## 2.1 Introduction

Properties of estimators are divided into two categories; small sample and large (or infinite) sample. These properties are defined below, along with comments and criticisms. Four estimators are presented as examples to compare and determine if there is a "best" estimator.

## 2.2 Finite Sample Properties

The first property deals with the mean location of the distribution of the estimator.

**P.1 Biasedness** - The bias of on estimator is defined as:

$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta,$

where $\hat{\theta}$ is an estimator of θ, an unknown population parameter. If $E(\hat{\theta}) = \theta$, then the estimator is unbiased. If $E(\hat{\theta}) \neq \theta$ then the estimator has either a positive or negative bias. That is, on average the estimator tends to over (or under) estimate the population parameter.

A second property deals with the variance of the distribution of the estimator. Efficiency is a property usually reserved for unbiased estimators.

**P.2 Efficiency** - Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be unbiased estimators of θ with equal sample sizes[1]. Then, $\hat{\theta}_1$ is a more efficient estimator than $\hat{\theta}_2$ if $\text{var}(\hat{\theta}_1) < \text{var}(\hat{\theta}_2)$.

Restricting the definition of efficiency to unbiased estimators, excludes biased estimators with smaller variances. For example, an estimator that always equals a single number (or a constant) has a variance equal to zero. This type of estimator could have a very large bias, but will always have the smallest variance possible. Similarly an estimator that multiplies the sample mean by [n/(n+1)] will underestimate the population mean but have a smaller variance. The definition of efficiency seems to arbitrarily exclude biased estimators.

One way to compare biased and unbiased estimators is to arbitrarily define a measuring device that explicitly trades off biasedness with the variance of an estimator. A simple approach

[1] Some textbooks do not require equal sample sizes. This seems a bit unfair since one can always reduce the variance of an estimator by increasing the sample size. In practice the sample size is fixed. It's hard to imagine a situation where you would select an estimator that is more efficient at a larger sample size than sample size of your data.

is to compare estimators based on their mean square error. This definition, though arbitrary permits comparisons to be made between biased and unbiased estimators

**P.3 Mean Square Error -** The mean square error of an estimator is defined as

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= Var(\hat{\theta}) + [Bias(\hat{\theta})]^2 \end{aligned}$$

The above definition arbitrarily specifies a one to one tradeoff between the variance and squared bias of the estimator. Some, (especially economists) might question the usefulness of the MSE criteria since it is similar to specifying a unique preference function. There are other functions that yield different rates of substitution between the variance and bias of an estimator. Thus it seems that comparisons between estimators will require specifications of an arbitrary preference function.

Before proceeding to infinite sample properties some comments are in order concerning the use of biasedness as a desirable property for an estimator. In statistical terms, unbiasedness means that the expected value of the distribution of the estimator will equal the unknown population parameter one is attempting to estimate. Classical statisticians tend to state this property in frequency statements. That is, on average $\hat{\theta}$ is equal to $\theta$. As noted earlier, when defining probabilities, frequency statements apply to a set of outcomes but do not necessarily apply to a particular event. In terms of estimators, an unbiased estimator may yield an incorrect estimate (that is $\hat{\theta} \neq \theta$) for every sample but on average yield a correct or unbiased estimator (i.e. $E(\hat{\theta}) = \theta$). A simple example will illustrate this point.

Consider a simple two outcome discrete probability distribution for a random variable X where

$$X = \begin{array}{cc} X_i & P(Xi) \\ \hline \mu + 5 & 0.5 \\ \mu - 5 & 0.5 \end{array}$$

It is easy to show that $E(X) = \mu$ and $var(X) = 25$.

To make this example more interesting assume that X is a random variable describing the outcomes of a radar gun used by a police officer to catch drivers exceeding the speed limit. The

radar gun either records the speed of the driver as 5-mph too fast or 5-mph too slow[2]. Suppose the police officer takes a sample equal to one. Clearly the estimator from the radar gun will be incorrect since it will either be 5-mph too high or 5-mph too low. Since the estimator overstates by 5-mph half the time and understates by 5-mph the other half of the time, the estimator is unbiased even though for a single observation it is always incorrect.

