

basic bootstrap in R

Andrew Robinson
Department of Mathematics and Statistics
University of Melbourne
Parkville, Vic. 3010
A.Robinson@ms.unimelb.edu.au

November 9, 2005

1 Introduction

The bootstrap provides an estimate of the empirical sampling distribution of many useful statistics, and functions of statistics. The estimate of the empirical sampling distribution can be examined directly for quantities of interest, such as standard errors or confidence intervals of parameters of interest.

For example, the sample mean is used to estimate the population mean in simple random sampling.

$$\hat{\mu}_v = \bar{v} = \frac{1}{n} \sum_{i=1}^n v_i \quad (1)$$

One or both of two measures of certainty are commonly deployed for the estimate of the population mean: the standard error, and the 95% confidence interval. The standard error is typically estimated by:

$$s_{\bar{v}} = \frac{s_v}{\sqrt{n}} \quad (2)$$

where s_v is the standard deviation of the sample. Then an approximate 95% confidence interval for the population mean is:

$$95\% \text{ C.I.} = \bar{v} \pm s_{\bar{v}} \times t_{0.975, n-1} \quad (3)$$

These quantities can be easily computed in R. For an example, we have 36 observations of point-level¹ above-ground volume in units of $m^3 ha^{-1}$ for the Priest River Experimental Forest in northern Idaho. The data are in a dataframe called `pref.point`; the variable of interest is `vol.m3.ha`.

¹I emphasize point level because the measurement is made, quite literally, at a point, and we can treat the population as though it were infinite.

```

> pref.tree <- read.csv("../data/pref.csv")
> pref.tree$vol.m3 <- pref.tree$vol.bf/12 * 0.0283168466
> pref.tree$TF.ha <- pref.tree$TF.ac * 2.47105381
> pref.tree$vol.m3.ha <- pref.tree$vol.m3 * pref.tree$TF.ha
> pref.point <- aggregate(x = list(vol.m3.ha = I(pref.tree$vol.m3.ha/5)),
+   by = list(stratum = pref.tree$stratum, cluster = pref.tree$cluster),
+   FUN = sum)
> SRS.v.bar <- mean(pref.point$vol.m3.ha)
> cat(SRS.v.bar, "m^3/ha")

213.1505 m^3/ha

> n <- length(pref.point$vol.m3.ha)
> SRS.v.se <- sd(pref.point$vol.m3.ha)/sqrt(n)
> cat(SRS.v.se, "m^3/ha")

18.51100 m^3/ha

> SRS.v.bar + SRS.v.se * qt(0.975, df = n - 1) * c(-1, 1)

[1] 175.5711 250.7298

```

The inferential use of estimates from equations (2) and (3) is predicated on large-sample theory, such as the Central Limit Theorem and Slutsky's Theorem (see, e.g., [Casella and Berger, 1990](#)). We ordinarily use large-sample theory to tell us how we can reasonably treat the sampling distribution of the parameter of interest. For example the sampling distribution of a mean can be reasonably treated as though it were normal.

The utility of these theories in any given situation depends on the circumstances. Generally, we will believe in them if the sample is big enough. What that means in practice is anyone's guess. If we do not wish to believe in the large-sample theory, then we can try to estimate the sampling distribution of the quantities of interest using the bootstrap. The bootstrap replaces (or supplements) assumptions based on large-sample theory with an estimated snapshot of the sampling distribution, warts and all.

A caveat is important: the bootstrap also relies upon its own large-sample theory, and it has concomitant strengths and weaknesses. Theoretical expositions suggest that the convergence of the bootstrap large-sample theory is faster than the classical large-sample theory in optimal situations ([Hall, 1992](#)). The bootstrap can ameliorate the problems associated with inference from small samples, but it will not eliminate them. Consequently, it is important to recognize that clever use of the bootstrap can lead to estimators with significantly better properties than naive use of the bootstrap.

