2. With a random sample of size n from $\mathsf{Poisson}(\lambda)$, find the bias of $\overline{X}^2$ as an estimator for λ^2.

3.2 Rao-Cramér Lower Bound and Efficiency

We have been comparing estimators to one another on the basis of precision. The next question to ask is whether there is a fixed standard to which to compare individual estimators, that is, a best possible precision. This question is answered in the following theorem, the *Rao-Cramér inequality.* We omit the proof on the grounds that it is based on a not particularly instructive trick. A proof is available on page 331 of our textbook, and other elementary statistics textbooks.

Theorem 3.2.1 (Rao-Cramér Lower Bound)**.** *Let $X_1, \ldots, X_n$ be iid with common pdf (or pmf) $f(x;\theta)$ such that the set $\{x : f(x|\theta) > 0\}$ does not depend on θ. Let $Y = g(X_1, \ldots, X_n)$ be a statistic with mean $E(Y) = E\{g(X_1, \ldots, X_n)\} = k(\theta)$. Then*

$$\mathrm{Var}(Y) \geq \frac{\{k'(\theta)\}^2}{nI(\theta)}, \tag{3.1}$$

where

$$I(\theta) = E\left\{\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2\right\}.$$

*We call $I(\theta)$ the **Fisher information**.*

Corollary 3.2.2. *If $Y = g(X_1, \ldots, X_n)$ is an unbiased estimator of θ, (meaning $k(\theta) = \theta$ in the Theorem 3.2.1), then the Rao-Cramér inequality becomes*

$$\mathrm{Var}(Y) \geq \frac{1}{nI(\theta)}. \tag{3.2}$$

Definition 3.2.3 (Information matrix*)**.** *Suppose the pdf of a distribution is $f(x|\theta_1, \ldots, \theta_k)$, where the parameter vector $(\theta_1, \ldots, \theta_k)$ lies in an open subset of $\mathbb{R}^k$. Suppose the set $\{x : f(x|\theta_1, \ldots, \theta_k)\}$ does not depend on the parameter vector $(\theta_1, \ldots, \theta_k)$. The **Fisher information matrix** $I(\theta_1, \ldots, \theta_k)$ is defined as the $k \times k$ matrix with (i, j) element equal to*

$$I(\theta_1, \ldots, \theta_k)_{i,j} = \mathrm{Cov}\left[\frac{\partial}{\partial\theta_i}\log f(x|\theta_1, \ldots, \theta_k), \frac{\partial}{\partial\theta_j}\log f(x|\theta_1, \ldots, \theta_k)\right].$$

The quantity on the right side of the above equations (3.1) and (3.2) is referred to as the *Rao-Cramér lower bound.* The formula for the bound is complicated enough to merit explanation. First, X is a random variable with density f. In the denominator $f(X;\theta)$ is a function of X ; hence it is a random variable as well. So is $\log f(X;\theta)$, and in turn so is the partial derivative of this function with respect to θ. The square of this partial is yet another function of X, so that it makes sense to take its expectation, which is defined as the integral (or sum) of $[\partial \log f(x;\theta)/\partial\theta]^2$ times the density (or pmf) f over the state space of X. Normally, properties of expectation will shortcut the computation so that it is not necessary to evaluate an integral to find the Rao-Cramér (R-C) bound.

Before proceeding to the examples. let us derive an equivalent expression for the denominator of the R-C lower bound that is often easier to compute than the one in the above inequality. Specifically, let us show that

$$E\left[\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right] = -E\left[\left(\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right)\right] \tag{3.3}$$

In full form, we have

$$\begin{aligned} E\left[\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2\right] &= \int \left(\frac{\partial \ln f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta)dx \\ &= \int \left(\frac{\partial f(x;\theta)}{\partial\theta}/f(x;\theta)\right)^2 \cdot f(x;\theta)dx \end{aligned} \tag{3.4}$$

