BST 676 – Spring 2011 – Dr. Charnigo

Unit III: Evaluating Point Estimators

a. Motivating Case Study #1: What if the sample variance entailed division by n?

In the first week of your first methods course, you learned to divide by $(n-1)$ when computing the sample variance,

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

where $X_1, X_2, \ldots, X_n$ are independently and identically distributed with mean $\mu \in (-\infty, \infty)$ and variance $\sigma^2 \in (0, \infty)$. You were told, either that same week or perhaps a few weeks later, that the rationale for division by $(n-1)$ was to ensure that S^2 would be an unbiased estimator of the population variance σ^2. Thus, while a given data set might yield a numerical value for S^2 in excess of σ^2, or a numerical value for S^2 in deficit of σ^2, there would be no systematic tendency toward overestimation or underestimation across repeated sampling.

To gain further insight, suppose that somehow we knew the population mean μ and could calculate the quantity

$$V^2 := \frac{\sum_{i=1}^n (X_i - \mu)^2}{n}.$$

By linearity of the expectation operator we would have

$$E[V^2] = n^{-1} \sum_{i=1}^n E[(X_i - \mu)^2] = n^{-1} \sum_{i=1}^n Var[X_i] = n^{-1} n\sigma^2 = \sigma^2.$$

We can show (do so at home, for practice) that $\sum_{i=1}^n (X_i - a)^2$ is minimized when $a := \bar{X}$. Hence, we would have $\sum_{i=1}^n (X_i - \bar{X})^2 \leq \sum_{i=1}^n (X_i - \mu)^2$, and then by monotonicity of the expectation operator

$$E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}\right] \leq E\left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{n}\right] = \sigma^2. \quad (1)$$

This suggests that re-defining S^2 with division by n would lead to systematic underestimation of σ^2 across repeated sampling.

However, relation (1) does not clarify whether we should divide by $(n-1)$

or $(n-2)$ or $(n-3)$, etc., to obtain an unbiased estimator. To see why division by $(n-1)$ yields an unbiased estimator, write

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

Then we see that

$$E[\sum_{i=1}^{n}(X_i - \bar{X})^2] = E[\sum_{i=1}^{n} X_i^2] - nE[(\bar{X})^2] =$$

which implies that

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

This explains the specific choice of $(n-1)$ for the divisor of S^2.

On the other hand, Slutsky's Theorem #2 yields

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n+a)} = \frac{(n-1)}{(n+a)} \times S^2 \xrightarrow{P} 1 \times \sigma^2 = \sigma^2$$

for any constant $a \in (-\infty, \infty)$. Putting $a := 0$, we see that σ^2 is still consistently estimated even with a divisor of n rather than $(n-1)$. Is there any reason to consider a divisor of n?

Suppose that $X_1, X_2, \ldots, X_n$ are normally distributed. In this case, $\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$. As such,

$$Var[\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2] = 2(n-1)$$

and hence

$$Var[\sum_{i=1}^{n}(X_i - \bar{X})^2/(n+a)] =$$

Thus, dividing by n instead of $(n-1)$ reduces the variance in estimating σ^2.

So, there is a tradeoff. We can have an unbiased estimator with greater variance or a biased estimator with smaller variance. Which is preferable?

We will revisit this case study at the end of Unit III.

b. Motivating Case Study #2: Sample mean and sample median as competing estimators

Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed with probability density function $f(x;\theta)$, where $\theta \in \Theta \subset \mathbb{R}$ is both the mean and median of the distribution. That is,

$$\int_{\mathbb{R}} x\ f(x;\theta)\ dx = \theta \quad \text{and} \quad \int_{-\infty}^{\theta} f(x;\theta)\ dx = 1/2 = \int_{\theta}^{\infty} f(x;\theta)\ dx.$$

Suppose, moreover, that $f(x;\theta)$ is continuous and positive in a neighborhood of θ. This latter requirement rules out, for example,

$$f(x;\theta) := 1_{\{x\in[\theta-1,\theta-1/2]\cup[\theta+1/2,\theta+1]\}}.$$

However, two examples of probability density functions meeting all of the above requirements are the normal distribution with mean θ (and variance 1)

$$f(x;\theta) := (2\pi)^{-1/2}\exp[-(x-\theta)^2/2] \tag{2}$$

and the double exponential distribution with mean θ (and variance 2)

$$f(x;\theta) := \exp[-|x-\theta|]/2. \tag{3}$$

One can show (using tools beyond the scope of BST 676) that

$$n^{1/2}(M_n - \theta) \xrightarrow{L} N\left(0, \frac{1}{4f(\theta;\theta)^2}\right), \tag{4}$$

where M_n denotes the sample median and $f(\theta;\theta)$ denotes the probability density function $f(x;\theta)$ evaluated at $x = \theta$.

