Chapter 11:

Estimators and Intervals*

Until this point, the two quantities of interest have been the estimator and the confidence interval. Thus far, we have been using estimators without justification. It *seems* logical that the sample mean would be the correct estimator of the population mean, but is it? It turns out that a better question is "*When* is it?"

The $(1-\alpha)100\%$ confidence interval is supposed to contain the true population value $(1-\alpha)100\%$ of the time. However, most of the exact tests of the previous three chapters are conservative; that is, the confidence intervals contain the true population value *more than* $(1-\alpha)100\%$ of the time. This means the true α level is smaller than expected, which reduces the power of the test.

In this chapter, we experiment with different estimators and different confidence intervals to determine when each is preferred.

An opinion poll performed prior to the 2012 US Presidential election found that 370 of the 900 respondents stated they would vote for Barack Obama; 340 for Mitt Romney; and 290 would not vote. What is the best estimate of the proportion of Americans voting for Mitt Romney? What is the best 95% confidence interval for that quantity?

Thus far, we have used estimators without explaining why we used them. While it does seem logical that the sample mean is the best estimator of the population mean and that the sample proportion is the best estimator of the population proportion, we have done no mathematics to support that conclusion.

The purpose of this chapter is to explore different estimators and different methods for determining which estimator is appropriate in a given situation. However, before we can determine which estimator is better, we need to define what we mean by "better."

11.1: Selecting Estimators

Ideally, we want an estimator to always equal the true population value. Since we cannot have that (randomness in life and all that), we have to decide what we want to emphasize from our estimator. We may want it unbiased. We may want it to have minimum variability. We may want it to optimize some combination of these two.

population value

For several reasons, including the mathematics, statisticians tend to prefer estimators that are concentrated around the true population parameter. We prefer our estimator to be close to the true value most frequently; that is, we want the average distance between the estimator and the parameter to be small.

horseshoes

Writing this in mathematical terms, if we define τ as the population parameter we are estimating and T as the estimator we are using, then we want to minimize $\mathbb{E}\left[(T-\tau)^2\right]$. This quantity is called the **mean square error** (MSE). The estimator with the smaller mean square error is usually the preferred estimator.

MSE

As T is a function of the data, it is a random variable. As τ is a population parameter, it is *not* a random variable. We can rewrite the MSE definition to show that it is the sum of the variance of the estimator and the square of the bias of the estimator.

random variable

bias

Before we prove this, let us introduce the two Oni Equations, which will help us in calculating the mean square error.

Theorem 11.1 (The Oni Equations)**.** *Let X be a random variable. Then*

$$\mathbb{E}[aX + c] = a\mathbb{E}[X] + c \tag{11.1}$$

$$\mathbb{V}[aX + c] = a^2\mathbb{V}[X] \tag{11.2}$$

Proof. These proofs start with the definitions and use algebra to achieve the conclusion. Without loss of generality, let us assume X is continuous.

$$\begin{aligned}
\mathbb{E}[aX + c] &= \int_S \Big(ax + c\Big)\, f(x)\, dx \\
&= \int_S \Big(ax\, f(x) + c\, f(x)\Big)\, dx \\
&= \int_S ax\, f(x)\, dx + \int_S c\, f(x)\, dx \\
&= a\int_S x\, f(x)\, dx + c\int_S f(x)\, dx \\
&= a\mathbb{E}[X] + c
\end{aligned}$$

The second equation follows in a similar manner

$$\begin{aligned}
\mathbb{V}[aX + c] &= \int_S \Big((ax + b) - \mathbb{E}[aX + b]\Big)^2\, dx \\
&= \int_S \Big(ax + b - a\mathbb{E}[X] - b\Big)^2\, dx \\
&= \int_S \Big(ax - a\mathbb{E}[X]\Big)^2\, dx \\
&= \int_S \Big(a(x - \mathbb{E}[X])\Big)^2\, dx \\
&= a^2\int_S \Big(x - \mathbb{E}[X]\Big)^2\, dx \\
&= a^2\mathbb{V}[X]
\end{aligned}$$

If X is discrete, replace integration with summation and reach the same conclusion.

It is easy to show that the first equation holds for a linear combination of multiple random variables, $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$. Also, if the random variables X and Y are independent, then the second equation also holds for their linear combination, $\mathbb{V}[aX + bY + c] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y]$. □

Now, with the theorem proven, we can show that the mean square error can be written as the sum of the estimator's variance and the square of its bias.

Mean Square Error

Theorem 11.2 (The Mean Square Error)**.** *Let T be an estimator of a population statistic τ. The mean square error of T is the sum of its variance and the square of its bias:*

$$MSE(T) = \mathbb{V}[T] + bias^2(T)$$

Proof. Let us be given an estimator T. By definition, the mean square error of T is

$$\text{MSE}(T) := \mathbb{E}\left[(T-\tau)^2\right]$$

Expanding the square inside the expectation gives

$$= \mathbb{E}\left[T^2 - 2T\tau + \tau^2\right]$$

By the first Oni Equation (Theorem 11.1), this is

$$= \mathbb{E}\left[T^2\right] - 2\tau\mathbb{E}[T] + \tau^2$$

Adding and subtracting $\mathbb{E}[T]^2$ gives

$$= \mathbb{E}\left[T^2\right] - \mathbb{E}[T]^2 + \mathbb{E}[T]^2 - 2\tau\mathbb{E}[T] + \tau^2$$

By definition, this reduces to

$$= \mathbb{V}[T] + \Big(\mathbb{E}[T] - \tau\Big)^2$$

Thus, $\text{MSE}(T) = \mathbb{V}[T] + \text{bias}^2(T)$, as needed. □

By itself, the mean squared error is of little use. However, it does allow us to compare estimators to determine which is the better. Again, the estimator with a smaller MSE is the preferred estimator.

compare

Warning: *While no estimator is* always *the best, some are* never *the best. Such estimators are termed* ***inadmissible****.*

Example 11.1: Let $X \sim Bin(n, \pi)$. Thus far, we have used $\hat{\pi} = \frac{X}{n}$ as our estimator of the population proportion, π. It is a good estimator; it is unbiased. However, there is a second estimator of the population proportion, $T_1 = \frac{X+1}{n+2}$. Which of these two estimators is preferred?

