Chapter 3: Point Estimation

In this chapter we explore several methods to derive (point) estimators of a parameter and then discuss how to evaluate an estimator.

1 Introduction

Reading Assignment: Section 7.1 of the textbook.

Let $f(x \mid \theta)$ be the p.d.f. or p.m.f. where a sample $\mathbf{X} = (X_1, \ldots, X_n)$ is drawn. That is, $X_i \sim f(x \mid theta)$ and $X_1, \ldots, X_n$ are independent of each other.

The goal of point estimation is to estimate the value of θ based on the observed sample. In other words, the goal is to determine which value of θ is the one that generates this sample.

Definition 3.1.1 A statistic $W(\mathbf{X})$ is called a point estimator of θ if its value $W(\mathbf{x})$, after $\mathbf{X} = \mathbf{x}$ is observed, is used to estimate the value of θ. We call the random quantity $W(\mathbf{X})$ a point "estimator" and its observed value $W(\mathbf{x})$ a point "estimate" to distinguish the two. ■

There are several guiding principles to construct a point estimators and we will introduce three major approaches: the method of moment, the maximum likelihood method, and the Bayes method. Sometimes, we are also interested in estimating a function of a parameter, $\tau(\theta)$, so these approaches will be extended to $\tau(\theta)$ as well.

2 Method of Moments Estimators

Reading Assignment: Section 7.2.1 of the textbook and Handout 3.

The simplist approach to derive a point estimator is to apply the substitution principle which is the essence of the method of moment. To be more specific, we have seen that the sample mean from $X_1, \ldots, X_n$ could be used to estimate the population mean $\mu = E(X_i)$. By this token, if we let $Y_i = X_i^j$, then the sample mean based on $Y_1, \ldots, Y_n$ would be a reasonable estimate of $E(Y_i) = E(X_i^j) \equiv \mu_j$, which is the j-th moment of X_i. That is, the sample jth moment, defined as: $m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$, would be a reasonable estimate of the jth population moment μ_j.

What if the parameter θ is neither μ nor any of the μ_j? A general approach is to express θ as a function of the moments and then replace these moments with their sample estimator to get an estimate for θ. This leads to the method of moment described below.

Definition 3.2.1 (Method of Moments Estimator) Let $X_1, \ldots, X_n$ be a random sample from a population with parameter indexed by θ. Suppose that $\mu_1(\theta), \ldots, \mu_K(\theta)$ are the first K moments of the population we are sampling from. Thus,

$$\mu_j(\theta) = E(X_1^j \mid \theta).$$

Define the jth sample moment m_j as:

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j, \quad \text{for } j \geq 1. \tag{1}$$

To apply the method of moments to the problem of estimating $q(\theta)$, a function of θ, we need to be able to express q as a function g of the first K moments. Thus, suppose

$$q(\theta) = g(\mu_1(\theta), \ldots, \mu_K(\theta)).$$

The *method of moments* prescribes that we estimate $q(\theta)$ by $g(m_1, \ldots, m_K)$. ■

Example 3.2.1 Suppose that

$$q(\theta) = Var(X) = \sigma^2 = \mu_2(\theta) - \mu_1^2(\theta).$$

The moment estimator of σ^2 would be

$$\hat{\sigma}^2 = m_2 - m_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{2}$$

The last equality is a consequence of $2\overline{X}\sum_{i=1}^{n} X_i = 2n\overline{X}^2$. Note that this moment estimator resembles the sample variance but differs by a factor $\frac{n}{n-1}$.

Further notice that we have not used any property of a model so the above derivation works for any model, whether it is parametric or nonparametric. In this sense the method of moment approach in this example is a nonparametric method. ■

When the model is parametric, one should revert to the model structure when deriving a method of moment estimator. For instance, if the data $X_1, \ldots, X_n$ come from a $N(\mu, \sigma^2)$ population, then the method of moments estimates of μ and σ^2 are still $\overline{X}$ and $\hat{\sigma}^2$. (Why?) Thus, the normality assumption has not changed the method of moment estimator for $\hat{\sigma}^2$. Below we give another example to find the method of moment estimator.

Example 3.2.2 Suppose that $X_1, \ldots, X_n$ are indicators of a set of Bernoulli trials with probability of success θ. Since $\mu_1(\theta) = \theta$ the method of moments leads to the natural estimator $\overline{X}$, which is also the sample proportion of successes. To estimate the population variance $q(\theta) = \theta(1-\theta)$ we are led by the first moment to the estimate $\overline{X}(1-\overline{X})$. At a first glance it is not clear whether $\overline{X}(1-\overline{X})$ is equal to $\hat{\sigma}^2$ in (2). But, it turns out to be the same (Check this out!) even though we have derived it independently from a special formula that applies only to the Bernoulli family. ■

There are often several methods of moment estimates for the same $q(\theta)$. For example, if we are sampling from a Poisson population with parameter θ, then θ is both the population mean and the population variance. The method of moments can lead to either the sample mean or the sample variance. Other method of moments estimates for θ can also be constructed. The issue of which moment estimator is preferred will be addressed in a later topic.

What are the advantages of the method of moments?

1. They generally lead to procedures that are easy to compute and are therefore valuable as preliminary estimates.

2. If the sample size is large, these estimates are likely to be close to the value estimated (consistency). In fact it converges at the $\sqrt{n}$ rate, a gold standard for a parametric estimator to have.

These two properties are especially valuable when, for example, the MLE (maximum likelihood estimator) needs to be estimated numerically as in the Gamma distribution case to be illustrated in the next section (Example 3.3.5). The method of moments estimator could serve as an initial estimate for the MLE algorithm and it can be shown that just one iteration could already lead to an estimate which is asymptotically equivalent to the MLE in many scenarios (because the initial estimate is already $\sqrt{n}$ -consistent).

The main difficulty with these methods is that they do not provide a unique estimate. Another drawback of the method of moments is that unlike the MLE they often do not lead to efficient estimator. The issue of efficiency will be addressed later but below we provide an example where the method of moment estimators may have undesirable properties.

Example 3.2.3 Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $[0, \theta]$. Since $\mu = \mu_1 = \theta/2$, the method of moment estimator for $\theta = 2\mu_1$ is $2\overline{X}$. This is clearly not a good estimator if $X_{(n)} = max\ X_i$, the largest order statistics, is greater than $2\overline{X}$, since in this model θ is always at least as large as $X_{(n)}$. ■

Another example where the method of moment estimator has undesirable property is provided in example 7.2.2 of the textbook. However, even with all the shortfalls, the method of moments have stood the test of time due to its computational simplicity. Moreover, it is a reliable initial value in the algorithm to find the MLE or any efficient estimator (to be defined later). When the sample size n is large, the method of moment estimator is already very close to the true parameter, so often one iteration suffices to lead to an efficient estimator.

Another useful property of the method of moment is to find a well known distribution to approximate the sampling distribution of a statistic (such as the distribution of the test statistic under the null hypothesis). This can be done by "moment matching" as in the Satterwaite approximation in Example 7.2.3. Such moment matching approaches has been very useful and widely employed before when computing power was limited. Nowadays, bootstrap methods have prevailed and taken over many of the elegant theory and methods. But those interested in the intellectual merits of an approach are encouraged to explore the Satterwaite approximation method in Example 7.2.3.