On the other hand, we already know from the Central Limit Theorem that, if the variance of the distribution σ^2 is positive and finite,

$$n^{1/2}(\bar{X}_n - \theta) \xrightarrow{L} N(0, \sigma^2). \tag{5}$$

Expressions (4) and (5) raise a couple of questions:

1. Is M_n a consistent estimator of θ? We cannot apply the Weak Law of Large Numbers to obtain such a conclusion since M_n cannot be written as a sum of independent and identically distributed quantities.

2. Assuming that M_n is somehow shown to be a consistent estimator of θ, are there any circumstances under which we are better off using M_n than $\bar{X}_n$?

We will revisit this case study at the end of Unit III.

c. Unbiased estimation (Cf. pp. 379-398 of Larsen and Marx)

Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed with probability density or mass function $f(x;\theta)$, where $\theta \in \Theta \subset \mathbb{R}$. We already know that an unbiased estimator $\hat{\theta}$ is one for which $E_\theta[\hat{\theta}] = \theta$. However, there may be several unbiased estimators available to us.

For example, if θ is a mean parameter, then both $\hat{\theta}_1 := X_1$ and $\hat{\theta}_2 := \bar{X}$ are unbiased estimators. Which estimator shall we choose? Assuming that the underlying distribution has positive finite variance σ^2, we have

$$Var_\theta[\hat{\theta}_2] = \sigma^2/n < \sigma^2 = Var_\theta[\hat{\theta}_1]$$

at any fixed sample size $n > 1$, so $\hat{\theta}_1$ has unnecessarily large variance. Thus, $\hat{\theta}_1$ is less likely to be close to θ than is $\hat{\theta}_2$. Moreover, $\hat{\theta}_2$ is consistent for θ as $n \to \infty$, while $\hat{\theta}_1$ is not consistent (prove this at home, for practice). Hence, $\hat{\theta}_2$ is a better estimator than $\hat{\theta}_1$. However, we have not shown that $\hat{\theta}_2$ is the best available estimator, only that it is better than $\hat{\theta}_1$. Might there exist an unbiased estimator $\hat{\theta}_3$ with even smaller variance than $\hat{\theta}_2$?

In fact, we can pose a general question: is there a way to tell whether we have found an unbiased estimator with the smallest possible variance?

The Cramer-Rao lower bound helps us to address this question when the probability density or mass function has the exponential family form

$$f(x;\theta) = a(x)\exp[b(\theta) + c(x)d(\theta)].$$

First we need to introduce some notation. We define the Fisher information from n observations as

$$\begin{aligned} J_n(\theta) &:= nE_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] \\ &= nVar_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] \qquad (6) \\ &= -nE_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]. \qquad (7) \end{aligned}$$

Assuming for convenience that $X_1, X_2, \ldots, X_n$ are continuous, equality (6) arises because

$$\begin{aligned} E_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] &= \int_{\mathbb{R}}\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right] f(x;\theta)\,dx \\ &= \int_{\mathbb{R}}\frac{\partial}{\partial\theta}f(x;\theta)\,dx \\ &= \frac{d}{d\theta}\int_{\mathbb{R}} f(x;\theta)\,dx \\ &= 0. \end{aligned}$$

The proof of equality (7) is similar but more tedious.

The Cramer-Rao lower bound says that, for any unbiased estimator $\hat{\theta}$ of θ, we have

$$Var_\theta[\hat{\theta}] \geq \frac{1}{J_n(\theta)}.$$

Hence, if we exhibit an unbiased estimator $\hat{\theta}$ whose variance is $1/J_n(\theta)$, we can be assured that there is no other unbiased estimator with smaller variance. An unbiased estimator whose variance attains the Cramer-Rao lower bound is called efficient. An unbiased estimator whose variance is less than or equal to that of any other unbiased estimator is called best unbiased, even if its variance does not attain the Cramer-Rao lower bound.

Remark #1. As hinted above, there are situations in which no unbiased estimator attains the Cramer-Rao lower bound. In these situations, or when the probability density or mass function is not of the exponential family form, the detection of a best unbiased estimator is more challenging and relies on the identification of a "complete statistic" (beyond the scope of BST 676).

Remark #2. One can show that a best unbiased estimator is essentially unique, in the sense that two best unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ must satisfy $P_\theta(\hat{\theta}_1 = \hat{\theta}_2) = 1$.