Solution: To compare these two estimators, let us calculate their respective mean square errors. First, for the usual estimator, the so-called Wald estimator (1939):

Wald Estimator

$$MSE(\hat{\pi}) = \mathbb{V}[\hat{\pi}] + \text{bias}^2(\hat{\pi})$$

Calculating the first term of the right hand side, we have

$$\mathbb{V}[\hat{\pi}] = \mathbb{V}\left[\frac{X}{n}\right]$$

By the second Oni Equation, this is

$$= \frac{1}{n^2}\mathbb{V}[X]$$

As $X \sim Bin(n, \pi)$, $\mathbb{V}[X] = n\pi(1-\pi)$, which leads to

$$\begin{aligned}\mathbb{V}[\hat{\pi}] &= \frac{1}{n^2}n\pi(1-\pi), \text{ which simplifies to} \\ &= \frac{\pi(1-\pi)}{n}\end{aligned}$$

Next, we calculate the bias

$$\begin{aligned}\text{bias}(\hat{\pi}) &= \mathbb{E}[\hat{\pi}] - \pi \\ &= \mathbb{E}\left[\frac{X}{n}\right] - \pi\end{aligned}$$

By the first Oni Equation, this is

$$\begin{aligned}&= \frac{\mathbb{E}[X]}{n} - \pi \\ &= \frac{n\pi}{n} - \pi \\ &= 0\end{aligned}$$

Because the bias is zero, the Wald estimator is **unbiased**. Thus, its mean squared error is just its variance

$$\text{MSE}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$$

> ***Note***: Note that this is a function of the population parameters. While we know n, we do not know π. However, we usually have at least a vague idea of the value of π before we begin our experiment.

Now, let us calculate the mean squared error of the new estimator:

$$\begin{aligned}\text{MSE}(T_1) &= \mathbb{V}[T_1] + \text{bias}^2(T_1) \\ \mathbb{V}[T_1] &= \mathbb{V}\left[\frac{X+1}{n+2}\right]\end{aligned}$$

By the second Oni Equation, this is

$$\begin{aligned}&= \frac{1}{(n+2)^2}\mathbb{V}[X+1] \\ &= \frac{1}{(n+2)^2}\mathbb{V}[X]\end{aligned}$$

As $X \sim Bin(n, \pi)$, this is

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

Notice that this is always less than the variance of the Wald estimator.

Next, we calculate the bias of T_1:

$$\begin{aligned}\text{bias}(T_1) &= \mathbb{E}[T_1] - \pi \\ &= \mathbb{E}\left[\frac{X+1}{n+2}\right] - \pi \\ &= \frac{n\pi+1}{n+2} - \pi\end{aligned}$$

In general, this does not equal zero. As such, this estimator is biased. The only time this estimator will be unbiased is when $\pi = 0.500$ (See Extension Problem 6). Thus, the mean square error of estimator T_1 is

biased

$$\text{MSE}(T_1) = \frac{n}{(n+2)^2}\pi(1-\pi) + \left(\frac{n\pi+1}{n+2} - \pi\right)^2$$

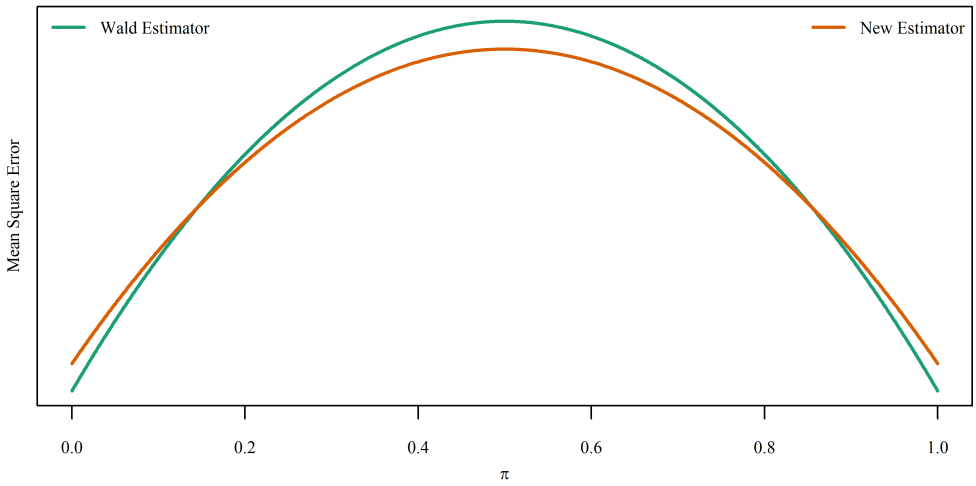

Figure 11.1: *Graphic of the mean square error for two estimators of π. Here, $n = 50$. Note that neither estimator always has the lowest mean squared error. Thus, our estimator of choice will depend on our estimate of π.*

The question remains: Which of the two estimators has the lower mean squared error? The answer is that it depends on n and π. While we do know n, we do not know the value of π. However, we usually have some idea of its approximate value. ◇

Example 11.2: To make this more concrete, let us randomly select and ask $n = 50$ people whether they support the president's agenda. Across the United States, this number should be somewhere between 30% and 60%, thus we have a very rough estimate of π. With this, which of the two estimators is better for this situation?

