6

Statistical Bootstrap Methods

6.1 The Jackknife and Pseudovalues

6.1.1 introduction

In this chapter, we will describe the use of the bootstrap to generate estimates of bias, standard errors, and confidence intervals around parameter estimates. We will also include a short treatment on the jackknife, which predated the bootstrap as a method for generating estimates of bias and standard errors around parameter estimates. While the jackknife and bootstrap share some common uses, they are based on very different principles of resampling.

6.1.2 Parameter estimation and Bias

A random sample from a population may not necessarily be representative of the whole population (in the sense of reflecting all of the properties of the sampled population). For example, a sample from a normal random deviate (N(0,1), i.e., mean = zero, variance = 1) could have values almost solely from one arm of the bell-shaped curve (Figure 6.1), and in this way the sample would not represent the full range of possible values.

Such a sample (Figure 6.1) could well be random (and indeed was) but might also come about because the sampling is not strictly random. Truly random samples can be difficult to arrange; for example, the sample in Figure 6.1 only arose after repeated trials. But without the samples being truly random, there is a possibility that the sample could produce biased estimates of such population parameters as the mean and variance or other parameters with less well-known behaviour. There can also be other sources of bias. With the example above, a visual inspection is enough to inform us that our sample is likely to give poor estimates of such statistics. But what can be done when we are unsure of the exact probability density function of the population from which we have a sample? One obvious answer is to take a larger sample, or replicate samples, but this is frequently not possible for reasons of time, funding, or circumstances.

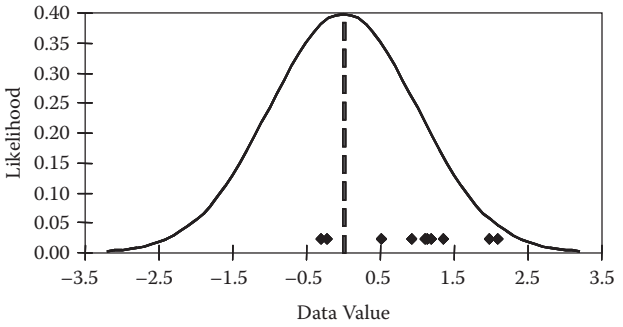

Fig ur e 6.1
The solid line shows the expected distribution from a standard random normal deviate of mean = 0 and variance = 1. The expected mean of samples from this distribution is shown as the dashed line. However, the diamonds are a random sample of ten having a sample mean of 0.976 and variance of 0.642.

An early solution to the investigation of bias, termed the half-sample approach (because it split the original data randomly into two equal-sized groups), gave unreliable estimates of the statistic of interest and its variance (Hinkley, 1983). Quenouille (1956) produced a relatively sophisticated extension of this half-sample idea in an effort to estimate bias. Tukey (1958) went a step further by generalizing Quenouille's particular approach, calling it the jackknife, and recommending it be used to produce estimates of parameters along with approximate confidence limits.

The jackknife method is relatively straightforward. Assuming one has a random sample (independent and identically distributed (iid)) of n values x_1, x_2, …, x_n, the sample mean is

$$\bar{x} = \frac{\sum x_i}{n} \tag{6.1}$$

This is also an unbiased estimate of the mean of the population from which the sample was taken. The jackknife methodology subsets the data sequentially by calculating the required statistic with one original value missing each time. In a data set of n values there will be n subsets of $n - 1$ data points (Table 6.1). Thus, for the sample mean minus the jth value,

$$\bar{x}_{-j} = \frac{\left[\left(\sum_{1}^{n} x_i\right) - x_j\right]}{(n-1)} \tag{6.2}$$

where x_j is the jth observation out of the n values. Equations 6.1 and 6.2 are related thus:

TABLe 6.1

An Illustration of the Generation and Use of Jackknife Samples and Replicate Observations

	Original	Data Series Systematically Minus One Observation				
	1		1	1	1	1
	2	2		2	2	2
	3	3	3		3	3
	4	4	4	4		4
	5	5	5	5	5	
Average	3	Mean-j	3.5	3.25	3	2.75	2.5
StDev	1.581	StDev-j	1.291	1.708	1.826	1.708	1.291
n	5		4	4	4	4	4
PV Mean	3	PV for Mean	1	2	3	4	5
PV Mean	1.647	PV for StDev	2.742	1.074	0.603	1.074	2.742

Note: The pseudovalues when estimating the mean are simply the original data values, while those when estimating the standard deviation are not directly related to the data values. The mean is an unbiased estimator of a sample mean, hence the mean of the pseudovalues for the mean equals the original sample mean. The sample is not taken from a normal distribution and contains some consequent bias when estimating the standard deviation. This is reflected in the difference between the sample StDev (1.581) and the Mean of the StDev pseudovalues (1.647) (Example Box 6.1).

$$(n-1)\bar{x}_{-j} = \left(\sum_{1}^{n} x_i\right) - x_j = n\frac{\left(\sum_{1}^{n} x_i\right)}{n} - x_j \tag{6.3}$$

which, when we convert the sum of x over n to the average of x, gives us

$$(n-1)\bar{x}_{-j} = n\bar{x} - x_j \tag{6.4}$$

and then, rearranging:

$$x_j = n\bar{x} - (n-1)\bar{x}_{-j} \tag{6.5}$$

which is simply a way of showing how the x_j values relate to the sample mean and the mean with the x_j values removed. Equation 6.5 is important because it is the basis for generating the pseudovalues upon which the jackknife calculations are based; we thus go one step further than a simple resample.