Also we can write

$$
\begin{aligned}
& E\left[\left(\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right)\right] \\
= & \int \left(\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2}\right) f(x;\theta)dx \\
= & \int \frac{\partial}{\partial\theta}\left(\frac{\partial f(x;\theta)}{\partial\theta}/f(x;\theta)\right)\cdot f(x;\theta)dx \\
= & \int \left[\left(f(x;\theta)\frac{\partial^2 f(x;\theta)}{\partial\theta^2} - \left(\frac{\partial f(x;\theta)}{\partial\theta}\right)^2\right)/f(x;\theta)^2\right]\cdot f(x;\theta)dx \\
= & \int \frac{\partial^2 f(x;\theta)}{\partial\theta^2}dx - \int\left(\frac{\partial f(x;\theta)}{\partial\theta}/f(x;\theta)\right)^2\cdot f(x;\theta)dx \\
= & \ 0 - \int\left(\frac{\partial f(x;\theta)}{\partial\theta}/f(x;\theta)\right)^2\cdot f(x;\theta)dx \qquad (3.5)
\end{aligned}
$$

The integral of $\partial^2 f(x;\theta)/\partial\theta^2$ in the next to last line vanishes because, interchanging derivative with integral, it is the second derivative of the integral of f , which is the constant 1. Combining this result with (3) yields the desired equality.

Definition 3.2.4. *Let $Y = g(X_1, X_2, ..., X_n)$ be an unbiased estimator of a parameter θ. The statistic Y is called an efficient estimator of θ if and only if the variance of Y attains the Rao-Cramér lower bound.*

Definition 3.2.5. *The efficiency of an unbiased estimator $Y = g(X_1, X_2, ..., X_n)$ of θ is the ratio*

$$\frac{\textit{R-C lower bound}}{\mathrm{Var}(Y)}$$

Example 3.2.6. *Show that the sample mean $\overline{X}$ of a random sample of size n from the $N(\mu;\sigma^2)$ distribution has the smallest possible standard error among all unbiased estimators of μ.*

Since by the theorem no estimator can have a variance smaller than the R-C lower bound, if we show that the variance of $\overline{X}$ achieves the bound, then no other estimator can do strictly better. It will then follow that the standard error of $\overline{X}$ is the smallest possible.

Solution. We have

$$\begin{aligned}
f(x|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \\
\Longrightarrow \log f(x|\mu) &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}, \\
\frac{\partial \log f(x|\mu)}{\partial \mu} &= -\frac{1}{2\sigma^2} \cdot 2(x-\mu) \times (-1) = \frac{x-\mu}{\sigma^2}.
\end{aligned}$$

To finish the solution we now have two options, corresponding the two forms of the R-C lower bound. Note that for both of the forms the above $f(x|\mu)$ should be changed to $f(X|\mu)$ since we need to take expectation with respect to the random variable X.

Method 1.

$$E\left[\left(\frac{\partial \log f(X|\mu)}{\partial \mu}\right)^2\right] = E\left[\left(\frac{X-\mu}{\sigma^2}\right)^2\right] = \frac{1}{\sigma^4} E\left[(X-\mu)^2\right] = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Method 2. We first compute the second order derivative.

$$\frac{\partial^2 \log f(x|\mu)}{\partial^2 \mu} = \frac{\partial}{\partial \mu}\left(\frac{x-\mu}{\sigma^2}\right) = -\frac{1}{\sigma^2},$$

then use the other form of the R-C bound.

$$E\left[\left(\frac{\partial \log f(X|\mu)}{\partial \mu}\right)^2\right] = -E\left[\left(\frac{\partial^2 \log f(X|\mu)}{\partial \mu^2}\right)\right] = -E\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2}.$$

Therefore we have

$$\frac{1}{nI(\theta)} = \frac{\sigma^2}{n}.$$

Since $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$, the estimator $\overline{X}$ achieves the R-C lower bound.

Example 3.2.7 (for the tutorial). *Let n independent replications of the experiment be performed. Construct random variables $X_1, X_2, ..., X_n$ such that*

$$X_i = \begin{cases} 1 & \textit{if the event occurs on the } i\textit{th replication,} \\ 0 & \textit{otherwise.} \end{cases}$$

Then $X_1, X_2, ..., X_n$ is a random sample from the Bernoulli(p) *distribution. The sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n = \widehat{p}$, is the proportion of times that the event occurred in the sample, an intuitively appealing estimator of p. Since $E[X_i] = p$, $\widehat{p}$ is unbiased for p . Let us try to verify that $\widehat{p}$ is a best estimator by comparing its variance to the R-C lower bound.*

Example 3.2.8 (for the tutorial)**.** *Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with $\alpha = 4$ and $\beta = \theta > 0$.*

a. *Find the moment estimator of θ.*
b. *Find the maximum likelihood estimator for θ.*
c. *Find the Fisher information $I(\theta)$.*
d. *Show that the MLE, $\widehat{\theta}$, of θ is an efficient estimator of θ.*