Solution: Figure 11.1 shows a plot of the mean square error for these two estimators when $n = 50$. Note that neither is always better than the other. For extreme values of π, the Wald estimator has a lower mean square error and is preferred to the new estimator. For middling values of π, the opposite is true. According to the graphic, if we expect π to be between 15 and 85%, we should use the new estimator. As we estimated π between 30% and 60%, we should use T_1, the Agresti-Coull (1998) estimator over the usual one, even though it is biased. ◇

middling

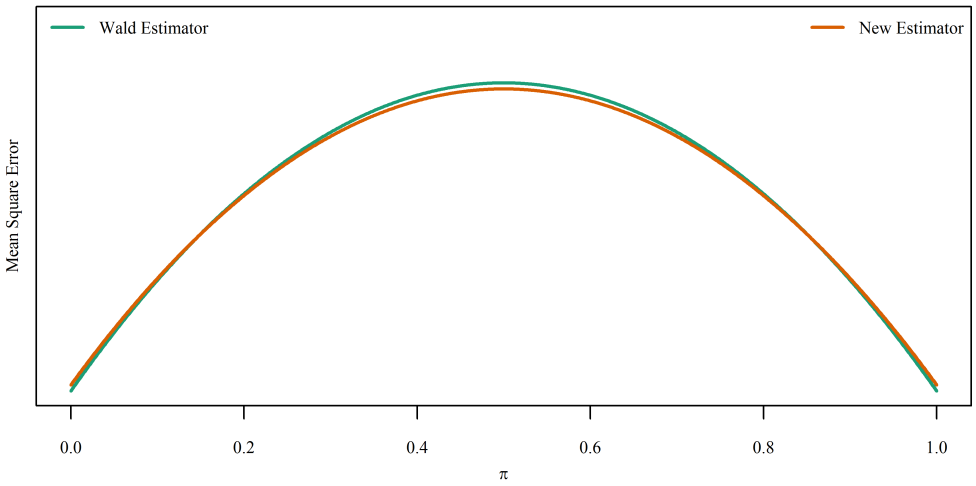

Figure 11.2: *Graphic of the mean square error for two estimators of π. Here, $n = 200$. Note that neither estimator always has the lowest mean squared error. Thus, our estimator of choice will depend on our estimate of π.*

The steps to produce graphics like Figure 11.1 are straight-forward. First, set the known value of n. Then, create a vector of several values of π across its entire domain. Third, calculate the mean square errors. Finally, plot the results, realizing that the range of the mean square errors are not necessarily the same.

Here is the code used to produce a graphic similar to Figure 11.1.

```
n  = 50
p  = seq(0,1,length=1000)
MSE0 = p*(1-p)/n
MSE1 = n/(n+2)^2*p*(1-p) + ( (n*p+1)/(n+2) - p )^2

plot(p,MSE0, type="n", xlim=c(0,1), ylim=c(0,0.005), las=1,
    yaxt="n", xlab=expression(pi),ylab="Mean Square Error",
    col=3, lwd=2 )
lines(p,MSE1, col=2,lwd=2)
```

Figure 11.1 shows that the choice of estimator depends on π. To see that it also depends on n, let us ask $n = 200$ people. Figure 11.2 shows the relative mean square error of the two estimators. Note that the general rule that one should use $\hat{\pi}$ for extreme values of π and T_1 for middling values still

middling

holds. The crossing points change, however. When $n = 200$, one should use T_1 over $\hat{\pi}$ when π is between 0.1460 and 0.8540.

Note: While the range for which the Agresti-Coull estimator is better than the Wald estimator is a function of the sample size n, there is little change in the range. When $n = 10$, the range is from 0.1381 to 0.8619. When $n = 1,000$, the range is from 0.1463 to 0.8537.

Example 11.3: Let us be given that $X_i \overset{\text{iid}}{\sim} \mathcal{U}(0, b)$, for $i = \{1, 2, \ldots, n\}$. In other words, we are collecting a sample of size n from a Uniformly-distributed random variable whose minimum is 0 and whose maximum is unknown. We want to estimate b. To do this, we can use either $T_1 = 2\overline{X}$ or $T_2 = \max X_i$. Which of these two estimators is preferred, T_1 or T_2?

This example has historical significance. In a war, it is helpful to know how many tanks your enemy has. In World War II, Allied statisticians were given the job of devising an estimator for this number. Thankfully, the Germans were very orderly with their tanks' serial numbers, which reduced the problem to determining b.

Solution: At this point, we do not have the exact distribution of $\overline{X}$. However, we can estimate it using the Central Limit Theorem (Appendix C) assuming n is large enough (30 should work here). From that theorem, we have $\overline{X} \dot{\sim} \mathcal{N}\left(\frac{b}{2}, \frac{b^2}{12n}\right)$. And so, we have the approximate distribution of T_1:

$$T_1 \dot{\sim} \mathcal{N}\left(b, \frac{b^2}{3n}\right)$$

With this, we can estimate the mean square error of T_1:

$$\begin{aligned} \text{MSE}(T_1) &= \mathbb{V}[T_1] + \text{bias}^2(T_1) \\ \mathbb{V}[T_1] &= \frac{b^2}{3n} \\ \text{bias}(T_1) &= \mathbb{E}[T_1] - b \\ &= b - b \end{aligned}$$

And, the mean squared error of estimator T_1 is

$$\text{MSE}(T_1) = \frac{b^2}{3n}$$

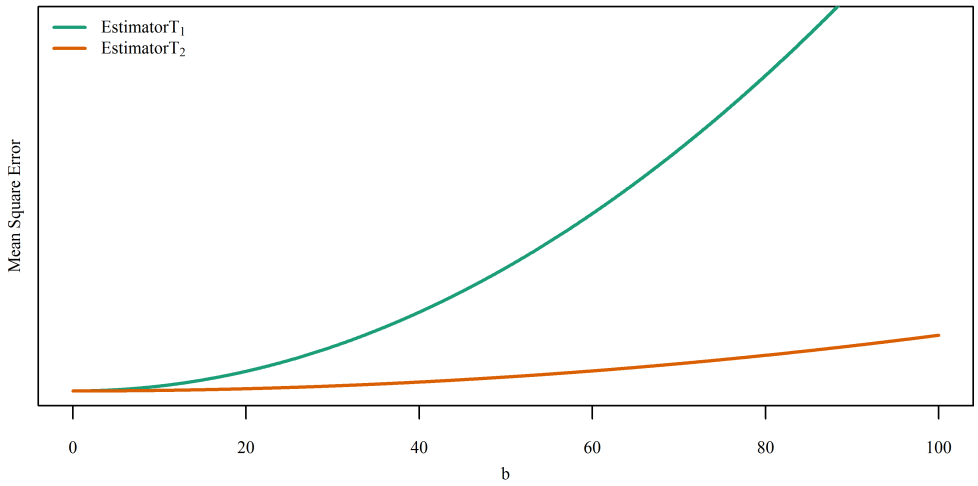

Figure 11.3: *Graphic of the mean square error for two estimators of b. Here, $n = 50$. Note that estimator T_2 always has the lower mean squared error. Thus, of the two, T_1 is not admissible.*

