PLSC 502: "Statistical Methods for Political Research"

Estimation and Properties of Estimators
October 11, 2016

Random Variables, Take Two

One way to think of a random variable is as made up of two parts: a systematic component and a random part. E.g:

$$X_i = \mu + u_i$$

This implies something about u:

$$u_i = X_i - \mu$$

Q: What is the expected value of u?

A:

$$\begin{aligned} \mathrm{E}(u) &= \mathrm{E}(X - \mu) \\ &= \mathrm{E}(X) - \mathrm{E}(\mu) \\ &= \mathrm{E}(X) - \mu \\ &= \mu - \mu \\ &= 0 \end{aligned}$$

This makes sense. Think of what a mean is: the number such that, if it is subtracted from each value of X in the sample, the sum of those differences will be zero...

What about the variance of X and u?

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}[(X - \mu)^2] \\ &= \mathrm{E}(u^2) \end{aligned}$$

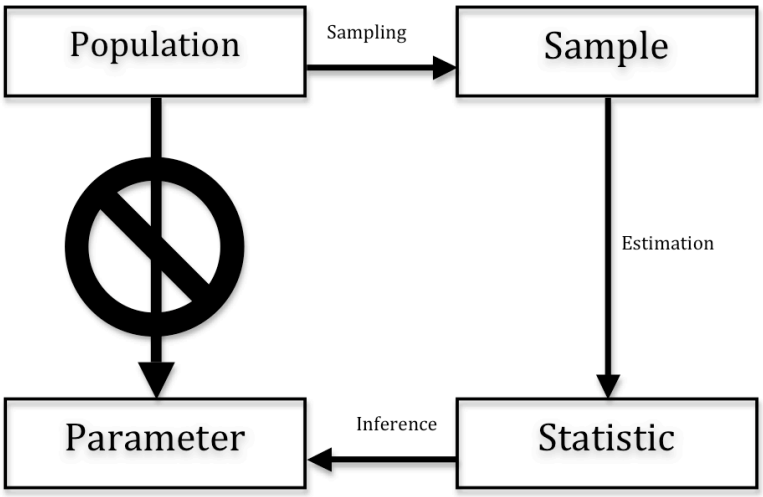
$$\begin{aligned} \mathrm{Var}(u) &= \mathrm{E}[(u - E(u))^2] \\ &= \mathrm{E}[(u - 0)^2] \\ &= \mathrm{E}(u^2) \end{aligned}$$

Essentially, this is simply showing that, if we define a random variable as composed of a fixed part and a random part, then:

- That variable will have a population mean (i.e. $\mathrm{E}(X)$) equal to μ, and
- The variance of X is equal to the variance of u.

Estimation: Concepts

You remember this figure:



We talked about the first step – *sampling* – last week. Today and going forward, we'll be focusing on estimation and inference. FYI, I'll usually use greek letters for population parameters (like θ), and letters with "hats" ($\hat{\theta}$) for specific data-based *estimates* of those parameters.

When we take a sample, we're drawing "realizations" from the underlying distribution of that variable. This means that, since X is a random variable, then ***any estimate we make is also a random variable***. (As I mentioned last time, the preceding statement is arguably the most important thing you should take away from this entire class...).

A Simple Example

Consider the lowly mean (μ)...

- We want to know μ for the population, but
- we only have data on a sample of N observations from the population, so
- we use these data to *estimate* the mean.

How do we do that? Well...

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1}$$

Now, recalling that each $X_i = \mu + u_i$, then we can write:

$$\begin{aligned}
\bar{X} &= \frac{1}{N}\sum_{i=1}^{N}(\mu + u_i) \\
&= \frac{1}{N}\sum_{i=1}^{N}(\mu) + \frac{1}{N}\sum_{i=1}^{N}(u_i) \\
&= \frac{1}{N}(N\mu) + \frac{1}{N}\sum_{i=1}^{N}(u_i) \\
&= \mu + \bar{u}
\end{aligned}$$

What does this mean?

- *The estimate of the mean is itself a random variable*,
- For different samples, we'll get different values of $\bar{u}$ (the sample-based "average" of the stochastic component of X), and correspondingly different estimates of the mean,
- *All* of this (stochastic) variation is again due to the random component of X.

We could show in analogous fashion that the usual estimate of the variance σ^2 (usually denoted s^2) is also a random variable...