[left blank for Example 3.2.8]

Example 3.2.9. *Find the Fisher information $I(\theta)$ of the distribution $\theta x^{\theta-1}$, $0 < x < 1$, $\theta > 0$, zero elsewhere.*

Example 3.2.10 (for the tutorial). *Let* $X_1, \ldots, X_n$ *be a random sample from* $\mathsf{Poisson}(\lambda)$.

a. *Find the moment estimator of* λ.
b. *Find the maximum likelihood estimator for* λ.
c. *Find the Fisher information* $I(\lambda)$.
d. *Show that the MLE* $\widehat{\lambda}$, *is an efficient estimator of* λ.

Exercise 3.2

1. Find the Fisher information $I(p)$ of the distribution $\mathsf{Geometric}(p)$.
2. Find the Fisher information $I(\theta)$ of the distribution $\theta(1-x)^{\theta-1}$, $0 < x < 1$, $\theta > 0$, zero elsewhere.
3. Let $X_1, \dots, X_n$ be a random sample from $N(0, \sigma)^2$. Find the MLE $\widehat{\sigma^2}$ for σ^2. Show that $\widehat{\sigma^2}$ is unbiased. Find the variance of $\widehat{\sigma^2}$. Find the R-C lower bound for σ^2.
4. Let $X_1, \dots, X_n$ be a random sample taken from density
$$f(x|\theta) = \theta(x+1)^{-(1+\theta)}, \quad 0 < x < \infty, \quad \theta > 0, \quad \text{zero elsewhere.}$$
Find the C-R lower bound for the variance of all the unbiased estimators of $1/\theta$.
5. Find the Fisher information $I(\theta)$ of the distribution
$$f(x|\theta) = (\theta^2 + \theta)x^{\theta-1}(1-x), \quad x \in [0,1], \quad \theta > 0, \quad \text{zero elsewhere.}$$

3.3 Consistency and Limit Distributions

Although some bias may be acceptable in an estimator, we would like the bias to tend to 0 as the sample size, n, tends to ∞. In addition we would like the variance to tend to 0 as n tends to ∞. These requirements are related to the idea of consistency.

Definition 3.3.1 (Convergence in Probability)**.** *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. We say that X_n converges in probability to X if for all $\varepsilon > 0$*
$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0,$$
We denote this convergence by $X_n \xrightarrow{P} X$.

Theorem 3.3.2 (some properties)**.** *We have the following facts.*

(i) If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$.

(ii) If $X_n \xrightarrow{P} X$, then $aX_n \xrightarrow{P} aX$, where a is a constant.

(iii) If $X_n \xrightarrow{P} X$ and g is a continuous function, then $g(X_n) \xrightarrow{P} g(X)$.

(iv) If $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y$, *then* $X_nY_n \xrightarrow{P} XY$.

Theorem 3.3.3 (Weak Law of Large Numbers)**.** *Let* $X_1, \ldots, X_n$ *be a random sample from a distribution that has mean* μ *and positive variance* $\sigma^2 < \infty$. *Let* $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. *Then*

$$\overline{X}_n \xrightarrow{P} \mu\,.$$

Definition 3.3.4 (consistency)**.** *Let* $X_1, \ldots, X_n$ *be a random sample from some distribution family with parameter* θ *and let* $\widehat{\theta}$ *denote an estimator for* θ. *We say* $\widehat{\theta}$ *is (weakly*[1]*)* **consistent** *if*

$$\widehat{\theta} \xrightarrow{P} \theta.$$

An estimator which is not consistent will rarely be acceptable, except occasionally on the grounds of ease of computation, robustness, etc.; fortunately most "sensible-looking" estimators are consistent.

It may be difficult to prove consistency using the definition above, but it turns out that a sufficient (though not necessary) condition for consistency is that $\text{bias}(\widehat{\theta}) \to 0$ and $\text{Var}(\widehat{\theta}) \to 0$ as $n \to \infty$.

To prove this, we first define the mean square error (MSE) of $\widehat{\theta}$. As we shall see, $\text{MSE}(\widehat{\theta})$ is property of an estimator $\widehat{\theta}$ which takes account of both its bias and variance.

