An Introduction to the Bootstrap

KEYWORDS:
Teaching;
Standard error;
Confidence interval;
Minitab;
Bias;
Mean square error.

Roger W. Johnson
South Dakota School of Mines & Technology, USA.
e-mail: rwjohnso@taz.sdsmt.edu

Summary
This article presents bootstrap methods for estimation, using simple arguments. Minitab macros for implementing these methods are given.

◆ INTRODUCTION ◆

In the eighteenth century stories of *The Adventures of Baron Munchausen* by Rudolph Erich Raspe (Raspe 1785), the Baron apparently falls to the bottom of a deep lake. Just when it looks like all is lost, he saves himself by picking himself up by his own bootstraps. Likewise bootstrap methods in statistics seem to accomplish the impossible. These computationally intensive methods, brought to prominence through the pioneering work of Bradley Efron, are commonly used by statistics professionals and are beginning to work their way into elementary, even algebra-based statistics texts (e.g. Stout *et al.* 1999). In this article I present bootstrap methods for estimating standard errors and producing confidence intervals. Bootstrap methods are more flexible than classical methods which may be analytically intractable or unusable because of a lack of the appropriate assumptions being satisfied. When classical methods may reasonably be used, however, we will typically see that bootstrap methods give quite similar results. The presentation that follows is based on details that appear in Efron and Tibshirani (1986, 1993) and Rice (1995), and includes short Minitab macros to come up with the desired estimates. Related articles that have appeared in this journal are those of Ricketts and Berry (1994), Reeves (1995) and Taffe and Garnham (1996).

◆ BOOTSTRAP ESTIMATES OF STANDARD ERROR ◆

For concreteness, let us consider an example to illustrate bootstrap estimates of standard error. Figure 1 displays a histogram of the 40 interarrival times between 41 consecutive vehicles passing by a fixed point near junction 13 of the M1 motorway in Bedfordshire, England. These cars were travelling northwards on the M1 in the late evening on Saturday 23 March 1985 (see Hand *et al.* 1994, p. 3).

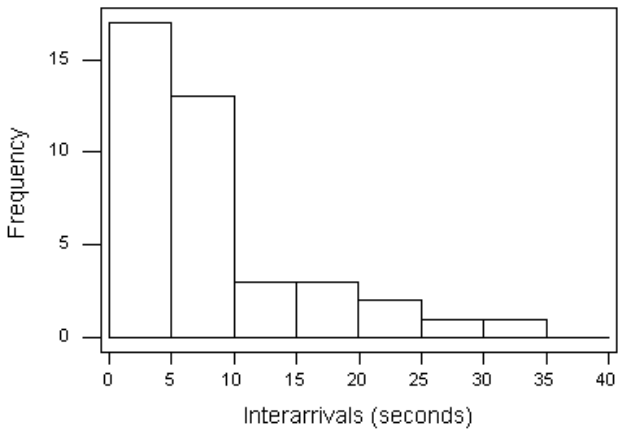

Fig 1. M1 motorway interarrival times

For this data set of $n = 40$ values we find a sample mean of $\bar{x} = 7.80$ seconds and sample standard deviation of $s = 7.87$ seconds. For future reference note that this data set is fitted well by an exponential density, $f(x) = \lambda e^{-\lambda x}$, with $\lambda = 1/7.80$ (this is both the method of moments and the maximum likelihood estimate of λ).

The standard error of $\overline{X}$, $\mathrm{SE}(\overline{X})$, is given by

$$\mathrm{SE}(\overline{X}) \equiv \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation. Estimating σ by s, we have

$$\mathrm{SE}(\overline{X}) \approx \frac{s}{\sqrt{n}} = \frac{7.87}{\sqrt{40}} = 1.24 \text{ seconds}$$

so that 1.24 seconds is an estimate of the standard error of $\overline{X}$. In layperson's terms we believe the population mean, μ, to be about $\bar{x} = 7.80$ seconds, give or take around 1.24 seconds. The population of interarrivals, not yet carefully defined, may be

thought of as the collection of all interarrivals during the same general time during the week in 1985 under similar (e.g. weather) conditions.

Although we do not ordinarily proceed in this way, here is another way of estimating the standard error of $\overline{X}$ which relies only on $\sigma_{\overline{X}} \approx s_{\overline{X}}$:

(i) Sample $n = 40$ interarrivals from our population and compute $\bar{x}$.
(ii) Repeat (i) a moderate to large number, B, of times to come up with estimates $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_B$.
(iii) Use the standard deviation of the B estimates in (ii) to estimate the standard error.