The jackknife estimate of the mean is simply the mean of these x_j values, known as pseudovalues, which is clearly the same as the original unbiased mean estimate:

EXAMPLE BOX 6.1

A simple jackknife example. The five jackknife samples are in columns C to G. The average and standard deviation of all columns is estimated in rows 8 and 9 with the count in row 10 (copy C8:C12 across to column G and into column B). These are used with Equation 6.5 to create the required pseudovalues for the mean and standard deviation in rows 11 and 12. By putting =average(C11:G11) and =average(C12:G12) into D14 and D15 we generate the jackknife estimates of the two parameters. Try altering the data in column B to see the impact on the estimates.

	A	B	C	D	E	F	G
1		Original	JK1	JK2	JK3	JK4	JK5
2		1		1	1	1	1
3		2	2		2	2	2
4		3	3	3		3	3
5		4	4	4	4		4
6		5	5	5	5	5	
7							
8	**Mean**	3	=average(C2:C6)	3.25	3	2.75	2.5
9	**StDev**	1.5811	=stdev(C2:C6)	1.708	1.826	1.708	1.291
10	**Count**	5	=count(C2:C6)	4	4	4	4
11	**PV Mean**		=($B$10*$B8)–(C$10*C8)	2	3	4	5
12	**PV StDev**		=($B$10*$B9)–(C$10*C9)	1.074	0.603	1.074	2.742
13							
14	Mean of pseudovalue Mean Values			3			
15	Mean of pseudovalue StDevs			1.647			

$$\tilde{x} = \frac{\sum x_j}{n} \tag{6.6}$$

As Manly (1991, p. 25) said, "Obviously, this is not a useful result if the sample values are known in the first place. However, it is potentially useful in situations where the population parameter being estimated is something other than a sample mean." These other parameters might be a standard error or measure of kurtosis or skewness, or some other statistic or model parameter.

Analogous to the jackknife estimate of the mean, the jackknife estimate of standard error of the parameter using the pseudovalues would be either option in Equation 6.7 (i.e., the standard deviation of the pseudovalues divided by root *n*). Thus,

$$\tilde{s}_{jack} = \sqrt{\frac{\sum\left(x_j - \tilde{x}\right)^2}{(n-1)}} \Big/ \sqrt{n} \quad \text{or} \quad \sqrt{\frac{\sum\left(x_j - \tilde{x}\right)^2}{(n-1)n}} \tag{6.7}$$

If other population parameters, such as the variance, were estimated, then the equivalents to the x_j pseudovalues would not necessarily be known. It is, however, possible to generalize the jackknife procedure to such population parameters. Given a population parameter, θ, which is estimated as a function of a sample of n values of x_i,

$$\hat{\theta} = f(x_1, x_2, ..., x_n) \tag{6.8}$$

Just as before, there are n estimates of this θ with the jth observation removed, and just like Equation 6.4, we can define the set of pseudovalues θ_j (Efron and Tibshirani, 1993) to be

$$\theta_j = n\hat{\theta} - (n-1)\hat{\theta}_{-j} \tag{6.9}$$

These θ_j values act in the same manner as the x_j values when estimating the mean. To produce the jackknife estimate of the parameter θ,

$$\tilde{\theta} = \frac{\sum \theta_j}{n} \tag{6.10}$$

If θ is the sample mean, then the pseudovalues are exactly the same as the x_j values, but with other parameters they become rather different from the original data (Table 6.1). If we extend this special case, then we can calculate the jackknife estimate of the standard error of the jackknife replicates as

$$\tilde{s}_{jack} = \sqrt{\frac{\sum\left(\theta_j - \tilde{\theta}\right)^2}{(n-1)n}} \tag{6.11}$$

This is treating the n pseudovalues as independent data values. Efron and Tibshirani (1993, p. 145) state: "Although pseudo-values are intriguing, it is not clear whether they are a useful way of thinking about the jackknife." Partly, I suggest they are saying this because they are advocates of using bootstrapping for producing better estimates of such parameters.

Some people suggest using such jackknife parameter estimates along with their standard errors to produce jackknife confidence intervals:

$$\tilde{\theta} \pm t_{n-1}\tilde{s}_{jack} \tag{6.12}$$

where t_{n-1} is the percentile value of the t distribution (e.g., 95% value) with $n-1$ degrees of freedom. Efron and Tibshirani (1993) suggest that confidence intervals produced in this way are not significantly better than cruder intervals based on asymptotic standard errors. It is thought that uncertainty over exactly how many degrees of freedom are involved in jackknife standard errors is part of the problem with this approach.

6.1.3 Jackknife Bias estimation

The jackknife procedure was originally introduced in an effort to estimate and remove bias from parameter estimates. The sample parameter value estimate, $\hat{\theta}$ is compared with the mean of the jackknife pseudovalues of the θ statistic; thus, where the jackknife replicates are defined as a function of the jackknife sample data (note the missing x_j value),

$$\tilde{\theta}_{-j} = f\left(x_1, x_2, .., x_{j-1}, x_{j+1}, .., x_n\right) \tag{6.13}$$

In addition, each jackknife replicate has its complementary pseudovalue as in Equation 6.9, so that the jackknife estimate of bias is defined as

$$\tilde{b}_{jack} = (n-1)\left(\tilde{\theta}_m - \hat{\theta}\right) \tag{6.14}$$

where the mean of the jackknife pseudovalues is

$$\tilde{\theta}_m = \frac{\sum \tilde{\theta}_j}{n} \tag{6.15}$$

Of course, with these formulations one will see that when $\hat{\theta} = \bar{x}$ the estimator is unbiased (assuming a truly random sample).