A good, basic description can be found in [Efron and Tibshirani \(1993\)](#)

2 Implementation

We start with the simplest situation: a bootstrap estimate of the standard error of the sample mean, in R. Firstly, we need a function that computes the mean of the sample, using an index.

```
> boot.mean <- function(x, index) {
+   mean(x[index])
+ }
```

Surficially this seems like it must be more complicated than is necessary, and for this problem, it is. However, for more complicated situations this syntax lends itself to very easy generalization. The bootstrap function will efficiently call this function many times, sending the same data each time, along with a randomly generated index, which permutes the sample. We call it as follows:

```
> require(boot)
```

```
[1] TRUE
```

```
> pref.SRS.boot <- boot(pref.point$vol.m3.ha, boot.mean, R = 1999)
> pref.SRS.boot
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP
```

```
Call:
```

```
boot(data = pref.point$vol.m3.ha, statistic = boot.mean, R = 1999)
```

```
Bootstrap Statistics :
```

```
      original      bias  std. error
t1* 213.1505 -0.3193268   18.11971
```

The resulting object contains the output of the bootstrap call. We can examine it, as above, and plot it, as per Figure 1.

We're looking for approximate normality of the realizations in order for our bootstrap estimate to be reliable. If we are satisfied, then an estimate of the standard error is the standard deviation of the estimated bootstrap means.

```
> sd(pref.SRS.boot$t)
```

```
[1] 18.11971
```

This is the key bootstrap point. We have replaced an estimate that was based on an assumption about the underlying distribution with an estimate that is computed directly from a simulated underlying distribution. Much theory exists that says that this is a defensible idea in many cases.

We can then either use that directly to compute a confidence interval, or we can use one of a number of other strategies - which will be shown shortly. The following approach assumes that the sampling distribution is normal, and uses the familiar 1.96 as a scale multiplier.

```
> boot.ci(pref.SRS.boot, type = "norm")
```

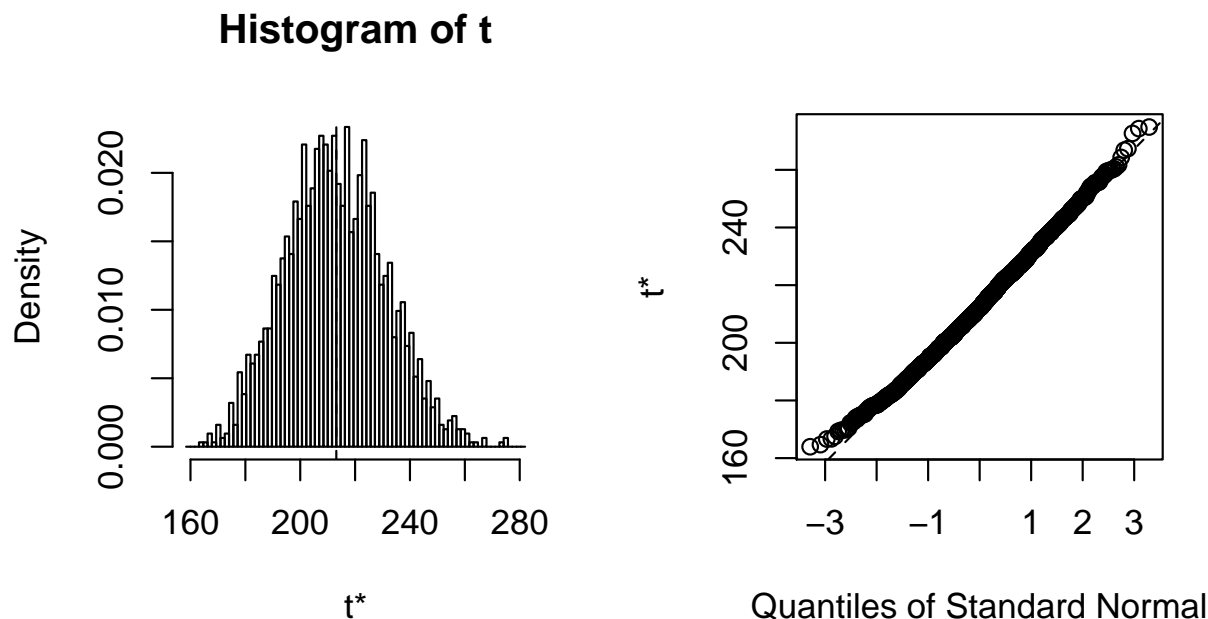


Figure 1: Diagnostic graphical output for the bootstrap object, obtained by calling `plot(pref.SRS.boot)`.

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 1999 bootstrap replicates

CALL :

```
boot.ci(boot.out = pref.SRS.boot, type = "norm")
```

Intervals :

Level Normal

95% (178, 249)

Calculations and Intervals on Original Scale

We can also deploy some graphical diagnostics to assess the effect of individual units upon our estimates. The `jack.after.boot()` function presents for each observation the empirical distribution of the simulations that *omit* that observation. It is here presented in Figure 2. Figure 2 identifies rows 4 and 12 as having a strong influence on the results. This observation is clarified by Figure 3, which provides a labelled normal quantile plot of the original sample.