We close this section with a case study (see Handout 3).

Summary: The method of moments abides by a substitution principle, which is a special case of the *empirical substitution* principle (i.e. express an unknown quantity as a function of the c.d.f and then obtain an estimate of this unknown quantity by substituting the c.d.f with its empirical distribution).

It has the benefit that the MOM estimators are simple, easy to compute, and useful as preliminary estimates. However, they may not be your final choice of estimator due its lack of efficiency.

3 Maximum Likelihood Estimators

Reading Assignment: Section 7.2.2 of the textbook.

The maximum likelihood approach is by far the most popular approach for deriving estimates and it obeys the likelihood principle discussed in Section 2.3.

Let $X_1, \ldots, X_n$ be a random sample from $f(x \mid \theta)$, where $\theta \in \Theta$. Then the joint p.d.f. or p.m.f. of $(X_1, \ldots, X_n)$ is

$$f_{\mathbf{X}}(\mathbf{x} \mid \theta) = f_{\mathbf{X}}(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta). \tag{3}$$

where $x_1, \ldots, x_n$ are dummy/generic variables and could use $a_1, \ldots, a_n$ instead.

If $X_1 = x_1, \ldots, X_n = x_n$ is observed, let $\mathbf{x} = (x_1, \ldots, x_n)$ be the observed data. A key question is what is the likelihood that the observed data $\mathbf{x}$ comes from a distribution with parameter θ within a family $\{f(x \mid \theta) : \theta \in \Theta\}$.

Recall the definition of a likelihood function in Section 1.3.

$$L(\theta \mid \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{4}$$

Definition 3.3.1 (Maximum Likelihood Estimator) If $\hat{\theta} = g(\mathbf{x})$ is such that $L(\hat{\theta} \mid \mathbf{x}) = \max_{\theta \in \Theta} L(\theta \mid \mathbf{x})$, then $\hat{\theta}$ is called the ML estimate of θ based on the sample $\mathbf{x} = (x_1, \ldots, x_n)$.

Clearly, $\hat{\theta}$ involves $\mathbf{x} = (x_1, \ldots, x_n)$ (because $X_i = x_i$ is observed) and its values is called the maximum likelihood estimate. The statistic $\hat{\theta}\,(X_1, \ldots, X_n)$ is called the ML estimator of θ. ■

Note: In the definition above, $X_1, \ldots, X_n$ need not be i.i.d. as long as they have a joint p.d.f or p.m.f. $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n \mid \theta)$ that can take the place of $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$ in (4).

3.1 Derivation of the MLE

There two general strategies.

1. First check if $L(\theta \mid \mathbf{x})$ is a monotone function of θ. If yes and it is a non-decreasing function, the MLE is the largest possible value of θ. Otherwise, if L is a non-increasing function of θ then the MLE is the smallest possible value of θ. (See Example 3.3.1 below for an illustration.)

2 If $L(\theta \mid \theta)$ is not monotone but is a smooth function of θ, set $\dfrac{dL(\theta \mid \mathbf{x})}{d\theta} = 0$ or $\dfrac{d \log L(\theta \mid \mathbf{X})}{d\theta} = 0$, then check $\dfrac{d^2 dL(\theta \mid \mathbf{x})}{d\theta^2} < 0$ or $\dfrac{d^2 \log dL(\theta \mid \mathbf{x})}{d\theta^2} < 0$.

Because most of the time, the log likelihood is easier to work with (why?), from now on we denote $\log L(\theta \mid \mathbf{x}) = \log f_{\mathbf{X}}(\mathbf{x} \mid \theta) =: l(\theta)$ (Note that $(x_1, \ldots, x_n)$ is implicitly in $l(\theta)$.)

Example 3.3.1 $X_1, \ldots, X_n \sim U(0, \theta), \theta > 0$.

$f(x \mid \theta) = \dfrac{1}{\theta}, 0 \leq x \leq \theta,$ or $f(x \mid \theta) = \dfrac{1}{\theta} 1_{[0 \leq x \leq \theta]}(x)$.

Then $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \dfrac{1}{\theta^n}$, if $0 \leq x_i \leq \theta, \forall x_i$.

This is a decreasing function in θ. $\Rightarrow \hat{\theta}$ is the smallest possible value of θ. Since $\theta \geq \max(x_i) \Rightarrow \hat{\theta} = \max(x_i) = x_{(n)}$. Note that $\hat{\theta}(X_1, \ldots, X_n) = \max(X_i) = X_{(n)}$ is the estimator, which is a rule (or formula).

Caveat: If in the definition of the uniform distribution we use the domain $0 < x < \theta$ instead of $0 \leq x \leq \theta$, then the MLE does NOT exist ! However, this is a nuisance as $P(X = 0) = P(X = \theta) = 0$, so one could just use the version of uniform on $[0, \theta]$. ■

Example 3.3.2 $X_1, \ldots, X_n \sim$ Poisson(θ)

$$f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_i x_i}}{\prod_i x_i!}$$

$$l(\theta) \equiv \log f_{\mathbf{X}}(\mathbf{x} \mid \theta) = -n\theta + (\sum_i x_i) \log \theta - \sum_i \log(x_i!)$$

$$\frac{\partial l(\theta)}{\partial \theta} = -n + \sum_i x_i \frac{1}{\theta} = 0 \Rightarrow \theta = \frac{\sum_i x_i}{n} = \bar{x}$$

Question: Do you expect this or is surprised? since $E(X_i) = \theta$ for Poisson this is not surprising at all.

Likewise, you can solve the MLE for exponential distributions, $Exp(\theta)$, $N(\theta, \sigma^2)$, etc.

Continue on the Poisson case, if we know that $\theta > 3$ then (Check carefully)

$$\text{MLE} = \begin{cases} \bar{X} & \text{if } \bar{X} > 3 \\ 3 & \text{otherwise} \end{cases}$$

■

Example 3.3.3 (MLE may not be unique!) $X_1, \ldots, X_n \sim U(\theta, \theta + 1)$, i.i.d.

$$f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \begin{cases} 1 & \text{if } \theta \leq x_i \leq \theta + 1 \\ 0 & \text{otherwise} \end{cases}$$

$\Rightarrow \theta \leq x_{(1)}, \theta \geq x_{(n)} - 1 \Rightarrow x_{(n)} \leq \hat{\theta} \leq x_{(1)}$, so any value in $[x_{(n)} - 1, x_{(1)}]$ can be MLE.