Exactly what one does once one has an estimate of bias is not clear. The usual reason for its calculation is to correct $\hat{\theta}$ so that it becomes less biased:

$$\bar{\theta} = \hat{\theta} - \tilde{b}_{jack} \tag{6.16}$$

where $\bar{\theta}$ is the bias-corrected estimate of $\hat{\theta}$. However, Efron and Tibshirani (1993, p. 138) finish a discussion of bias estimation thus:

> To summarize, bias estimation is usually interesting and worthwhile, but the exact use of the bias estimate is often problematic…. The straightforward bias correction [Equation 6.16] can be dangerous to use in practice, due to high variability in [the bias estimate]. Correcting the bias may cause a larger increase in the standard error…. If the bias is small compared to the estimated standard error [Equation 6.11], then it is safer to use $\hat{\theta}$ than $\bar{\theta}$. If bias is large compared to standard error, then it may be an indication that the statistic $\hat{\theta}$ is not an appropriate estimate of the parameter θ.

This appears to suggest one should never apply Equation 6.16 in practice! It is useful to see if there is bias, but if detected, it appears it is best not to try to correct it (presumably one should try again to obtain an unbiased sample if the bias is too large).

6.2 The Bootstrap

6.2.1 The Value of Bootstrapping

There are a number of analytical strategies for producing estimates of parameters with confidence intervals from samples that have an unknown probability distribution. Efron and LePage (1992) state the general problem thus:

> We have a set of real-valued observations $x_1, \ldots, x_n$ independently sampled from an unknown probability distribution F. We are interested in estimating some parameter θ by using the information in the sample data with an estimator $\hat{\theta} = t(x)$. Some measure of the estimate's accuracy is as important as the estimate itself; we want a standard error of $\hat{\theta}$ and, even better, a confidence interval on the true value θ.

Since Efron (1979) first discussed the idea of bootstrapping it has become relatively popular, even trendy, at least among statisticians; three reasons have been suggested (Kent et al., 1988):

1. Elegance: The principle behind the bootstrap, that of resampling from the empirical distribution function (as represented by a sample), instead of the actual probability density function, is simple and elegant, and yet very powerful.
2. Packaging: The catchy name *bootstrap* makes it easy for people to recognize the product, though the potential for confusion exists now that parametric resampling, as in classical Monte Carlo simulations, is sometimes included in the term *bootstrapping*.
3. Ease of use (Kent et al., 1988, p. 355): "for the practitioner there is the hope of a fairly automatic and clear methodology that can be used without the need for any thought."

The last reason seems frightening and I would suggest that this latter idea be discouraged vigorously in any analyses; one should always think deeply about one's analyses. I imagine Kent was only implying that such analyses could become routine. Bootstrap resampling is a general form of resampling in that it is resampling with replacement to produce bootstrap samples

Original Sample		Bootstrap Samples					
3		8	5	3	8	3	
5		5	8	11	9	3	
7	→	11	5	5	3	11	...etc.
8		7	8	7	7	9	
9		8	7	7	3	8	
11		8	11	11	9	11	
Mean 7.17		7.83	7.33	7.33	6.5	7.5	

Fig ur e 6.2
An original sample of six numbers with their average. From this are drawn five bootstrap samples, each with a separate average. It is clear that with a sample of size n there are n^n possible bootstrap combinations. Replacement implies it is easily possible for a single observation to appear more than once in the bootstrap sample; it is also possible that some original observations will not occur in the bootstrap samples. The average of the five bootstrap replicates is 7.292 (Example Box 6.2).

of size n. When one resamples with replacement there are many possible arrangements (n^n in fact) of the available data (Figure 6.2, Example Box 6.2).

If one had a sample from a known normal distribution, then there would be no advantage to using a bootstrap method for estimating the standard error of a parameter; "normal" theory would be best. However, in situations where the sampled population cannot be adequately represented by a normal distribution, and especially where the underlying population distribution is unknown, bootstrapping becomes most useful.

6.2.2 empirical versus Theoretical Probability Distributions

In fact, given a sample from a population, the nonparametric, maximum likelihood estimate of the population distribution is the sample itself. Expressed precisely, if the sample consists of n observations (x_1, x_2, x_3, …, x_n), the maximum likelihood, nonparametric estimator of the population distribution is the probability function that places probability mass $1/n$ on each of the n observations x_i. Take note that this is not saying that all values have equal likelihood; instead, it implies that each observation has equal likelihood. One expects, some of the time, to obtain the same or similar values in different observations if the population distribution being sampled has a mode.