Definition 3.3.5.

$$\text{MSE}(\widehat{\theta}) = E\left[\left(\widehat{\theta} - \theta\right)^2\right].$$

It is easy to see that

$$\begin{aligned}
\text{MSE}\left(\widehat{\theta}\right) &= E\left[\left\{\left(\widehat{\theta} - \overline{\theta}\right) + \left(\overline{\theta} - \theta\right)\right\}^2\right] \quad (\text{where} \quad \overline{\theta} := E\left[\widehat{\theta}\right]) \\
&= E\left[\left(\widehat{\theta} - \overline{\theta}\right)^2\right] + E\left[\left(\overline{\theta} - \theta\right)^2\right] + 2\left(\overline{\theta} - \theta\right) E\left[\left(\widehat{\theta} - \overline{\theta}\right)\right] \\
&= \text{Var}\left(\widehat{\theta}\right) + \left[\text{bias}\left(\widehat{\theta}\right)\right]^2 + 0.
\end{aligned}$$

[1]Strong consistency corresponds to convergence with probability 1. We will not expand the topic here in this subject. Details are available in the chapter 5 of our textbook, and any statistics textbook.

If bias $\to 0$ and variance $\to 0$ as $n \to \infty$, then MSE $\to 0$ as $n \to \infty$.

Theorem 3.3.6. *If $MSE(\widehat{\theta}) \to 0$ as $n \to \infty$, then $\widehat{\theta}$ is a consistent estimator for θ.*

Example 3.3.7. *Let us estimate the mean of the normal distribution $N(\mu, \sigma^2)$. Consider whether the following two estimators are consistent: $\overline{X}$, and X_1.*

Solution. Since bias$(\overline{X}) = E[\overline{X} - \mu] = 0$, MSE$(\overline{X}) = \text{Var}(\overline{X}) = \frac{\sigma^2}{n} \to 0$ as the sample size n tends to ∞. So $\overline{X}$ is consistent. $P\left[|X_1 - \mu| > \varepsilon\right]$ is the same regardless of n, and certainly does not tend to 0 as $n \to \infty$ for any $\varepsilon > 0$. Hence X_1 is not a consistent estimator for μ.

Example 3.3.8 (for the tutorial)**.** *Let $X_1, \ldots, X_n$ be a random sample from* Bernoulli(p)*, where $p \in (0,1)$ is unknown. Let $Y = \sum_{i=1}^{n} X_i$.*

(a) Show that $\widehat{p} = Y/n$ is an unbiased estimator of p.
(b) Compute the variance of $\widehat{p}$. Is $\widehat{p}$ a consistent estimator for p?

Definition 3.3.9 (Convergence in Distribution)**.** *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Let F_{X_n} and F_X be, respectively, the cdfs of*

X_n *and* X. *Let* $C(F_X)$ *denote the set of all points where* F_X *is continuous. We say that* X_n *converges in distribution to* X *if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \qquad \textit{for all} \quad x \in C(F_X).$$

We denote this convergence by $X_n \xrightarrow{D} X$.

Theorem 3.3.10. *We have the following facts.*

(i) *If* $X_n \xrightarrow{P} X$, *then* $X_n \xrightarrow{D} X$.

(ii) *If* $X_n \xrightarrow{D} b$ *where* b *is a constant, then* $X_n \xrightarrow{P} b$.

(iii) *If* $X_n \xrightarrow{D} X$ *and* $Y_n \xrightarrow{P} 0$, *then* $X_n + Y_n \xrightarrow{D} X$.

(iv) *If* $X_n \xrightarrow{D} X$ *and* g *is a continuous function, then* $g(X_n) \xrightarrow{D} g(X)$.

(v) *Let* X_n,X, A_n, B_n *be random variables and let* a *and* b *be constants. If* $X_n \xrightarrow{D} X$, $A_n \xrightarrow{P} a$, *and* $B_n \xrightarrow{P} b$, *then* $A_n + B_n X_n \xrightarrow{D} a + bX$.

Theorem 3.3.11 (Δ-Method)**.** *Suppose that*

$$\sqrt{n}(X_n - \theta) \xrightarrow{D} N(0, \sigma^2)$$

and $g(x)$ *is differentiable at* θ *and* $g'(\theta) \neq 0$. *Then*

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2(g'(\theta)^2)).$$

Theorem 3.3.12 (Central limit theorem)**.** *Let* $X_1, \ldots, X_n$ *be a random sample from a distribution that has mean* μ *and positive variance* σ^2. *Then*

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0, 1).$$

Theorem 3.3.13 (Properties of MLE)**.** *Suppose* $\widehat{\theta}$ *is the MLE of* θ. *Under fairly general conditions, the MLE possesses the following good properties:*

- *Let* $\tau = \tau(\theta)$ *be a parameter of interest. Then* $\widehat{\tau} = \tau(\widehat{\theta})$ *is the MLE of* $\tau = \tau(\theta)$.