For reference, the mean square error of T_2 is $\frac{2b^2}{(n+1)(n+2)}$. Figure 11.3 plots the two mean square error functions for $n = 50$. Note that the second estimator is always better than the first in terms of mean squared error. Thus, estimator T_1 is inadmissible to T_2.

inadmissible

Furthermore, as b increases, the advantage of the second estimator over the first increases. Again, one should *never* use the T_1 estimator; it is inadmissible. ◇

11.1.1 Estimating the Mean Square Error*

Frequently, one cannot directly calculate the mean square error: the distribution of the test statistic may be unknown, or the formula may be too complicated. In such cases, one can estimate the mean square error using simulation.

Monte Carlo

Example 11.4: Let our data come from a Normal distribution; that is, let $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Thus far, we have used the sample mean $\overline{x}$ to estimate the center of the distribution, μ. However, since the Normal distribution is symmetric, the median $\widetilde{x}$ will also work. (Why?) Which of the two estimators will have a lower mean square error?

Solution: While we do have an exact distribution for the sample mean, we do not have one for the sample median.[1] So, let us use simulation to estimate the mean square errors of two estimators.

mean square error

Recall that the mean square error is $\mathbb{E}\left[(T-\tau)^2\right] = \mathbb{V}[T] + \text{bias}^2(T)$. To estimate it, we draw a random sample of size n from the given distribution, calculate the test statistic, repeat these two steps B times, and calculate $\mathbb{E}\left[(T-\mu)^2\right]$. This final value is the mean square error.

```
set.seed(5)
mu = 1
n  = 10
B  = 10000

Xbar = numeric()
Xmed = numeric()

for(i in 1:B) {
  X = rnorm(n,m=mu)        # random draw
  Xbar[i] = mean(X)        # mean estimate
  Xmed[i] = median(X)      # median estimate
}
mean( (Xbar-mu)^2 )        # MSE of the mean estimator
mean( (Xmed-mu)^2 )        # MSE of the median estimator
```

As written, this script will take quite a while. Loops take longer to perform than vector arithmetic. Because of this, the following code runs much faster, albeit at the expense of readability.

```
set.seed(5)
mu=1; n=10; B=10000        # initialization

Y = rnorm(n*B,m=mu)        # all random draws
X = matrix(Y,nrow=B)       # put into matrix form

Xbar = apply(X,1,mean)     # mean estimates
Xmed = apply(X,1,median)   # median estimates

mean( (Xbar-mu)^2 )        # MSE of mean estimator
mean( (Xmed-mu)^2 )        # MSE of median estimator
```

or speeded

Note that we sped things up a bit by drawing the entire sample at once and applying the estimator to each sample. Also note that we combined the first

[1] The Central Limit Theorem can give us an approximate distribution for the sample median, $\widetilde{X} \dot{\sim} \mathcal{N}(\mu, \sigma^2 = \frac{1.57}{n})$. This converges quickly, so n does not have to be 'large.'

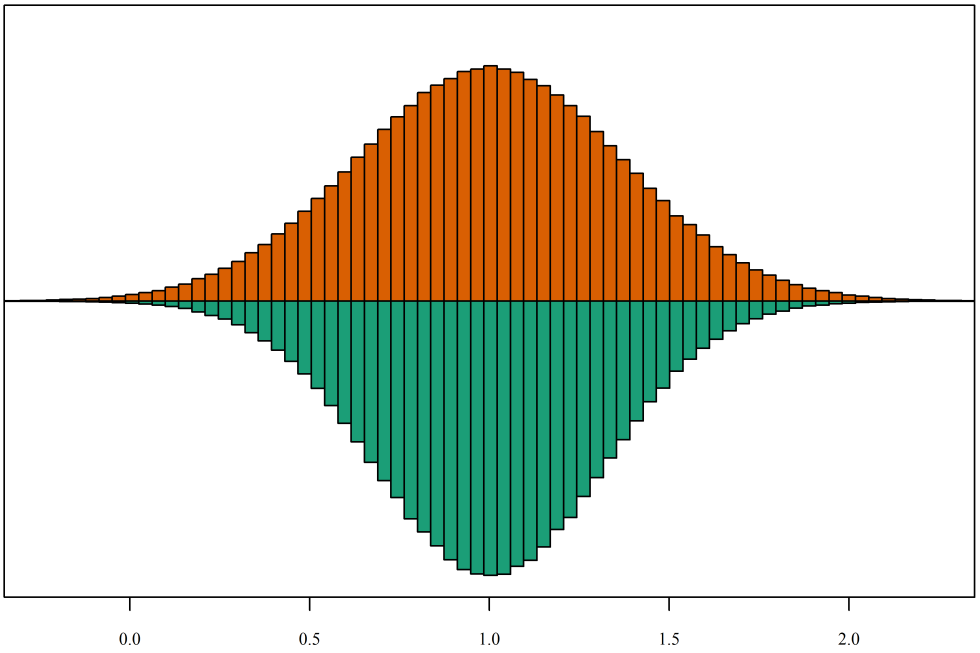

Figure 11.4: *Comparison of the median estimator (top) and the mean estimator (bottom). Both estimate the center of this Normal distribution ($\mu = 1$).*

three lines into one by using the semicolon. This is useful in making your code footprint a bit smaller. **readability**

The estimated mean square error for the mean estimator is approximately 0.10; for the median estimator, approximately 0.14. Thus, the mean square error of the median estimator is about 40% higher than that of the mean estimator. Add to this the longer calculation time for the median and it is little wonder that we use the sample mean instead of the sample median to estimate the population mean when the data is Normally distributed. ◇ **result**

Figure 11.4 shows histograms of the mean estimator and of the median estimator for this example; the mean estimator is on the bottom. Note that the mean estimator is more concentrated around the true population center of 1.0 than is the median estimator. **empirical pdf** **concentration**

Example 11.5: In the previous example, we relied heavily on the data being distributed Normally. What if this were not the case? Let us suppose the

data follows a Cauchy distribution and we still want to estimate the center of the population. Should we use the sample mean or the sample median to estimate the center of the population?