Properties of Estimators

There are, in theory, a lot of different estimators... Which one(s) we might choose to use depends critically on their *properties*. There are two general types of properties of estimators:

Small-Sample Properties

- These properties hold irrespective of the size of the sample on which the estimate is based.
- In other words, in order for a estimator to have these properties, they must hold for all possible sample sizes.

Large-Sample (Asymptotic) Properties

- These are properties which hold only as the sample size increases to infinity.
- In practical terms, it means that to receive the benefits of these properties, "more is better" (at least as far as sample size goes).

As a running example, consider an abstract population parameter θ; this might be a mean, a correlation, whatever. Assume we estimate it with a sample of N observations; call this generic estimator $\hat{\theta}$.

Unbiasedness

We generally prefer that estimators be "accurate;" that is, that they reflect the population parameter as closely as possible. I.e. we would prefer that

$$\mathrm{E}(\hat{\theta}) = \theta.$$

If this property holds, then we say an estimator is *unbiased*. That is, **an unbiased estimator is one for which its expected value is the population parameter**. That, in turn, means that we can think of the *bias* in an estimator as:

$$B(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta \tag{2}$$

That suggests that another way to say that an estimator is unbiased is if $B(\hat{\theta}) = 0$.

But how do we know if an estimate is unbiased? Sometimes we can *prove* or demonstrate it. Consider again the mean:

$$\begin{aligned} \mathrm{E}(\bar{X}) &= \mathrm{E}(\mu + \bar{u}) \\ &= \mathrm{E}(\mu) + \mathrm{E}(\bar{u}) \\ &= \mu + 0 \\ &= \mu \end{aligned}$$

So the sample mean is an unbiased estimate of the population mean.

In other cases, it can be very difficult, or even impossible, to show that an estimator is unbiased. Moreover, there might be many, many unbiased estimators for a particular population parameter...

Example: Consider a sample of two observations X_1 and X_2, and a generalized estimator

$$Z = \lambda_1 X_1 + \lambda_2 X_2.$$

Note that

$$\begin{aligned} \mathrm{E}(Z) &= \mathrm{E}(\lambda_1 X_1 + \lambda_2 X_2) \\ &= \mathrm{E}(\lambda_1 X_1) + \mathrm{E}(\lambda_2 X_2) \\ &= \lambda_1 \mathrm{E}(X_1) + \lambda_2 \mathrm{E}(X_2) \\ &= \lambda_1 \mu + \lambda_2 \mu \\ &= (\lambda_1 + \lambda_2)\mu \end{aligned}$$

So long as $(\lambda_1 + \lambda_2) = 1.0$, then $\mathrm{E}(Z) = \mu$ and the estimator is unbiased.

- This means that there are in principle an infinite number of unbiased estimators.
- We could extend this to N observations: So long as the sum of the "weights" add up to 1.0, the estimate is unbiased.

So why do we always use the "mean" (i.e. give equal weights $\frac{1}{N}$ to each observation)? The answer has to do with...

Efficiency

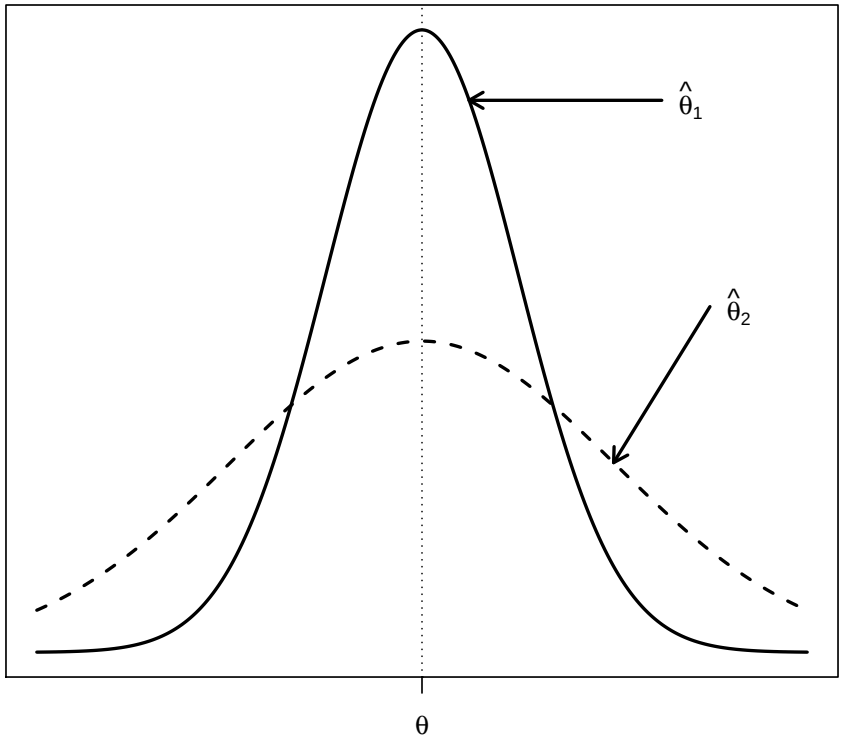
In addition to preferring an estimator that is unbiased, we also want one that has smaller, rather than larger, variance.