Remark #3. The Cramer-Rao lower bound generalizes to accommodate situations in which we wish to estimate not θ itself but rather some continuously differentiable function of θ denoted $\tau(\theta)$. Let $W(\mathbf{X})$ be an unbiased estimator of $\tau(\theta)$. Then the Cramer-Rao lower bound for $W(\mathbf{X})$ is

$$Var_\theta[W(\mathbf{X})] \geq \frac{[\tau'(\theta)]^2}{J_n(\theta)}.$$

Example #1. Put $f(x;\theta) := (2\pi)^{-1/2}\exp[-(x-\theta)^2/2]$, where $\theta \in \Theta := \mathbb{R}$. Then

$$\log f(x;\theta) = -\frac{1}{2}\log[2\pi] - (x-\theta)^2/2,$$

so that

$$\frac{\partial}{\partial\theta}\log f(x;\theta) = (x-\theta) \quad \text{and} \quad \frac{\partial^2}{\partial\theta^2}\log f(x;\theta) = -1.$$

Hence $J_n(\theta) = n$. Since $Var_\theta[\bar{X}] = 1/J_n(\theta) = 1/n$, we conclude that $\bar{X}$ is the unique best unbiased estimator of the mean parameter θ for normally distributed data.

Example #2. Put $f(x;\theta) := \frac{1}{2\theta}1_{\{x\in[0,2\theta]\}}$, where $\theta \in \Theta := (0,\infty)$. The Cramer-Rao lower bound does not apply because the support set depends on θ. However, this example is worth pursuing because it demonstrates a way to prove that an unbiased estimator is not best: simply exhibit another unbiased estimator with smaller variance.

Since $E_\theta[X_1] = \theta$ and $Var_\theta[X_1] = \theta^2/3$, the sample mean $\bar{X}$ is unbiased with variance $\theta^2/(3n)$. On the other hand, put $Z := \max_{\{i\in\{1,2,\ldots,n\}\}} X_i$ and consider $\hat{\theta} := (n+1)Z/(2n)$. For any $z \in [0,2\theta]$ we have (Why?)

$$P_\theta\left(Z \le z\right) = \left(\frac{z}{2\theta}\right)^n, \quad \text{so that} \quad f_Z(z;\theta) = \frac{nz^{n-1}}{(2\theta)^n}.$$

Hence,

$$E_\theta[Z] = \int_0^{2\theta} z\,\frac{nz^{n-1}}{(2\theta)^n}\,dz = \frac{n(2\theta)^{n+1}}{(n+1)(2\theta)^n} = \frac{2n\theta}{(n+1)} \tag{8}$$

$$\text{and} \quad E_\theta[Z^2] = \int_0^{2\theta} z^2\,\frac{nz^{n-1}}{(2\theta)^n}\,dz = \frac{n(2\theta)^{n+2}}{(n+2)(2\theta)^n} = \frac{4n\theta^2}{(n+2)}.$$

Thus,

$$Var_\theta[Z] = \frac{4n\theta^2}{(n+2)} - \frac{4n^2\theta^2}{(n+1)^2} = 4n\theta^2\frac{1}{(n+1)^2(n+2)} < 4\theta^2\frac{1}{(n+1)^2}. \tag{9}$$

Using (8) and (9) we see that $\hat{\theta}$ is unbiased with variance less than θ^2/n^2, which is strictly less than $\theta^2/(3n)$ when $n > 3$. Therefore, the sample mean is not the best unbiased estimator of the mean parameter θ for uniformly distributed data when $n > 3$. In particular, an unbiased estimator is available whose variance decreases as n^{-2} rather than n^{-1}.

Example #3. Put $f(x;\theta) := \theta^{-1}\exp[-x\theta^{-1}]$ for $x \in (0,\infty)$, where $\theta \in \Theta := (0,\infty)$. Recall that $E_\theta[X_1] = \theta$ and $Var_\theta[X_1] = \theta^2$. Suppose that $n \geq 3$.

We have

$$\log f(x;\theta) = -\log\theta - x\theta^{-1},$$

so that

$$\frac{\partial}{\partial\theta}\log f(x;\theta) = -1/\theta + x/\theta^2.$$

We have

$$Var_\theta[-1/\theta + X_1/\theta^2] = Var_\theta[X_1]/\theta^4 = 1/\theta^2,$$

so that $J_n(\theta) = n/\theta^2$. Since $Var[\bar{X}] = 1/J_n(\theta) = \theta^2/n$, we conclude that $\bar{X}$ is the unique best unbiased estimator of the mean parameter θ for exponentially distributed data.

What if we wish to estimate $\tau(\theta) := 1/\theta$? Since $\tau'(\theta) = -1/\theta^2$, the Cramer-Rao lower bound is $[-1/\theta^2]^2/J_n(\theta) = 1/[n\theta^2]$.

Our first instinct may be to try estimating $1/\theta$ by $1/\bar{X} = n/\sum_{i=1}^n X_i$. However, we can show (do so at home, for practice; use Jensen's inequality) that $E_\theta[1/\bar{X}] > 1/\theta$.