The implication, first suggested by Efron (1979), is that when a sample contains or is all the available information about a population, why not proceed *as if* the sample really *is* the population for purposes of estimating the sampling distribution of the test statistic? That is, apply Monte Carlo procedures, sampling with replacement but from the original sample itself, as if it were a theoretical distribution. Sampling with replacement is consistent

EXAMPLE BOX 6.2

A simple bootstrap example (see Figure 6.2). One can generate a bootstrap sample in Excel using the vlookup function (or the offset function). This entails generating random integers that match the numbers in an index list as in column A. The vlookup function uses these index numbers to return the data value next to each respective index. There is nothing to stop the same integer from arising, which will lead to sampling with replacement. In C7 put the function =vlookup(trunc (rand()*6)+1,$A$7:$B$12,2,false). Use the Excel help to understand this function. The trunc(rand()*6)+1 term will generate random integers between 1 and, in this case, 6. Try typing just this term elsewhere in the sheet and examining its performance each time you press F9 (i.e., recalculate the sheet). Copy C7 down to C12. Then copy C7:C12 across to IV7:IV12. Copy B1 across to IV1. The four averages in C2:C5 are bootstraps with different numbers of replicates (20 to 254). Keep pressing F9 to generate new random numbers and hence new bootstrap samples, and examine the differences between the observed mean and the bootstrap means. Which sample size provides the better estimates? The option of duplicating the parameter calculation either up or down the sheet (perhaps use hlookup instead) can be fast but will generate large worksheets. Plot B2:B5 against A2:A5.

	A	B	C	D	E	F	G
1	**n**	=average(B7:B12)	5.833	8.833	7.667	7.500	7.833
2	20	=average(C1:V1)	=B1–B2				
3	50	=average(C1:AZ1)	=B1–B3				
4	100	=average(C1:CX1)	=B1–B4				
5	254	=average(C1:IV1)	=B1–B5				
6	**Index**	**Data**	**BS1**	**BS2**	**BS3**	**BS4**	**BS5**
7	1	3	8	11	9	5	3
8	2	5	3	7	9	8	9
9	3	7	5	8	5	7	8
10	4	8	7	7	7	9	11
11	5	9	3	9	11	5	8
12	6	11	9	11	5	11	8

with a population that is essentially infinite. Therefore, we are treating the sample as representing the total population.

In summary, bootstrap methods are used to estimate a parameter of an unknown population by summarizing many parameter estimates from replicate samples derived from replacing the true population by one estimated from the population (the original sample from the population).

6.3 Bootstrap Statistics

Standard error is a general term for the standard deviation of a summary statistic. So one may have the standard deviation of a sample and a standard error of a sample mean, but one could not have a standard error of a sample. With a sample from a normally distributed variable, one can estimate the standard error of, for example, the mean of a sample of n observations, analytically by using:

$$se_{\bar{x}} = \frac{StDev}{\sqrt{n}} = \sqrt{\frac{\sum \left(x_i - \bar{x}\right)^2}{(n-1)n}} \tag{6.17}$$

Alternatively, one could take a large number of independent samples from the same population and find the standard deviation of the means of these samples. This latter process is exactly what one does for a bootstrap estimate of the standard error for any parameter. A general approach for producing the bootstrap estimation of the standard error of any parameter θ is as follows:

1. Generate b independent bootstrap samples x_1, x_2, x_3, …, x_b, each consisting of n data values drawn randomly with replacement from the n values in the original sample (the empirical distribution). Efron and Tibshirani (1993) recommend b to be at least in the range of 25 to 200 for estimating a standard error.
2. Calculate the bootstrap replicate of the parameter or statistic $\hat{\theta}_b$ for each of the b bootstrap samples, x_b. The statistic must be a continuous function of the data (Equation 6.18).

$$\hat{\theta}_b = f\left(x_b\right) \tag{6.18}$$

3. Estimate the standard error se_θ of the parameter θ by calculating the standard deviation of the b bootstrap replicates (note we are using a standard deviation of multiple samples to estimate a standard error; don't confuse the concepts).

$$se_\theta = \sqrt{\frac{\sum \left(\hat{\theta}_b - \bar{\theta}_b\right)^2}{b-1}} \tag{6.19}$$

where $\bar{\theta}_b$ is the mean of the bootstrap replicates of θ, which is the bootstrap estimate of the statistic θ:

$$\bar{\theta}_b = \frac{\sum \hat{\theta}_b}{b} \tag{6.20}$$

Such estimates would be called *nonparametric bootstrap* estimates because they are based upon an empirical distribution instead of a theoretical probability distribution. It would be unusual today to use the bootstrap to estimate a standard error. More commonly, the bootstrap is used to generate percentile confidence intervals around parameter estimates for which other methods of obtaining confidence intervals would present difficulties.

6.3.1 Bootstrap Standard errors

Efron and Tibshirani (1993) use a bootstrap to estimate the standard error of a correlation coefficient (a notoriously difficult thing to do). With their example Efron and Tibshirani (1993) suggest that a sample size of between 25 and 200 should be sufficient to calculate standard errors with adequate precision. We have repeated this exercise to determine the repeatability of their results. The example used here is the correlation between catches of tiger prawns and king prawns in the Australian Northern Prawn Fishery across the years 1976 to 1987 (Figure 6.3). The tiger prawns constitute the target in this fishery and the king prawns are a bycatch species.

