V. Properties of estimators

A. Definitions & Desiderata

1. model
2. estimator
3. sampling errors and sampling distribution
4. unbiasedness
5. low sampling variance
6. low mean squared error (MSE)
   i. definition of MSE
   ii. implied loss function on errors
   iii. relationship to variance and bias
7. evocative simple examples

B. Finite sample (exact) properties (also called “small sample properties,” although equally valid for large samples)

1. relative efficiency
2. best linear unbiasedness (BLUness)
3. efficiency – definition and relationship to BLUness

C. Large Sample or Asymptotic properties

1. limiting distribution vs. limit of sampling distribution
2. asymptotic variance
3. asymptotic mean & asymptotic unbiasedness
4. consistency
   i. Consistency through vanishing limit of MSE (Squared Error Consistency)
   ii. probability limit (plim)
      - definition & relation to consistency
      - theorems for manipulating/evaluating probability limits
      - illustrative examples
5. asymptotic efficiency – definition

D. Minimum MSE linear estimator
V.A. Definitions & Desiderata

1. model

\[ y = \text{random variable we seek to investigate using sample data } y_1 \ldots y_N \]

-- since \( y \) is random, the most we can know about it is its probability distribution

\textbf{model} = everything we assume we already know about the distribution of \( y \) and about how it was sampled to obtain \( y_1 \ldots y_N \).

for example: in previous section of the course model was \( y_i \sim \text{NIID}(\mu, \sigma^2) \) with \( \sigma^2 \) given.

Let \( \theta \) be the name we use for one of the unknown parameters that specifies the distribution of \( y \).

-- for example, \( \mu \) in previous section

-- if there are \( k \) unknown parameters, use \( \theta_1, \theta_2, \ldots \theta_k \)

2. estimator

\[ \hat{\theta} \quad \hat{\theta} \]

-- typically called \( \theta \) or \( \theta( y_1 \ldots y_N ) \)

or \( \tilde{\theta} \) if we need symbol for second estimator of \( \theta \)

\[ \hat{\theta} \quad \hat{\theta} \quad \hat{\theta} \quad \hat{\theta} \]

or \( \theta_1, \theta_2, \ldots \theta_k \) if there are \( k \) unknown parameters

-- a function of the sample data (and perhaps other things)

\textbf{does not} depend on \( \theta \) (or any other unknown parameter)

\[ \bar{y} \text{ depends on } y_1 \ldots y_N \text{ but } \textbf{not} \text{ on } \mu \]
3. sampling error

\[ ^\wedge \theta - \theta \]

= error made by estimator

sampling distribution of \[ ^\wedge \theta \] = distribution of \[ ^\wedge \theta \]

sampling error distribution = distribution of \[ ^\wedge \theta - \theta \]

4. unbiasedness

\[ ^\wedge \text{bias}(\theta) \text{ defined to be } E\{ ^\wedge \theta \} - \theta \]

\[ ^\wedge \theta \text{ is unbiased if and only if } \text{bias}(\theta) = 0 \]

importance: the bias(\theta) is typically known only when it is zero. Known bias is necessary for doing inference on \theta, since inference requires standardizing \theta to zero mean:

\[ E\{ ^\wedge \theta - ^\wedge \theta - \text{bias}(\theta) \} = 0 \]

5. low sampling variance

e.g. length of 95% confidence interval for \theta is proportional to var(\theta)