Figure 3 was constructed as follows:

```
normalized <- qqnorm(pref.point$vol.m3.ha, plot.it=F)
plot(normalized$x, normalized$y, type="n", main="Normal Q-Q Plot",
      ylab="Sample Quantiles", xlab="Theoretical Quantiles")
```

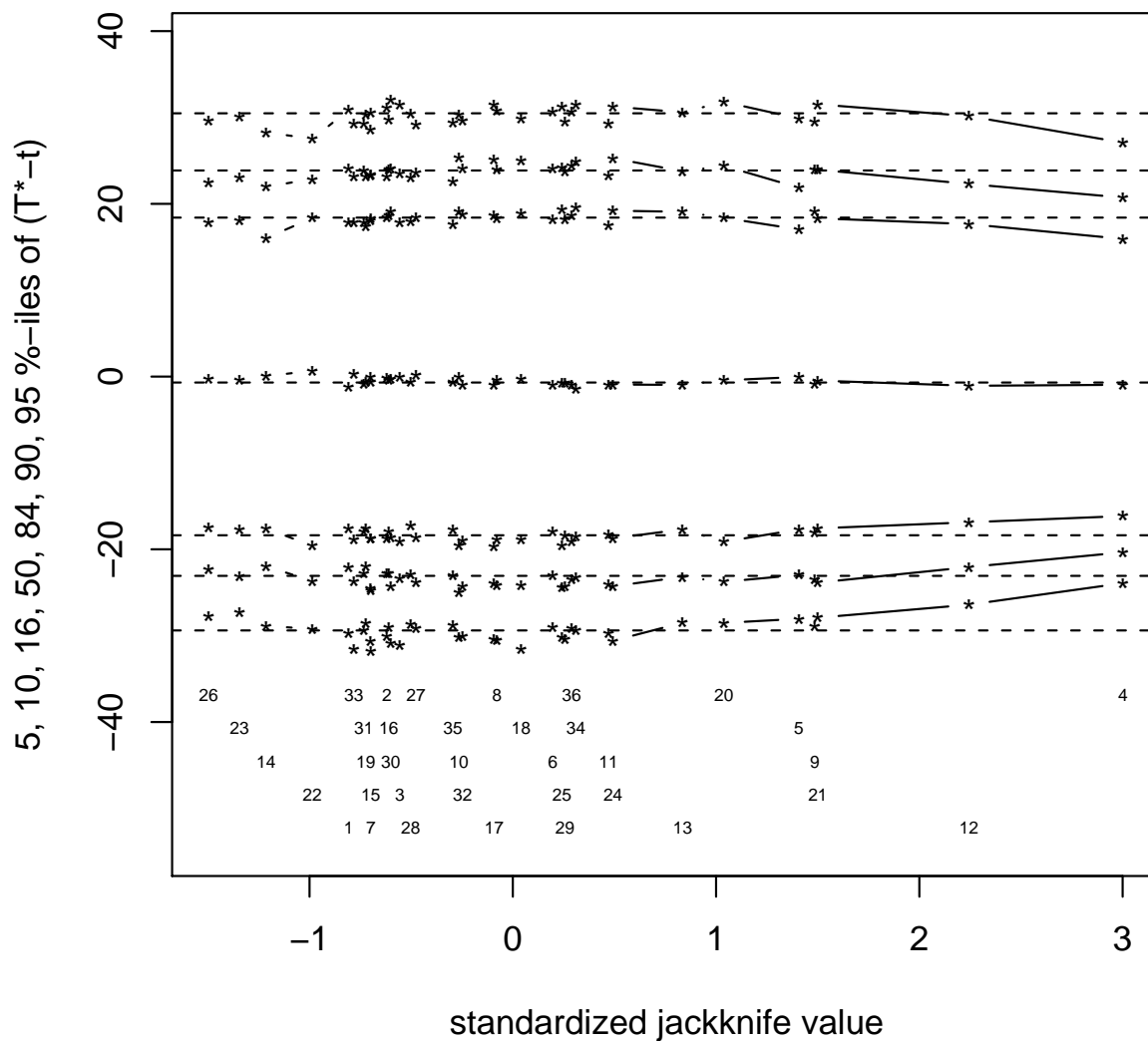


Figure 2: Diagnostic graphical output of jackknifing the bootstrap object, obtained by calling `jack.after.boot(pref.SRS.boot)`.

```
qqline(pref.point$vol.m3.ha, col="darkgray")
text(x = normalized$x, y = normalized$y, labels=1:36,
     cex=0.85)
```