$\Rightarrow$The principle of MLE breaks down here! ■

3.2 MLE for Multivariate θ

If $\underset{\sim}{\theta} = (\theta_1, \dots, \theta_k)$ is a vector, e.g. $N(\mu, \sigma^2)$, set $\dfrac{\partial L}{\partial \theta_i} = 0 \;\; \forall i$, and check the Hessian matrix

$$H = \left(\frac{\partial^2 l(\theta)}{\partial \theta_i \partial \theta_j} \right) \text{ or } \left(\frac{\partial^2 f_{\mathbf{X}}(\mathbf{x} \mid \theta)}{\partial \theta_i \partial \theta_j} \right)$$

which needs to be negative definite. (i.e. all eigenvalues are negative $\Leftrightarrow \underset{\sim}{a}^T H \underset{\sim}{a} < 0, \forall \underset{\sim}{a} \neq 0.$)

Example 3.3.4 $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, i.i.d.

$$\begin{aligned}
& f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{-\frac{\sum_i (x_i - \mu)^2}{2\sigma^2}} \\
\Rightarrow \quad & l(\theta) = \log f_{\mathbf{X}}(\mathbf{x} \mid \theta) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 \\
\Rightarrow \quad & \frac{\partial l(\theta)}{\partial \mu} = -\frac{1}{\sigma^2} \sum_i (x_i - \mu)(-1) = 0 \\
\Rightarrow \quad & \sum_i (x_i - \mu) = 0 \\
\Rightarrow \quad & \hat{\mu} = \bar{x}
\end{aligned}$$

Let $\sigma^2 = \omega$, $\dfrac{\partial L(\theta)}{\partial \sigma^2} = -\dfrac{n}{2}\dfrac{1}{\omega} + \dfrac{1}{2\omega^2} \sum_i (x_i - \mu)^2 = 0.$

Set $\mu = \bar{x} \Rightarrow \omega = \frac{1}{n} \sum_i (x_i - \bar{x})^2 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2.$

$\Rightarrow$ MLE of $\theta = (\mu, \sigma^2)$ is $\hat{\theta}(\mathbf{X}) = (\bar{X}, \frac{1}{n} \sum_i (X_i - \bar{X})^2)$, which is the same as the moment estimator.

What if we want MLE of σ?

- Could redo the above without ω and work directly with σ. If go through the above drill again you will find out that

$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_i (X_i - \bar{X})^2} = \sqrt{\hat{\sigma}^2} \Rightarrow$ This is called the invariance principle of MLE.

Note in the derivation above, we need to check the Hessian matrix.

$$H = \begin{vmatrix} -\frac{n}{\sigma^2} & \frac{1}{\sigma^4} \sum_i (x_i - \mu) \\ \frac{1}{\sigma^4} \sum_i (x_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_i (x_i - \mu)^2 \end{vmatrix} = \begin{vmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4} \end{vmatrix}$$

Hence eigen values are $-\frac{n}{\sigma^2} < 0$, $-\frac{n}{2\sigma^4} < 0 \Rightarrow$ H is negative definite. ■

A second example of multivariate θ is left as a homework problem below.

HW 3.1 A Pareto distribution has c.d.f of the form

$$F(x \mid \theta_1, \theta_2) = 1 - \left(\frac{\theta_1}{x}\right)^{\theta_2}, \theta_1 \le x, \theta_1 > 0, \theta_2 > 0.$$

Find the MLE for (θ_1, θ_2). ■

3.3 Properties of the MLE

1. **Invariance principle of the MLE**

If $\hat{\theta}$ is the MLE of θ and $\eta = \tau(\theta)$, $\Rightarrow \tau(\hat{\theta})$ is the MLE of η.

Proof: Assume first that τ is a one-to-one fucntion, $\Rightarrow \theta = \tau^{-1}(\eta)$.
If we reparametrize the likelihood function by η, then the *log* likelihood function of η is $l^*(\eta) = l(\tau^{-1}(\eta)) \Rightarrow$ MLE of η is the one that $\max_\tau l(\tau^{-1}(\eta))$. Since $\max_\tau l(\tau^{-1}(\eta)) = \max_\theta l(\theta)$, the maximum of $l^*(\eta)$ is attained at $\eta = \tau(\hat{\theta})$.
When τ is not a one-to-one function there may be more than one value of θ that satisfies $\tau(\theta) = \eta$ for a given value η, so we need to redefine the *log* likelihood function l^* as

$$l^*(\eta) = \sup_{\{\theta:\tau(\theta)=\eta\}} l(\theta).$$

The value of η that maximizes $l^*(\eta)$ will corresponds to those θ that maximizes $l(\theta)$. Thus, the maxima of l^* and l is the same and we have

$$\begin{aligned} l^*(\hat{\eta}) &= \sup_\eta \sup_{\{\theta:\tau(\theta)=\eta\}} l(\theta) \\ &= \sup_\theta l(\theta) \\ &= l(\hat{\theta}), \end{aligned}$$

where the second equality is not obvious but correct. Since

$$\begin{aligned} l(\hat{\theta}) &= \sup_{\{\theta:\tau(\theta)=\tau(\hat{\theta})\}} l(\theta) \\ &= l^*(\tau(\hat{\theta})), \end{aligned}$$

we have shown that $l^*(\hat{\eta}) = l^*(\tau(\hat{\theta}))$ and $\tau(\hat{\theta}))$ is the MLE of $\tau(\theta)$. ■

The invariance property is a very attractive one as it alleviates us from tedious calculations. A question is whether the method of moments also share the invariance property. The answer is yes and the proof is left as a homework problem.

HW 3.2 Show that the method of moments MOM) estimator has the invariance property that: If $\hat{\theta}$ is a MOM estimator of θ then $\tau(\hat{\theta})$ is a MOM of $\tau(\theta)$. ■

2. MLE is usually efficient (best in some sense to be discussed later).

3. MLE is consistent (it approaches the target as $n \to \infty$).

4. Problems with MLE

a) It may not exist (e.g. $U(0,\theta)$, or $0 < x < \theta$) or may not be unique (e.g. $U(\theta, \theta+1)$)

b) MLE may exist but cannot be solved explicitly and only numerically, i.e. no explicit or closed form solution. An example is provided below and in fact, it is rather the norm that MLEs generally do not have a closed form solution. The computation could be daunting for a complex system.

Example 3.3.5 Let $X_1, \ldots, X_n \sim Gamma(\alpha, \beta)$, where α is the shape parameter, β the scale parameter, and

$$f(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, x > 0$$

$$f_{\mathbf{X}}(\mathbf{x} \mid \alpha, \beta) = \frac{1}{\Gamma^n(\alpha)\beta^{n\alpha}} \left(\prod_i x_i\right)^{\alpha-1} e^{-\sum_i x_i/\beta}$$

$$\Rightarrow \quad l(\alpha, \beta) = \log f_{\mathbf{X}}(\mathbf{x} \mid \alpha, \beta) = -n \log \Gamma(\alpha) - n\alpha \log \beta + (\alpha - 1) \log \left(\prod_i x_i\right) - \frac{\sum_i x_i}{\beta}.$$

If α is known, $\frac{\partial l(\alpha,\beta)}{\partial \beta} = -\frac{n\alpha}{\beta} + \frac{\sum_i x_i}{\beta^2} \Rightarrow \beta = \frac{\sum_i x_i}{n\alpha}$, so the MLE is $\hat{\beta} = \frac{\sum_i x_i}{n\alpha}$.