6. low Mean Squared Error or MSE

\[ \text{var}(\hat{\theta}) = \text{dispersion of } \hat{\theta} \text{ around its population mean} = E\{(\hat{\theta} - E\{\hat{\theta}\})^2\} \]

\[ \text{MSE}(\hat{\theta}) = \text{dispersion of } \hat{\theta} \text{ around true value of } \theta = E\{(\hat{\theta} - \theta)^2\} \]
If the way we care about the errors made by the estimator \( \hat{\theta} \) (or if the costs or losses associated with errors in estimating \( \theta \)) can be expressed as:

\[
\text{Loss}(\hat{\theta} - \theta) = \{\text{positive constant}\} (\hat{\theta} - \theta)^2 \quad \text{(squared error loss function)}
\]

then MSE(\( \hat{\theta} \)) is proportional to the expected loss or cost associated with using \( \hat{\theta} \). Note that

a. other reasonable loss functions are available, such as

\[
\text{Loss}(\hat{\theta} - \theta) = \{\text{positive constant}\} (\hat{\theta} - \theta)^2 \quad \text{for positive errors}
\]

\[
= \{\text{different positive constant}\} (\hat{\theta} - \theta)^2 \quad \text{for negative errors}
\]

which penalizes over-estimates and under-estimates unequally and

\[
\text{Loss}(\hat{\theta} - \theta) = \{\text{positive constant}\} |\hat{\theta} - \theta| \quad \text{(absolute value of error loss function)}
\]

which still penalizes over-estimates and under-estimates equally, but no longer give quickly increasing weight to errors which are larger in magnitude. But these loss functions are not commonly used because they are tremendously harder to deal with than the squared error loss function. So low MSE is almost always our key criterion for what constitutes a "good" estimator.

Evidently, \( \text{var}(\hat{\theta}) = \text{MSE}(\hat{\theta}) \) for unbiased estimators. More specifically:\(^1\)

---

\(^1\)Note that this derivation uses the fact that bias(\( \theta \)) is fixed – i.e., non-random.
\[ \text{MSE}(\hat{\theta}) = E\{(\hat{\theta} - \theta)^2\} \]
\[ = E\{(\hat{\theta} - E\{\hat{\theta}\} + E\{\hat{\theta}\} - \theta)^2\} \]
\[ = E\{[\hat{\theta} - E\{\hat{\theta}\}] + [E\{\hat{\theta}\} - \theta]\]^2\} \]
\[ = E\{[\hat{\theta} - E\{\hat{\theta}\}]^2 + 2[ E\{\hat{\theta}\} - \theta][\hat{\theta} - E\{\hat{\theta}\}] + [E\{\hat{\theta}\} - \theta]^2\} \]
\[ = E\{[\hat{\theta} - E\{\hat{\theta}\}]^2 + 2\text{bias}(\hat{\theta})[\hat{\theta} - E\{\hat{\theta}\}] + \text{bias}(\hat{\theta})^2\} \]
\[ = E\{[\hat{\theta} - E\{\hat{\theta}\}]^2\} + 2\text{bias}(\hat{\theta}) E\{\hat{\theta} - E\{\hat{\theta}\}\} + [\text{bias}(\hat{\theta})]^2 \]
\[ = E\{[\hat{\theta} - E\{\hat{\theta}\}]^2\} + 2\text{bias}(\hat{\theta}) \quad 0 \quad + [\text{bias}(\hat{\theta})]^2 \]
\[ = \text{var}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2 \]

7. Evocative simple examples

Recall that unbiasedness (or known bias) is necessary in order to standardize \( \hat{\theta} \) to zero mean for use in inference, but this result indicates that a "good" estimator should have both low sampling variance and low |bias|. For example, suppose that \( y_i \sim \text{iid}[\mu, \sigma^2] \) for \( i = 1, 2 \) and consider the following estimators:

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Bias</th>
<th>Variance</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\theta}_A = 59 )</td>
<td>59 - ( \mu )</td>
<td>0</td>
<td>((59 - \mu)^2)</td>
</tr>
<tr>
<td>( \hat{\theta}_B = \hat{y}_1 )</td>
<td>0</td>
<td>( \sigma^2 )</td>
<td>( \sigma^2 )</td>
</tr>
<tr>
<td>( \hat{\theta}_C = \hat{y}_1 + \hat{y}_2 )</td>
<td>( \mu )</td>
<td>2 ( \sigma^2 )</td>
<td>2 ( \sigma^2 + \mu^2 )</td>
</tr>
<tr>
<td>( \hat{\theta}_D = \frac{1}{2} \hat{y}_1 + \frac{1}{2} \hat{y}_2 )</td>
<td>0</td>
<td>((\frac{1}{4} + \frac{1}{4})\sigma^2 = \frac{1}{2} \sigma^2 )</td>
<td>( \frac{1}{2} \sigma^2 )</td>
</tr>
<tr>
<td>( \hat{\theta}_E = \alpha \hat{y}_1 + (1-\alpha) \hat{y}_2 )</td>
<td>0</td>
<td>((2\alpha^2 - 2\alpha + 1)\sigma^2 )</td>
<td>((2\alpha^2 - 2\alpha + 1)\sigma^2 )</td>
</tr>
</tbody>
</table>

\[^2\text{Yes, you should be able to compute the bias, sampling variance, and MSE for these estimators.}\]
a. $\hat{\theta}_A$ illustrates that low sampling variance is not in-and-of-itself that great a property.

b. $\hat{\theta}_B$ illustrates that unbiasedness is not in-and-of-itself that great a property. This estimator is unbiased, but if $N > 1$ we can surely find better estimators – i.e., estimators with lower MSE.

c. $\hat{\theta}_C$ uses more of the sample data than does $\hat{\theta}_B$ but is clearly and inferior estimator.

d. $\hat{\theta}_D = \text{sample mean over first two observations}$ is clearly the best estimator so far; unless $\mu$ happens to be close to 59! It is unbiased like $\hat{\theta}_B$ and (because it uses the additional observation sensibly) also has lower sampling variance (and hence lower MSE) than $\hat{\theta}_B$.

B. Finite sample (exact) properties

1. Relative Efficiency

Since both $\hat{\theta}_D$ and $\hat{\theta}_B$ are unbiased, they can be compared using the concept of "relative efficiency":

**Relative Efficiency** (Definition):

$\hat{\theta}$ is said to be "relatively efficient" compared to another estimator, $\tilde{\theta}$, if and only if

1. both estimators are unbiased
   and
2. $\text{var}(\hat{\theta}) \leq \text{var}(\tilde{\theta})$

Clearly, $\hat{\theta}_D$ is efficient relative to $\hat{\theta}_B$, but the concept simply does not apply to comparisons involving biased estimators like $\hat{\theta}_A$ or $\hat{\theta}_C$. 
Why the focus on unbiased estimators?
1. unbiasedness usually easy to get
2. known bias necessary for inference
3. (foreshadowing) requiring unbiasedness places enough structure on the problem that an estimator can in some cases be shown to have the property of being efficient relative to all other unbiased estimators (such an estimator is called "efficient"); in contrast, we are typically not able to find estimators which have MSE no larger than that of any other estimator.
2. Best Linear Unbiasedness (BLUness)

Now consider $\hat{\theta}_E = \alpha y_1 + (1-\alpha) y_2$. This estimator has two properties by construction:

1. $\hat{\theta}_E$ is a **linear estimator** – that just means that it is a linear function of the sample data, $y_1$ and $y_2$.

and

2. $\hat{\theta}_E$ is **unbiased** by construction.

In fact, as $\alpha$ varies, $\hat{\theta}_E$ yields all of the possible unbiased linear estimators of $\theta$ using just the first two sample observations. Since $\hat{\theta}_E$ is unbiased, $\text{MSE}(\hat{\theta}_E) = \text{var}(\hat{\theta}_E)$. Let the function $V(\alpha)$ evaluate the $\text{var}(\hat{\theta}_E)$ as a function of $\alpha$:

$$V(\alpha) = 2\alpha^2\sigma^2 - 2\alpha\sigma^2 + \sigma^2$$

Since $V(\alpha)$ is a quadratic function of $\alpha$ with a positive coefficient on $\alpha^2$, a graph of $V(\alpha)$ versus $\alpha$ is an upward-opening parabola with a unique global minimum at $\alpha^* = \frac{1}{2}$:
Note that slope of this graph at $\alpha = \alpha_o$ is the derivative $dV(\alpha)/d\alpha$ evaluated at $\alpha = \alpha_o$ where:

$$\frac{dV(\alpha)}{d\alpha} = 4\alpha \sigma^2 - 2\sigma^2$$

from which we see that the slope is negative for all $\alpha < \sigma^2/2$ – this implies that increasing $\alpha$ always lowers the $\text{var}(\theta_E)$, so long as $\alpha < \sigma^2/2$. Since $dV(\alpha)/d\alpha$ is positive for all values of $\alpha > \sigma^2/2$, increasing $\alpha$ beyond $\sigma^2/2$ increases (worsens) the $\text{var}(\theta_E)$. Therefore, $\alpha^*$, the optimal value for $\alpha$, is $\frac{1}{2}$, corresponding to $\theta_E = \bar{y}$.

This result implies that $\theta_E$ with $\alpha = \alpha^* = \frac{1}{2}$ yields the minimum variance unbiased linear estimator of $\mu$ when the sample size is 2. We could state this as:

If $y_i \sim \text{iid}[\mu, \sigma^2]$ for $i = 1, \ldots, N$ and $N = 2$, then $\bar{y}$ is the BLU estimator for $\mu$.

more generally:

**Best Linear Unbiased Estimator (BLU) (Definition):**

$\theta$ is said to be the "Best Linear Unbiased" or BLU estimator for $\theta$ if and only if

1. $\theta$ is unbiased

and

2. $\theta$ is linear (i.e., $\theta$ is a linear function of the sample data)

and

3. $\text{var}(\theta) \leq \text{var}(\tilde{\theta})$ where $\tilde{\theta}$ is any linear unbiased estimator of $\theta$.

The proof given above (that $\bar{y}$ is BLU for $\mu$ when $y_i \sim \text{iid}[\mu, \sigma^2]$ for $i = 1, \ldots, N$ and $N = 2$) can be easily extended to larger sample sizes:
Let $\tilde{y}(y_1, y_2, \ldots, y_N; d_1, d_2, \ldots, d_N)$ be the linear estimator defined by the weights $d_1 \ldots d_N$:

$$
\tilde{y}(y_1, y_2, \ldots, y_N; d_1, d_2, \ldots, d_N) = \sum_{i=1}^{N} d_i y_i
$$

and note that

$$
y_i \sim \text{iid}\mu, \sigma^2$$

implies that

$$
\mathbb{E}\{\tilde{y}\} = \mu \sum_{i=1}^{N} d_i
$$

$$
\text{var}\{\tilde{y}\} = \sigma^2 \sum_{i=1}^{N} d_i^2
$$

so that $\tilde{y}$ is unbiased for $\mu$ if and only if the weights $d_1 \ldots d_N$ sum to 1. Thus, the set of weights that make $\tilde{y}$ BLU must be the solution to the minimization problem:

choose $d_1 \ldots d_N$ to minimize

$$
\sum_{i=1}^{N} d_i^2
$$

subject to the constraint that

$$
\sum_{i=1}^{N} d_i = 1
$$

since this problem is symmetric in $d_1 \ldots d_N$, the solution must be that $d_i^* = 1/N$ for all $i$, which implies that the BLU estimator of $\mu$ is $\bar{y}$.
3. Efficiency

a. Definition and relationship to BLUness

BLUness is a very nice property for an estimator to have. It says that, out of the class of linear unbiased estimators, this one is the best. If one is willing and able to further assume that the $y_i$ are gaussian, it is possible to show that $\bar{y}$ has an even stronger optimality property called "efficiency":

Efficiency (Definition):

\[ \hat{\theta} \] is said to be "efficient" for $\theta$ if and only if

1. $\hat{\theta}$ is unbiased

and

2. $\text{var}(\hat{\theta}) \leq \text{var}(\tilde{\theta})$ where $\tilde{\theta}$ is any unbiased estimator of $\theta$.