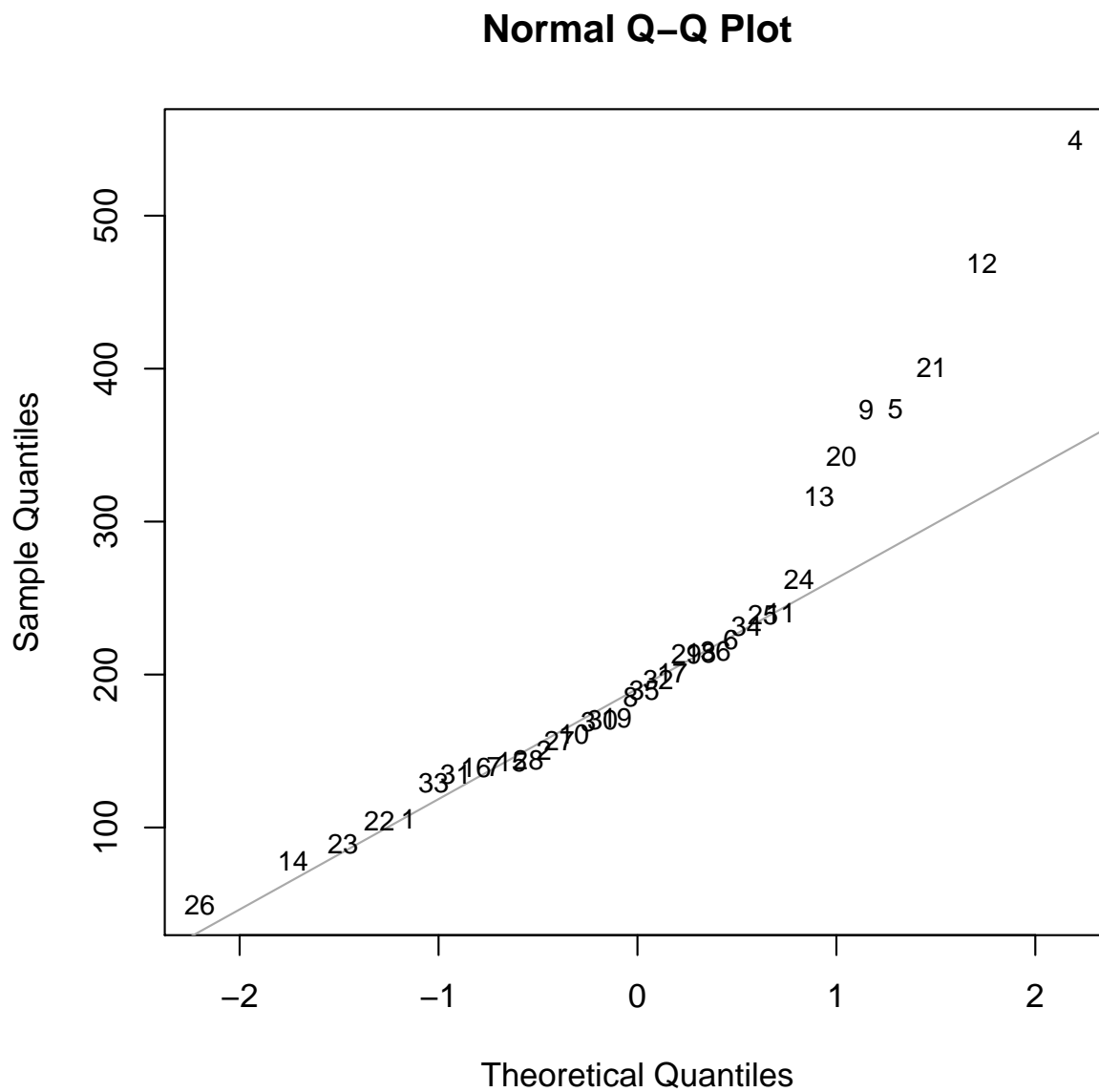


Figure 3: Normal Q-Q plot of the PREF point-level volume data, replacing the points with the row numbers of the observations.