However, there is no explicit form to solve α even if β is known, because

$$\frac{\partial l(\alpha, \beta)}{\partial \alpha} = -\frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} - n \log \beta + \log \left(\prod_i x_i\right) = 0$$

$$\Rightarrow \quad \sum_i \log x_i = \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + n \log \beta,$$

which has no explicit solution.

Thus, numerical method, such as the Newton-Raphson method, needs to be employed to solve the MLE for α. We illustrate how to do this in the next subsection but close this example but pointing out that a good initial values is important for the success of any iterated algorithm. ■

5. **Numerical computation for MLE and in general**

Let $f(\theta)$ be a real-value function of a real variable and we want to solve $f(\theta) = 0$.

i) Let θ_0 be an initial guess of the solution.

ii) Newton's method is a iterative algorithm that replaces θ_0 by $\theta_1 = \theta_0 - \frac{f(\theta_0)}{f'(\theta_0)}$

iii) Repeat this by updating $\theta_{k+1} = \theta_k - \frac{f(\theta_k)}{f'(\theta_k)}$ until θ_{k+1} converges.

Motivation of Newton's method: Taylor's expansion

$$\begin{aligned}
& f(\theta) = f(\theta_0) + f'(\theta_0)(\theta - \theta_0) + \ldots \text{(smaller order than } \theta - \theta_0\text{)} \\
\Rightarrow\quad & f(\theta) \approx f(\theta_0) + f'(\theta_0)(\theta - \theta_0) \\
\Rightarrow\quad & \theta - \theta_0 \approx -\frac{f(\theta_0)}{f'(\theta_0)} \\
\Rightarrow\quad & \theta \approx \theta_0 - \frac{f(\theta_0)}{f'(\theta_0)}
\end{aligned}$$

(- insert the geometric interpretation here -)

Here a good initial value θ_0 is important so that the algorithm converges and converges fast. The method of moment estimator is a good choice for θ_0 if it is readily available.

In Example 3.3.5. a MoM for α can be derived easily when β is known (Check this out!) , so we could use it as the initial α-value for the above Newton-Raphson (N-R) algorithm. It can be shown (but not in this course) that one iteration will lead to an estimator that is asymptotically equivalent to the MLE if the initial estimate is $\sqrt{n}$- consistent.

Example 3.3.5 Continued In reality, neither α nor β might be available. So we have to estimate both of them. The above Newton-Raphson algorithm can be extended to multivariate θ using the multivariate version of Taylor expansion. Luckily for the Gamma family the MoM estimators for both are easy to derive. (AGAIN - please check this!) So at least we have good starting values for both parameters in the N-R algorithm.

Even luckier for us, there is a special method that's much simpler than the two-dim N-R algorithm and it works for the two-parameter Gamma family. We describe it below.

Profile Likelihood Method Let $\theta = (\alpha,\beta)$ and $f(x \mid \theta), \theta \in \Theta$, be any family of distributions,

If for any given α the MLE of β exist and is denoted by β_α, the β_α is a function of α and we denote it by $h(\alpha)$,, i.e. $\beta_\alpha = h(\alpha)$. Now plug it into the log likelihood function $l(\alpha,\beta)$ to arrive at $l^*(\alpha) = l(\alpha, h(\alpha))$, which is called the "log profile likelihood". This l^* now contains only the parameter α, hence is easier to to locate its argmax numerically (or explicitly). It can be easily shown, but requires a moment of thought, that

$$max_\alpha l^*(\alpha) = max_\alpha l(\alpha, h(\alpha)) = max_\alpha l(\alpha,\beta_\alpha) = max_\alpha max_\beta l(\alpha,\beta).$$

Hence if $\hat{\alpha}^*$ maximizes $l^*(\alpha)$, then $(\hat{\alpha}^*, h(\hat{\alpha}^*))$ maximizes $l(\alpha,\beta)$. ■

The profiles likelihood approach, if feasible (that is, when a close-form solution for β_α exists if the value of α is given), simplifies the MLE approach. It also produces a legitimate standard error estimate for $\hat{\alpha}^*$ (to be elaborated later). This is especially attractive when β is high or infinite dimensional. An example is the Cox Proportional hazard model when the baseline hazard function is unknown and modeled nonparametrically.

Summary: The MLE approach abides by the likelihood principle discussed in Chapter 2 (Section 3). It enjoys the invariance principle of point estimation, which is also shared by the method of moments. It typically leads to efficient estimator (to be elaborated later) hence has been the prevailing method of estimation. Its main drawback is on the computational front, numerical procedure is often required to compute the MLE and it can get very complicated when the model is complex.

In fact, the derivation of the likelihood itself could be nontrivial. An example is the linear mixed-effects model for repeated measurements or longitudinal data. For example, let $Y_i(t) = \beta_0 + \beta_1 t + b_i + e_i(t)$, be the longitudinal response function of subject i which follows a linear time trend with a subject-specific random effects b_i and b_i and e_i are independent. Try to find out the likelihood function if you observe $Y_{ij} \equiv Y_i(t_{ij})$, $j = 1, \ldots, n_i$, for the ith subject and $i = 1, \ldots, n$. First try the simplest case when $e_{ij} \equiv e_i(t_{ij})$ are independent (across both i and j), $e_{ij} \sim N(0, \sigma^2$ and $b_i \sim N(0, \sigma_b^2)$. Next try a general case when the normality assumption on e_{ij} are replaced by another parametric family.

Even if you can write down the likelihood function, the numerical solution of the MLEs of $\theta = (\beta_0, \beta_1, \sigma, \sigma_b)$ typically relies on the EM-algorithm.

4 Bayes Estimators

Reading Assignment: Section 7.2.3. of the textbook.

This is a branch of statistics that is called "Bayesian Statistics". People who abide by this principle are called "Bayesian", who view the parameter θ as a random variable rather than a fixed value as we've discussed so far for the MLE and MOM.

- Bayesians view θ as a random variable because they believe they have some idea where θ is more likely to be, even before any data are observed.

4.1 Prior and Posterior Distribution

Definition 3.4.1 (Prior Distribution) A prior distribution of θ is a distribution defined on Θ such that $\theta \sim \pi(\cdot)$. Such a prior distribution is chosen before an experiment is carried out (i.e. before the data were collected) and reflects the experimenter's belief. A sample is then taken from a population indexed by θ with p.d.f. or p.m.f. $f(x \mid \theta)$ and the prior distribution is updated with the sample information. ■

If a random sample is observed, Bayesians regard $X_1, \ldots, X_n$ as conditional independent and identically distributed given the value of θ. That is, the joint p.d.f. of $\mathbf{X} = (X_1, \ldots, X_n)$ given θ is $f_{\mathbf{X}|\theta}\,(\mathbf{x} \mid \theta) = \prod_i f(x_i \mid \theta)$. The joint p.d.f. of $\mathbf{X}$ and θ is thus $f_{\mathbf{X},\theta}\,(\mathbf{x} \mid \theta)\,\pi(\theta)$ and the marginal p.d.f. or p.m.f. of $\mathbf{X}$ is $f_{\mathbf{X}}(\mathbf{x}) = \int_\Theta f_{\mathbf{X},\theta}\,(\mathbf{x} \mid \theta)\,\pi(\theta) d\theta$.