Put $Z := \sum_{i=1}^n X_i$. We can show (do so at home, for practice; use moment generating functions) that Z has the gamma distribution with shape parameter n and scale parameter θ. Noting that $n \geq 3$ and $1/\bar{X} = n/Z$, we find that

$$\begin{aligned}
E_\theta[1/\bar{X}] &= n\int_0^\infty z^{-1}\frac{z^{n-1}}{\theta^n\Gamma[n]}\exp[-\theta^{-1}z]\,dz \\
&= \frac{n}{(n-1)\theta}\int_0^\infty \frac{z^{n-2}}{\theta^{n-1}\Gamma[n-1]}\exp[-\theta^{-1}z]\,dz \qquad (10)\\
&= \frac{n}{(n-1)\theta}. \qquad (11)
\end{aligned}$$

Line (10) holds because $\Gamma[n] =$

Line (11) holds because We also find that

$$\begin{aligned}
E_\theta[(1/\bar{X})^2] &= n^2\int_0^\infty z^{-2}\frac{z^{n-1}}{\theta^n\Gamma[n]}\exp[-\theta^{-1}z]\,dz \\
&= \frac{n^2}{(n-1)(n-2)\theta^2}\int_0^\infty \frac{z^{n-3}}{\theta^{n-2}\Gamma[n-2]}\exp[-\theta^{-1}z]\,dz \\
&= \frac{n^2}{(n-1)(n-2)\theta^2}.
\end{aligned}$$

Hence,

$$Var_\theta[1/\bar{X}] = \frac{n^2}{(n-1)(n-2)\theta^2} - \frac{n^2}{(n-1)^2\theta^2} = \frac{n^2}{\theta^2(n-1)^2(n-2)}.$$

Thus, $(n-1)/(n\bar{X})$ is an unbiased estimator of $1/\theta$ with variance $1/[(n-2)\theta^2]$.

Although the variance is greater than the Cramer-Rao lower bound, one can use complete statistics (beyond the scope of BST 676) to show that $(n-1)/(n\bar{X})$ is the best unbiased estimator of $1/\theta$. Thus, the Cramer-Rao lower bound is not attainable when we seek to estimate $1/\theta$, even though it was attainable when we sought to estimate θ.

What we have encountered in this example is a rather general phenomenon. If $\nu_1(\theta)$ and $\nu_2(\theta)$ are two functions of θ, not linearly related, then the Cramer-Rao lower bound cannot be attained in the unbiased estimation of both. To see how this relates to the above example, put $\nu_1(\theta) := \theta$ and $\nu_2(\theta) := 1/\theta$. Unless $\tau(\theta)$ has the form $k_1 + k_2\theta$ for real constants k_1, k_2, we need not waste our time looking for an unbiased estimator of $\tau(\theta)$ that attains the Cramer-Rao lower bound in the above example.

d. Mean square error

<u>Example #3, continued</u>. If we use $U(\mathbf{X}) := (n-2)/(n\bar{X})$ to estimate $1/\theta$, then

$$Var_\theta[U(\mathbf{X})] = (n-2)/[(n-1)^2\theta^2] < 1/[n\theta^2].$$

This does not contradict the Cramer-Rao lower bound since $U(\mathbf{X})$ is not an unbiased estimator of $1/\theta$. However, this does raise the question of whether an unbiased estimator must be pursued at all costs. Is a slightly biased estimator with smaller variance preferable?

Of course, we must decide what makes an estimator preferable. One criterion is mean square error, which we define as

$$E_\theta[(\hat{\theta} - \theta)^2] = \{E_\theta[\hat{\theta}] - \theta\}^2 + Var_\theta[\hat{\theta}] \tag{12}$$

if θ is the target. The equality in (12) may be established by writing $\hat{\theta} - \theta = \hat{\theta} - E_\theta[\hat{\theta}] + E_\theta[\hat{\theta}] - \theta$ and expanding the square on the left side (do so at home, for practice). The quantity $E_\theta[\hat{\theta}] - \theta$ is called the bias and represents the tendency of $\hat{\theta}$, if any, toward systematic overestimation or underestimation of θ.

More generally, we define mean square error as

$$E_\theta[(W(\mathbf{X}) - \tau(\theta))^2] = \{E_\theta[W(\mathbf{X})] - \tau(\theta)\}^2 + Var_\theta[W(\mathbf{X})]$$

when some function $\tau(\theta)$ is the target. The mean square error is partitioned into the squared bias of $W(\mathbf{X})$ in estimating $\tau(\theta)$ plus the variance of $W(\mathbf{X})$.

