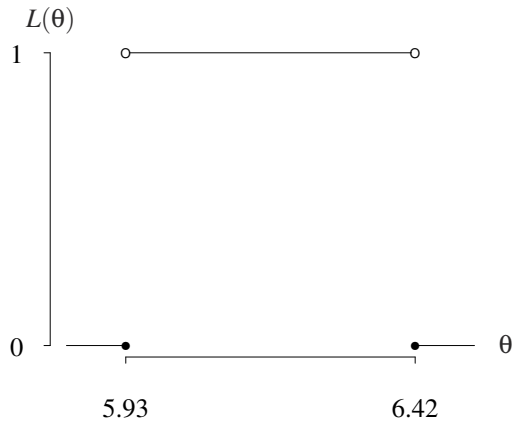

Figure 2.13: The likelihood function for $n = 3$ observations from a $U(\theta, \theta + 1)$ population.

2.4 Properties of Point Estimators

The two previous sections have introduced two techniques for finding a point estimator of an unknown population parameter θ from a random sample $X_1, X_2, \ldots, X_n$: the method of moments and maximum likelihood estimation. But the techniques only define the point estimator—they do not give us any information about the quality of the estimator. This section introduces several properties that reflect the quality of a point estimator.

Data which was not previously collected (due to the resources required to do so) is now being collected automatically, which has resulted in huge data sets. The notion of "big data" has arisen in the past few years as the ability to capture and store large amounts of data automatically and cheaply has increased. Most of the data captured throughout history has been collected in just the past few years. So how can these massive data sets be summarized? One way is with a statistical graphic. A second way is with some carefully selected statistics that capture the essence of the data values. These statistics should be selected so that they reduce the data set in a manner that keeps the information of interest and discards the information that is not of interest.

It is too much to expect that a point estimator $\hat{\theta}$ would always match the population parameter θ exactly. Since point estimates are statistics, they are also random variables. This means that point estimates will vary from one sample to the next. But some evidence that $\hat{\theta}$ will be reasonably close to θ would certainly be reassuring. This section presents some of the thinking and analysis along these lines.

The subsections presented here outline properties of point estimators that can be useful when comparing two or more point estimators. These properties help a statistician decide which is the best to use in a particular statistical setting. The properties outlined here are unbiasedness, efficiency, sufficiency, and consistency.

Unbiased estimators

An important property of point estimators is whether or not their expected value equals the unknown parameter that they are estimating. If θ is considered the target parameter, then we want $\hat{\theta}$ to be *on target*, as defined formally next.

Definition 2.1 Let $\hat{\theta}$ denote a statistic that is calculated from the sample $X_1, X_2, \ldots, X_n$. Then $\hat{\theta}$ is an *unbiased estimator* of the unknown parameter θ defined on the parameter space Ω if and only if, for all $\theta \in \Omega$,

$$E\left[\hat{\theta}\right] = \theta.$$

If $\hat{\theta}$ is not unbiased, that is, $E\left[\hat{\theta}\right] \neq \theta$ for some $\theta \in \Omega$, then $\hat{\theta}$ is a *biased estimator* of θ.

Determining whether a point estimator $\hat{\theta}$ is an unbiased estimator of θ typically arises in the following setting. The random sample $X_1, X_2, \ldots, X_n$ is drawn from the probability distribution described by $f(x)$, which defines a parametric distribution with a single unknown parameter θ. Knowing that $E\left[\hat{\theta}\right] = \theta$ is valuable information in that the sampling distribution of $\hat{\theta}$ is centered on the target value. Although the specific value of $\hat{\theta}$ for a specific data set $x_1, x_2, \ldots, x_n$ might fall above θ or below θ, knowing that the mean value of $\hat{\theta}$ is θ assures us that our (metaphorical) arrow, the point estimator $\hat{\theta}$, is pointing at the center of the target (the true value of the unknown parameter θ).

We have now effectively partitioned the set of all point estimators $\hat{\theta}$ for the unknown parameter θ into two sets: unbiased estimators and biased estimators. The Venn diagram in Figure 2.14 illustrates this partition. All other factors being the same, one would always prefer an unbiased estimate over a biased estimate. But as subsequent examples will show, the decision is not always that clear-cut. We begin with an example of determining whether or not a point estimate is unbiased.