Definition 3.4.2 The posterior distribution of θ after we have observed $\mathbf{x}$ is the distribution of the variable θ given the observed value $(x_1, \ldots, x_n)$ of $\mathbf{X}$. We denote the posterior p.d.f. or p.m.f. as $\pi_{\theta|\mathbf{x}}\,(\theta \mid \mathbf{x})$. ■

It then follows that the posterior p.d.f. or p.m.f. takes the form

$$\pi_{\theta|\mathbf{x}}(\theta \mid \mathbf{x}) = \frac{f_{\mathbf{X}|\theta}\,(\mathbf{x} \mid \theta)\,\pi(\theta)}{\int_\Theta f_{\mathbf{X},\theta}\,(\mathbf{x} \mid \theta)\,\pi(\theta) d\theta}$$

- A consequence here is that the distribution of θ has changed from a "prior" to "posterior" one after we have seen the data.
- Note that

  $$\pi_{\theta|\mathbf{x}}\,(\theta \mid \mathbf{x}) \propto f_{\mathbf{X},\theta}\,(\mathbf{x} \mid \theta)\,\pi\,(\theta),$$

  where $\propto$ means proportional as a function of "θ".

  and $\int \pi_{\theta|\mathbf{X}}\,(\theta \mid \mathbf{x}) d\theta = 1$ (?), so $f_{\mathbf{X}}(\mathbf{x})$ is the normalizing constant.

Example 3.4.1 Assume that each person has a probability θ to catch a cold. Let

$$X_i = \begin{cases} 1, & \text{if the } i\text{th random sampled person catches a cold} \\ 0, & \text{otherwise} \end{cases}$$

$\Rightarrow X_i \mid \theta \sim \text{Bin}(1, \theta), \forall i$, where θ=Pr($X_i = 1$) and $f_{\mathbf{X}|\theta}(\mathbf{x} \mid \theta) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$.
Since $0 \le \theta \le 1$, a question is what could be a prior distribution of θ? e.g. U(0,1) and Beta(α,β)

Let's consider priors that are Beta(α,β) as they have nice properties and U(0,1) is a special case.

Recall a Beta random variable, Beta(α,β), is a continuous random variable on [0,1] with p.d.f.

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 \le x \le 1, \alpha,\beta > 0,$$

where $B(a, b) = \displaystyle\int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

For prior θ, this implies

$$\pi(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, 0 \le \theta \le 1, \alpha,\beta > 0.$$

$$\begin{array}{rcl} \Rightarrow f_{\mathbf{X},\theta}(\mathbf{x}, \theta) & = & \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} \dfrac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ & \propto & \theta^{\sum_i x_i+\alpha-1}(1-\theta)^{n-\sum_i x_i+\beta-1} \\ & \text{resembles} & Beta\,(\sum_i x_i + \alpha, n - \sum_i x_i + \beta) \end{array}$$

$\Rightarrow \pi_{\theta|\underset{\sim}{X}}(\theta \mid \underset{\sim}{x}) \sim Beta(\sum_i x_i + \alpha, n - \sum_i x_i + \beta)$. ■

- This property that the posterior distribution is also a beta distribution (like the prior distribution) is called "Beta distribution is a conjugate prior family for a sample from a Bernoulli distribution". i.e. the family of Beta distribution is a closed under the sampling plan from a Bernoulli distribution.

- α and β are called "prior hyper parameters", "$\sum_i x_i + \alpha$ and $n - \sum_i x_i + \beta$" are called posterior hyper parameters.

Definition 3.4.3 (Conjugate Prior) Let $\mathcal{F}$ be the class of p.d.f.s or p.m.f.s $f(x \mid \theta)$. A class Π of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the class for all $f \in \mathcal{F}$, all priors in Π, and all $\mathbf{x}$ in the sample space. ■

Remarks:

1. Note that conjugate priors are just convenient priors, they are not necessarily better or more correct than other priors.

2. The choice of prior distribution is controversial!

- Bayesian use "subjective" prior distribution, meaning that two Bayesians may come up with two different priors for θ.

- Non-Bayesians (Frequentists) believe that θ can be assigned a prior distribution only when there is extensive previous information almost the relative frequency with which θ has taken its possible values, so any two would agree on the "correct" prior distribution to be used.

- Baysians retorted that considering θ a fixed value is equivalent to choose a prior distribution that is degenerate (meaning $\text{Var}(\theta) = 0$) at the constant θ!

3. Saving grace - The posterior distributions are often quite similar regardless of the choice of the prior, if there are a lot of data (i.e. n is large).

4. GOOD NEWS! You have learned how to calculate the posterior distribution when X and θ are both discrete in 200A. This is the Bayes Theorem!

A list of conjugate priors is provided in Handout 4.

4.2 Bayes Estimator

Both the MLE and MOM estimate follow a certain rule/criteria. What are they?

MLE - maximize likelihood, MOM - use moments.

The Bayes estimates rely on a target (criterion), the loss function $\mathcal{L}(\theta, a)$, where θ is parameter value and a is the estimated value.

e.g. $\mathcal{L}(\theta, a) = (\theta - a)^2$, square error loss;

$\mathcal{L}(\theta, a) = |\theta - a|$, absolute error loss.

Ideally, we want to minimize $\mathcal{L}(\theta, a)$ for a given $\mathbf{x}$, but θ is random, so we minimize $E(\mathcal{L}(\theta, a) \mid \mathbf{x})$. Note here that the expected value is taken w.r.t. the conditional distribution of θ given $\mathbf{X} = \mathbf{x}$.

Definition 3.4.1 The Bayes estimator for the model $\{f(x \mid \theta) : \theta \in \Theta\}$ with prior distribution $\pi(\theta)$ (i.e. $\theta \sim \pi(\theta)$) w.r.t. a loss function. $\mathcal{L}(\theta, a)$ is the value a (which dependson $\mathbf{x}$) that $\min_a E(\mathcal{L}(\theta, a) \mid \mathbf{x})$. ■

Theorem 3.4.1 (a) The Bayes estimator for the square error loss function is $E(\theta \mid \mathbf{x})$, the posterior mean.

(b) The Bayes estimator for the absolute error loss function is the median of the posterior distribution.

Proof: (a) follows from the fact that $E[(X-a)^2]$ is minimized by $a = E(X)$ and the expectation can be replaced by a conditional expectation. (b) follows from the fact that $E[\mid X-a \mid]$ is minimized by a = median of the distribution of X and the expectation can be replaced by a conditional expectation. Details will be done in discussion section. ■

Example 3.4.1 (continued) $X_1,\ldots,X_n \sim Bin(1,\theta)$, $\theta \sim Beta\ (\alpha,\beta)$, conjugate prior, $\theta \mid \underset{\sim}{X} \sim Beta\ (\sum_i X_i+\alpha, n-\sum_i X_i+\beta)$.