Solution: The steps are the same. In fact, let us use the previous code and make a single change. Can you spot it? Why would that line need to be changed?

```
set.seed(5)
mu=1; n=10; B=10000              # initialization

Y = rcauchy(n*B,location=mu)     # all random draws
X = matrix(Y,nrow=B)             # put into matrix form

Xbar = apply(X,1,mean)           # mean estimates
Xmed = apply(X,1,median)         # median estimates

mean( (Xbar-mu)^2 )              # MSE of mean estimator
mean( (Xmed-mu)^2 )              # MSE of median estimator
```

Running this code gives a mean square error of 1910 for the mean estimator and 0.346 for the median estimator. As such, the *median* should be used as the estimator of the population's center in this case. ◇

outlier

> ***Note***: The Cauchy distribution is useful in modeling data that has more outliers than expected under the Normal model. Thus, if your data has outliers, you may wish to use the median as an estimate of the center. In future chapters, this leads to median regression (Chapter 12).

Warning: *The Cauchy distribution has neither a mean nor a variance. It is symmetric and bell-shaped. It has a median. However, by the definition of expected value, a Cauchy random variable has no mean. This is why using the sample mean to estimate the center of the Cauchy distribution is problematic.*

Example 11.6: In this final MSE example, let us suppose the data follows an Exponential distribution and we still want to estimate the center of the population. Should we use the sample mean or the sample median to estimate the center of the population?

Solution: Again, the program needs only one line to be changed. Line 3 should now be `Y = rexp(n*B,rate=1)`. Running the altered code gives a mean square error of 0.1013 for the mean estimator and 0.1588 for the median estimator. As such, the *mean* should be used as the estimator of the population's center in this case. ◇

Note: It should *not* be surprising that MSE values depend on the value of the population parameter λ. When λ is greater than (approximately) 0.9, the mean is the better estimator. Otherwise, the median is the better estimator.

Using algebra, is can be shown that the mean should be used when $\lambda > 0.9$. Recall that as lambda decreases, the variance of the distribution increases while the skew remains constant (Appendix B.5). This causes more values to appear as though they are outliers. This presence of "outliers" explains why the median is the preferred estimator of the two.

11.2: Coverage and Intervals

In the previous section, we discussed methods to determine appropriate point estimators. The decision is based on the mean square error (MSE) of each estimator; the one with the smaller MSE is preferred.

Now that we have an estimator of the population parameter, we need to construct confidence intervals. By definition, a $(1-\alpha)100\%$ confidence interval should contain the population parameter $(1-\alpha)100\%$ of the time when the experiment is repeated.

confidence interval

coverage

It is rare that a confidence interval will have this exact coverage rate. When the distribution of the test statistic is discrete, the coverage will often be too large, which means the Type I Error rate is smaller than the expected α. While this may seem like a good consequence, smaller Type I Error rates correspond to *larger* Type II Error rates. Thus, the tests will have lower power than they could have if the Type I Error rate were exactly α. Such a test is termed **conservative** if the Type I Error rate is less than α.

Type I Error rate

conservative

In the previous section, we wanted estimators that had smaller mean square errors. In this section, we want confidence intervals with coverage rates closest to $(1-\alpha)100\%$.

While it is possible to mathematically calculate the coverage rate for some distributions and statistics, such is rarely the case. In lieu of exactly calculating the coverage, one frequently estimates it using Monte Carlo.

Monte Carlo

Estimating the coverage rate is straight-forward: generate a sample of the random variable, calculate the endpoints of the confidence interval based on that sample, and determine if the population parameter is between those endpoints.

Example 11.7: Let us return to estimating the population proportion. To wit: Let $X \sim Bin(n, \pi)$. If n is large enough, X has an approximate Normal distribution, $X \dot{\sim} \mathcal{N}\Big(n\pi, n\pi(1-\pi)\Big)$. Standardizing and inverting gives the endpoints of our confidence interval:

$$\hat{\pi} \pm Z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$$

While this is a nice interval, it is worthless; we do not now the value of π. What value can we use in its place? Over the years, several values have been recommended, including $\pi_0 = \hat{\pi} = \frac{X}{n}$, $\pi_1 = \frac{X+1}{n+2}$, $\pi_2 = \frac{X+2}{n+4}$, and $\pi_3 = \frac{1}{2}$.

Under the assumptions of this problem, which of the four estimators is the best in terms of the resulting coverage rate?

Solution: While we do have the ability to explicitly calculate the endpoints, let us use simulation. Explicit calculation is rarely useful; simulation is flexible. Here is the R code for determining the coverage rate for π_0:

```
set.seed(5)
B   = 1e6; n = 10; p = 0.50      # initialize
X   = rbinom(B,size=n,prob=p)    # random draw
pi0 = X/n                        # estimator

lcl = pi0+qnorm(0.025)*sqrt( pi0*(1-pi0)/n )
ucl = pi0-qnorm(0.025)*sqrt( pi0*(1-pi0)/n )

mean( lcl<p & p<ucl )            # coverage rate
```

seed

Line 1 sets the random number seed so our results will match. Line 2 initializes the values for the number of experiments (B), the sample size in each experiment (n), and the success probability for each experiment (p). In Line 3, R draws a sample of size B from the assumed distribution of X. Line 4 calculates the value of π_0 for each of the B experiments. Lines 5 and 6 calculate

assumed

the endpoints of the 95% confidence interval. Line 7 estimates the coverage rate. The '&' character is the logical AND symbol. Thus, `lcl<p & p<ucl` is TRUE if p is within the confidence interval; that is, it is TRUE if p is inside the interval (`lcl`, `ucl`). This gives the coverage rate.