- *Consistency:* $\widehat{\theta} \xrightarrow{P} \theta_0$, *where* θ_0 *is the true parameter.*

- *Asymptotic normality:*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{D} N\left(0, \frac{1}{I(\theta_0)}\right)$$

and

$$\sqrt{n}(\tau(\widehat{\theta}) - \tau(\theta_0)) \xrightarrow{D} N\left(0, \frac{\tau'(\theta_0)^2}{I(\theta_0)}\right),$$

where $I(\theta)$ *is the Fisher information as defined in the next section.*

Example 3.3.14 (for the tutorial)**.** *Let* $X_1, \ldots, X_n$ *be a random sample from a* $N(0, \theta)$ *distribution. We want to estimate the standard deviation* $\sqrt{\theta}$. *Find the constant* c *so that* $Y = c\sum_{i=1}^{n} |X_i|$ *is an unbiased estimator of* $\sqrt{\theta}$ *and determine its efficiency.*

Example 3.3.15 (for the tutorial)**.** *If $X_1, \ldots, X_n$ is a random sample with pdf*

$$f(x;\theta) = \begin{cases} \frac{3\theta^3}{(x+\theta)^4}, & 0 < x < \infty, 0 < \theta < \infty \\ 0, & otherwise \end{cases}$$

Show that $Y = 2\overline{X}$ is an unbiased estimator of θ and determine its efficiency.

Example 3.3.16 (Example 3.2.8, continued)**.** *Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with $\alpha = 4$ and $\beta = \theta > 0$. Find the asymptotic distribution of $\sqrt{n}(\widehat{\theta} - \theta)$.*

3.4 Supplementary Notes: Variance, Covariance, and Linear Transform

In this section we systematically review the topics of variance, covariance, and linear transform. The students are suggested to review by themselves the topics of vector and matrix in linear algebra.

A *vector* is a list of numbers. The length of the list is called the *dimension* of the vector. Each element of a vector is called a *coordinate*. The totality of all the vectors of dimension n with real coordinates is denoted by $\mathbb{R}^n$. A *matrix* is a table of real numbers. The size is called the *dimension* of the matrix. The totality of all the matrices with m rows and n columns is denoted by $\mathbb{R}^{m\times n}$.

Example 3.4.1. *The following vectors* $\mathbf{a}$, $\mathbf{b}$, *and* $\mathbf{c}$ *are of dimensions 2, 3, and 4 respectively.*

$$\mathbf{a} = \begin{pmatrix} 1.22 \\ -6 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \pi \\ 0 \\ 100 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} \sqrt{2} \\ 0 \\ -2 \\ 150 \end{pmatrix}.$$

The following matrices A *and* B *are of dimensions* 3×2 *and* 4×4 *respectively.*

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Note that matrices of type B *is referred to as identity matrix. The transpose of a matrix and a vector is defined by*

$$\mathbf{b}^T = (\pi \quad 0 \quad 100), \qquad A^T = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{pmatrix}.$$

The (i, j) entry of a matrix A is usually written as A_{ij}. The i'th coordinate of a vector $\mathbf{c}$ is written as $\mathbf{c}_i$. The product of matrix and vector, and the product of matrix and matrix are summarized below. Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times r}$, and $\mathbf{c} \in \mathbb{R}^n$. One has

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}, \quad (A\mathbf{c})_j = \sum_{k=1}^{n} A_{jk}\mathbf{c}_k.$$

Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be a vector of random variables. Let $\mathbf{a} = (a_1, \ldots, a_n)^T$ be a vector of constants. Consider the statistic

$$T = \mathbf{a}^T\mathbf{X} = \sum_{i=1}^{n} a_i X_i.$$

Theorem 3.4.2. *Let $T = \sum_{i=1}^n a_i X_i$ and assume $E(X_i)$ exists for every i. We have*

$$E(T) = \sum_{i=1}^{n} a_i E(X_i).$$

Recall that $\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$, and therefore $\text{Cov}(X, X) = \text{Var}(X)$. Since $E(X + Y) = E(X) + E(Y)$, one has

$$\begin{aligned} \text{Var}(X + Y) &= E[(X + Y - E(X) - E(Y))^2] \\ &= E[(X - E(X))^2] + 2E[(X - E(X))(Y - E(Y))] + E[(Y - E(Y))^2] \\ &= \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y). \end{aligned}$$