Think of this as "reliability:" in repeated samples, an unbiased estimator which is more likely to achieve the true population value (i.e. which has smaller variance) is said to be more *efficient* than one with larger variance.[1]

Note that:

- A fully efficient estimator *must be unbiased*;
- It's possible that a biased estimator has smaller variance, but that doesn't make it efficient.

[1]Note that another way of talking about efficiency is as an estimator where the variance of an estimator is equal to the reciprocal of the Fisher information from the sample; we'll talk about this a bit more when we get to likelihood theory and the like...



In the figure, θ_1 is more efficient than θ_2.

In the context of our example above, consider the variance of the possible estimator Z:

$$\begin{aligned} \text{Var}(Z) &= \text{Var}(\lambda_1 X_1 + \lambda_2 X_2) \\ &= (\lambda_1^2 + \lambda_2^2)\sigma^2 \end{aligned}$$

We want to know what combination minimizes this variance. Since we know that $\lambda_1 + \lambda_2 = 1.0$, we can rewrite:

$$\begin{aligned} \lambda_1^2 + \lambda_2^2 &= \lambda_1^2 + (1 - \lambda_1)^2 \\ &= \lambda_1^2 + (1 - 2\lambda_1 + \lambda_1^2) \\ &= 2\lambda_1^2 - 2\lambda_1 + 1 \end{aligned}$$

We then minimize (how?):

$$\begin{aligned} 4\lambda_1 - 2 &= 0 \\ \lambda_1 &= 0.5 \end{aligned}$$

So the (equally-weighted) sample average has the smallest variance of all the possible unbiased estimators for the population mean. That is, it is *efficient*.

Q: How do we know if an estimator is most efficient? (i.e., more efficient than any other estimator)?

A: We usually don't.

- It can be difficult to determine efficiency, but
- Knowing is made easier if we restrict our "search" to linear estimators – that is, estimators which are some linear combinations of the sample values.
- If an estimator is the unbiased linear estimate with minimum variance, we say it is BLUE ("Best" Linear Unbiased Estimator).

Mean Squared Error

Occasionally we may decide to "trade off" some bias in favor of gains in efficiency. This is the idea behind the "mean squared error" (MSE) criterion for choosing an estimator. Mathematically, the MSE is:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \text{E}[(\hat{\theta} - \theta)^2] \\ &= \text{E}[B(\hat{\theta})^2] \end{aligned} \tag{3}$$

that is, the MSE is the same as the expected squared bias of an estimator. It can be shown that

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [B(\hat{\theta})^2],$$

that is, that the MSE is equal to the sum of the variance of $\hat{\theta}$ plus the square of its bias. (To understand why, recall from last week that $\text{E}(X^2) = \sigma^2 + \mu^2$).

That means that the MSE criterion "balances" bias and efficiency; a biased estimator with a small variance might easily have a smaller MSE than an unbiased estimator (with a much larger variance). At the same time, among unbiased estimators, the MSE will always be lowest for the efficient estimator, if one exists (since, by definition, unbiasedness means that $B(\hat{\theta}) = [B(\hat{\theta})]^2 = 0$).

MSE and Choice of Estimators: A Silly Example

Consider a silly example: We want to estimate μ, the expected value / "mean" value of X in some population, with a sample size of N. One alternative is the mean ($\bar{X}$), as defined above, which has:

- $B(\bar{X}) = 0$ (because the mean is an unbiased estimator of μ), and
- $\text{Var}(\bar{X}) = \sigma^2/N$, where σ^2 is the variance of X. This means that
- $\text{MSE}(\bar{X}) = \sigma^2/N + (0)^2 \; = \; \sigma^2/N$.

Now consider a lazy, alternative estimator, λ, defined as

$$\lambda = 6 \tag{4}$$

In other words, this estimator says that its guess of the expectation of X is always equal to six. Period. This is in many respects a rotten estimator, but let's see how it does in terms of MSE.