Efficiency is a much stronger optimality property than BLUness – the efficient estimator is the minimum variance (and minimum MSE) estimator out of the class of all unbiased estimators, linear or not.

Efficiency is a more meaningful working definition of "best estimator" than is BLUness, but it is typically much harder to prove. People have dealt with this difficulty in 3 ways:

1. confront the difficulty ---\> Cramer-Rao Theorem

2. lower our sights and content ourselves with the best unbiased estimator out of the class of linear estimators ---\> settle for BLUness

3. lower our sights by contenting ourselves with showing that our estimator has nice asymptotic (large sample) properties

\[ (\text{in terms of variance and MSE}) \]
BLUness is (relatively) easy to show (and its validity does not hinge on the correctness of a distributional (typically gaussianity) assumption. On the other hand, the BLU estimator is not necessarily the best unbiased estimator – there may well be unbiased nonlinear estimators with smaller variance and MSE than that of the BLU estimator. Asymptotic properties (as we shall see shortly) are usually fairly easy to obtain (indeed, in many real world settings they are the only properties we can obtain), but in most econometric settings it is difficult to be confident that our sample sizes are large enough for these properties to be meaningful.

B.3.c. [SKIP for now] Likelihood function of a random sample

The likelihood function for the sample \( y_1, y_2, \ldots, y_N \) is just the joint density function for observing this n-tuple of observations. Since \( y_1, y_2, \ldots, y_N \) are here assumed to be a random sample, they are individually iid(\( \theta_1, \theta_2 \))

- Let the density function for the ith observation (\( y_i \)) be called \( f(y_i; \theta_1, \theta_2) \). (The form of this density function does not depend on i because the \( y_i \) are identically distributed.)

- Then, because the individual observations are independently distributed, the likelihood of observing the particular sample \( (y_1, y_2, \ldots, y_N) \) is just the product of the individual density functions:

\[
\text{likelihood}(y_1, y_2, \ldots, y_N; \theta_1, \theta_2) = f(y_1; \theta_1, \theta_2) \cdot f(y_2; \theta_1, \theta_2) \cdots f(y_N; \theta_1, \theta_2)
\]

**Example #1:**

Suppose that we have a die for which the probability of landing with \( k \) spots up is \( p_m \) (\( m = 1, 2, \ldots, 6 \)) and that we throw the die three times, obtaining three independent realizations \( (y_1, y_2, y_3) \) of the random variable

\( y = \) number of spots up on die

(This presumes that we throw the die in an ordinary way so that the result of each throw is independent of the others.)

Then the probability of observing

\[
y_1 = 2 \quad y_2 = 4 \quad \text{and} \quad y_3 = 1
\]
is just the product

\[ p_2 \times p_4 \times p_1 \]

and, more generally, the likelihood of observing \( y_1 = i \) \( y_2 = j \) and \( y_3 = k \)
(i.e., the probability of state \{i, j, k\} occurring) is

\[ \text{likelihood}(y_1 = i, y_2 = j, y_3 = k; \theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = p_i \times p_j \times p_k \]

{Here the five parameters \((\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)\) which suffice to specify the
distribution of \( y \) might be \( p_1 \ldots p_5 \) – i.e., \( \theta_1 = p_1, \theta_2 = p_2, \ldots \theta_5 = p_5 \). There is
no need for a \( \theta_6 \) since

\[ p_6 = 1 - (p_1 + p_2 + p_3 + p_4 + p_5) \]

Example #2:

Suppose that \( y_i \sim \text{NIID}(\theta_1, \theta_2) \) for \( i = 1, 2, \) and 3. Then

\[ \text{likelihood}(y_1, y_2, y_3; \theta_1, \theta_2) = f^N(y_1; \theta_1, \theta_2) \times f^N(y_2; \theta_1, \theta_2) \times f^N(y_3; \theta_1, \theta_2) \]

where \( f^N(y_i; \theta_1, \theta_2) \) is just the usual gaussian density function

\[ f^N(y_i; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2\theta_2} [y_i - \theta_i]} \]

hence the joint density function (likelihood function) for the 3 NIID(\( \theta_1, \theta_2 \)) observations, \( y_1, y_2, \) and \( y_3 \) is

\[ \text{Likelihood}(y_1, y_2, y_3; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2\theta_2} [y_1 - \theta_1]} \times \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2\theta_2} [y_2 - \theta_1]} \times \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2\theta_2} [y_3 - \theta_1]} \]
Now lets make this expression more explicit by assuming that \( \text{var}(y) = \theta_2 = 1 \) and that \( \text{E}\{y\} = \theta_1 = 8 \):

\[
\text{Likelihood}(y_1, y_2, y_3 ; 8, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_1 - 8)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_2 - 8)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_3 - 8)^2}
\]

Notice that the likelihood function is largest when \( y_1 = y_2 = y_3 = 8 \), in which case the likelihood function equals

\[
\left( \frac{1}{\sqrt{2\pi}} \right)^3 = .0635
\]

But when any one of the three observations departs substantially from the population mean of 8 it causes the corresponding factor to become tiny (in magnitude) and hence causes the value of the likelihood function to be tiny (in magnitude). For example, if \( y_1 = y_3 = 8 \) and \( y_2 = 12 \), the likelihood function equals

\[
= \left( \frac{1}{\sqrt{2\pi}} \right)^3 e^{-\frac{(12 - 8)^2}{2}}
\]

\[
= .0635 e^{-8}
\]

\[
= .0635 \times .0003355
\]

\[
= .0000213
\]

It is much easier to work with the log-likelihood function, however, since it converts this extended product into a sum.
logarithm refresher:

\[
\log(x) = b \quad \leftrightarrow \quad e^b = x
\]

\[
\log(x y z) = \log(x) + \log(y) + \log(z)
\]

\[
\log(x^a) = a \log(x)
\]

\[
\log(e^a) = a
\]

\[
d\log\{g(x)\}/dx + \{1/g(x)\} \frac{dg(x)}{dx}
\]

So that:

\[
L(y_1, y_2, \ldots, y_N; \theta_1, \theta_2) = \text{log-likelihood}(y_1, y_2, \ldots, y_N; \theta_1, \theta_2)
\]

\[
= \log\{f(y_1; \theta_1, \theta_2) f(y_2; \theta_1, \theta_2) \ldots f(y_N; \theta_1, \theta_2)\}
\]

\[
= \log\{f(y_1; \theta_1, \theta_2)\} + \log\{f(y_2; \theta_1, \theta_2)\} + \ldots + \log\{f(y_N; \theta_1, \theta_2)\}
\]

so that

Log-Likelihood Function for a Gaussian Random Sample

\[
L(y_1, y_2, \ldots, y_N; \theta_1, \theta_2) = \sum_{i=1}^{N} \log[f(y_i; \theta_1, \theta_2)]
\]

If \( y_i \sim \text{NIID}(\theta_1, \theta_2) \) for \( i = 1 \ldots N \), then

\[
\text{likelihood}(y_1, y_2, y_3; \theta_1, \theta_2) = f^N(y_1; \theta_1, \theta_2) \times f^N(y_2; \theta_1, \theta_2) \times f^N(y_3; \theta_1, \theta_2)
\]

and hence the log-likelihood function, \( L((y_1, y_2, y_3; \theta_1, \theta_2)) \), is

\[
L(y_1, y_2, \ldots, y_N; \theta_1, \theta_2) = \sum_{i=1}^{N} \log[f^N(y_i; \theta_1, \theta_2)]
\]