If X and Y are independent, $\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E[X - E(X)]E[Y - E(Y)] = 0$. Therefore, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. In general, we have

Theorem 3.4.3. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)$ be two random vectors. Let $T = \sum_{i=1}^n a_i X_i$ and $W = \sum_{i=1}^m b_i Y_i$. If $E(X_i^2) < \infty$, and $E(Y_j^2) < \infty$, $i = 1, \ldots, n$, and $j = 1, \ldots, m$, then*

$$\text{Cov}(T, W) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, Y_j).$$

$$
\begin{aligned}
\mathrm{Cov}(T,W) &= E\left[\sum_{i=1}^{n}\sum_{j=1}^{m}(a_iX_i - a_iE(X_i))(b_jY_j - b_jE(Y_j))\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_ib_jE[(X_i - E(X_i))(Y_j - E(Y_j))] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_ib_j\mathrm{Cov}(X_i, Y_j).
\end{aligned}
$$

Corollary 3.4.4. *Let* $T = \sum_{i=1}^{n} a_iX_i$. *If* $E(X_i^2) < \infty$, $i = 1, \ldots, n$, *then*

$$
\mathrm{Var}(T) = \mathrm{Cov}(T,T) = \sum_{i=1}^{n} a_i^2\mathrm{Var}(X_i) + 2\sum_{i<j} a_ia_j\mathrm{Cov}(X_i, X_j).
$$

Corollary 3.4.5. *If* $X_1, \ldots, X_n$ *are* ***independent*** *random variables with finite variance, then*

$$
\mathrm{Var}(T) = \sum_{i=1}^{n} a_i^2\mathrm{Var}(X_i).
$$

3.5 Supplementary Notes: Order Statistics

In this section we systematically review the topic of order statistics.

Definition 3.5.1. *Let* $X_1, \ldots, X_n$ *be a random sample. We arrange them in ascending order of magnitude, and denote the sorted values as* $X_{(1)} \leq \ldots \leq X_{(n)}$. *That is, let* $X_{(1)}$ *be the smallest of these* X_i, *let* $X_{(2)}$ *be the next* X_i *in order of magnitude,* $\ldots$, *and* $X_{(n)}$ *the largest of* X_i. *The obtained statistics* $X_{(1)}, \ldots, X_{(n)}$ *are referred to as the order statistics.*

We will mainly focus on the continuous distribution in this section. For discrete distribution, the theory obtained here applies to $X_{(1)}$ and $X_{(n)}$, i.e., the smallest and largest order statistics, but not others. This is because of a technical difficulty, that is for continuous distribution, $X_i = X_j$ for any (i, j) with probability zero (think, why?), and this is in general not true for discrete distributions.

Theorem 3.5.2. *Let* $X_1, \ldots, X_n$ *be a random sample from a distribution with density* *function* $f(x)$ *and cumulative distribution function (cdf)* $F(x)$. *Let* $X_{(1)} = \min\{X_1, \ldots, X_n\}$ *and* $X_{(n)} = \max\{X_1, \ldots, X_n\}$. *Then the marginal density functions* $f_{X_{(1)}}$ *and* $f_{X_{(n)}}$, *for* $X_{(1)}$ *and* $X_{(n)}$ *respectively, are*

$$\begin{aligned} f_{X_{(n)}} &= nF(t)^{n-1} f(t), \\ f_{X_{(1)}} &= n[1 - F(t)]^{n-1} f(t). \end{aligned}$$

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

$$\begin{aligned} F_{X_{(n)}}(t) = P(X_{(n)} \leq t) &= P(\max_{1 \leq i \leq n}\{X_i\} \leq t) = P(X_1 \leq t, \cdots, X_n \leq t) \\ &= \prod_{i=1}^{n} P(X_i \leq t) = \left(\int_{-\infty}^{t} f(x)dx\right)^n, \\ f_{X_{(n)}}(t) &= \frac{d}{dt} F_{X_{(n)}}(t) = nF(t)^{n-1} f(t). \end{aligned}$$

The remaining part of the proof is left as an exercise.