(a) For $L(\theta,a)=(\theta-a)^2$, the Bayes estimator is

$$\begin{aligned}
\hat{\theta} &= \frac{\alpha+\sum_i X_i}{\alpha+\beta+n}, \quad \text{(the mean of } Beta(\alpha,\beta) \text{ is } \frac{\alpha}{\alpha+\beta}) \\
&= \frac{\bar{X}+\alpha/n}{1+(\alpha+\beta)/n} \approx \bar{X} \to \mu, \text{as } n\to\infty \\
&= \frac{1}{1+(\alpha+\beta)/n}\bar{X} + \frac{(\alpha+\beta)/n}{1+(\alpha+\beta)/n}\frac{\alpha/n}{(\alpha+\beta)/n} \\
&= \frac{1}{1+(\alpha+\beta)/n}\cdot\bar{X} + \frac{(\alpha+\beta)/n}{1+(\alpha+\beta)/n}\cdot\frac{\alpha}{\alpha+\beta}
\end{aligned}$$

where $\bar{X}$ is the MLE, $\frac{\alpha}{\alpha+\beta}$ is the prior mean.

This implies that the Bayes estimator is closed to $\bar{X}$ and not sensitive to the prior distribution (which is good!). Interestingly, the Bayes estimator shrinks the MLE towards the prior estimate mean.

(b) For $L(\theta,a)=|\theta-a|$, the Bayes estimator is the median of a Beta distribution with $Beta(\sum_i X_i+\alpha, n-\sum_i X_i+\beta)$, i.e. $\hat{\theta}=m$, s.t. $\displaystyle\int_0^m \frac{t^{\sum_i X_i+\alpha}(1-t)^{n-\sum_i X_i+\beta-1}}{Beta\ (\sum_i X_i+\alpha, n-\sum_i X_i+\beta)}dt = 0.5$.

This requires numerical methods to evaluate! ■

HW 3.3 If we want to estimate $q(\theta)=\text{Var}(X_i)=\theta(1-\theta)$ in Example 3.4.1 and use the square error loss function, i.e. $L(q(\theta),a)=(q(\theta)-a)^2$, the Bayes estimator is $\hat{\theta}=E(q(\theta)\mid \mathbf{x})=\int\theta(1-\theta)\pi(\theta\mid\mathbf{x})d\theta$.

(a) Find the Bayes estimator of θ.

(b) Show that it tends to $\bar{X}(1-\bar{X})$

(c) Does the invariance principle hold in this case? ■

A list of conjugate priors for several parametric families is provided in Handout 4.

4.3 Minimal sufficiency Revisited

Proposition If a sufficient statistic T (or jointly sufficient statistic) exists, and MLE $\hat{\theta}$ (or Bayes estimator) exists, then $\hat{\theta}$ (MLE or Bayes statistic) depends on the observations only through T.

Proof: By Factorization Theorem, $L(\theta \mid \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x} \mid \theta) = g(T(\mathbf{x_n}), \theta)h(\mathbf{x_n})$.

$$\begin{aligned} \max L(\theta) \quad &\Leftrightarrow \quad \max g(T(\mathbf{x_n}), \theta) \\ &\Rightarrow \quad \text{The MLE } \hat{\theta} \text{ is a function of } T(\mathbf{x_n}) \end{aligned}$$

The same proof applies to a vector $T = (T_1, \ldots, T_k)$.

The results also hold for Bayes estimators because $f_{\Theta|\mathbf{X}}(\theta \mid \mathbf{x_n}) \propto f_{\mathbf{X}}(\mathbf{x_n} \mid \theta)\pi(\theta) \propto g(T(\mathbf{x_n}), \theta)\pi(\theta) \Rightarrow$ $\hat{\theta}$ must be a function of $T(\mathbf{x_n})$. ■

Corollary If MLE $\hat{\theta}$ (or Bayes estimator) is sufficient $\Rightarrow \hat{\theta}$ is minimal sufficient.

Proof: By the proposition above, $\hat{\theta}$ is a function of any sufficient statistic. Since it is sufficient $\Rightarrow$ it is minimal sufficient. ■

Example 3.4.2 $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. MLE=$(\bar{X}, \frac{1}{n}\sum_i (X_i - \bar{X})^2)$ is sufficient. $\Rightarrow$ it is minimal sufficient. $\Rightarrow (\sum_i X_i, \sum_i X_i^2)$ is minimal sufficient. (why ?) ■

Example 3.4.3. $X_1, \ldots, X_n \sim U(\theta_1, \theta_2), \theta_1 \le x_{(1)} \le x_{(n)} \le \theta_2$.
$f_n(\mathbf{x_n}, \theta) = \left(\frac{1}{\theta_2 - \theta_1}\right)^n 1_{(-\infty, \theta_2)}(x_{(n)}) 1_{(\theta_1, \infty)}(x_{(1)})$ decreases in θ_2 and increases in θ_1.
$\Rightarrow$ MLE $(\hat{\theta}_1, \hat{\theta}_2) = (X_{(1)}, X_{(n)})$. By Factorization Theorem, $(X_{(1)}, X_{(n)})$ is jointly sufficient.
$\Rightarrow$ MLE=$(X_{(1)}, X_{(n)})$ is minimal sufficient. ■

Summary: We learn in this section the Bayesian approach for point estimation, which treats the parameter θ as a random variable. The prior and posterior distribution for θ represents a Bayesian's believe of what the distribution of θ should be before and after the sample $\mathbf{x}$ is observed.

The Bayes estimate is determined according to a choice of loss function and may vary with the loss function.

Pros : 1). The Bayes estimator is not sensitive to the prior distribution if n is large.

2). The Bayes approach is "Bayesian coherent" which means the subjective probability should be such as to ensure self-consistent betting behavior.

3). The Bayes approach leads automatically to uncertainty quantification while the frequentist approach requires additional effort.

Cons : 1). It might be sensitive to the prior for small or moderate n.

2). It requires a prior distribution for θ, which may not be easy if θ is a vector, as there are not many multivariate distribution for θ, and even if there are, the computation can be challenging! (Conjugate prior is rare in the multivariate case.)

3). The invariance principle that was enjoyed by the MOM and MLE does not hold for Bayes estimator.

5 Properties of Estimators

Reading Assignment: Sections 7.3.1 and 7.3.4 of the textbook.

We have learned so far several methods to derive an estimator so the next question is which method is preferred. To answer this, we first need to decide the criteria to evaluate an estimator and to compare estimators. After this we will learn how to improve an estimator if the circumstance affords it.