To determine if the bootstrap standard error estimate improved with increasing numbers of bootstrap replicates, more and more bootstrap replicate samples can be made for each of which the correlation coefficient is calculated (Figure 6.4, Example Box 6.3). The standard error estimates differ between series and with *b*, the number of bootstrap replicates, but all estimates where *b* is 200 or greater achieve an acceptable precision. A sample size of twenty-five would appear to be rather low, but two hundred and

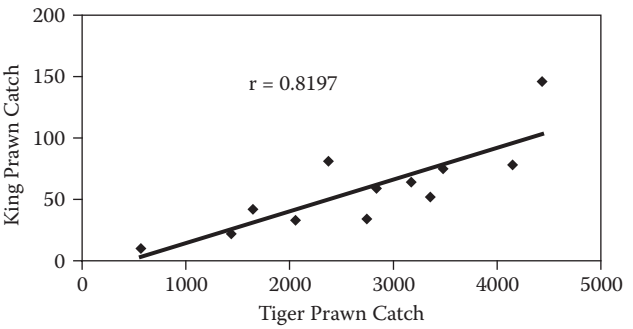

Fig ur e 6.3
Scatterplot relating Australian Northern Prawn Fishery tiger prawn catches, as tonnes, to king prawn catches, between the years 1976 and 1987. The linear regression is king = 0.02585 and tiger = 11.4826.

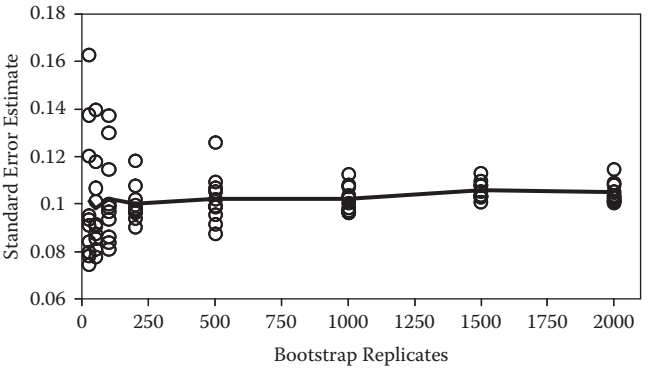

Fig ur e 6.4
Ten estimates of the standard error around the correlation coefficient at each of eight different numbers of bootstrap replicates (25, 50, 100, 200, 500, 1,000, 1,500, 2,000). The solid line is the mean of the estimates. Use Example Box 6.3 to generate an equivalent graph. The overall average standard error was approximately 0.102.

more is clearly enough. Hinkley (1988) suggests that the minimum number of bootstrap samples will depend on the parameter being estimated, but that it will often be one hundred or more. With the large increases in the relative power of computers in recent years, the number of replicate bootstraps to conduct should no longer be an issue for most problems.

6.3.2 Bootstrap r eplicates

Beyond using the bootstrap to estimate standard errors, it can also provide an estimate of an empirical frequency distribution (Figure 6.5) of possible values for the statistic or parameter in question. The distribution of correlation coefficients derived from the bootstrap replicates is clearly nonnormal with a significant skew.

So far, we have considered exactly what a bootstrap sample is and how it is produced. We have also seen how to determine a bootstrap estimate of a parameter (the average of the bootstrap replicates) and its standard error (standard deviation of bootstrap replicates). However, one of the major areas of research on bootstrapping has been the consideration of different approaches to calculating statistical confidence intervals for estimated parameters (DiCiccio and Romano, 1988; Romano, 1988). With nonnormal populations, these can be difficult to fit in a valid or unbiased way. Using standard errors to generate confidence intervals always leads to symmetrical intervals. If the distribution of the parameter of interest is not symmetric, such confidence intervals would be invalid. The bootstrap provides us with a direct approach that can give rise to excellent approximate confidence intervals. All these direct methods rely on manipulating in some way the empirical frequency distribution of the bootstrap replicates (Figure 6.5).

EXAMPLE BOX 6.3

Bootstrap standard errors around a correlation coefficient between tiger and king prawn catches (cf. Figure 6.3). We will use vlookup again, but this time, because we need to return data pairs, we will have a separate column of random index values (column D) and use them in both vlookup functions in columns E and F, e.g., in E5, =vlookup(D5,$A$5:$C$16,2,false), and in F5 put =vlookup(D5,$A$5:$C$16,3,false). Note the only change is the column from which to return data. Copy these down to row 16. Each press of F9 will renew the random numbers in column D, which will lead to new data pairs in columns E and F. In C2 put =correl(B5:B16,C5:C16), which is the original correlation, and in F2 put =correl(E5:E16,F5:F16), which is the bootstrap sample correlation. Record a macro through Tools/Macro/Record New Macro, and call it Do_Boot. While in absolute references copy F2, change to relative references, and paste values into G5. Stop recording. Place a button on the worksheet from the Forms toolbar and assign Do_Boot to it. Edit the macro (via Alt-F11) and make the changes listed in italics to make the macro functional. Change the number of bootstrap replicates by altering C1. Note the answers and determine a reasonable number of replicates to estimate the standard error of approximately 0.102 in G3 (using Equation 6.19). You will need to run each set a number of times (Yr is the year of catch, Tig is tiger prawn catches in tonnes, and King is king prawn catches).

