5. Assessing the Performance of Estimators

5.1 Different methods of Estimation

So far we have seen two general (parametric) model-based methods for estimating a population quantity (the method of moments and the methods of maximum likelihood), in which we find an expression for the population quantity in terms of the unknown parameter θ, use the principle in question to estimate θ, and finally plug the estimate $\hat{\theta}$ into the expression to give an estimate for the population quantity. Often there may also be a direct 'non-parametric' alternative, in which we simply use the relevant sample quantity to estimate the population value.

For example, say we wanted to estimate the population median for a population which has a Uniform$(0, \theta)$ distribution where θ is unknown, using a random sample with values $x_1, \ldots, x_n$. The parametric methods use the fact that the population median for this distribution is $\theta/2$.

- The method of moments estimates θ by $\hat{\theta}_{mom} = 2\bar{x}$ and so estimates the population median by $\hat{\theta}_{mom}/2 = \bar{x}$.
- The method of maximum likelihood estimates θ by $\hat{\theta}_{mle} = x_{(n)}$, where $x_{(n)} = \max\{x_1, \ldots, x_n\}$ is the largest value in the sample, and thus estimates the population median by $\hat{\theta}_{mle}/2 = x_{(n)}/2$.
- The non-parametric method estimates the population median by the sample median.

For a given set of sample data, the three methods will result in three different estimates. The questions are which estimate (or method of estimation) is best, and how can we compare the methods when we don't actually know the true value of the quantity we want to estimate.

If we do not know the true value we are trying to estimate, we cannot usefully compare methods of estimation using only the resulting numerical estimates from a single sample.

5.2 Repeated sampling, and sampling distributions

The main way we compare methods of inference is to see how they perform under *repeated sampling*. That is, we imagine future hypothetical samples of the same size from the same distribution, and examine how well each method performs in the long run. In probability language, we treat the sample as a collection of random variables $X_1, X_2, \ldots, X_n$, regard the estimators as functions of these random variables, and look at the distributions of these estimators. These estimator distributions are called *sampling distributions*, to help distinguish from the original population distribution.

A good estimator is one whose sampling distribution is concentrated close to the true value of the quantity it is trying to estimate. A poor estimate is one where the sampling distribution is either very spread out, or is concentrated around the wrong value.

In some cases, we can use methods like those featured in the Probability course to calculate a sampling distribution theoretically. For example, if $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, we know that $\hat{\mu}_{\text{mom}} = \overline{X}$, and that $\overline{X} \sim N(\mu, \sigma^2/n)$. We will see more examples later. A more general, but empirical, approach is to use simulation.

5.3 Evaluating sampling distributions of estimators using simulation

In statistics, simulation is the process of artificially generating a data set that has the same properties as a set of independent observations from a given probability distribution.

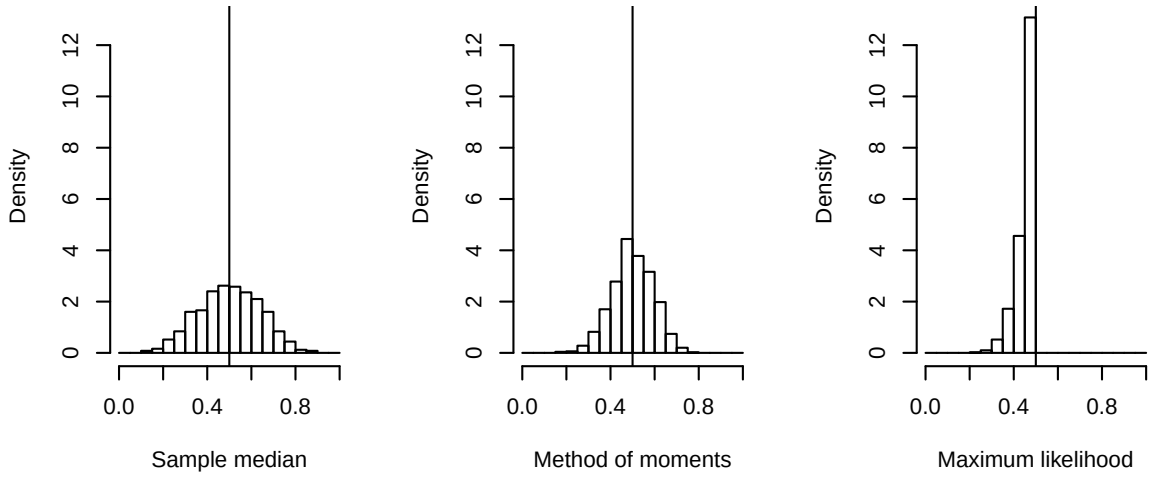
Simulation-based procedures for evaluating a method of estimation work by replacing the above idea of hypothetical future samples and probability calculations, with actual simulated numerical samples and numerical calculations. Thus, for a particular type of population distribution $f(x;\theta)$, we take particular values for the parameter θ and the sample size n. Then we generate a number (say B) of artificial data sets, each of which looks like a simple random sample of this fixed size n from this particular distribution with this particular value of θ, and calculate an estimate for each data set, giving a total of B different estimates. The idea is that this process represents, say, the experience of a single statistician repeatedly using the method a total of B times in similar statistical circumstances or, alternatively, the overall experience of B independent statisticians each using the method once.