5.1 Unbiasedness

Definition 3.5.1 Let $\hat{\theta} = \delta(X_1, \ldots, X_n)$ be a statistic, $\hat{\theta}_n$ is called an unbiased estimator of θ, if and only if $E_\theta(\hat{\theta}) = \theta, \forall\theta$.

$$\begin{array}{cccc} \text{Bias}(\hat{\theta}) = & E_\theta(\hat{\theta}) & - & \theta \\ & \downarrow & & \downarrow \\ & \text{aiming on the average at} E(\hat{\theta}) & & \text{target} \end{array}$$

■

Example 3.5.1 $E(\bar{X}) = \mu \Rightarrow \bar{X}$ is an unbiased estimator of μ. $m_k(\theta) = \frac{1}{n}\sum_i X_i^k$ is unbiased estimator of $\mu_k = E(X_i^k)$. However, MoM estimator may be biased. e.g. The MoM of σ^2 is

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_i X_i^2 - (\bar{X})^2 \\ &= \frac{1}{n}\left[\sum_i X_i^2 - n(\bar{X})^2\right] \\ &= \frac{1}{n}\sum_i (X_i - \bar{X})^2 \end{aligned}$$

This is also the MLE for $N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$.

Question: Is it unbiased?

We have learned in Chapter 1 that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \Rightarrow E(S^2) = \sigma^2$.
Since $\hat{\sigma}^2 = \frac{n-1}{n}S^2 \Rightarrow E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$,
$\Rightarrow \text{Bias}(\hat{\sigma}^2) = -\frac{\sigma^2}{n}$.
$\Rightarrow$ It always underestimate θ but the bias decreases to zero as $n \to \infty$.

Note, however, $\frac{1}{n}\sum_i (X_i - \mu)^2$ is an unbiased estimator of σ^2. (Why?)

HW 3.4 Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from $U(0, \theta)$. Find the bias of the MLE for θ. ■

Remarks 1. Unbiased estimators may not be unique.

Example 3.5.2 $X_1, \ldots, X_n \sim Pois(\theta) \Rightarrow \mu = \sigma^2 = \theta.$

Here both $\bar{X}$ and S^2 are unbiased.

Moreover, there are millions/gazillion of unbiased estimators (or more appropriately, there are infinitely many unbiased estimators) in this case because

$\alpha\bar{X} + (1-\alpha)\sigma_1^2$ is unbiased for any $-\infty < \alpha < \infty$.

Question: Which one is the best?

Answer: The one with the minimal variance! (To be continued in the next section.) ■

2. Unbiased estimator may not exist!

Example 3.5.3 Three is no unbiased estimator for $\sqrt{p}$ in $Bin(n, p)$.

Proof: We will prove by contradiction.

If an unbiased estimator exists then it must hold for $n = 1$.

If $\hat{\theta} = g(X_1)$ is unbiased $\Rightarrow E_\theta(\hat{\theta}) = g(1)p + g(0)(1-p) = \sqrt{p}, \forall p.$

Note that the RHS is $\sqrt{p}$ but the LHS is a linear function of p. Since p is arbitrary, this cannot hold and we have a contradiction. ■

3. Unbiased estimator may not be reasonable. It is not allways the truth

Example 3.5.4 Let $X = \#$ failures till the first success. $\Rightarrow P(X = k) = (1-\theta)^k\theta, k = 0, 1, 2, \ldots.$

Let's first see what an unbiased estimator for θ looks like. Again, we focus on the case $n = 1$.

If $\hat{\theta} = g(X_1)$ is unbiased $\Rightarrow E_\theta(\hat{\theta}) = \sum_{k=0}^{\infty} g(k)(1-\theta)^k\theta = \theta.$

$\Rightarrow \sum_{k=0}^{\infty} g(k)(1-\theta)^k = 1, \forall\theta.$ You are putting all your faith in the first trial to be heads g(0)=1

$\Rightarrow g(0) = 1$ and $g(k) = 0, \forall k \geq 1.$

$\Rightarrow$ We claim $\theta = 1$ if the 1st trial is a "success", and we claim $\theta = 0$ if the 1st trial is not a "success".

Thus, this unbiased estimator is not reasonable. (Why?)

* Note that once you have derived an unbiased estimator for θ We can now come up with an unbiased estimator of θ based on a sample of size n by using $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$. ■

Morale: Unbiasedness might be a good property but do not insist on it!

HW 3.5 Consider the family of $Pois(\lambda)$.

(a) Derive an unbiased estimator for $e^{-2\lambda}$ using a random sample of size n and show that it is unique.

(b) Explain why the unbiased estimator in (a) is unreasonable. ■

5.2 Measure of Quality of Estimators

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased and $\text{Var}_\theta(\hat{\theta}_1) \leq \text{Var}_\theta(\hat{\theta}_2), \forall \theta$.

$\Rightarrow \hat{\theta}_1$ is better than $\hat{\theta}_2$ since it tends to be closer to θ.

Question: How do we compare two estimators that are not necessarily unbiased?
One may have a larger bias and the other larger variance.

Answer: We compare them based on the mean square error (MSE).

$\text{MSE}(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2]$.

We claim $\hat{\theta}_1$ is better than $\hat{\theta}_2 \Leftrightarrow \text{MSE}(\hat{\theta}_1) \leq \text{MSE}(\hat{\theta}_2), \forall \theta$.

Proposition $\text{MSE}(\hat{\theta}) = [\text{Biase}(\hat{\theta})]^2 + \text{Var}_\theta(\hat{\theta})$

Proof:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E_\theta[(\hat{\theta} - \theta)^2] \\
&= E_\theta[(\hat{\theta} - E(\hat{\theta}) + \underbrace{E(\hat{\theta}) - \theta}_{\text{bias}})^2] \\
&= E_\theta\{(\theta - E_\theta(\hat{\theta}))^2 + [\text{Bias}(\hat{\theta})]^2 + 2(\hat{\theta} - E_\theta(\hat{\theta}))\text{Bias}\} \\
&= \underbrace{E_\theta(\hat{\theta} - E_\theta(\hat{\theta}))^2}_{\text{Var}_{\theta(\hat{\theta})}} + [\text{Bias}]^2 + \text{Bias} \cdot 2\, \underbrace{E_\theta(\hat{\theta} - E_\theta(\hat{\theta}))}_{=0}.
\end{aligned}$$

■

Example 3.5.6 Under $N(\mu, \sigma^2)$, consider $\hat{\sigma}_0^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$ (which is both the MoM estimator and the MLE); and $\hat{\sigma}_1^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$, which is the sample variance S^2.

We know that $\hat{\sigma}_0^2$ is biased from Example 3.5.1 but $\text{MSE}(\hat{\sigma}_0^2) \leq \text{MSE}(\hat{\sigma}_1^2)$.

The proof is provided in Examples 7.3.3 and 7.3.4 in the textbook (page 331). The key step is to show that $E_{\mu,\sigma}[(S^2 - \sigma^2)^2] = \text{Var}_{\mu,\sigma}(S^2) = \frac{2\sigma^4}{n-1}$ for a normal r.v.

Note, however, that this fact only holds for a normal family and MSE($\hat{\sigma}_0^2$) $\leq$ MSE($\hat{\sigma}_1^2$) does not hold in general. Moreover, caution must be exercised when declaring that $\hat{\sigma}_0^2$ is a better estimate than $\hat{\sigma}_1^2$ because the MSE is not necessarily the most sensible criterion to compare scale estimates. See Remark 4 below for a further discussion. ■

Remarks 1. If both $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased $\Rightarrow$ MSE($\hat{\theta}_i$) = $\text{Var}_\theta(\hat{\theta}_i)$, so the MSE criterion falls back to our earlier criterion based on the variance comparison.