	A	B	C	D	E	F	G
1	TRIALS		500				
2	Original r		0.8197	Do_Boot		0.9735	=average(g5:g2004)
3	Original Data pairs						=stdev(G5:G2004)
4	**Yr**	**Tig**	**King**	**Bootstrap Index**	**Tig**	**King**	**Bootstraps**
5	76	566	10	=trunc(rand()*12)+76	558	2.81	0.930064
6	77	1437	22	=trunc(rand()*12)+76	578	3.03	0.833836
7	78	1646	42	8	661	3.43	0.90485
8	79	2056	33	5	666	3.44	0.679865
9	80	3171	64	2	635	3.3	0.708266
10	81	2743	34	8	661	3.43	0.857945
11	82	2838	59	12	575	2.74	0.341376
12	83	4434	146	8	661	3.43	0.726493
13	84	4149	78	12	575	2.74	0.876033
14	85	3480	75	9	651	3.36	0.75519
15	86	2375	81	8	661	3.43	0.948527
16	87	3355	52	8	661	3.43	0.771965

continued

EXAMPLE BOX 6.3 (continued)

Some of the changes needed could be recorded as a separate macro and then copied into this one (e.g., the cell clearing lines just below the Dim statements). The MsgBox will soon lose its novelty and can be turned off by converting it to a comment with a '. Try running this a few times for fifteen replicates without the screen updating turned off (comment out the necessary statements). Plot the bootstrap replicate samples of column E against F and observe how they change.

```
Sub Do_Boot()
'
' Do_Boot Macro
'
Dim i As Integer, b As Integer
Dim start As Double, endtime As Double
'
Range("G5").Select
Range(Selection, Selection.End(xlDown)).Select
Selection.ClearContents
Range("G6").Select
Application.ScreenUpdating = False
  start = Timer
  b = Range("C1").Value     ' get the number of replicates
  For i = 1 To b
    ActiveSheet.Calculate
    Range("F2").Select
    Selection.Copy
    ActiveCell.Offset(2 + i, 1).Range("A1").Select
    Selection.PasteSpecial Paste:=xlPasteValues
  Next i
  Range("A1").Select
  endtime = Timer
  Application.ScreenUpdating = True
  MsgBox Format(endtime - start, "##0.000"), vbOKOnly,
    "Bootstrap Complete"
End Sub
```

6.3.3 Parametric Confidence intervals

Where the parameter being estimated is expected to exhibit a normal distribution of expected values (here the central limit theorem may play a part), confidence intervals around the parameter may be obtained from the usual:

$$\mathrm{CI} = \theta \pm t_{n-1,\alpha/2} se_{\theta} \tag{6.21}$$

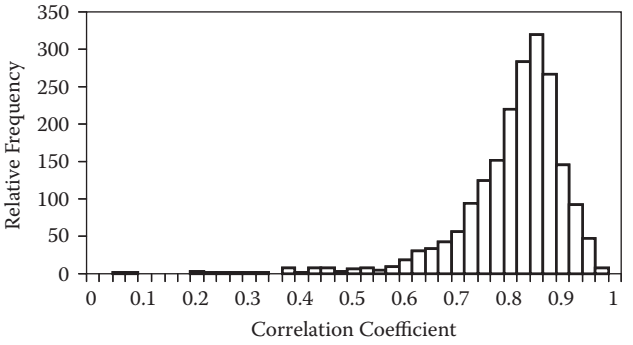

Fig ur e 6.5
Frequency distribution of bootstrap replicate values of the correlation coefficient. The observed value was 0.8197. The median value of the illustrated frequency distribution was 0.8423, and the bootstrap estimate of the correlation coefficient was 0.8191, a difference of 0.0006 from the sample estimate (this provides an estimate of any bias).

where θ is the sample parameter estimate, $t_{n-1,\alpha/2}$ is the student's t distribution value for $n-1$ degrees of freedom (where n is b, the number of bootstrap replicates), and $\alpha/2$ is the percentage confidence limits desired; with $\alpha = 0.05$, $100(1-\alpha)\%$ provides the 95% intervals, where $\alpha/2 = 2.5$ and 97.5. Because the number of replicates in bootstrapping is likely to be high, instead of using the t distribution many statisticians simple replace it with the z value (e.g., 1.96 for the 95% confidence intervals). The bootstrap estimates of standard error can be used in this way to generate confidence intervals. In the example we have been considering the observed correlation coefficient was 0.8197 and the standard error was approximately 0.102 (Figure 6.4). If we used a normal approximation, we would expect the confidence intervals to be 0.8197 ± 1.96 x 0.102 = 0.6198 and 1.0196. Clearly, with a correlation coefficient a value greater than 1 is nonsense and illustrates the problem of using normal theory to estimate confidence intervals for parameters with expectations from nonsymmetrical distributions.

6.3.4 Bootstrap estimate of Bias

If the sample estimate is biased, it is possible to remove this bias before adding and subtracting the requisite z value to find the confidence intervals:

$$\theta - \left(\bar{\theta}_b - \theta\right) \pm z_{\alpha/2} se_\theta \tag{6.22}$$

or, equivalently,

$$\left(2\theta - \bar{\theta}_b\right) \pm z_{\alpha/2} se_\theta \tag{6.23}$$

which is the sample estimate minus the bootstrap bias plus or minus the standard normal distribution for $\alpha/2$ times the bootstrap standard error estimate (standard deviation of the bootstrap replicates).

Of course, if the statistic being considered were far from normal, then Equation 6.23 would produce erroneous results. The example of the correlation coefficients (Figure 6.5) is one where a statistic has an expected distribution that is far from normal. If the estimates are corrected for bias prior to estimating the confidence intervals, the distortion on the confidence intervals becomes slightly greater. Using Equation 6.23 we have $(2 \times 0.8197 - 0.8191) \pm 1.96 \times 0.102 = 0.6204$ and 1.0202. Thus, the upper limit becomes slightly further removed from 1 (the maximum possible value for a correlation coefficient).