Returning to the question of whether $U(\mathbf{X})$ is preferable to $(n-1)/(n\bar{X})$ for estimating $1/\theta$, the reciprocal of the mean parameter for exponentially distributed data, we have that the bias of $U(\mathbf{X})$ is

$$E_\theta[U(\mathbf{X})] - \frac{1}{\theta} = \frac{(n-2)}{(n-1)\theta} - \frac{1}{\theta} = \frac{-1}{(n-1)\theta}.$$

Thus, the mean square error of $U(\mathbf{X})$ is

$$\frac{1}{(n-1)^2\theta^2} + \frac{(n-2)}{(n-1)^2\theta^2} = \frac{1}{(n-1)\theta^2}.$$

This is a smaller mean square error than the $1/[(n-2)\theta^2]$ obtained by the unbiased estimator $(n-1)/(n\bar{X})$. So, one can argue in favor of using the slightly biased estimator $U(\mathbf{X})$.

On the other hand, mean square error may not always be an appropriate criterion for judging between estimators. If $\tau(\theta)$ is constrained to lie in $(0, \infty)$, then underestimation of the target may be much worse than overestimation of the target by an equal amount. In this case, perhaps we should consider a criterion that is more sensitive to underestimation. One such criterion is

$$E_\theta\left[\frac{W(\mathbf{X})}{\tau(\theta)} - 1 - \log\left(\frac{W(\mathbf{X})}{\tau(\theta)}\right)\right]. \tag{13}$$

Returning to our example on estimating $1/\theta$ for exponentially distributed data, suppose that $W(\mathbf{X})$ has the form $c/\bar{X}$ for some constant c. Then criterion (13) is

$$E_\theta\left[\frac{c\,\theta}{\bar{X}} - 1 - \log c - \log\left(\frac{\theta}{\bar{X}}\right)\right] = \frac{cn}{(n-1)} - 1 - \log c - E_\theta\left[\log\left(\frac{\theta}{\bar{X}}\right)\right].$$

Whatever that last expected value may be, it does not play any role in identifying the c that minimizes (13). Using calculus we can show (do so at home, for practice) that the minimizing c is $(n-1)/n$. So, within the class of estimators having the form $c/\bar{X}$, the unbiased estimator is preferred by criterion (13) over the biased estimator $U(\mathbf{X})$ with smaller mean square error.

e. Consistency (Cf. pp. 406-410 of Larsen and Marx)

Suppose that $\theta \in \Theta \subset \mathbb{R}$ is an unknown parameter describing how data $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)^T$ arise. We say that $W(\mathbf{X}_n)$ is a consistent estimator of $\tau(\theta)$ if

$$W(\mathbf{X}_n) \xrightarrow{P} \tau(\theta)$$

as $n \to \infty$. Four techniques for proving consistency are as follows.

Technique #1. Suppose that $W(\mathbf{X}_n) = \psi(\hat{\eta})$ and $\tau(\theta) = \psi(\eta)$ for some continuous function ψ. If $\hat{\eta}$ is already known to be a consistent estimator of η, then $W(\mathbf{X}_n)$ is a consistent estimator of $\tau(\theta)$. (Why?)

As an example, suppose that

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x;\theta) := \theta x^{\theta-1} 1_{\{x \in (0,1)\}}$$

for some $\theta \in \Theta := (0, \infty)$. Let $\psi(t) := t/(1-t)$ for $t \in (0,1)$, and let $\eta := \theta/(\theta+1)$. Then, by the Weak Law of Large Numbers,

$$\hat{\eta} := \bar{X}_n \xrightarrow{P} E[X_1] = \theta/(\theta+1) = \eta.$$

Thus, $W(\mathbf{X}_n) := \psi(\bar{X}_n) = \bar{X}_n/(1-\bar{X}_n)$ is consistent for $\tau(\theta) := \psi(\eta) = \theta$.

Technique #2. Suppose that $U(\mathbf{X}_n)$ is another estimator already known to be consistent for $\tau(\theta)$. If

$$W(\mathbf{X}_n) - U(\mathbf{X}_n) \xrightarrow{P} 0 \quad \text{or} \quad W(\mathbf{X}_n)/U(\mathbf{X}_n) \xrightarrow{P} 1,$$

then $W(\mathbf{X}_n)$ is also consistent for $\tau(\theta)$. (Why?)