If B is large, the values of the estimates generated in these B repeated independent experiments should give a good indication of the sampling distribution, and hence the overall performance of the method. Moreover, we can get a good idea of the relative strengths and weaknesses of different methods of estimation by comparing their overall performances on the same B data sets.

5.4 Exploring the performance of different methods – histograms

One way of exploring the performance of different methods is just to plot a histogram of the B estimates obtained in the simulation study.

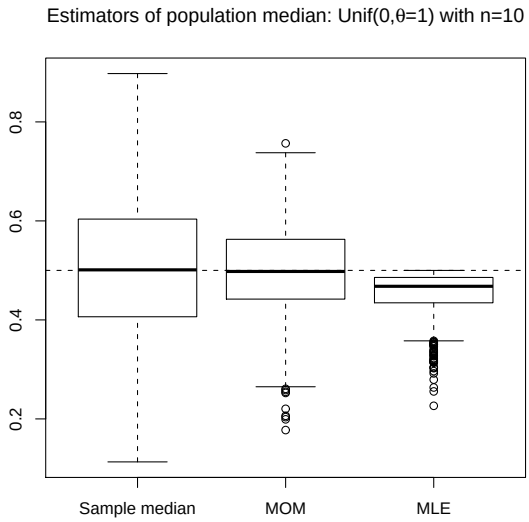
To see how this works, consider again the problem of estimating the population median for the Uniform$(0,\theta)$ distribution. The histograms below were constructed by simulating $B = 1000$ samples, each of size $n = 10$, from a Uniform distribution with $\theta = 1$ (so the true population median was $\theta/2 = 0.5$); computing the sample median, the method of moments estimate ($\bar{x}$) and the maximum likelihood estimate ($\max\{x_1,\ldots,x_n\}/2$) for each sample; and then plotting a histogram of the resulting 1000 estimates produced by each method. For this example, the differences in the shape of the histograms are particularly striking; however in other cases several methods may give identical or roughly similar estimates (e.g. they each give exactly the same form of estimates for the population mean when the population has an Exponential distribution or a Normal distribution).



5.5 Graphical summary of performance – boxplots

Parallel boxplots are a convenient graphical method for comparing visually the sampling distributions of several different estimators, and focussing in particular on the median of each as a measure of the centre of the distribution and the upper and lower hinges as a measure of spread.

The boxplots below correspond to the histograms in §5.4. Clearly the estimates produced by the sample median and the method of moments are both centred on the true population median value of 0.5 but fairly widely spread about this value (sample median more than mom). The maximum likelihood estimates are centred on a value just below 0.5 and so the method slightly, but consistently, underestimates the true value. However the narrow spread of mle values means that for most samples the mle is still nearer the true value than either the mom estimate or the sample median.



5.6 Numerical summary of performance – bias, variance and mean squared error

We can also use numerical rather than graphical summaries of performance, to compare the average value of our estimator(s) with the true value. Let $\hat{\theta}$ be an estimator of an unknown parameter θ. We define two key properties of its sampling distribution:

- bias$(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)$ and we say $\hat{\theta}$ is unbiased if bias$(\hat{\theta}) = 0$.
- mse$(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)^2]$ and you can check that mse($\hat{\theta}$) = Var$(\hat{\theta})$+ bias$(\hat{\theta})^2$.

In some cases we can derive explicit analytic expressions for the mean, variance and mean square error of an estimator. However, in more complicated cases we may have to evaluate these numerically from repeated samples. Say we want to estimate θ or, more generally, a function $\tau(\theta)$ whose true value is τ and our simulation study produces B estimates $\hat{\tau}_1, \ldots, \hat{\tau}_B$ with mean $\bar{\tau} = \sum_{i=1}^{B} \hat{\tau}_i / B$ and variance $\sum_{i=1}^{B} (\hat{\tau}_i^2 - \bar{\tau})^2 / (B-1)$.

Bias

The difference between the estimated value $\hat{\tau}_i$ and the true value τ represents the error in the estimate for the ith sample. Thus the average error is just $\bar{\tau} - \tau$, and the size and sign of this average error is an indicator of the bias in the method – i.e. by how much the method consistently under-estimates or over-estimates the true value.