The bias of λ, $B(\lambda)$, is:

$$\begin{aligned} B(\lambda) &= \text{E}(\lambda - \mu) \\ &= \text{E}(6) - \text{E}(\mu) \\ &= 6 - \mu \end{aligned}$$

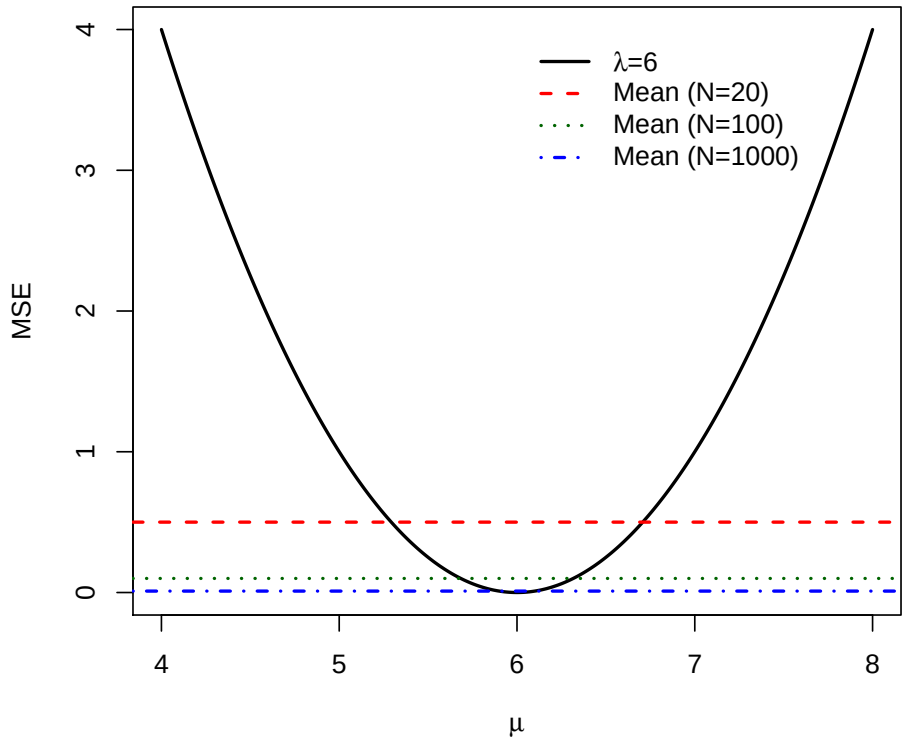
while its variance is:

$$\begin{aligned} \text{Var}(\lambda) &= \text{Var}(6) \\ &= 0 \end{aligned}$$

Thus, the MSE of λ is:

$$\begin{aligned} \text{MSE}(\lambda) &= \text{Var}(\lambda) + [B(\lambda)]^2 \\ &= 0 + (6 - \mu)^2 \\ &= 36 - 12\mu + \mu^2 \end{aligned}$$

We can compare this MSE graphically – for different values of the population mean μ – to the MSE of our "usual" estimator of the mean $\bar{X}$ with three different sample sizes:



The black line is the MSE of λ, expressed as a function of the "true" population mean μ. The other colored lines are the MSEs for $\bar{X}$, under the assumption that $\sigma^2 = 10$ and $N = \{20, 100, 1000\}$, respectively.

Notice several things:

- The MSE of λ is actually quite good if $\mu \approx 6$; thus, in some circumstances, the MSE for λ will be smaller than that for $\bar{X}$, even though $\bar{X}$ is both unbiased and clearly a "better" estimator (see the figure). However,
- ...it gets much worse as μ gets farther away from six. Since we don't know whether $\mu = 6$ or not (or else, why would we bother estimating it?), this is not a desirable property.
- Relatedly, our estimator λ doesn't "improve" in MSE terms if we add more data to our sample (that is, as $N \to \infty$). In contrast,
- The MSE of $\bar{X}$ drops considerably as N increases, and does so irrespective of the "true" value of μ (see the colored lines in the figure).

All of this is designed to underscore the point that, while MSE can be a fine way to choose among estimators, it shouldn't be applied uncritically.

Large-Sample Properties

Both unbiasedness and efficiency are *small-sample* properties of estimators, that hold irrespective of sample size.

In contrast, *large-sample properties* are properties of estimators that hold only as the sample size increases without limit.