As an example, suppose that

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x;\theta) := \theta^x (1-\theta)^{1-x} 1_{\{x \in \{0,1\}\}}$$

for some $\theta \in \Theta := [0,1]$. Let $U(\mathbf{X}_n) := \sum_{i=1}^n X_i/n$. Then, by the Weak Law of Large Numbers,

$$U(\mathbf{X}_n) \xrightarrow{P} E[X_1] = \theta.$$

Put $W(\mathbf{X}_n) := (\sum_{i=1}^n X_i + a)/(n+b)$ for some known reals a and b with $b \geq a \geq 0$. This is the Bayesian posterior mode corresponding to the prior $p(\theta) \propto \theta^a(1-\theta)^{b-a}$. By Slutsky's Theorem #2, we have

$$W(\mathbf{X}_n) - U(\mathbf{X}_n) = \left(a - b\sum_{i=1}^n X_i/n\right) \times \frac{1}{(n+b)} \xrightarrow{P} (a - b\theta) \times 0 = 0,$$

so $W(\mathbf{X}_n)$ is also consistent for $\tau(\theta) := \theta$.

Technique #3. Suppose we already know that

$$n^\alpha(W(\mathbf{X}_n) - \tau(\theta)) \xrightarrow{L} Y,$$

where $\alpha \in (0, \infty)$ and Y is some random variable. Then we may conclude that $W(\mathbf{X}_n)$ is consistent for $\tau(\theta)$. To see this, note that

$$W(\mathbf{X}_n) - \tau(\theta) = n^{-\alpha} \times n^\alpha(W(\mathbf{X}_n) - \tau(\theta)) \xrightarrow{L} 0 \times Y = 0$$

by Slutsky's Theorem #3. Technique #2 with $U(\mathbf{X}_n) := \tau(\theta)$ then provides the desired result.

Stay tuned for an example in the next section on large sample behavior of maximum likelihood estimators.

Technique #4. Suppose that the mean square error

$$E\left[\{W(\mathbf{X}_n) - \tau(\theta)\}^2\right] \to 0$$

as $n \to \infty$. For any fixed $\epsilon \in (0, \infty)$, Chebychev's Inequality yields

$$P\left[|W(\mathbf{X}_n) - \tau(\theta)| \geq \epsilon\right] \leq E\left[\{W(\mathbf{X}_n) - \tau(\theta)\}^2\right]/\epsilon^2. \tag{14}$$

Since the right side of (14) tends to 0 as $n \to \infty$, the left side must also tend to 0. This implies that $W(\mathbf{X}_n)$ converges in probability to $\tau(\theta)$, which is the definition of consistency.

As an example, suppose that

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x; \theta) := \theta^{-1} \exp[-x\theta^{-1}] 1_{\{x \in (0,\infty)\}}.$$

We already know that $W(\mathbf{X}_n) := (n-1)/(\bar{X}_n n)$ is an unbiased estimator of $\tau(\theta) := 1/\theta$ with variance $1/[(n-2)\theta^2]$. Variance and mean square error are the same for unbiased estimators, so the mean square error of $(n-1)/(\bar{X}_n n)$ equals $1/[(n-2)\theta^2]$. Since $1/[(n-2)\theta^2]$ tends to 0 as $n \to \infty$, $(n-1)/(\bar{X}_n n)$ is a consistent estimator of $1/\theta$.

The preceding example raises a point that is worth stating explicitly: An unbiased estimator whose variance tends to 0 as $n \to \infty$ is consistent.

f. Large sample behavior of maximum likelihood estimators

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, where $\theta \in \Theta \subset \mathbb{R}$. Under "regularity conditions" on the family $\{f(x;\zeta) : \zeta \in \Theta\}$, we have

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{L} N(0, 1/J_1(\theta)), \tag{15}$$

where $\hat{\theta}_n$ is the maximum likelihood estimator and $J_1(\theta)$ is defined as in formula (6) with n in that formula set to 1.

The regularity conditions are somewhat cumbersome, but that is the price to be paid for a result as general as (15).

Condition #1. The parameter space Θ contains an open interval around θ.

Condition #2. The support set of $X_1, X_2, \ldots, X_n$ does not depend on θ.

Condition #3. First and second partial derivatives of $f(x;\zeta)$ with respect to ζ exist and are continuous for all x in the support set. Moreover, the integrals (for continuous data) or summations (for discrete data) of these partial derivatives over the support set equal 0.

Condition #4. There exists a function $K(x)$ with $E_\theta[K(X)] < \infty$ such that $\left|\frac{\partial^2}{\partial\zeta^2} \log f(x;\zeta)\right| \le K(x)$ for all ζ in some open interval around θ.

Condition #5. The Fisher information $J_1(\theta)$ is finite and positive.

Condition #6. If $\zeta \in \Theta$ satisfies $f(x;\theta) = f(x;\zeta)$ for all x in the support set, then $\theta = \zeta$.

Condition #7. With probability approaching 1 as $n \to \infty$, the maximum likelihood estimator $\hat{\theta}_n$ is the unique root of the first derivative of the log likelihood.

When (15) is applicable, two other useful results can be harvested immediately.