6.4 Bootstrap Confidence Intervals

6.4.1 Percentile Confidence intervals

Given b bootstrap replicate samples we can generate b bootstrap estimates of the parameter of interest θ_b. An estimate of the $100(1 - \alpha)\%$ confidence limits around the sample estimate of θ is obtained from the two bootstrap estimates that contain the central $100(1 - \alpha)\%$ of all b bootstrap estimates (Efron, 1979). Thus, the 97.5 and 2.5 percentiles of the two thousand bootstrap estimates of the correlation coefficient, summarized in Figure 6.5, estimate the confidence intervals around the correlation coefficient (Figure 6.6).

With two thousand bootstrap replicates the 95% confidence intervals values would be found at the 50th and at the 1950th position in the sorted bootstrap estimates. In this case, the intervals are between 0.5446 and 0.9567. The bootstrap estimate of the correlation coefficient was very slightly to the left of the sample estimate (0.8191 vs. 0.8197), which shows that the bias in the estimate was small. Note, in this case, the sample estimate is also close to the median value of the bootstrap estimates (Figure 6.6).

In general terms, one would not wish to attempt to fit percentile bootstrap confidence intervals with less than one thousand bootstrap replicates. Even with two thousand replicates the histogram is not as smooth as one could desire for it to be used as a probability density function. However, there is a limit to the precision with which it is useful to estimate these simple bootstrap percentile confidence intervals because they are only ever approximate.

6.4.2 Bias-Corrected Percentile Confidence intervals

Often the bootstrap estimate of the parameter of interest differs somewhat from the observed parameter estimate, indicating that there may be evidence of bias. The validity of any confidence intervals should be improved if they

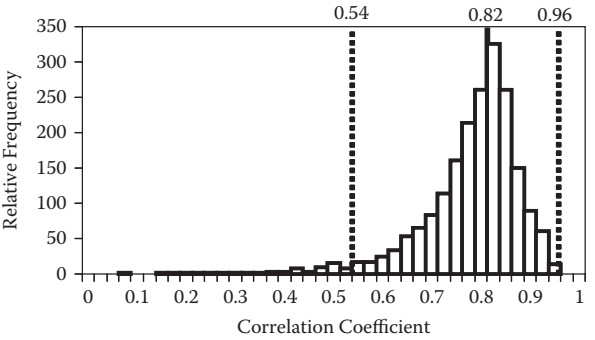

Fig ur e 6.6
Two thousand separate bootstrap estimates of a correlation coefficient grouped into classes of 0.025 with the observed correlation indicated as a solid line and the bootstrap percentile confidence intervals indicated as dotted lines. The 95% bootstrap confidence intervals were at the 51st (2,000 × 0.025) and the 1950th (2,000 × 0.975) sorted bootstrap replicate. The bootstrap estimate of r was 0.8191, while the median was 0.8423. Unlike normal confidence intervals, these confidence intervals are clearly asymmetric and do not suffer from suggesting an upper limit greater than 1. In Excel these can be found using =percentile(g5:g2004,0.025) and =percentile(g5:g2004,0.975).

take into account any bias present. Fitting bias-corrected percentile confidence intervals is slightly different than fitting usual bootstrap percentile intervals around a parameter (Fletcher and Webster, 1996). The difference lies in having first to determine exactly which percentiles should be used after removing any bias that arises because the observed parameter value is not the median of the distribution of bootstrap estimates. Thus, with a slightly biased sample, one might use percentiles such as 0.0618 and 0.992, or perhaps 0.0150 and 0.960 instead of 0.025 and 0.975.

The determination of bias-corrected percentiles is relatively simple (Efron, 1987; Efron and Tibshirani, 1993, provide a justification). After generating the bootstrap sample estimates, one determines F, the fraction of bootstrap replicates that are smaller than the original parameter estimate. Thus, if 56% of the bootstrap replicates were smaller than the original parameter estimate, F would be equal to 0.56. From this fraction, a constant z_0 is calculated to be the probit transform of F:

$$z_0 = \Phi^{-1}(F) \tag{6.24}$$

where Φ^{-1} is the inverse, standard cumulative normal distribution (=norminv (0.56,0,1) in Excel). From Equation 6.24, the appropriate percentiles for the 95% confidence intervals are calculated by the following:

$$P_{lower} = \Phi(2z_0 - 1.96)$$
$$P_{upper} = \Phi(2z_0 + 1.96) \tag{6.25}$$

where Φ is the cumulative normal distribution function. The 1.96 is, of course, the critical value from the inverse normal curve for the 95% confidence intervals. This can be altered appropriately for alternative intervals (e.g., the value would be 1.6449 for 90% intervals). After using Equation 6.25, one would then determine the bias-corrected percentile confidence intervals by taking the values from the bootstrap distribution that align with the calculated upper and lower percentiles.

For an example where 39.8% of bootstrap replicates are smaller than the original parameter estimate, this gives $F = 0.398$ so that $z_0 = -0.2585$, leading to $P_{\text{lower}} = \Phi(-2.4771) = 0.662\%$, and $P_{\text{upper}} = \Phi(1.4429) = 92.54\%$ (using =normdist(1.4429,0,1,true) in Excel). Thus, from one thousand ordered values in a bootstrap distribution the 95% confidence intervals would be the 6th and 925th values instead of the 25th and 975th values (using =percentile(g5:g2004,0.925)). If there is no bias at all, then $F = 0.5$, which would imply a $z_0 = 0$, which would finally lead to the lower and upper bounds being the default of 2.5 and 97.5% (Example Box 6.4).