Mean squared error
Since positive and negative errors cancel out in calculating the average error, a better measure may be the average of the squared errors, i.e. $\sum_{i=1}^{B}(\hat{\tau}_i - \tau)^2/B$. You can check that $\sum_{i=1}^{B}(\hat{\tau}_i - \tau)^2 = \sum_{i=1}^{B}(\hat{\tau}_i - \bar{\tau})^2 + B(\bar{\tau} - \tau)^2$, so that for B large the average squared error is just the variance + (average error)^2.

For our Uniform distribution example, the quantity τ that we want to estimate is the population median, and true value in our simulation study is $\tau = 0.5$. The mean, average error, variance and average squared error of the 1000 sample medians are given in the table below, together with the corresponding quantities for the 1000 method of moments estimates of the population median and the 1000 maximum likelihood estimates.

The table confirms numerically the impression from the graphical summary – the size of the average error is larger for the method of maximum likelihood than for the other two methods, but there is a much smaller spread of estimates than using the method of moments, which in turn has a smaller spread than the non-parametric method using the sample median. Here the variance term dominates in calculating the average squared error, so in this case the mle method has the smallest average squared error, then the method of moments, then the estimate based on the sample median.

Methods of estimating the population median	average error (estimates bias)	spread (estimates variance)	average squared error (estimates mse)
sample median	0.00497	0.01861	0.01863
mom	0.00376	0.00798	0.00800
mle	−0.04403	0.00168	0.00362

Returning to the case of estimating μ using a simple random sample from $N(\mu, \sigma^2)$ mentioned in §5.2, here we can calculate these key quantities analytically: the bias is 0, and the variance and mse are both σ^2/n.

5.7 Simulation using R: assessing a single estimator

The steps in a simulation study like the one above are

- generate $n \times B$ numbers (representing independent observations from the given distribution),
- arrange the values in B groups of n (representing the B simple random samples of size n)
- calculate the relevant estimate(s) for each sample
- analyse the results, numerically or graphically as required.

The following example shows how we might use **R** to investigate the performance of the method of moments when the observations come from the Unif(0,1) distribution, the method of moments estimate is $2\bar{x}$ and the quantity we want to estimate is the parameter θ, which here has true value $\theta = 1$. It uses $B = 1000$ samples each of size $n = 10$, and so requires the generation of $n \times B = 10000$ simulated observations.

```
xvalues      <- runif(10000)
xsamples     <- matrix(xvalues,nrow=1000)
sample.mean  <- apply(xsamples, 1, mean)
theta.mom    <- 2*sample.mean
hist(theta.mom)
boxplot(theta.mom)
true.theta   <- 1
mean(theta.mom - true.theta)
var(theta.mom)
mean( (theta.mom - true.theta)^2 )
```

The example introduces three new commands, `runif(), matrix()` and `apply()` – the commands `hist(), boxplot(), mean()` and `var()` were introduced in §1. Note that I have not shown the `>` prompt at the start of each line; also the names used for variables and arrays are just one I chose for clarity, and you could replace the names `xvalues, xsamples, sample.mean, theta.mom` and `true.theta` by others of your own choice.

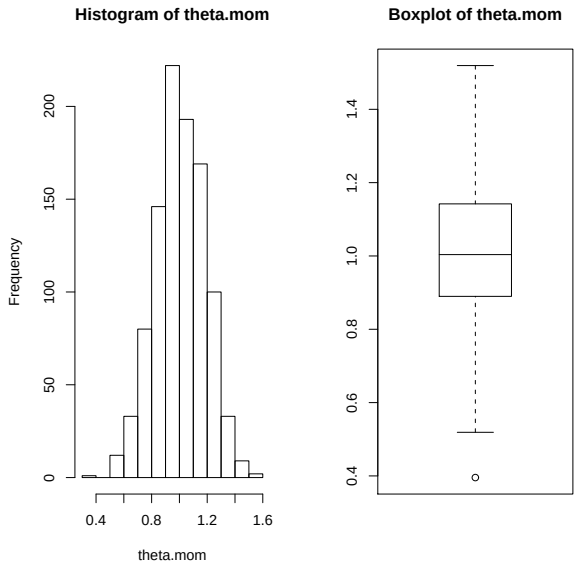
Let us look at these commands in turn:

Generating (random) numbers
`xvalues <- runif(10000)`
simulates 10000 independent *r*andom *unif*orm values from the standard Unif(0,1) distribution and assigns them to a vector which I have called `xvalues`. For observations from the Unif(0,1) distribution we just have to specify the number of observations required; to simulate observations from a general Uniform distribution we also need to specify the parameters of the distribution, e.g. for Unif(-1,2) we use `runif(10000,min =-1,max=2)` or more simply `runif(10000,-1,2)`.