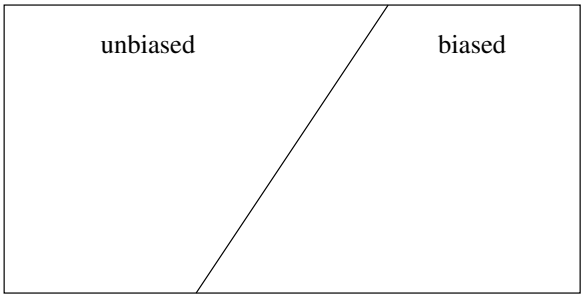

Figure 2.14: Venn diagram of unbiased and biased point estimators.

Example 2.16 Let $X_1, X_2, \ldots, X_n$ denote a random sample from a Bernoulli(p) population, where p is an unknown parameter satisfying $0 < p < 1$. Is the maximum likelihood estimator of p unbiased?

A statistician might encounter a data set of this kind in a political poll involving two candidates for elected office. Since the data is drawn from a Bernoulli population, each data value is a 0 or 1 corresponding to the respondent's preference for one of two candidates for a particular office. The Bernoulli distribution has probability mass function

$$f(x) = p^x(1-p)^{1-x} \qquad x = 0, 1.$$

So the likelihood function is

$$L(p) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\Sigma_{i=1}^{n} x_i}(1-p)^{n-\Sigma_{i=1}^{n} x_i}.$$

The log likelihood function is

$$\ln L(p) = \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p).$$

The score is the partial derivative of the log likelihood function with respect to p, which is

$$\frac{\partial \ln L(p)}{\partial p} = \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} x_i\right).$$

When the score is equated to zero,

$$\frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} x_i\right) = 0,$$

and solved for p, the resulting maximum likelihood estimate is

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

which is the sample mean. In the context of sampling from a Bernoulli population, the maximum likelihood estimate also has the interpretation as the *fraction of successes*. This is an intuitively appealing estimator because the true probability of success is being estimated by the observed fraction of successes. The second derivative of the log likelihood function is negative at the maximum likelihood estimator for all possible data sets, except for the extreme cases of all zeros or all ones, so we are assured that $\hat{p}$ maximizes the log likelihood function. We now return to the original question concerning whether $\hat{p}$ is an unbiased estimator of p. The expected value of the maximum likelihood estimator is

$$E[\hat{p}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\cdot(np) = p$$

because $\sum_{i=1}^{n} X_i$ has the binomial(n, p) distribution with expected value np. So $\hat{p}$ is an unbiased estimator of p.

This example allows some insight into the sampling distribution of $\hat{p}$ because $\hat{p}$ is the ratio of a binomial(n, p) random variable and n. This is a "scaled binomial" random variable. Using the random variable Y to denote $\hat{p}$ for notational clarity, the probability mass function of the maximum likelihood estimator $Y = \hat{p}$ is

$$f_Y(y) = \binom{n}{ny} p^{ny}(1-p)^{n-ny} \qquad y = 0, \frac{1}{n}, \frac{2}{n}, \ldots, 1.$$

This probability mass function is plotted in Figure 2.15 for $n = 50$ and $p = 3/5$. This scenario corresponds to a political poll of a random sample of 50 voters from a large population for a candidate having the support of 60% of the electorate. The sample of 50 voters is large enough so that the central limit theorem has kicked in and the probability mass function is roughly bell shaped. Could such a poll give a point estimate which incorrectly concludes that the candidate with support of 60% of the electorate will not

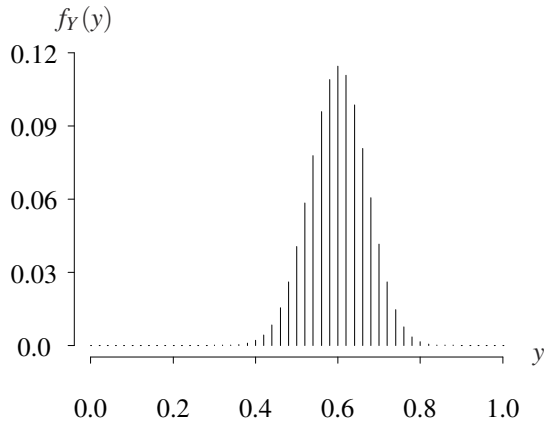

Figure 2.15: Maximum likelihood estimator probability mass function for $n = 50$ and $p = 3/5$.