It is important in this procedure to produce estimates $\bar{x}_i$ with a sample size identical to our original sample size of n. If, in step (ii), we used a sample size less than n, then the procedure would tend to overestimate the error in our original estimate of 7.80 seconds; likewise, if in step (ii) we used a value more than n, then the procedure would tend to underestimate the error in our original estimate of 7.80 seconds.

A bootstrap method of estimating the standard error of $\overline{X}$ now involves a modification of the above procedure. In particular, *use the sample as an approximation of our population. Specifically, take samples with replacement of size n from the* ***data*** *to approximate samples of size n from the population.* If you think that this is akin to 'lifting yourself by your bootstraps' you are not alone! Here, then, is a bootstrap method for estimating the standard error of $\overline{X}$:

(a) Sample $n = 40$ interarrivals *with replacement from the original data* and compute $\bar{x}$.
(b) Repeat step (a) a moderate to large number of times, B, to come up with 'bootstrap' estimates $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_B$.
(c) Use the standard deviation of the B estimates in step (b) to estimate the standard error.

(To find the standard error of a statistic other than the sample average, follow this same recipe but compute that statistic rather than the average.) Carrying out this bootstrap method using the Minitab macros in figure 2 with $B = 200$, we get an estimated standard error of 1.23 – a value nearly identical to the value of 1.24 obtained above. The value of $B = 200$ is usually sufficiently large for bootstrap standard error estimates. Efron and Tibshirani, in fact, indicate that 'very seldom are more than $B = 200$ replications needed for estimating a standard error' (1993, p. 52) and that the variance of the bootstrap standard error estimate is roughly $c_1/n^2 + c_2/(nB)$, where the constants c_1 and c_2 depend on the underlying population but not on n or B (p. 272).

This bootstrap method may be used with even smaller sized data sets than that given above. Loosely speaking, however, the bootstrap idea of approximating the population by the sample becomes more questionable as the sample size, n, decreases. As with other statistical procedures, our trust in the bootstrap will grow with increased sample size.

In the previous example there was, of course, no

File: sedriver.txt

```
noecho
erase c10
let k1 = n(c1)                  # the data must have previously been put in column c1
let k2 = 200                    # number of bootstrap samples, B
execute 'bootstrp.txt' k2
echo
let k3 = stdev(c10)             # se of mean
print k3
end
```

File: bootstrp.txt

```
sample k1 c1 c11;
replace.
let k20 = mean(c11)
stack c10 k20 c10
```

Use ***execute sedriver.txt*** at the Minitab prompt to run this bootstrap procedure

Fig 2. Bootstrap standard error code

need to estimate the standard error of $\overline{X}$ using the bootstrap method as we know $\mathrm{SE}(\overline{X}) = \sigma/\sqrt{n} \approx s/\sqrt{n}$. The bootstrap method of estimating the standard error of a statistic becomes valuable in those cases where we do not have a theoretical formula for the standard error of that statistic. There is, for example, no closed-form formula for the standard error of the sample median. Continuing the example involving interarrivals along the M1 motorway, the sample median of 5.0 is an estimate of the population median. To estimate the standard error of the sample median, use the bootstrap procedure given above, computing medians rather than averages. Doing so (replace the line ***let k20 = mean(c11)*** with ***let k20 = median(c11)*** in the Minitab code in figure 2) gives 0.68 as an estimate of the standard error of the sample median.

As a final example to illustrate the above bootstrap method of estimating standard errors, consider male mortality rate averaged over the years 1958–1964 for towns in England and Wales versus calcium (from Hand *et al.* 1994, pp. 5–6), shown as a scatter diagram in figure 3. The calcium concentration may be thought of as a measure of water hardness; the higher the calcium concentration, the harder the water. The correlation coefficient between male mortality and calcium concentration for the 61 data points shown is −0.655.