This code estimates the coverage rate as 89.0%, which is smaller than the expected coverage rate of 95%. This means we reject at more than twice the claimed rate of $\alpha = 0.05$.

Figure 11.5 illustrates this. The figure is a plot of 100 estimated confidence intervals. When the actual value of π is outside the estimated confidence interval, the interval is colored orange. In this example, 88 of the 100 intervals covered the population proportion ($\pi = 0.500$). Thus, the coverage rate is approximately 88%. Thus, the rejection rate is approximately 0.12, not $\alpha = 0.05$.

Extending this code to test the coverage rates of all four confidence interval estimators is clear. It just requires estimating the confidence limits for each method.

```
set.seed(5)
B = 1e6; n = 10; p = 0.50
X = rbinom(B, size=n,prob=p)

pi0 = X/n
  lcl = pi0+qnorm(0.025)*sqrt( pi0*(1-pi0)/n )
  ucl = pi0-qnorm(0.025)*sqrt( pi0*(1-pi0)/n )
  coverage0 = mean( lcl<p & p<ucl )

pi1 = (X+1)/(n+2)
  lcl = pi1+qnorm(0.025)*sqrt( pi1*(1-pi1)/n )
  ucl = pi1-qnorm(0.025)*sqrt( pi1*(1-pi1)/n )
  coverage1 = mean( lcl<p & p<ucl )

pi2 = (X+2)/(n+4)
  lcl = pi2+qnorm(0.025)*sqrt( pi2*(1-pi2)/n )
  ucl = pi2-qnorm(0.025)*sqrt( pi2*(1-pi2)/n )
  coverage2 = mean( lcl<p & p<ucl )

pi3 = 1/2
  lcl = pi3+qnorm(0.025)*sqrt( pi3*(1-pi3)/n )
  ucl = pi3-qnorm(0.025)*sqrt( pi3*(1-pi3)/n )
  coverage3 = mean( lcl<p & p<ucl )
```

This code produces estimates of the coverage rates for the four methods. The coverage rate for the π_0 method is 89.0%; the π_1 method, 97.8%; the π_2

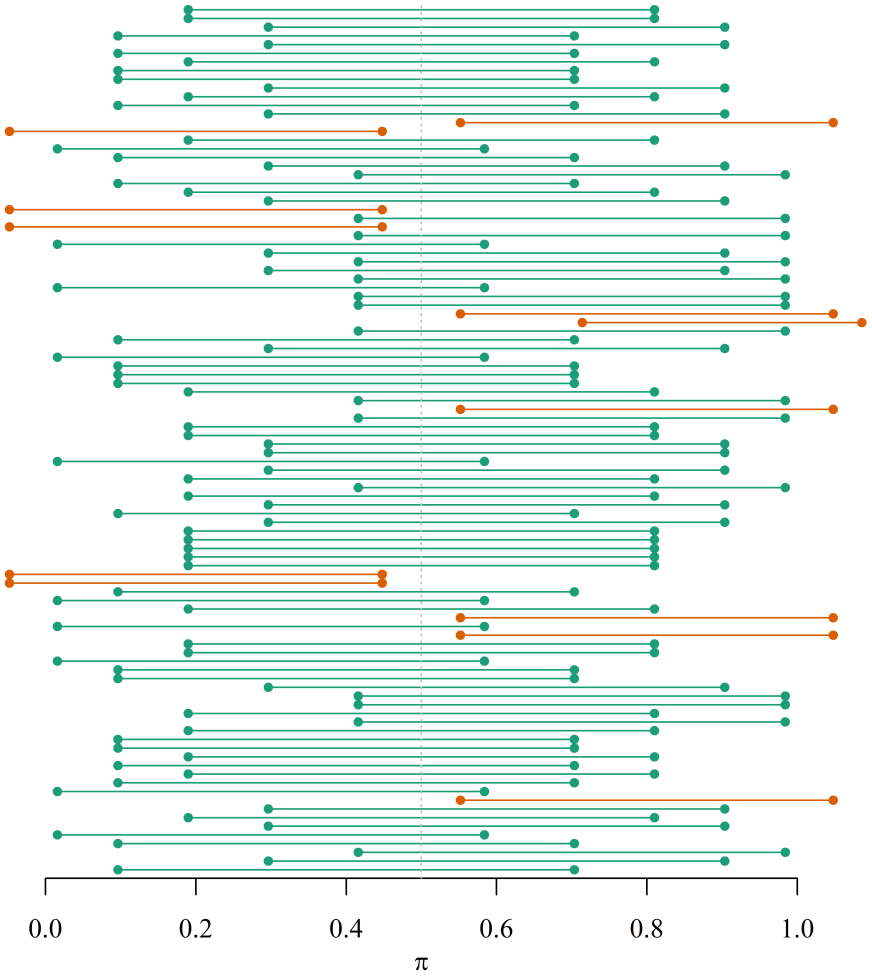

Figure 11.5: *Illustration of coverage rate for the example in the text. Note that 88 of the 100 displayed confidence intervals cover the population proportion* $\pi = 0.50$*. Thus, the coverage rate is* 88%*.*

method, 97.8%; and the π_3 method, 1. Of these, the middle two methods produce coverages closest to expected (95%).

Thus, either the π_1 or the π_2 method produces the better confidence intervals under these circumstances. ◇

As you would expect, the proper method depends on the unknown population parameter π. The above simulation assumed $\pi = 0.500$. To calculate the

Method:	π_0	π_1	π_2	π_3
$\pi = 0.01$	0.260	1.000	0.997	0.000
$\pi = 0.05$	0.782	0.997	0.984	0.000
$\pi = 0.10$	0.808	0.950	0.974	0.000
$\pi = 0.20$	0.946	0.930	0.964	0.000
$\pi = 0.45$	0.935	0.935	0.974	1.000
$\pi = 0.50$	0.957	0.957	0.957	1.000

Table 11.1: *Table of estimated coverage rates for the four methods discussed in the text and for several values of π. For this table, $n = 30$, $\alpha = 0.05$, and the number of replications is $B = 1,000,000$.*

coverage rates for other values of π and other values of n, just change the appropriate values in Line 2.