win the election? This probability is $P(Y \leq 0.5) \cong 0.09781$, which is calculated with the R statement `pbinom(25, 50, 3 / 5)`. If this value is unacceptably large, then a poll with a larger sample size n must be taken. The population variance of the maximum likelihood estimator $\hat{p}$ can also be easily calculated:

$$V[\hat{p}] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} V\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}.$$

The precision of $\hat{p}$ depends only on the sample size and the true value of p. Notice that the population variance of the maximum likelihood estimator $\hat{p}$ goes to zero as the sample size goes to infinity. For $n = 50$ and $p = 0.6$,

$$V[\hat{p}] = \frac{0.6(1-0.6)}{50} = 0.0048.$$

So the standard error of the estimate of p is

$$\sigma_{\hat{p}} = \sqrt{0.0048} \cong 0.06928.$$

A standard error of $\hat{p}$ of 7% in a political poll might be unacceptably high. This is why sample sizes in political polls are substantially higher than $n = 50$. A sample size of $n = 1000$, for example, reduces the standard error of the estimate of p when $p = 0.6$ to $\sigma_{\hat{p}} \cong 0.01549$.

In the previous example, we considered a specific parametric population distribution—the Bernoulli distribution—with a single unknown parameter p. But the notion of an unbiased estimator is more general. The next result applies to any population distribution, with only the rather mild restriction that the first two population moments must be finite. Recall from Theorem 1.1 that

$$E\big[\bar{X}\big] = \mu,$$

where $\bar{X}$ is the sample mean, μ is the finite population mean, and $X_1, X_2, \ldots, X_n$ are the values from a random sample from *any* population distribution. This result indicated that $\bar{X}$ is an unbiased

estimator of μ. The next result indicates that the sample variance S^2 is an unbiased estimator of the population variance σ^2.

Theorem 2.2 Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with finite population mean μ and finite population variance σ^2. The expected value of the sample variance is

$$E\left[S^2\right] = \sigma^2,$$

which means S^2 is an unbiased estimate of σ^2.

Proof The expected value of the sample variance is

$$\begin{aligned}
E\left[S^2\right] &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] \\
&= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i^2-2X_i\bar{X}+\bar{X}^2\right)\right] \\
&= \frac{1}{n-1}E\left[\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+n\bar{X}^2\right] \\
&= \frac{1}{n-1}E\left[\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right] \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}E\left[X_i^2\right]-nE\left[\bar{X}^2\right]\right) \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(V\left[X_i\right]+E\left[X_i\right]^2\right)-n\left(V\left[\bar{X}\right]+E\left[\bar{X}\right]^2\right)\right] \\
&= \frac{1}{n-1}\left[n\sigma^2+n\mu^2-n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \\
&= \sigma^2,
\end{aligned}$$

which, using Theorem 1.1, proves that S^2 is an unbiased estimator of σ^2. □

The fact that S^2 is an unbiased estimator of σ^2 for any population distribution is one of the most compelling reasons to use the $n-1$ in the denominator of S^2 in Definition 1.10. This result does not imply, however, that $E[S] = \sigma$. Even so, statisticians often use S to estimate σ, sometimes including an unbiasing constant for small values of the sample size n.

So far we have seen that $E\left[\hat{p}\right] = p$ for random sampling from a Bernoulli(p) population and $E\left[S^2\right] = \sigma^2$ for random sampling from any population with finite population mean and variance. It is helpful to define the bias explicitly in order to have a measure of the expected distance between a point estimator and its target value.

Definition 2.2 Let $\hat{\theta}$ denote a statistic that is calculated from the sample $X_1, X_2, \ldots, X_n$. The *bias* associated with using $\hat{\theta}$ as an estimator of θ is

$$B\left(\hat{\theta}, \theta\right) = E\left[\hat{\theta}\right] - \theta.$$

There is a subset of the biased estimators that is of interest. The classification is a bit of a consolation prize for biased estimators. Their redeeming feature is that although they are biased estimators for finite sample sizes n, they are unbiased in the limit as $n \to \infty$. These estimators are known as asymptotically unbiased estimators and are defined formally below.

Definition 2.3 Let $\hat{\theta}$ denote a statistic that is calculated from the sample $X_1, X_2, \ldots, X_n$. If

$$\lim_{n\to\infty} B(\hat{\theta}, \theta) = 0,$$

then $\hat{\theta}$ is an *asymptotically unbiased estimator* of θ.