First, by the delta method we have

$$n^{1/2}(\tau(\hat{\theta}_n) - \tau(\theta)) \xrightarrow{L} N(0, \tau'(\theta)^2/J_1(\theta)) \tag{16}$$

for any function τ with continuous first derivative. The interpretation of (16) is that, for large n, $\tau(\hat{\theta}_n)$ behaves like a normal random variable with mean $\tau(\theta)$ and variance $\tau'(\theta)^2/[nJ_1(\theta)] = \tau'(\theta)^2/J_n(\theta)$. Thus, we often say that $\tau(\hat{\theta}_n)$ is asymptotically unbiased and efficient.

Second, by applying Technique #3 from the preceding section on consistency, we see that $\tau(\hat{\theta}_n)$ is consistent for $\tau(\theta)$.

g. Resolution of motivating case studies

Suppose that $X_1, X_2, \ldots, X_n$ are independently and identically normally distributed with mean $\mu \in (-\infty, \infty)$ and variance $\sigma^2 \in (0, \infty)$. Consider estimating σ^2 by

$$W_a(\mathbf{X}) := \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n+a)} \tag{17}$$

for some $a \in (-\infty, \infty)$. Our first motivating case study posed the question of how to choose a.

We have

$$E[W_a(\mathbf{X})] = \frac{(n-1)\sigma^2}{(n+a)} = \sigma^2\left(1 - \frac{a+1}{n+a}\right),$$

so that the only choice of a for which $W_a(\mathbf{X})$ is unbiased is -1.

On the other hand, we have

$$Var[W_a(\mathbf{X})] = \frac{2(n-1)\sigma^4}{(n+a)^2},$$

showing that the variance of $W_a(\mathbf{X})$ is a strictly decreasing function of a.

The mean square error is

$$\sigma^4\left[\frac{2(n-1)}{(n+a)^2} + \frac{(a+1)^2}{(n+a)^2}\right]. \tag{18}$$

One idea is to find the a for which (18) is minimized. The calculus is a little easier if we work with the natural logarithm of (18),

$$4\log\sigma + \log[2(n-1) + (a+1)^2] - 2\log(n+a). \tag{19}$$

Differentiating (19) in a, we obtain

$$2(a+1)/[2(n-1) + (a+1)^2] - 2/(n+a),$$

whose unique root (verify this at home, for practice) is $a = 1$.

Therefore, dividing by n *plus* 1 rather than by n *minus* 1 yields the smallest mean square error! We can also see (verify this at home, for practice) that choosing $a = 0$ yields smaller mean square error than choosing $a = -1$.

Thus, there is some rationale for choosing a other than -1, if we do not insist upon having an unbiased estimator. Indeed, statistical inference often entails a tradeoff between bias and variance. Pages 2 through 6 of Lecture 2 from CPH 636 in Spring 2009 may be interesting further reading.

Yet, the criterion of mean square error is arguably inappropriate when we are estimating σ^2. There is a finite upper bound to how much we can underestimate σ^2, while there is not a finite upper bound to how much we can overestimate σ^2. Accordingly, mean square error is rather forgiving of underestimation. As noted earlier, the alternative criterion (13) is more sensitive to underestimation.

Using criterion (13), we have

$$\begin{aligned} & E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n+a)\sigma^2} - 1 - \log\left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n+a)\sigma^2}\right)\right] \qquad (20) \\ = & \ \frac{(n-1)}{(n+a)} - 1 - E\left[\log\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)\right] + \log(n+a) + 2\log\sigma. \end{aligned}$$

If we are clever, we may figure out some way to evaluate $E\left[\log\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)\right]$.

If we are very clever, we will realize that no such effort is necessary because this quantity does not depend on a and so is not relevant to our choice of a. Indeed, differentiating (20) in a yields

$$-\frac{(n-1)}{(n+a)^2} + \frac{1}{(n+a)},$$

whose unique root (verify this at home, for practice) is $a = -1$.

Hence, the unbiased estimator $W_{-1}(\mathbf{X})$ is preferred by criterion (13), notwithstanding its higher mean square error than $W_1(\mathbf{X})$ and $W_0(\mathbf{X})$.

Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed with probability density function $f(x;\theta)$, where $\theta \in \Theta \subset \mathbb{R}$ is both the mean and median of the distribution. Our second case study posed the question of whether the sample mean $\bar{X}_n$ or the sample median M_n should be used to estimate θ.

Assuming that $X_1, X_2, \ldots, X_n$ have positive finite variance σ^2, the Central Limit Theorem provides

$$n^{1/2}(\bar{X}_n - \theta) \xrightarrow{L} N(0, \sigma^2).$$

We also have

$$n^{1/2}(M_n - \theta) \xrightarrow{L} N\left(0, \frac{1}{4f(\theta;\theta)^2}\right),$$

where $f(\theta;\theta)$ denotes the probability density function $f(x;\theta)$ evaluated at $x = \theta$. Note that, by Technique #3 above, M_n is consistent for θ.