Arranging into random samples
`xsamples <- matrix(xvalues, nrow=1000)`
takes the one-dimensional data array `xvalues` of length 10000, and rearranges it into a matrix `xsamples` in which the elements of `xvalues` are rearranged into 1000 rows each with $10000/1000 = 10$ values. The 10 data values in each row represent a simple random sample of size $n = 10$ from a Unif(0,1) distribution and the 1000 rows represent the $B = 1000$ independent repeated samples.

Calculating sample statistics
`sample.mean <- apply(xsamples,1,mean)`
The `apply` command works as follows: let x_{ij} denote the ijth element of the matrix `xsamples`, then for each value of the `1`st subscript `apply(xsamples, 1, mean)` applies the command `mean` to the set of values that share that `1`st subscript (i.e. to each row in turn). Thus `sample.mean` has 1000 entries (since there are 1000 rows), and these values represent the means of 1000 independent random samples, each of size 10, from a Unif(0,1) distribution. Similar commands can be used for other sample statistics, e.g. `apply(xsamples, 1, median)` or `apply(xsamples, 1, max)`

The commands produce a histogram and boxplot similar to the ones shown below, together with summary numerical values of the average error, variance and averaged squared error of the estimates. Note the bell shaped symmetrical distribution about the true value of $\theta = 1$.



5.8 Simulation using R: comparing several estimators

Once we have generated our random samples we can easily use **R** to compare different methods of estimation.

Consider again the example introduced in §5.1, where we wanted to estimate the population median $\tau = \theta/2$ for a population which has a Uniform$(0, \theta)$ distribution, where θ is unknown. Recall that the non-parametric method estimate is the sample median, the method of moments estimate is $\bar{x}$ and the mle is $\max\{x_1, \ldots, x_n\}/2$.

The following commands extend those in §5.7 to allow us to compare graphically the performance of these three methods of estimation, producing a boxplot like that in §5.5. Again, we focus on the case when the true value is $\theta = 1$, so the true value of the median is $\tau = 0.5$, and simulate $B = 1000$ samples each of size $n = 10$ from the Unif(0,1) distribution.

```
xsamples       <- matrix(runif(10000),nrow=1000)
sample.median  <- apply(xsamples,1,median)
sample.mean    <- apply(xsamples,1,mean)
sample.max     <- apply(xsamples,1,max)
tau.nonparam   <- sample.median
tau.mom        <- sample.mean
tau.mle        <- sample.max/2
boxplot(tau.nonparam,tau.mom,tau.mle)
true.tau       <- 0.5
abline(h=true.tau,lty=2)
```

Notes on these and other possible commands:

`xsamples <- matrix(runif(10000),nrow=1000)`
– here we have combined into one the two separate lines we used before, generating the raw values and formatting them as 1000 samples of size 10. Where this doesn't sacrifice readibility, this kind of thing makes the **R** code more compact, and is a good idea.

`boxplot(tau.nonparam,tau.mom,tau.mle)`
plots all three boxplots in a single figure. You could produce annotations similar to those shown

in §5.5, by using the subcommand `names` to add labels to each plot and the subcommand `main` to add an overall title, as in the command below. I've displayed the prompts to show how the command stretches over three lines.

```
> boxplot(tau.nonparam,tau.mom,tau.mle,
+ names = c("sample median","mom","mle"),
+ main="Estimators of the population median ")
```

```
abline(h=true.tau,lty=2)
```

Here the `abline()` command which we met in §2 is used to add a horizontal line to the boxplot at the true value of τ to make comparison easier. The `lty=2` gives a dashed line (type 2) rather than an ordinary line (type 1).

The extra commands below will produce a comparison plot of histograms of the three estimators, similar to that in §5.4. The `par(mfrow = c(1,3))` sets the graphics to produce a page with three plots aligned in one row (and `par(mfrow = c(1,1))` resets the graphics to one plot per page); the `xlim()` and `ylim()` commands specify the x and y plotting range explicitly rather than using default values, ensuring that all three plots are on the same scale for easier comparison.

```
par(mfrow = c(1,3))
hist(tau.nonparam,xlim=c(0,1),ylim=c(0,350))
hist(tau.mom,xlim=c(0,1),ylim=c(0,350))
hist(tau.mle,xlim=c(0,1),ylim=c(0,350))
par(mfrow = c(1,1))
```

5.9 Disadvantages of simulation-based methods for evaluating the sampling distribution

The disadvantages of simulation-based methods are that each simulation only provides information about one particular situation and gives no direct information about what would happen for:

- other sample sizes n
- other values of true parameter θ
- other types of population distribution $f(x;\theta)$
- other methods of estimation

Also, the numerical accuracy of estimates of quantities like the bias is limited by the finite size of B, the number of samples.

Therefore, as mentioned in §5.2, we use probability theory to find sampling distributions whenever this is possible.