All unbiased estimators are necessarily asymptotically unbiased. But only some of the biased estimators are asymptotically unbiased. To this end, we subdivide the biased portion of the Venn diagram from Figure 2.14 to include asymptotically unbiased estimators in Figure 2.16.

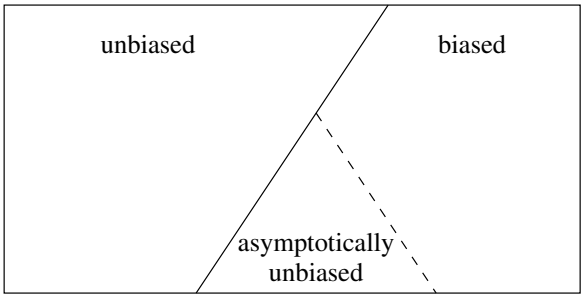

Figure 2.16: Venn diagram of unbiased, biased, and asymptotically unbiased point estimators.

Example 2.17 Let $X_1, X_2, \ldots, X_n$ denote a random sample from a $U(0, \theta)$ population, where θ is a positive unknown parameter. Classify the following point estimators of θ into the categories given in Figure 2.16 and select the best estimator:

- $2\bar{X}$,
- $3\bar{X}$,
- $X_{(n)}$,
- $(n+1)X_{(n)}/n$,
- $(n+1)X_{(1)}$,
- 17,

where $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$ and $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$.

When faced with a real data set, we oftentimes have to choose a point estimator from a set of potential point estimators such as this. The purpose of this example is to investigate the properties of these six point estimators.

- The first point estimator, $2\bar{X}$, is the method of moments estimator. The derivation was given in Example 2.2. Since the population mean of the $U(0, \theta)$ distribution

is $\theta/2$ and

$$E\left[2\bar{X}\right] = 2E\left[\bar{X}\right] = 2 \cdot \frac{\theta}{2} = \theta,$$

via Theorem 1.1, the method of moments estimator is classified as an unbiased estimator.

- The point estimator $3\bar{X}$ is classified as a biased estimator because

$$E\left[3\bar{X}\right] = 3E\left[\bar{X}\right] = 3 \cdot \frac{\theta}{2} = \frac{3}{2}\theta.$$

This estimator overestimates the population parameter θ on average. The positive bias is

$$B\left(\hat{\theta}, \theta\right) = B\left(3\bar{X}, \theta\right) = E\left[3\bar{X}\right] - \theta = \frac{3}{2}\theta - \theta = \frac{\theta}{2}.$$

- The point estimator $X_{(n)}$ is the maximum likelihood estimator. The derivation, after some minor manipulation of the objective function or the support of the population distribution, was given in Example 2.10. Using an order statistic result, the expected value of $X_{(n)}$ is

$$E\left[X_{(n)}\right] = \frac{n\theta}{n+1}.$$

The maximum likelihood estimator underestimates the population parameter θ, on average, because $E\left[X_{(n)}\right]$ is less than θ. Since the expected value is not equal to θ for finite values of n, this estimator is biased. The bias is

$$B\left(\hat{\theta}, \theta\right) = B(X_{(n)}, \theta) = E\left[X_{(n)}\right] - \theta = \frac{n\theta}{n+1} - \theta = -\frac{\theta}{n+1}.$$

This estimator should be classified as asymptotically unbiased, however, because

$$\lim_{n\to\infty} B\left(\hat{\theta}, \theta\right) = \lim_{n\to\infty} \left(-\frac{\theta}{n+1}\right) = 0.$$

- The point estimator $(n+1)X_{(n)}/n$ was presented in Example 2.10 as a modification of the maximum likelihood estimator that included an unbiasing constant. The expected value of $(n+1)X_{(n)}/n$ is

$$E\left[\frac{n+1}{n} X_{(n)}\right] = \frac{n+1}{n} \cdot \frac{n\theta}{n+1} = \theta,$$

so this point estimator is classified as an unbiased estimator.

- The point estimator $(n+1)X_{(1)}$ is also an unbiased estimator of θ, that is,

$$E\left[(n+1)X_{(1)}\right] = \theta.$$

This can be seen by invoking an order statistic result and computing the appropriate expected value.