Table 11.1 provides estimates of coverage rates for several values of the population proportion π for each of the four methods discussed above. Here, $n = 30$. For all six values of π, the adjusted estimators are better than the usual estimator and the fixed estimator. Beyond that observation, there is not enough information as to whether π_1 or π_2 should be used.[2]

11.3: A Final Example

Let us now return to the opening example of this chapter:

> An opinion poll performed prior to the 2012 US Presidential election found that 370 of the 1000 respondents stated they would vote for Barack Obama; 340 for Mitt Romney; and 290 would not vote. What is the best estimate of the proportion of Americans voting for Mitt Romney? What is the best 95% confidence interval for that quantity?

Solution: We have several estimators available for proportions. Section 11.1 covers several of them. Since our sample proportion is 340/710 = 0.479 one

estimator

[2] As a side note, Agresti and Coull (1998) suggest a different estimator, the Wilson-Agresti-Coull estimator $\pi_{\text{WAC}} = \frac{X+0.5Z^2_{\alpha/2}}{n+Z^2_{\alpha/2}}$.

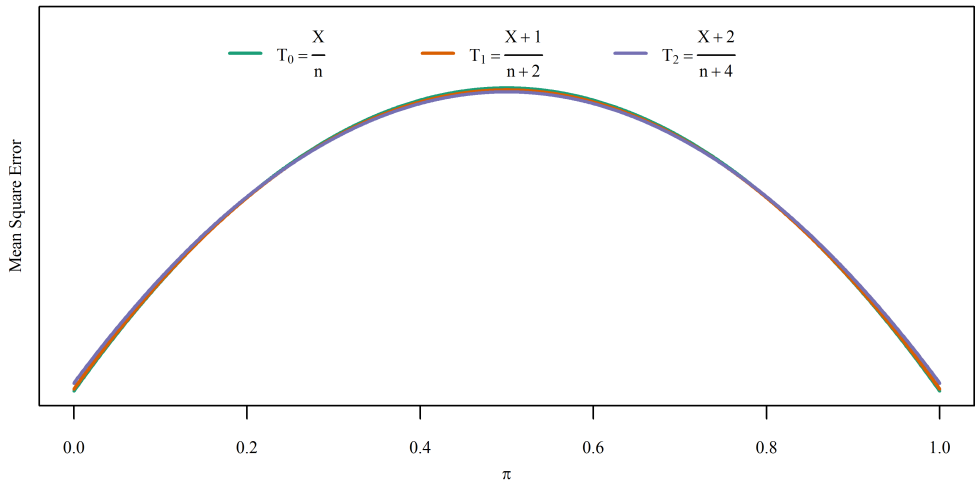

Figure 11.6: *Graphic of mean square error curves for the example in the text. The three curves are so close because the sample size and the number of successes is large.*

of the adjusted estimators seems to be appropriate here. Before we make our final decision, we need to determine which of the estimators has the lowest mean square error under these circumstances.

mean square error

Figure 11.6 plots the three mean square error curves. Note that there is little difference between them at this sample size. As such, none of the three would be inappropriate. For the sake of simplicity, which is often a consideration when explaining methods, let us use the Wald estimator, T_0.

large n

Using $T_1 = \frac{x+1}{n+2}$, our estimate of the proportion of voters supporting Mitt Romney is 0.479. Note that the other two estimators give the same answers (to three digits). This is because of the large sample size.

Next, we need to determine the best confidence interval to use of those we know. Section 11.2 provides the script to run (with appropriate changes). According to that script, the coverage rates for the three estimators are T_0 : 0.948, T_1 : 0.953, and T_2 : 0.953. Again note that there is not much difference among the estimators. This is also due to the large sample size.

As consistency is good, let us use the T_0 confidence interval. Accordingly, a 95% confidence interval for the level of support amongst voters for Mitt Romney is between 0.439 and 0.519. ◇

Note: In this case, we were able to match the estimator and the confidence interval. This will not always be the case. When working with 'exotic' estimators, the estimators and the confidence intervals may be from two different sources. There is no problem with this since the two are estimating different things.

11.4: Conclusion

In this chapter, we turned our attention to the estimators and the confidence intervals themselves. The preferred estimator of a population parameter has the lowest mean squared error. The preferred confidence interval has a coverage rate closest to the claimed rate, $(1-\alpha)100\%$.

mean squared error

coverage rate

While both mean square error and coverage rates can sometimes be calculated directly, simulation allows one to estimate the quantity in many more cases. Simulation requires the same general steps in each case. For estimating the mean square error: draw a random sample, calculate the estimate, calculate the mean square error, $\mathbb{E}\left[(T-\tau)^2\right]$.

Monte Carlo

For the coverage rate: draw a random sample, calculate the upper and lower confidence limits, calculate the proportion of confidence intervals containing (covering) the true population parameter.

Again, the preferred estimator and confidence interval is a function of the population parameter. As such, one cannot be positive which estimator (or confidence interval) is optimal. However, one usually has an estimate of the population parameter; thus, an appropriate estimator (and confidence interval) can be suggested.

In the next two chapters, we cover independence between two variables of the same type. The next chapter shows how to test for independence between two categorical variables; the following chapter, two quantitative variables. The first is called a chi-squared test. The second is called linear regression. While testing for independence between two categorical variables is usually an end itself, while we are testing independence between continuous variables we begin to realizing that we are modeling some unknown process.