Before coming up with a general answer to the question posed in our second case study, let us consider two examples.

First, suppose that $f(x;\theta) := (2\pi)^{-1/2}\exp[-(x-\theta)^2/2]$. We have $\sigma^2 = 1$ so that

$$n^{1/2}(\bar{X}_n - \theta) \xrightarrow{L} N(0,1).$$

This yields an approximate 95% confidence interval for θ of $\bar{X}_n \pm 1.96/\sqrt{n}$. If we have 64 observations, then the width of the confidence interval is 0.49. On the other hand, we have $f(\theta;\theta) = (2\pi)^{-1/2}$ so that

$$n^{1/2}(M_n - \theta) \xrightarrow{L} N(0,\pi/2).$$

This yields an approximate 95% confidence interval for θ of $M_n \pm 2.46/\sqrt{n}$. If we have 100 observations, then the width of the confidence interval is 0.49.

Thus, for normally distributed data with mean θ and variance $\sigma^2 = 1$, using the sample median to make an inference about θ rather than the sample mean is like throwing away 36% of your observations! Clearly, then, $\bar{X}_n$ is preferable to M_n. Actually, there is nothing special about $\sigma^2 = 1$; this was just a convenient choice for our example. Indeed, for normally distributed data, using the sample median rather than the sample mean is like throwing away 36% of your observations no matter the value of σ^2 (verify this yourself at home, for practice).

Second, suppose that $f(x;\theta) := \exp[-|x-\theta|]/2$. We have $\sigma^2 = 2$ so that

$$n^{1/2}(\bar{X}_n - \theta) \xrightarrow{L} N(0,2),$$

from which we obtain an approximate 95% confidence interval of $\bar{X}_n \pm 2.77/\sqrt{n}$. On the other hand, we have $f(\theta;\theta) = 1/2$ so that

$$n^{1/2}(M_n - \theta) \xrightarrow{L} N(0,1),$$

from which we obtain an approximate 95% confidence interval of $M_n \pm 1.96/\sqrt{n}$. Now the situation has been reversed. For double exponentially distributed data, using the sample mean to make an inference about θ rather than the sample median is like throwing away   % of your observations.

These two examples show that, in general, we cannot regard the sample mean as superior to the sample median or vice versa. However, we can enumerate some simple criteria by which to decide, in a given situation, whether the sample mean or the sample median is preferable.

1. Are they really estimating the same quantity? Many continuous distributions are not symmetric, and so their means differ from their medians. In that case, using the sample median to estimate the mean of the distribution or using the sample mean to estimate the median of the distribution does not make sense.

2. Is the density positive and continuous in a neighborhood of the median? Not all symmetric continuous distributions have densities that are positive and continuous in a neighborhood of the median. One example, already given earlier, is

$$f(x;\theta) := 1_{\{x\in[\theta-1,\theta-1/2]\cup[\theta+1/2,\theta+1]\}}.$$

Result (4) is inapplicable in these cases, so we default to the sample mean.

3. Is the variance finite? Not all symmetric continuous distributions have finite variances, even if they have finite means. One example is the θ-shifted T distribution on 2 degrees of freedom, which has a finite mean but an infinite variance. To see that the variance is infinite, note that the probability density function

$$f(x;\theta) \propto \left(\frac{1}{(x-\theta)^2+1}\right)^{3/2} \approx \frac{1}{x^3}$$

as $x \to \pm\infty$. This implies that $x^2 f(x;\theta) \approx 1/x$ as $x \to \pm\infty$, so that $\int_{\mathbb{R}} x^2 f(x;\theta)\, dx = \infty$. Result (5) is inapplicable in these cases, so we default to the sample median.

4. How large is the density at the median? If the first three criteria do not make the decision for us, then this one will. The variance of the target in (4) will be smaller than the variance of the target in (5), favoring the sample median, if $f(\theta;\theta) > 1/(2\sigma)$. The variance of the target in (4) will be greater than the variance of the target in (5), favoring the sample mean, if $f(\theta;\theta) < 1/(2\sigma)$.

To better understand the fourth criterion, first suppose that $f(\theta;\theta)$ is large. Then $X_1, X_2, \ldots, X_n$ are likely to take values close to θ. With an abundance of observations in proximity to the median, estimating the median precisely will be easy. Now suppose that $f(\theta;\theta)$ is small. Then $X_1, X_2, \ldots, X_n$ are unlikely to take values close to θ. With a dearth of observations in proximity to the median, estimating the median precisely will be difficult.