Suppose we increase the sample size to two. Now the distribution of the sample mean is:

| | | $X_i$ | $P(X_i)$ |
|---|---|---|---|
| | | $\mu + 5$ | 0.25 |
| $\overline{X} = (X_1 + X_2)/2$ | = | $\mu$ | 0.50 |
| | | $\mu - 5$ | 0.25 |

The radar gun will provide a correct estimate ( i.e. $P(\overline{X}) = \mu$) 50% of the time.

As we increase n, the sample size, the following points can be made. If n equals an odd number $\overline{X}$ can never equal $\mu$ since the number of (+5)'s cannot equal the number of (-5)'s. In the case where n is an even number, $\overline{X} = \mu$ only when the number of (+5)'s and (-5)'s are equal. The probability of this event declines and approaches zero as n becomes very large[3].

In the case when X is a continuous probability distribution it is easy to demonstrate that $P(\overline{X} = \mu) = 0$. A continuous distribution must have an area (or mass) under the distribution in order to measure a probability. The $P(|\overline{X} - \mu| < e)$ may be positive (and large) but $P(\overline{X} = \mu)$ must equal zero.

To summarize, unbiasedness is not a desirable property of an estimator since it is very likely to provide an incorrect estimate from a given sample. Furthermore, an unbiased estimator may have an extremely large variance. It's unclear how an unbiased estimator with a large variance is useful. To restrict the definition of efficiency to unbiased estimators seems arbitrary and perhaps not useful. It may be that some biased estimators with smaller variances are more helpful in estimation. Hence, the MSE criterion, though arbitrary, may be useful in selecting an estimator.

## 2.3 Infinite Sample Properties

---

[2] The size of the variance is arbitrary. A radar gun like a speedometer estimates velocity at a point in time. A more complex probability distribution (more discrete outcomes or continuous) will not alter the point that unbiasedness is an undesirable property.

[3] Assuming a binomial distribution where $\pi = 0.50$, the sampling distribution is symmetric around $X_i = n/2$, the midpoint. As n increases to (n+2), the next even number, the probability of $X_i = n/2$ decreases in relative terms by (n+1)/(n+2).

Large sample properties may be useful since one would hope that larger samples yield better information about the population parameters. For example, the variance of the sample mean equals $\sigma^2/n$. Increasing the sample size reduces the variance of the sampling distribution. A larger sample makes it more likely that $\bar{X}$ is closer to $\mu$. In the limit $\sigma^2/n$ goes to zero.

Classical statisticians have developed a number of results and properties when n gets larger. These are generally referred to as asymptotic properties and take the form of determining a probability as the sample size approaches infinity. The Central Limit Theorem (CLT) is an example of such a property. There are several variants of this theorem, but generally they state that as the sample size approaches infinity, a given sampling distribution approaches the normal distribution. The CLT has an advantage over the previous use in applying the limit to the frequency definition of a probability. At least in this case, the limit of a sampling distribution can be proven to exist, unlike the case where the limit of (K/N) is assumed to exist and approach P(A). Unfortunately, knowing the limit that all sampling distributions are normal may not be useful since all sample sizes are finite. Some known distributions (e.g. Poisson, Binomial, Uniform) may visually appear to be normal as n increases. However, if the sampling distribution is unknown, how does one know and determine how close a sampling distribution is to the normal distribution? Oftentimes, it is convenient to assume normality so that the sample mean is normally distributed[4]. If the distribution of $X_i$ is unknown, it's unclear how one describes the sampling distribution for a finite sample size and then assert that normality is a close approximation?

One of my pet peeves are instructors that assert normality for student test scores when there is a large (n>32) sample. Some instructors even calculate z-scores (with $\sigma$ unknown, how is it's value determined?) and make inferences based on the normal distribution (e.g. 95% of the scores will fall within $2\sigma$ of $\bar{X}$). Assuming students have different abilities, what if the sample scores are bimodal? The sampling distribution may appear normal, but for a given sample it seems silly to blindly assume normality[5].