6.4.3 Other Bootstrap Confidence intervals

We have only considered the simple bootstrap percentile and first-order bias-corrected bootstrap percentile confidence intervals, but many other algorithms exist for generating confidence intervals from bootstrap samples. When using confidence intervals (for example, 90% intervals), one would want them to fail to cover the parameter θ exactly 5% of the time in each direction. There have been many comparative studies conducted on the variety of methods of constructing confidence intervals around a parameter (Efron, 1992; Manly, 1997). Because they use data sets whose properties are known, these studies are able to indicate the strengths and weaknesses of the various methods available. There are two methods that usually perform rather better (generate confidence intervals that work—see Example Box 7.1 in Chapter 7, for an example of how to test for success) than the simple and first-order bias-corrected algorithms considered here; these are named the bootstrap *t* method and the accelerated bias-corrected percentile methods (Efron, 1987). Both of these extensions are more computer-intensive than the two we have considered here in that they both require the use of a jackknife for best performance (Efron and Tibshirani, 1993; Manly, 1997).

Manly (1997) provided a listing of published biological work using the bootstrap. Generally, in fisheries, the simple percentile confidence interval is most commonly used (Kimura and Balsinger, 1985; Sigler and Fujioka, 1988), though as the first-order bias-corrected percentile intervals involve only a small amount of extra work and the confidence intervals are generally improved, we would recommend that these be used instead. Further examples of the use of the bootstrap will be given when we consider other, more complex models, including the surplus production models that are the simplest stock assessment models available and dynamic age-structured models.

EXAMPLE BOX 6.4

Simple percentile and first-order bias-corrected bootstrap percentile confidence intervals. This example extends the worksheet created in Example Box 6.3. Column G contains one thousand bootstrap replicates of a correlation coefficient with the mean bootstrap estimate in G2 and the standard error estimate in G3. The additions are in columns I, J, and K. The simple percentile intervals are the easiest to implement; in K10 and K11 put the functions =percentile(G5:G1004, 0.975) and =percentile(G5:G1004, 0.025). The values that arise will depend upon the particular bootstrap replicates that are on the sheet, but they will be similar to that shown. J3 contains the estimate of *F*, the fraction of bootstrap replicates less than the observed value. J4 contains Equation 6.24. J5 and J6 contain the $2z_0 - 1.96$ and $2z_0 + 1.96$, and J7 and J8 contain Equation 6.25, the percentile points where the bias-corrected confidence intervals will be found. In the example sheet, the values were 0.9277 and 0.4140, which have been shifted to the left of the simple percentile intervals in K10 and K11. Run the macro a few times with one thousand bootstrap replicates and notice the impact on the different percentile confidence intervals. Increase the number of replicates to two thousand and modify the equations in J3, J4, and J10:K11. Does the increased number of replicates stabilize the estimated intervals?

	G	H	I	J	K
1					
2	0.81905		**Original r**	0.819691	
3	0.10225		**Fraction F**	=countif(G5:G1004,"<"&J2)	
4	Bootstraps		**norminv(F)**	=norminv(j3/1000,0,1)	
5	0.90599		**P_Lower**	=2*J4–1.96	
6	0.84306		**P_Upper**	=2*J4+1.96	
7	0.88260		**Percentile L**	=normdist(J5,0,1,true)	
8	0.80945		**Percentile U**	=normdist(J6,0,1,true)	
9	0.77037			**Bias–Corrected Intervals**	**Simple**
10	0.84627		**Upper95**	=percentile(G5:G1004,J8)	0.9567
11	0.86610		**Lower95**	=percentile(G5:G1004,J7)	0.5446

6.4.4 Balanced Bootstraps

In the process of running a bootstrap procedure the resampling with replacement means it is possible that not all observations in the original sample will be resampled the same number of times across however many replicates are used. While this is not seen as a problem when there are many thousands of bootstrap replicates, if there are only, say, one thousand, then a better

algorithm might be to use a balanced bootstrap. In the balanced bootstrap one first copies the values to be resampled the same number of times as there will be bootstrap replicates; then one resamples at random, without replacement, from within the multiple copies. This process will automatically mean that all observations occur an equal number of times while bootstrapping.

The balanced bootstrap does not appear to have been used very often. It would certainly be easier to use with a single series of data than with more complex models or with structured data. If a model had more than one series of data to be bootstrapped, a balanced design would still be possible but would require some sort of cross-tabulation between data series. For example, two-dimensional balance could be achieved using a classic orthogonal Latin square design. By balancing the use of each observation, the amount of resampling needed to estimate the expectation and other moments of a parameter's distribution, with acceptable accuracy, is reduced by up to fivefold (Hinkley, 1988). However, Hinkley (1988) reports that balanced bootstraps are not so effective for estimating bootstrap percentiles, especially for $p < 0.05$ or $p > 0.95$. This is an area deserving of further research.

6.5 Concluding Remarks

Bootstrapping offers the ability to generate confidence intervals around parameters and model outputs in situations that were previously impossible to approach. The ability to test for bias around a parameter estimate is valuable, but of most value is the ability to estimate the uncertainty around the parameter estimates by estimating standard errors or confidence intervals and the underlying frequency distribution of the parameter of interest.

Where it is valid to use parametric statistics, the best strategy is to use them. However, if one is uncertain what the underlying distribution of a parameter is, then bootstrapping is to be recommended. Fisheries data are generally so variable and uncertain that the debate over the optimum algorithm, while interesting, neglects the fact that the real limitation is most often in the quality of the available data.