11.5: End of Chapter Materials

11.5.1 R Functions

In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

Probability:

rcauchy(n) Returns n random numbers from the specified Cauchy distribution: `rcauchy(100, location=3, scale=6)` gives 100 random numbers drawn from a cauchy distribution with median 3 and scale (spread) 6.

rnorm(n) Returns n random numbers from the specified Normal distribution: `rnorm(100,m=3,s=6)` gives 100 random numbers drawn from a $\mathcal{N}(\mu = 3, \sigma = 6)$ distribution.

qnorm(p) Returns the value of x corresponding to the p-value provided according to the specified Normal distribution: `qnorm(0.95,m=5,s=1)` returns the x-value such that $\mathbb{P}[X < x] = 0.95$, where X is distributed as $\mathcal{N}(\mu = 5, \sigma = 1)$.

rexp(n) Returns n random numbers from the specified Exponential distribution: `rexp(100,r=3)` gives 100 random numbers drawn from an $\mathcal{E}xp(\lambda = 3)$ distribution.

Mathematics:

apply(X, MARGIN, FUN) Applies function `FUN` to the columns (`MARGIN=1`) or rows (`MARGIN=2`) of the matrix `X`. This is very useful in speeding up the script as loops tend to be slow.

matrix(x, nrow=R, byrow=FALSE) Creates a two-dimensional matrix out of the `x` vector. The matrix will have `R` rows, and the cells will be assigned in column order. If you wish to assign the cell values in row order, use `byrow=TRUE`.

11.5.2 Exercises and Extensions

This section offers suggestions on things you can practice from this chapter. Save the scripts in your Chapter 11 folder. For each of the following problems, please save the associated `R` script in the chapter folder as `ext0x.R`, where `x` is the problem number.

Summary:

1. Define the mean square error. For what is it used?
2. Let us suppose I have two estimators. The mean square error of the first is 0.4; of the second, 0.8. Which estimator should I use? Explain.
3. What is the definition of coverage rate? If we use $\alpha = 0.05$, what should the coverage rate equal?
4. If the actual coverage rate is greater than expected, what effect will that have on the power?

Theory:

5. Prove Theorem 11.1 in the case that X is discrete.
6. In Example 11.1, I stated that the bias of the T_1 estimator is not zero unless $\pi = 0.500$. Prove this.
7. In Example 11.4, we estimated the mean square error using simulation. Repeat the example, but estimate the mean square error directly using the approximate distribution of the sample median.
8. Directly calculate the mean squared error of the following estimators of π:
   a) $T_2 = \frac{X+2}{n+4}$, which is the Agresti-Coull estimator.
   b) $T_{\text{WAC}} = \frac{X+\frac{1}{2}1.96^2}{n+1.96^2}$, which is the Wilson-Agresti-Coull estimator for the case of $\alpha = 0.05$.
   c) $T_n = \frac{X+n}{2n}$, which does not have a name.
9. The above estimators all have the same form:

$$T_k = \frac{X + k}{n + 2k}$$

If we know $n = 100$ and $\pi \approx 0.40$, which value of k produces the lowest mean square error? Prove this through direct calculation.

10. Create a graphic like Figure 11.1 comparing all three estimators of Problem 8, above. Let $n = 10$.

Data:

11. I am curious about the proportion of water in Fort Loudoun Lake that has bacteria levels in excess of safety. To estimate this quantity, I collect 30 samples from around the lake. Of those samples 25 have bacteria levels in excess of safety. Estimate the best estimate of the population proportion. Estimate the best 95% confidence interval for the population proportion.

Monte Carlo:

12. Use simulation to estimate the mean square error for each estimator in Problem 8. Again, let $n = 10$.

13. Assume $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu = 0, 1)$, for $i = \{1, 2, \ldots, n\}$. Use simulation to estimate the coverage rate of the usual confidence interval for μ, $\overline{X} \pm Z_{\alpha/2}\sqrt{s^2/n}$, where s^2 is the sample variance. Use simulation to estimate the coverage rate of an unusual confidence interval for μ, $\widetilde{X} \pm Z_{\alpha/2}\sqrt{\frac{\pi}{2n}}$. If $n = 5$, which of the two confidence intervals is better? Note that $\pi = 3.1415\ldots$ in this problem.

14. Assume $Y_i \overset{\text{iid}}{\sim} \text{Cauchy}(\tilde{\mu} = 0, 1)$, for $i = \{1, 2, \ldots, n\}$. Use simulation to estimate the coverage rate of the confidence interval for $\tilde{\mu}$, $\overline{X} \pm Z_{\alpha/2}\sqrt{s^2/n}$, where s^2 is the sample variance. Use simulation to estimate the coverage rate of this other confidence interval for $\tilde{\mu}$, $\widetilde{X} \pm Z_{\alpha/2}\sqrt{\frac{\pi}{2n}}$. If $n = 5$, which of the two confidence intervals is better? Note that $\pi = 3.1415\ldots$ in this problem.

15. Let us assume $T_i \overset{\text{iid}}{\sim} \mathcal{E}xp(\lambda)$, where λ is not known, but is believed to be between 1 and 5. Create two estimators of λ. Either directly calculate the mean square error of each or estimate it using simulation. If $n = 10$, which of your two estimators is better?

16. Let us assume $T_i \overset{\text{iid}}{\sim} \mathcal{E}xp(\lambda)$, where λ is not known, but is believed to be between 1 and 5. Create two estimators of a confidence interval for λ. Using simulation, determine the coverage rate of your two intervals. If $n = 10$, which of the two is better?

11.5.3 References and Additional Readings

- Alan Agresti. (2002) *Categorical Data Analysis*, Second Edn. New York: Wiley-Interscience.
- Alan Agresti. (2007) *An Introduction to Categorical Data Analysis*. New York: Wiley Series in Probability and Statistics.
- Alan Agresti. (2010) *Analysis of Ordinal Categorical Data*, Second Edition New York: Wiley Series in Probability and Statistics.
- Alan Agresti and Brent A. Coull. (1998) "Approximate Is Better than 'Exact' for Interval Estimation of Binomial Proportions." *The American Statistician*, **52**(2): 119-126.
- Abraham Wald and Jacob Wolfowitz. (1939) "Confidence Limits for Continuous Distribution Functions." *The Annals of Mathematical Statistics*, **10**(2): 105-118.
- Edwin B. Wilson. (1927) "Probable Inference, the Law of Succession, and Statistical Inference." *Journal of the American Statistical Association*, **22**(158): 209-212.