## 2.4 An Example

[4] Most textbooks seem to teach that as long as n > 32, one can assume a normal sampling distribution. These texts usually point out the similarity of the t and z distribution when the degrees of freedom exceed 30. However, this is misleading since the t-distribution depends on the assumption of normality.

[5] The assumption of normality is convenient, but may not be helpful in forming an inference from a given sample.

The examples below will compare the usefulness of four estimators.[6] For convenience assume that $X_i \sim N(\mu, \sigma^2)$. Four estimators are specified as:

I. $\hat{\mu}_1 = \bar{X} = \sum X_i / n$

II. $\hat{\mu}_2 = \mu^*$

III. $\hat{\mu}_3 = w^* (\mu^*) + \bar{w} (\bar{X})$, where $\bar{w} = (1 - w^*) = n/(n + n^*)$

IV. $\hat{\mu}_4 = \bar{X}^* = \sum X_i / (n+1)$,

where $\mu^*$ and n* are arbitrarily chosen values (n* > 0). The first estimator is the sample mean and has the property of being BLUE, the best (most efficient) linear unbiased estimator. The second estimator picks a fixed location $\mu^*$, regardless of the observed data. It can be thought of as a prior location for $\mu$ with variance equal to zero. The third estimator is a weighted average of the sample mean and $\mu^*$. The weights add up to one and will favor either location depending on the relative size of n and n*[7]. The fourth estimator is similar to $\bar{X}$, except that the sum of the data are divided by (n + 1) instead of n.

The data are drawn from a normal distribution which yield the following sampling distributions:

I. $\hat{\mu}_1 \sim N[\mu, \sigma^2/n]$

II. $\hat{\mu}_2$ : $E(\hat{\mu}_2) = \mu^*$ and $var(\hat{\mu}_2) = 0$

III. $\hat{\mu}_3 \sim N[\, w^* \mu^* + \bar{w} \mu \,,\, \bar{w}^2 (\sigma^2/n) \,]$

IV. $\hat{\mu}_4 \sim N[\, (n/(n+1))\mu, (n/(n+1))^2 (\sigma^2/n) \,]$

From the above distributions it is easy to calculate the bias of each estimator:

---

6 The first three estimators are similar to ones found in Leamer and MM.

7 The value of n* can be thought of as a weight denoting the likelihood that $\mu^*$ is the correct location for $\mu$.

I. $\text{Bias}(\hat{\mu}_1) = E(\hat{\mu}_1) - \mu = \mu - \mu = 0$

II. $\text{Bias}(\hat{\mu}_2) = E(\hat{\mu}_2) - \mu = (\mu^* - \mu) \begin{matrix} > \\ = \\ < \end{matrix} 0 \text{ as } \mu^* \begin{matrix} > \\ = \\ < \end{matrix} \mu$

$\text{Bias}(\hat{\mu}_3) = E(\hat{\mu}_3) - \mu = (w^* \mu^* + \bar{w}\mu) - \mu$

$= (w^* \mu^* + (1-w^*)\mu) - \mu$

$= w^*(\mu^* - \mu) \begin{matrix} > \\ = \\ < \end{matrix} 0 \text{ as } \mu^* \begin{matrix} > \\ = \\ < \end{matrix} \mu$

$\text{Bias}(\hat{\mu}_4) = E(\hat{\mu}_4) - \mu = (n/(n+1))\mu - \mu = -\mu/(n+1) < 0 \text{ if } \mu > 0$

The sample mean is the only unbiased estimator. The second and third estimators are biased only if $\mu^* \neq \mu$ and may yield a very large bias depending on how far $\mu^*$ is from $\mu$. For the third estimator the size of n* relative to n, will also influence the bias. As long as $\mu^*$ receives some weight ($n^* > 0$), $\hat{\mu}_3$ will combine the sample data and the prior location and choose an estimate between $\mu^*$ and $\mu$, which is biased. The fourth estimator is biased so long as $\mu \neq 0$.