where \( f^N(y_i; \theta_1, \theta_2) \) is just the usual gaussian density function
\[
f^N(y_i; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi \theta_2}} e^{-\frac{1}{2\theta_2} [y_i - \theta_1]}
\]

so that

\[
\log\{f^N(y_i; \theta_1, \theta_2)\} = \log\left\{\frac{1}{\sqrt{2\pi \theta_2}} e^{-\frac{1}{2\theta_2} [y_i - \theta_1]}\right\} = \log\left\{\frac{1}{\sqrt{2\pi \theta_2}}\right\} + \log\left\{e^{-\frac{1}{2\theta_2} [y_i - \theta_1]}\right\} = \log\left\{2\pi \theta_2\right\}^{\frac{1}{2}} - \frac{1}{2\theta_2} [y_i - \theta_1] = -\frac{1}{2} \log\{2\pi \theta_2\} - \frac{1}{2\theta_2} [y_i - \theta_1]
\]

and hence,

\[
L(y_1, y_2, \ldots, y_N; \theta_1, \theta_2) = \sum_{i=1}^{N} \log\{f^N(y_i; \theta_1, \theta_2)\} = \sum_{i=1}^{N} \left\{-\frac{1}{2} \log\{2\pi \theta_2\} - \frac{1}{2\theta_2} [y_i - \theta_1] \right\} = -\frac{N}{2} \log\{2\pi \theta_2\} - \frac{1}{2\theta_2} \sum_{i=1}^{N} [y_i - \theta_1] = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(\theta_2) - \frac{1}{2\theta_2} \sum_{i=1}^{N} [y_i - \theta_1]
\]
D. Minimum MSE estimator

If one's loss function on estimation errors is not proportional to the square of the error, then one might prefer some other estimator over the efficient estimator. But suppose that MSE really is what one cares about, is it possible to find an estimator with a smaller MSE than that of the efficient estimator? It turns out that the answer to this question is both "yes" and "no":

Let \( \hat{\theta} \) be the efficient estimator for \( \theta \) and let

\[
\tilde{\theta}(k) = k \hat{\theta} \quad \text{and} \quad t = \frac{\theta}{\sqrt{\text{var}(\hat{\theta})}}
\]

where \( \tilde{\theta}(k) \) is an alternative estimator of \( \theta \) and \( k \) is some positive number. What value of \( k \) yields the \( \tilde{\theta}(k) \) with the smallest MSE?

\[
\text{bias}[\tilde{\theta}(k)] = E\{\tilde{\theta}(k)\} - \theta = E\{k \hat{\theta}\} - \theta = k E\{\hat{\theta}\} - \theta = k \theta - \theta = (k - 1) \theta
\]

\[
\text{var}[\tilde{\theta}(k)] = \text{var}[k \hat{\theta}] = k^2 \text{var}[\hat{\theta}]
\]

Thus,

\[
\text{MSE}[\tilde{\theta}(k)] = \text{var}[\tilde{\theta}(k)] + \left\{\text{bias}[\tilde{\theta}(k)]\right\}^2
\]

\[
= k^2 \text{var}[\hat{\theta}] + (k - 1)^2 \theta^2
\]

\[
= k^2 \text{var}[\hat{\theta}] + (k - 1)^2 \theta^2 \left(\frac{\text{var}[\hat{\theta}]}{\text{var}[\hat{\theta}]}\right)
\]

\[
= k^2 \text{var}[\hat{\theta}] + (k - 1)^2 \text{var}[\hat{\theta}] \left(\frac{\theta^2}{\text{var}[\hat{\theta}]}\right)
\]

\[
= k^2 \text{var}[\hat{\theta}] + (k - 1)^2 \text{var}[\hat{\theta}] t^2
\]

\[
= \text{var}[\hat{\theta}](k^2 + (k - 1)^2 t^2)
\]
Thus $k^*$, the value of $k$ which minimizes the $\text{MSE} \{ \tilde{\theta}(k) \}$, must satisfy