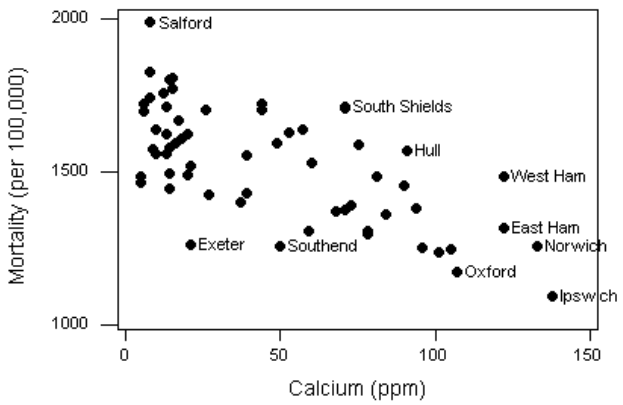

Fig 3. Mortality vs calcium

We can again use the bootstrap to estimate the standard error of the correlation coefficient. Obtain a bootstrap sample of 61 ordered pairs of mortality and calcium concentration by sampling with replacement from the 61 cities 61 times. Compute the correlation coefficient of this bootstrap sample. Repeat the entire process B times (200 times, say). Then estimate the standard error of the correlation coefficient as the standard deviation of the B bootstrapped correlation coefficient values. One such execution of this bootstrap method yielded the estimated standard error of 0.075 (Minitab macros to implement this are available from the author). If we were willing to assume that mortality and calcium have a bivariate Normal distribution – highly doubtful here as the scatter diagram is not strongly elliptical with its points concentrated toward the centre of the ellipse – then we could estimate the standard error of the correlation coefficient with the known approximate formula $(1 - r^2)/\sqrt{n-3} = 0.094$. One of the benefits of the bootstrap procedure is that no distributional assumptions are necessary to use it.

◆ BOOTSTRAP CONFIDENCE INTERVALS ◆

In this section we outline a bootstrap method for producing a confidence interval (see Rice 1995). As before, it is helpful to compare this method with a standard technique, so we start with the well-known case of estimating a mean with a 'large' sample size. Returning to the M1 motorway data, an approximate 95% confidence interval for the mean interarrival time μ, by the central limit theorem, is

$$\left(\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}\right) = (5.36, 10.24)$$

Now we give the rationale for a bootstrap confidence interval. Suppose we can find values c_1 and c_2 so that

$$P(\overline{X} - \mu \le c_2) = 0.975 \text{ and } P(\overline{X} - \mu \le c_1) = 0.025 \quad (1)$$

then $P(c_1 < \overline{X} - \mu < c_2) = 0.95$ or, with some algebra, $P(\overline{X} - c_2 < \mu < \overline{X} - c_1) = 0.95$ so that

$$(\overline{X} - c_2, \overline{X} - c_1) \quad (2)$$

is a 95% confidence interval for μ. By way of reminder, (2) is a random interval that we are using in an attempt to capture the fixed unknown value of μ. Returning to (1), the value of μ, of course, is unknown. We can estimate it however by the observed sample mean which, recall, is 7.80. Then (1) approximately becomes

$$p(\overline{X} \le c_2 + 7.80) = 0.975 \quad \text{and}$$
$$p(\overline{X} \le c_1 + 7.80) = 0.025 \quad (3)$$

That is, $c_2 + 7.80$ is approximately the 97.5th percentile of the distribution of $\overline{X}$ and $c_1 + 7.80$ is approximately the 2.5th percentile of the distribution of $\overline{X}$.

Here is what we could do to estimate c_1 and c_2 *if we were able to time travel back to 1985 and sample from the interarrival population*:

(1) Sample $n = 40$ interarrivals from our population and compute $\bar{x}$.
(2) Repeat step (1) a number of times, B, to come up with estimates $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_B$.
(3) Use the sample percentiles to estimate the desired population percentiles. With $B = 1000$, for example, sort the estimates above as $\bar{x}_{(1)} \le \bar{x}_{(2)} \le \cdots \le \bar{x}_{(999)} \le \bar{x}_{(1000)}$ and use $\bar{x}_{(25)}$ to estimate the 2.5th percentile of $\overline{X}$ (set this equal to $c_1 + 7.80$ and solve for c_1) and use $\bar{x}_{(975)}$ to estimate the 97.5th percentile of $\overline{X}$ (set this equal to $c_2 + 7.80$ and solve for c_2). Finally, report $(\bar{x} - c_2, \bar{x} - c_1)$ as the desired confidence interval.

Mirroring the earlier discussion for the standard error, the corresponding bootstrap method comes up with the estimates in steps (1) and (2) *not by sampling from the population, but by sampling with replacement n times from the* ***data***. Carrying out this bootstrap method on one occasion using the Minitab macro in figure 4 along with the macro bootstrp.txt in figure 2 with $B = 1000$ gave $\bar{x}_{(25)} = 5.525$ and $\bar{x}_{(975)} = 10.325$. The value of $B = 1000$ is considered reasonably large for bootstrap confidence intervals; see Efron and Tibshirani (1993, pp. 273–5). Consequently, setting $\bar{x}_{(25)} = c_1 + 7.80$ and $\bar{x}_{(975)} = c_2 + 7.80$ in (3) gives $c_1 = -2.275$ and $c_2 = 2.525$, so that the desired 95% bootstrap confidence interval is $(\bar{x} - c_2, \bar{x} - c_1) = (5.28, 10.08)$. Note that this is quite close to the standard calculation given at the beginning of the section.