In general, we have the following theorem, of which we skip the proof. The students are suggested to read the relative section in the textbook.

Theorem 3.5.3. *Let $X_1, \ldots, X_n$ be a random sample from a distribution with density function $f(x)$ and cumulative distribution function (cdf) $F(x)$. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the order statistics. The joint density of the order statistics is given by*

$$f_{\mathsf{ord}}(y_1, \ldots, y_n) = \begin{cases} n! f(y_1) f(y_2) \cdots f(y_n), & -\infty < y_1 < y_2 < \cdots < y_n < \infty, \\ 0, & \textit{otherwise.} \end{cases}$$

The marginal density of X_k is given by

$$f_{X_{(k)}}(t) = \frac{n!}{(k-1)!(n-k)!} [F(t)]^{k-1} [1 - F(t)]^{n-k} f(t).$$

3.6 Supplementary Notes: Simulation using R

Example 3.6.1. Recall: Example 2.3.17 The Weibull density with parameters λ and β is

$$f(t; \lambda, \beta) = \beta \lambda^\beta t^{\beta - 1} e^{-(\lambda t)^\beta}, \qquad t > 0$$

Assume that it is known that $\lambda = 2$. Let us try to explore the properties for the MLE of β via simulation.

- Generate observations from the true "unknown" model. For example, suppose the true value for β is $\beta_0 = 2$. Generate n i.i.d. samples from Weibull(2,1/2):

  X=rweibull(n, shape=2, scale = 1/2)

- Objective function (log-likelihood):

  $L(\beta) = n \log \beta + n\beta \log(2) + (\beta - 1) \sum_{i=1}^n \log X_i - \sum_{i=1}^n (2X_i)^\beta$.

  Score function:

  $0 = L'(\beta) = n\beta^{-1} + n \log(2) + \sum_{i=1}^n \log X_i - \sum_{i=1}^n (2X_i)^\beta \log(2X_i)$.

  Second order $L''(\beta) = -n\beta^{-2} - \sum_{i=1}^n (2X_i)^\beta \log^2(2X_i)$.

- Obtain MLE via Newton's method
- Repeat R times.

Output: R MLEs $\rightarrow$ Confidence Interval!

Take Home Questions: Other solvers?

Example 3.6.2. Following Example 1, assume that λ is also unknown.

- Generate observations from the true "unknown" model. For example, suppose the true value for β is $\beta_0 = 2$. Generate n i.i.d. samples from Weibull(2,1/2):

  X=rweibull(n, shape=2, scale = 1/2)

- Objective function (log-likelihood):

$$f(t; \lambda, \beta) = \beta \lambda^\beta t^{\beta-1} e^{-(\lambda t)^\beta}, \qquad t > 0$$

  $L(\beta) = n \log \beta + n\beta \log(\lambda) + (\beta - 1) \sum_{i=1}^n \log X_i - \sum_{i=1}^n (\lambda X_i)^\beta.$

  Score functions:

  $0 = \frac{\partial L(\beta,\lambda)}{\partial \beta} = n\beta^{-1} + n \log(\lambda) + \sum_{i=1}^n \log X_i - \sum_{i=1}^n (\lambda X_i)^\beta \log(\lambda X_i);$

  $0 = \frac{\partial L(\beta,\lambda)}{\partial \lambda} = n\beta\lambda^{-1} - \beta \sum_{i=1}^n (\lambda X_i)^{\beta-1} X_i.$

  Second order

  $\frac{\partial^2 L(\beta,\lambda)}{\partial \beta^2} = -n\beta^{-2} - \sum_{i=1}^n (\lambda X_i)^\beta \log^2(\lambda X_i);$

  $\frac{\partial^2 L(\beta,\lambda)}{\partial \lambda^2} = -n\beta\lambda^{-2} - \beta(\beta - 1) \sum_{i=1}^n (\lambda X_i)^{\beta-2} X_i^2;$

  $\frac{\partial^2 L(\beta,\lambda)}{\partial \beta \partial \lambda} = \frac{\partial^2 L(\beta,\lambda)}{\partial \lambda \partial \beta} = n\lambda^{-1} - \sum_{i=1}^n (\lambda X_i)^{\beta-1} X_i - \beta \sum_{i=1}^n (\lambda X_i)^{\beta-1} X_i \log(\lambda X_i);$

- Obtain MLE via Newton's method
- Repeat R times.

Take Home Questions: Initial values? Coordinate descent?