Properties of Estimators

BIOS 6611

CU Anschutz

Week 1

Notation

Notation

Notation:

- $\mathbf{x} = (x_1, \ldots, x_n)$ represents a sample of size n drawn from the population of interest
  - Bold represents a vector
- θ represents a population parameter
- $\hat{\theta}$ represents the estimate for that parameter
  - $\hat{\theta}$ is a function of the sample data.
  - For example $\hat{\theta}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$.
  - For brevity, we omit and just write $\hat{\theta}$.

“Good” Estimators

How do we evaluate if an estimator is “good”? How do we compare the performance of different estimators?

Properties of Estimators

Bias

$\hat{\theta}$ is **unbiased** if it's expected value is equal to the parameter of interest, i.e.,

$$E[\hat{\theta}] = \theta \tag{1}$$

The **bias** of an estimator is

$$Bias(\hat{\theta}) = E[\hat{\theta}] - \theta \tag{2}$$

An estimator that is unbiased is "right on target."

Efficiency

- When comparing two estimators, we say one is more **efficient** if its variance is smaller, i.e.,

$$Var(\hat{\theta}_1) > Var(\hat{\theta}_2) \tag{3}$$

- The **relative efficiency** of two estimators is

$$e(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} \tag{4}$$

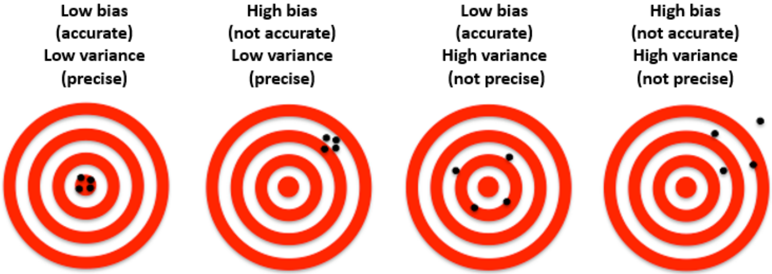
- $\hat{\theta}$ is an **efficient estimator** if its variance reaches the *Cramer-Rao lower bound*.
    - Basically, the variance is as small as it could possibly be.
    - Only unbiased estimators can reach this lower bound.
    - An estimator is **asymptotically efficient** if it reaches the Cramer-Rao lower bound as the sample size becomes infinitely large.
- An efficient estimator is precise.

Mean Square Error

Bias looks at the expected value of an estimator. Efficiency looks at the variance. The **mean square error (MSE)** looks at both:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + \{Bias(\hat{\theta})\}^2 \quad (5)$$

Measures both accuracy and precision of an estimator. Getting a small MSE often involves a trade-off between variance (precision) and bias (accuracy).



Asymptotic Consistency

$\hat{\theta}$ is **asymptotically consistent** (often we just say **consistent**, used synonymously) if

$$\hat{\theta}(x_1, \ldots, x_n) \underset{p}{\rightarrow} \theta \tag{6}$$

The more data you collect, a consistent estimator will be closer and closer to the true population parameter.

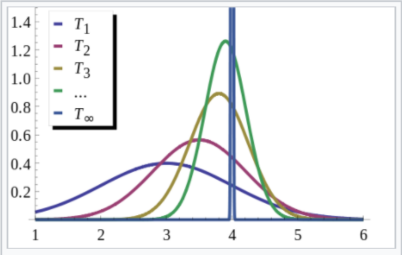


Figure 1: Source: Wikipedia

Using Simulations to Investigate Estimators

Estimating Bias from Simulations

- Say we simulate 100 data sets of sample size n. For each data set, calculate $\hat{\theta}$.
  - We generate the data with true knowledge of the parameter θ
- Estimate the bias as

$$\hat{Bias}(\hat{\theta}) = \frac{\sum_{j=1}^{100} \hat{\theta}_j}{100} - \theta \tag{7}$$

- Why does this work? The *Law of Large Numbers*!

$$\frac{\sum_{j=1}^{100} \hat{\theta}_j}{100} \to E[\hat{\theta}] \tag{8}$$

- The more simulations we generate and the larger the sample size, the better the estimate will be.

Estimating Variance from Simulations

To estimate the variance of the estimator, calculate the sample variance of the estimators across all simulations:

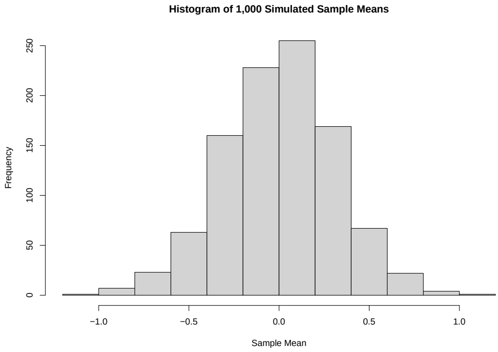
$$\hat{Var}(\hat{\theta}) = \frac{\sum_{j=1}^{100} \left(\hat{\theta}_j - \frac{\sum_{j=1}^{100} \hat{\theta}_j}{100} \right)^2}{100} \tag{9}$$

Alternatively, we can use functions in `R` to assist in our estimates from the simulations (e.g., `mean`, `var`).

Sample Mean Estimator Example

Let's check out a simulation study for the sample mean with 1,000 simulated samples of $n = 10$ from a standard normal distribution:

```
set.seed(6611)
simres <- sapply(1:1000, function(x) mean( rnorm(n=10) ))
hist(simres, main='Histogram of 1,000 Simulated Sample Means',
     xlab='Sample Mean')
```



Sample Mean Estimator Example cont.

```r
# average of 1000 sample means
avg_mean <- mean(simres); avg_mean
```

```
## [1] 0.003741583
```

```r
# variance of 1000 sample means
var_mean <- var(simres); var_mean
```

```
## [1] 0.09832073
```

```r
# bias (estimate minus true simulated mean)
avg_mean - 0
```

```
## [1] 0.003741583
```

```r
# mean squared error (var + bias^2)
var_mean + (avg_mean - 0)^2
```

```
## [1] 0.09833473
```