A bit of algebra shows that the 95% bootstrap confidence interval just given with $B = 1000$ can be expressed as $(2\bar{x} - \bar{x}_{(975)}, 2\bar{x} - \bar{x}_{(25)})$ (verify this). Generalizing the above chain of reasoning, a $100(1 - \alpha)\%$ bootstrap confidence interval for the parameter θ using the estimate $\hat{\theta}$ is

$$(2\hat{\theta} - \theta^*_{\text{Upper}}, 2\hat{\theta} - \theta^*_{\text{Lower}}) \qquad (4)$$

where θ^*_{Upper} is the $B(1 - \alpha/2)$ order statistic of the bootstrapped estimates and θ^*_{Lower} is the $B\alpha/2$ order statistic of the bootstrapped estimates.

If we want, for example, a 95% bootstrap confidence interval for the population median, run the Minitab macro in figure 4 replacing the line ***let k8 = mean(c1)*** with ***let k8 = median(c1)*** along with the macro bootstrp.txt in figure 2 replacing the line ***let k20 = mean(c11)*** with ***let k20 = median(c11)***. One such execution of this code with the M1 motorway data and $B = 1000$ gave the 25th smallest bootstrap median as 4.0 and the 975th smallest bootstrap median as 6.0. Consequently, the desired 95% bootstrap confidence interval, recalling the sample median to be 5.0, is $(2\hat{\theta} - \theta^*_{\text{Upper}}, 2\hat{\theta} - \theta^*_{\text{Lower}}) = (2(5.0) - 6.0, 2(5.0) - 4.0) = (4.0, 6.0)$.

```
File: cidriver.txt
noecho
erase c10
let k1 = n(c1)                    # the data must have previously been put in column c1
let k2 = 1000                     # number of bootstrap samples, B
execute 'bootstrp.txt' k2
sort c10 c11
let k3 = 0.95                     # desired confidence level
let k4 = round(k2*(1–k3)/2)
let k5 = round(k2*(1+k3)/2)
let k6 = c11(k4)
let k7 = c11(k5)                  # (k6,k7) is the percentile interval
let k8 = mean(c1)
let k10 = 2*k8–k7
let k11 = 2*k8–k6
print k10 k11                     # 100*k3% confidence interval (4) for mean
end
```

Use ***execute cidriver.txt*** at the Minitab prompt to run this bootstrap procedure

Fig 4. Bootstrap confidence interval code

Finally, returning to the problem of estimating the correlation between male mortality and calcium levels, executing analogous code (available from the author) with $B = 1000$ in one instance gave the 25th smallest bootstrap correlation as -0.776 and the 975th smallest bootstrap correlation as -0.490. Consequently, the desired 95% bootstrap confidence interval, recalling the sample correlation coefficient to be -0.655, is $(2\hat{\theta} - \theta^*_{\text{Upper}}, 2\hat{\theta} - \theta^*_{\text{Lower}}) = (2(-0.655) - (-0.490), 2(-0.655) - (-0.776)) = (-0.820, -0.534)$.

The above is but one (simple) method of using the bootstrap to come up with interval estimates; see Efron and Tibshirani (1993) for others. One other particularly simple interval, however, deserves a mention (and appears in Taffe and Garnham (1996)). In some cases the bootstrap distribution is symmetric about $\hat{\theta}$. Then $\hat{\theta} - \theta^*_{\text{Upper}} = -(\hat{\theta} - \theta^*_{\text{Lower}})$, and (4) simplifies as follows:

$$\begin{aligned}(2\hat{\theta} - \theta^*_{\text{Upper}}, 2\hat{\theta} - \theta^*_{\text{Lower}}) &= (\hat{\theta} + (\hat{\theta} - \theta^*_{\text{Upper}}), \hat{\theta} + (\hat{\theta} - \theta^*_{\text{Lower}})) \\ &= (\hat{\theta} - (\hat{\theta} - \theta^*_{\text{Lower}}), \hat{\theta} - (\hat{\theta} - \theta^*_{\text{Upper}})) \\ &= (\theta^*_{\text{Lower}}, \theta^*_{\text{Upper}})\end{aligned}$$