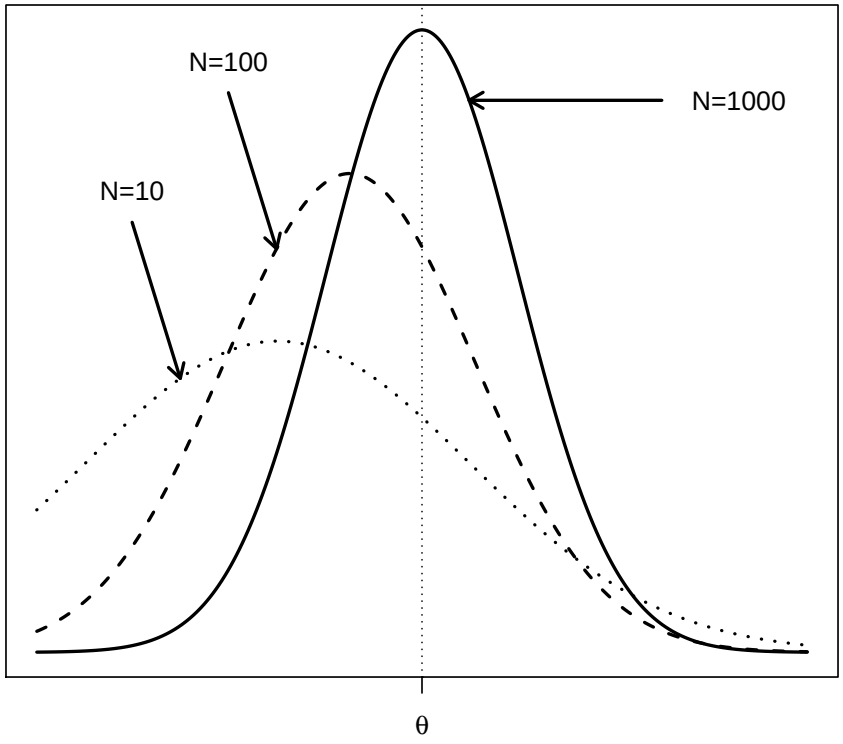
- Note that this is dependent on *sample size*, not on the "number" of samples drawn...
- Intuitively: what would you expect to happen as sample size gets larger?
    - The variance around the "true" value decreases (less possibility of drawing a "bad" sample).
- Eventually the sample size = population, and the estimate "collapses on the true value.

Consistency

Think of consistency as "asymptotic unbiasedness." Formally, an estimator is consistent if it converges in probability to its population value as N goes to infinity; we can write

$$\lim_{N \to \infty} \Pr[|\hat{\theta} - \theta| < \epsilon] = 1.0 \tag{5}$$

for an arbitrarily small $\epsilon > 0$. If we consider an estimator whose properties vary by sample size (say, $\hat{\theta}_N$), then $\hat{\theta}_N$ is consistent if $\mathrm{E}(\hat{\theta}_N) \to \theta$ as $N \to \infty$.



The idea of consistency underscores the point that some estimators can be biased in small samples, but get less and less so as the sample size increases. An example is the maximum-likelihood estimator of the variance:

$$s^2_{ML} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2 \tag{6}$$

This is a biased estimator of the population variance σ^2; the unbiased estimator is:[2]

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 \tag{7}$$

Asymptotically, however, it is easy to see that as $N \to \infty$, the bias disappears. Thus, s^2_{ML} is a *biased* but *consistent* estimator of σ^2 (as, in fact, all ML estimators are).

Asymptotic Efficiency

Asymptotic efficiency can be thought of as efficiency as $N \to \infty$. It is intuitive to think of this as the "speed" with which $\hat{\theta}$ "collapses" on θ: all else equal, we prefer one that does so

[2]Note as well that – among the class of estimators of the form $c\sum_{i=1}^{N}(X_i - \bar{X})^2$ – neither of these are the estimator with the smallest MSE, which is $s^2_{MSE} = \frac{1}{N+1}\sum_{i=1}^{N}(X_i - \bar{X})^2$.

faster (i.e. for smaller sample sizes) rather than more slowly. There's not much more need to go into this; it's a pretty straightforward idea.

General Issues

In general,

- We prefer estimators that have desirable small-sample properties.
    - We prefer unbiased to consistent estimators, and
    - We prefer efficient/BLUE estimators to asymptotically efficient ones.

but...

- We can't always figure out the small sample properties of certain estimators, and/or
- Our estimators with desirable small-sample properties may have other problems (e.g. computational cost).

As a result, we often have to choose among estimators that differ in their degree of desirable properties. And, as with all things, the significance of all this will become clearer once we start taking about OLS next semester.