2. In general, it is not possible to find the best estimator that minimizes the MSE (see Figure 7.3.1 on page 333 of the textbook) because the candidate pool is too large (the tradeoff between bias and variance is complicated). However, if we restrict the choice to a subclass then it may be possible to find the best estimator in this subclass. For instance, if we insist on using only unbiased estimator then it may be possible to find the best unbiased estimator. Likewise, it may be possible to find the best equivariant estimator for a class of transformation. We will explore this in section 3.7.

3. Here we use a certain type of measure/criterion the "square error" to compare estimates. It is possible to use other measure/criterion such as "absolute error" $E_\theta|\theta - \hat{\theta}|$ or $E_\theta(e^{(\hat{\theta}-\theta)^2})$, etc. That is, any loss functions $L(a, \theta)$ can be used as a criterion in addition to the square error loss function in MSE and an estimator $\hat{\theta}_1$ is better than $\hat{\theta}_2$ if $E_\theta(L(\hat{\theta}_1, \theta)) \leq E_\theta(L(\hat{\theta}_2, \theta)), \forall\theta$.

4. The MSE criterion is more suitable for a location parameter than a scale parameter because it penalizes equally for overestimation and underestimation, which makes sense for a location parameter but not for scale parameter which is greater than zero. (Why?)

Stein proposed to use $L(a, \sigma^2) = \frac{a}{\sigma^2} - 1 - \log\frac{a}{\sigma^2}$ to evaluate an estimate of σ^2.

5. If θ is a r.v. with a prior distribution $\pi(\theta)$ as in the Bayesian framework, then the Bayes' estimate is the one that minimizes the Bayes risk, $\int_\Theta E(L(a, \theta) \mid \mathbf{x})\pi(\theta)d\theta$.

That is, $\hat{\theta} = argmin_a \int_\Theta E(L(a, \theta) \mid \mathbf{x})\pi(\theta)d\theta$.

So, the Bayes estimator is by definition the best estimator w.r.t. to the Bayes risk.

Summary: An unbiased estimator aims at the correct target on the average so it is a good property to have for an estimator. But it may not exist or may lead to an unreasonable procedure so one must embrace it with cautious. In general, the performance of an estimator is evaluated through a criterion and MSE is a popular choice. However, other choices might be preferred if the target is not a location parameter.

6 Improving an Estimator

Reading Assignment: Section 7.3.3 (up to page 343) of the textbook.

We have learned that Sufficient statistic is a data reduction method. In this section, we show that it can be used to improve an estimator.

Rao-Blackwell Theorem For any estimator $W(X_1, \ldots, X_n)$ and any sufficient statistics T of θ, let $\phi(T) = E_\theta[W(\mathbf{X}) \mid T]$. Then

(a) $\phi(T)$ does not involve θ, $E_\theta(\phi(T)) = E_\theta(W(\mathbf{X})) \equiv \tau(\theta)$, and $\mathrm{Var}_\theta(\phi(T)) \le \mathrm{Var}_\theta(W)$.

$\Rightarrow$ MSE$(\phi(T)) \le$ MSE$(W), \forall \theta \in \Omega$, where MSE$(\phi(T)) = E(\phi(T) - \tau(\theta))^2$.

(b) The "=" in (a) holds iff $\phi(T) = W \Leftrightarrow W$ depends on $X_1, \ldots, X_n$ only through T,

i.e. W is a function of T.

Proof. We have already seen the proof of (a) in Exercise 6.36. Specifically, $\phi(T)$ does not involve θ because the distribution of $\mathbf{X} \mid T$ does not involve θ (by definition of sufficient statistic) so the conditional distribution of any function W of $\mathbf{X}$ given θ also does not involve θ. Moreover, for any two r.v.'s U and V,

$E(U) = E[E(U \mid V)]$, and Var $(U) = \mathrm{Var}[E(U \mid V)] + E[\mathrm{Var}(U \mid V)] \ge \mathrm{Var}[E(U \mid V)]$.

Let $U = W$ and $V = T$, then $\phi(T) = E_\theta(U \mid V)$ and Part (a) follows.

For (b), the "=" holds $\Leftrightarrow E[\mathrm{Var}(U \mid V)] = 0 \Leftrightarrow \mathrm{Var}(U \mid V) = 0$

$\Leftrightarrow U \mid V$ is a constant $\Leftrightarrow U$ is completely determined by $V \Leftrightarrow U$ is a function of V. ■

Thus, the R-B Theorem implies that $\phi(T)$ is an improvement of W as an estimator of $E_\theta(W)$ unless W itself is already sufficient. An interesting observation is that an improvement will occur if we replace T by any statistic in the proof even if it is not sufficient but the problem is that the resulted $\phi(T)$ may involve θ and hence is not a legitimate estimator. See Example 7.3.18 in the textbook for an illustration.

Definition 3.6.1 (Inadmissible, admissible, dominates)

A estimator δ is inadmissible for the risk function $R(a, \theta) \Leftrightarrow$ There exists another estimator δ_0 such that $R(\delta_0, \theta) \le R(\delta, \theta)$ for every $\theta \in \Omega$ and there is strict inequality for at least one value of $\theta \in \Omega$.

If so, δ_0 is said to dominate δ w.r.t. the risk $R(\delta, \theta)$. ■

R-B Theorem implies that any estimator $W(\mathbf{X})$ that is not a function of a sufficient statistic T must be inadmissible w.r.t. the MSE (because it can be improved or is dominated by $E(W(\mathbf{X}) \mid T)$).

Example 3.6.1 $X_1, \ldots, X_n \sim U(\theta_1, \theta_2)$.

If θ_1 is known, $\max(X_i) = X_{(n)}$ is sufficient for θ_2.

If θ_2 is known, $\max(X_i) = X_{(1)}$ is sufficient for θ_1.

If θ_1 and θ_2 are both unknown, $(X_{(1)}, X_{(n)})$ is jointly sufficient. Since they are also MLE, $\Rightarrow$ $(X_{(1)}, X_{(n)})$ is minimal sufficient.

Q: Is MoM admissible?

A: No. For example, $\theta_1 = 0, \theta_2 = \theta \Rightarrow$ MoM $= 2\bar{X}$, not a function of $X_{(n)}$. Had we known this, we would not use MoM for θ and we would use $E(2\bar{X} \mid X_{(n)}) = 2E(X_1 \mid X_{(n)}) = \frac{n+1}{n} X_{(n)}$ instead.

(See if you can derive the last equality but first take a guess why you might expect this.) ■

Remark $E(W(\mathbf{X}) \mid T)$ is usually hard to compute (e.g. Example 3.6.1 above), but the morale of the R-B Theorem is that we should only search for good estimators that are functions of a sufficient statistic!

In the next section, we learn how to show that an estimator is the best unbiased estimator and how to find a best unbiased estimator.

Summary Sufficient statistic can be used to improve an estimator as illustrated by the Rao-Blackwell Theorem, which also tell us to use only estimators that are functions of sufficient statistic.