The interval $(\theta^*_{\text{Lower}}, \theta^*_{\text{Upper}})$ is, unsurprisingly, called a *bootstrap percentile interval.* To illustrate, consider two of our earlier examples. For the mean interarrival time on the M1 we had $(2\bar{x} - \bar{x}_{(975)}, 2\bar{x} - \bar{x}_{(25)}) = (5.28, 10.08)$ as opposed to the percentile interval $(\bar{x}_{(25)}, \bar{x}_{(975)}) = (5.53, 10.33)$. The closeness of these two intervals is due to the near symmetry in the bootstrap distribution of $\overline{X}$ values. In the correlation example, however, we had $(2r - r^*_{(975)}, 2r - r^*_{(25)}) = (-0.820, -0.534)$, whereas $(r^*_{(25)}, r^*_{(975)}) = (-0.776, -0.490)$. The first of these two intervals, since it does not assume symmetry, should be favoured over the latter percentile interval.

Given the correspondence between confidence intervals and hypothesis tests, it should come as no surprise that there are bootstrap procedures for conducting hypothesis tests; see Efron and Tibshirani (1993) for more details. For further discussion on the use of computationally intensive methods for hypothesis testing see Edgington (1995) and Good (2000).

◆ FURTHER DETAILS ◆

When trying to assess the performance of an estimate $\hat{\theta}$ of θ we will, in general, be concerned with the bias, $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$ of $\hat{\theta}$ as well as with $\text{SE}(\hat{\theta})$ In fact, a measure of the typical deviation of $\hat{\theta}$ from θ is the root mean square error, $\sqrt{\text{MSE}(\hat{\theta})}$, where

$$\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = \left\{\text{Bias}(\hat{\theta})\right\}^2 + \left\{\text{SE}(\hat{\theta})\right\}^2$$

Consequently, if we can estimate the bias as well as the standard error of an estimate, we can determine an estimate of the root mean square error. Fortunately, the bias can also be estimated by a bootstrap procedure (before continuing, do you see how?). In particular, reasoning as we have before, we can use the approximation

$$\begin{aligned}\text{Bias}(\hat{\theta}) &= E(\hat{\theta}) - \theta \\ &\approx \text{Average of Bootstrap Estimates} - \hat{\theta}\end{aligned}$$

The average of the bootstrap estimates can be obtained by inserting the lines ***let k9 = mean(c10)***, ***print k9,*** just before the end statement in the sedriver.txt macro. In the given illustrative examples involving the sample mean, sample median and sample correlation as estimates of their corresponding population parameters, little evidence of bias was seen (indeed, we know the sample mean is unbiased for the population mean).

The bootstrap procedures discussed here, one for the standard error of an estimate $\hat{\theta}$ of θ, and one for producing confidence intervals for θ, are *nonparametric bootstrap procedures*. In each case the bootstrap samples are obtained by repeated samples with replacement from the data. Alternatively, with *parametric bootstrap procedures* the original data can be used to fit a probability model and our samples can be drawn from it. To illustrate, the exponential density $f(x) = \lambda e^{-\lambda x}$ with $\lambda = 1/7.80$ fits the M1 motorway data well. Consequently, bootstrap samples can be obtained by taking random samples of size n from this fitted density and then computing the relevant statistic (e.g. mean, median) as before. To generate a particular random sample from an exponential distribution the *inverse cdf* method may be used. Specifically, generate a uniform random number U between 0 and 1. Then $-\ln(U)/\lambda = -7.80\ln(U)$ will have the desired exponential distribution.

Acknowledgement

Thanks are due to the referee for comments that led to an improved presentation.

References

Edgington, E. (1995). *Randomization Tests* (3rd revised edn). New York: Marcel Dekker.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, **1**(1), 54–77.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.

Good, P. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses* (2nd edn). New York: Springer.

Hand, D., Daly, F., Lunn A., McConway, K. and Ostrowski, E. (eds) (1994). *A Handbook of Small Data Sets*. London: Chapman & Hall.

Raspe, R.E. (1785). *The Adventures of Baron Munchausen*.

Reeves, J. (1995). Resampling stats. *Teaching Statistics*, **17**(3), 101–3.

Rice, J. (1995). *Mathematical Statistics and Data Analysis* (2nd edn), pp. 271–2.

Ricketts, C. and Berry, J. (1994). Teaching statistics through resampling. *Teaching Statistics*, **16**(2), 41–4.

Stout, W., Travers, K. and Marden, J. (1999). *Statistics: Making Sense of Data* (2nd edn). Rantoul, Illinois: Möbius Communications.

Taffe, J. and Garnham, N. (1996). Resampling, the bootstrap and Minitab. *Teaching Statistics,* **18**(1), 24–5.