In a similar manner one can compare the variance of the four estimators. Since $\hat{\mu}_2$ is a constant it has the smallest possible variance equal to zero. For comparison purposes, we calculate the ratio of the variances for two estimators.

I. $\text{Var}(\hat{\mu}_4)/\text{Var}(\hat{\mu}_1) = [(n/(n+1))^2 (\sigma^2/n)/(\sigma^2/n)] = (n/(n+1))^2 < 1$

II. $\text{Var}(\hat{\mu}_3)/\text{Var}(\hat{\mu}_1) = [(\bar{w})^2 (\sigma^2/n)/(\sigma^2/n)] = (\bar{w})^2 < 1$

$\text{Var}(\hat{\mu}_3)/\text{Var}(\hat{\mu}_4) = [(\overline{w})^2 \ (\sigma^2/n)/ \ (n/(n+1)^2 \ (\sigma^2/n)]$

$= [(n/(n+n^*)]^2/[\ n/(n+1)]^2$

$= [(n+1)/(n+n^*)]^2 < 1$ if $n^* > 1$

In terms of overall rankings we have

$\text{Var}(\hat{\mu}_1) > \text{Var}(\hat{\mu}_4) > \text{Var}(\hat{\mu}_3) > \text{Var}(\hat{\mu}_2) = 0$

As noted earlier, the sample mean has the smallest variance when compared with unbiased estimators, but has a larger variance when compared to simple biased estimators. If we use the MSE criterion to compare estimators we have:

I. $\text{MSE}(\hat{\mu}_1) = \text{Bias}(\hat{\mu}_1)^2 + \text{Var}(\hat{\mu}_1)$

$= \ 0 \ + \ \sigma^2/n$

$= \ \sigma^2/n$

$\text{MSE}(\hat{\mu}_2) = (\mu^* - \mu)^2$

$\text{MSE}(\hat{\mu}_3) = w^{*2}(\mu^* - \mu)^2 \ + (1-w^*)^2 \ (\sigma^2/n)$

$= (\sigma^2/n)[\ (1-w^*) + w^{*2}(\mu^* - \mu)^2/((\sigma^2/n) \ ]$

IV. $\text{MSE}(\hat{\mu}_4) = = (-\mu/(n+1))^2 + (n/(n+1))^2 \ (\sigma^2/n)$

Making the comparisons relative to $\hat{\mu}_1$ we have,

I. $MSE(\hat{\mu}_1)/MSE(\hat{\mu}_1) = 1$

II. $MSE(\hat{\mu}_2)/MSE(\hat{\mu}_1) = (\mu^* - \mu)^2/(\sigma^2/n) = Z^{*2}$

III. $MSE(\hat{\mu}_3)/MSE(\hat{\mu}_1) = [\ 1 + (n^*/n)^2\ Z^{*2}\ ]/[\ 1 + (n^*/n)^2\ ]$

IV. $MSE(\hat{\mu}_4)/MSE(\hat{\mu}_1) = (n/(n+1))^2 + Z_o^2/(n+1)$

Figure 3.6.1 graphs the relative MSE of the first three estimators. The fixed location estimator, $\hat{\mu}_2$, dominates the sample mean as long as $\mu^*$ is within one standard deviation of $\mu$. The third estimator will dominate the sample mean by a wider margin. The graph of $MSE(\hat{\mu}_3)/MSE(\hat{\mu}_1)$ assumes a weight of 40% for the prior location ($w^* = 0.40$). The estimator $\hat{\mu}_3$ will dominate the sample mean within an area of two standard deviations. In other words, if one is confident that a prior location can be selected within two standard deviations of an unknown population parameter (and $w^* > 0.40$), an estimator that incorporates the sample mean and the prior fixed location will do a better job of estimation