- The point estimator $\hat{\theta} = 17$ is quite bizarre. The statistician simply ignores the data values $X_1, X_2, \ldots, X_n$ and pulls 17 out of thin air as the estimate of θ. The expected value of $\hat{\theta}$ is $E\left[\hat{\theta}\right] = E[17] = 17$, which is not θ (unless θ just happens to be 17), so this estimator is classified as a biased estimator. Recall from Definition 2.1 that an estimator $\hat{\theta}$ is an unbiased estimator of θ if $E\left[\hat{\theta}\right] = \theta$ for *all* $\theta \in \Omega$.

We now know that three of the six suggested point estimators are unbiased. The results of our analysis are summarized in the Venn diagram in Figure 2.17.

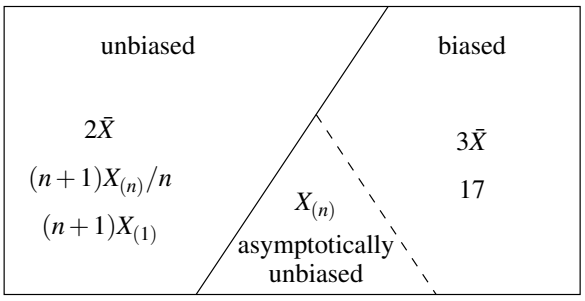

Figure 2.17: Venn diagram of several point estimates of θ for a $U(0, \theta)$ population.

Now to the more difficult question: which is the best of the six point estimators? This is a purposefully vague question at this point, so the question will be addressed from several different angles. The choice between the point estimators boils down to which point estimator will perform best for a higher fraction of data sets drawn from a $U(0, \theta)$ population than the other point estimators. This does not imply, of course, that the point estimator selected will be the best for every data set. We begin by plotting the sampling distributions of the three unbiased estimators to gain some additional insight. This can only be done for specific values of n and θ, so let's arbitrarily choose $n = 5$ and $\theta = 10$. For this choice, the probability density functions of $2\bar{X}$, $6X_{(5)}/5$, and $6X_{(1)}$ are plotted in Figure 2.18. APPL was used to calculate the probability density functions. The sampling distributions of $2\bar{X}$, $6X_{(5)}/5$, and $6X_{(1)}$ reveal vastly different shapes even though all three have expected value $\theta = 10$. The probability density function of $2\bar{X}$ is bell shaped (via the central limit theorem) and symmetric about $\theta = 10$; the probability

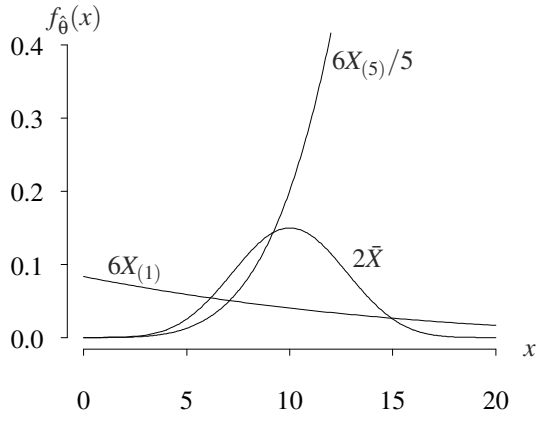

Figure 2.18: Sampling distributions of $2\bar{X}$, $6X_{(5)}/5$, and $6X_{(1)}$ when $n = 5$ and $\theta = 10$.

density functions of $6X_{(5)}/5$ and $6X_{(1)}$ are skewed distributions. Since the support of the population is $(0, 10)$, the support of $2\bar{X}$ is $(0, 20)$, the support of $6X_{(5)}/5$ is $(0, 12)$, and the support of $6X_{(1)}$ is $(0, 60)$. Figure 2.18 reveals that $6X_{(1)}$ has a significantly larger population variance than the other two unbiased estimators, so it is probably the weakest candidate of the three unbiased estimators.

Since the population variance of the point estimators plays a critical role in analyzing the sampling distributions of the three unbiased estimators, it is worthwhile calculating the population means and population variances of all six of the estimators. The values are summarized in Table 2.5. Notice that the three unbiased estimators, $2\bar{X}$, $(n+1)X_{(n)}/n$, and $(n+1)X_{(1)}$ all collapse to the same estimator when $n = 1$; the point estimator is just double the single observation. Choosing the point estimator with the smallest population variance is not appropriate here because this would result in choosing the strange point estimate $\hat{\theta} = 17$. Instead, it is advantageous to choose the unbiased estimator with the smallest variance. Using this criteria, $(n+1)X_{(n)}/n$ has the smallest population variance of the three unbiased estimators for samples of $n = 2$ or more observations.