$$0 = \frac{d \text{MSE}[\tilde{\theta}(k)]}{dk} \bigg|_{k=k^*} = \frac{d}{dk} \left[ \text{var}[\hat{\theta}] \left[ (k^*)^2 + (k^* - 1)^2 t^2 \right] \right]$$

$$= \text{var}[\hat{\theta}] \frac{d}{dk} \left[ (k^*)^2 + (k^* - 1)^2 t^2 \right]$$

$$= \text{var}(\hat{\theta}) \left[ 2k^* + 2(k^* - 1)t^2 \right]$$

$$= 2\text{var}(\hat{\theta}) \left[ k^* + (k^* - 1)t^2 \right]$$

$$= 2\text{var}(\hat{\theta}) \left[ (1 + t^2)k^* - t^2 \right]$$

which implies that

$$k^* = \frac{t^2}{1 + t^2} < 1.$$ 

Therefore, it is always optimal (i.e. it lowers the MSE) to **bias** the estimator a bit toward zero. However, it must be noted that $\tilde{\theta}(k^*)$ is not in fact a feasible estimator since the expression for $k^*$ involves $t$ which depends on $\theta$. 
E. The Zen of Econometrics

We have now gone to considerable trouble to see that $\bar{y}$ is a good (in some senses, optimal) estimator of $\mu$ when $y_i \sim \text{NIID}(\mu, \sigma^2)$. But what if our actual sample observations are clearly not gaussian?

For example, consider the problem of estimating the mean size of a U.S. firm, as quantified by its annual sales revenues. The empirically observed distribution of firm sizes looks more like the graph of a $\chi^2$ density function than it does like the graph of a gaussian density function:

However, since the size of a firm can be viewed as the product of a large number of (more or less) independent factors, the empirically observed distribution of the logarithm of the firm sizes looks reasonably gaussian.

Therefore there are two different ways to approach the problem of estimating the mean size of a U.S. firm:

1. Try to figure out the most efficient estimator for the (population) mean of a random sample drawn from a log-normally distributed population. Then one must still face the problem of calculating its sampling distribution so that you can do inference. These problems can be dealt with to some degree, but even their approximate solution is difficult and complicated.

or

2. Since it is reasonable to suppose that the observations on log(firm sales revenue) are NIID, use the sample mean of the logarithms of the observed sales revenue figures to efficiently estimate the population mean of the logarithm of firm sales revenue. Similarly, use the methods we have covered to obtain a 95% confidence interval for the population mean of the
logarithm of firm sales revenue. (This interval will be centered around the sample mean of the logarithms of the observed sales revenue data.) Letting \( e^{c_{95\text{lower}}} \) and \( e^{c_{95\text{upper}}} \) denote the upper and lower endpoints of this 95% confidence interval for the population mean of the logarithm of firm sales revenue, one can immediately infer that the interval \([ e^{c_{95\text{lower}}}, e^{c_{95\text{upper}}} ]\) is a 95% confidence interval for the population mean of the of sales revenue itself. (This interval will not be centered around the exponential of the sample mean of the logarithms of the observed sales revenue data, however.) Or, more reasonably, one can – and, I would argue, typically should – decide that the data itself is telling us that the size of the firm ought to be quantified by the logarithm of its annual sales revenues, in which case the sample mean of the logarithms of the observed sales revenue figures is precisely the estimator one wants.

This example suggests a completely different way of looking at the theoretical results we have obtained. Instead of seeing them as arbitrary special cases where everything worked out nicely and resolving to always assume that our data is NIID (whether it is or not) so that we can "use" these results, this example suggests that we could more gracefully observe that our statistical machinery for using data to obtain knowledge about the world "works" best (i.e., most simply and effectively) when we have framed the problem in such a way that the sample data is NIID and act accordingly.