Point estimate $\hat{\theta}$	$E\left[\hat{\theta}\right]$	$V\left[\hat{\theta}\right]$	Categorization
$2\bar{X}$	θ	$\dfrac{\theta^2}{3n}$	unbiased
$3\bar{X}$	$\dfrac{3\theta}{2}$	$\dfrac{3\theta^2}{4n}$	biased
$X_{(n)}$	$\dfrac{n\theta}{n+1}$	$\dfrac{n\theta^2}{(n+2)(n+1)^2}$	asymptotically unbiased
$\dfrac{(n+1)X_{(n)}}{n}$	θ	$\dfrac{\theta^2}{n(n+2)}$	unbiased
$(n+1)X_{(1)}$	θ	$\dfrac{n\theta^2}{n+2}$	unbiased
17	17	0	biased

Table 2.5: Population means and variances of the six point estimators of θ.

But the unbiased estimator with the smallest population variance is not the only criteria that can be used to select the preferred estimator. The R code below conducts a Monte Carlo simulation with 10,000 random samples of size $n = 5$ from a $U(0, \theta)$ population when $\theta = 10$. All six point estimators are calculated for each sample, and the point estimator that lies closest to θ is identified and tabulated. Finally, the fraction of times that each estimator is closest to $\theta = 10$ is printed.

```
set.seed(8)
n = 5
theta = 10
nrep = 10000
theta.hat = numeric(6)
count = numeric(6)
```

```
for (i in 1:nrep) {
  x = runif(n, 0, theta)
  theta.hat[1] = 2 * mean(x)
  theta.hat[2] = 3 * mean(x)
  theta.hat[3] = max(x)
  theta.hat[4] = (n + 1) * max(x) / n
  theta.hat[5] = (n + 1) * min(x)
  theta.hat[6] = 17
  index = which.min(abs(theta.hat - theta))
  count[index] = count[index] + 1
}
print(count / nrep)
```

The results of the Monte Carlo simulation are given in Table 2.6 for sample sizes $n = 5$, $n = 50$, and $n = 500$. The entries give the fractions of the simulations giving the closest estimator to the true parameter value $\theta = 10$. As expected, the column sums of the entries in the table equal 1. When $n = 5$, even the maligned $\hat{\theta} = 17$ is the closest to $\theta = 10$ for two of the 10,000 random samples. The reader is encouraged to imagine what type of data set would lead to this awful estimator outdoing the other estimators. Table 2.6 shows that, by a somewhat narrow margin, the unbiased estimator $(n+1)X_{(n)}/n$ dominates the other estimators for the sample sizes considered here.

Point estimate	$n = 5$	$n = 50$	$n = 500$
$2\bar{X}$	0.1765	0.0912	0.0328
$3\bar{X}$	0.1323	0.0000	0.0000
$X_{(n)}$	0.3178	0.3749	0.3905
$(n+1)X_{(n)}/n$	0.3262	0.5275	0.5762
$(n+1)X_{(1)}$	0.0470	0.0064	0.0005
17	0.0002	0.0000	0.0000

Table 2.6: Monte Carlo simulation results for a $U(0,10)$ population.

In summary, based on $\hat{\theta} = (n+1)X_{(n)}/n$ being (a) an unbiased estimate, (b) the unbiased estimate with the smallest population variance, and (c) the estimate that is most likely to be the closest to the population value of θ for several sample sizes in a Monte Carlo experiment, we conclude that $\hat{\theta} = (n+1)X_{(n)}/n$ is the best of the six point estimators. It carries the additional bonus that all of the data values are necessarily less than $\hat{\theta}$, which is a desirable property for this particular population distribution.

Keep in mind that although the Monte Carlo simulation used 10,000 replications, the statistician only sees a single data set of n observations and must choose among the six candidate point estimators. We have chosen the unbiased estimator $(n+1)X_{(n)}/n$ as the best of the six because it performs the best on average.

This example has brought up three issues concerning point estimators that will be addressed in the paragraphs that follow.