3 Improvements

We now consider the construction of confidence intervals for a population parameter. As mentioned earlier, non-naive use of the bootstrap can lead to estimates that have better properties than naive use will. For example, transforming the statistic of interest so that it is a pivot results in bootstrap simulations with much better properties.

The `boot.ci` function provides up to 5 different non-parametric confidence interval estimates for any given alpha value. The basic interval structure is:

$$\hat{\mu} + s_{\hat{\mu}} \times t_{\alpha/2}, \quad \hat{\mu} + s_{\hat{\mu}} \times t_{1-\alpha/2} \quad (4)$$

norm the usual normal confidence interval (Eqn 4), with bootstrap-based bias correction for the estimates of the mean ($\hat{\mu}$) and standard error $s_{\hat{\mu}}$.

basic as per **norm**, but replace the $s_{\hat{\mu}} \times t$ values with the corresponding quantiles of the mean-shifted bootstrap sampling distribution.

stud as per **norm**, but replace t with the corresponding quantiles of the bootstrap sampling distribution of the studentized statistic.

perc directly use the quantiles of the estimated sampling distribution.

bca bias corrected and accelerated percentile intervals.

Furthermore, the accuracy of the first three bootstrap estimators can improve for certain datasets if a variance-stabilizing transformation is first employed.

In order to studentize the simulations it is necessary to obtain an estimate of the variance of the estimate within each realization. We could do this by adding another bootstrap, or by using an approximation. Since we are computing the mean, we can use the corresponding non-parametric delta method estimate². This function must be computed and reported as part of the bootstrap function, thus:

```
> boot.mean.2 <- function(x, index) c(mean(x[index]), (length(x) -
+   1) * var(x[index])/length(x)^2)
```

This function is then called as before, but now the bootstrap function can compute studentized values, which are approximate pivots.

```
> pref.SRS.boot.2 <- boot(pref.point$vol.m3.ha, boot.mean.2, R = 1999)
> pref.SRS.boot.2
```

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

```
boot(data = pref.point$vol.m3.ha, statistic = boot.mean.2, R = 1999)
```

²It's very similar to the usual non-parametric estimator of the variance of the mean

Bootstrap Statistics :

	original	bias	std. error
t1*	213.1505	-0.6306316	18.08040
t2*	333.1390	-13.1185919	97.13499

The bootstrap object now contains estimated means and variances. These can be graphed to assess stability (Figure 4).

```
scatter.smooth(pref.SRS.boot.2$t[,1], sqrt(pref.SRS.boot.2$t[,2]),
  main="", ylab="Standard Error", xlab="Bootstrap Mean")
```

Not very stable here. Since we know that these are volumes, and that they are skewed (and limited to positive numbers only) it's possible that a log transformation will be profitable. Figure 5 shows that this seems to be a reasonable strategy.

```
pref.SRS.boot.3 <- boot(I(log(pref.point$vol.m3.ha)), boot.mean.2, R=1999)
scatter.smooth(pref.SRS.boot.3$t[,1], sqrt(pref.SRS.boot.3$t[,2]),
  main="", ylab="Standard Error", xlab="Bootstrap Mean")
```

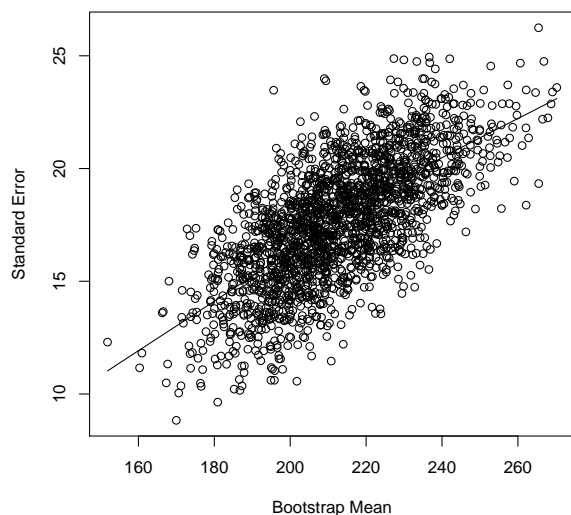


Figure 4: Scatterplot and lowess smooth of estimated standard errors against estimated means from each of 1999 bootstrap replicates.

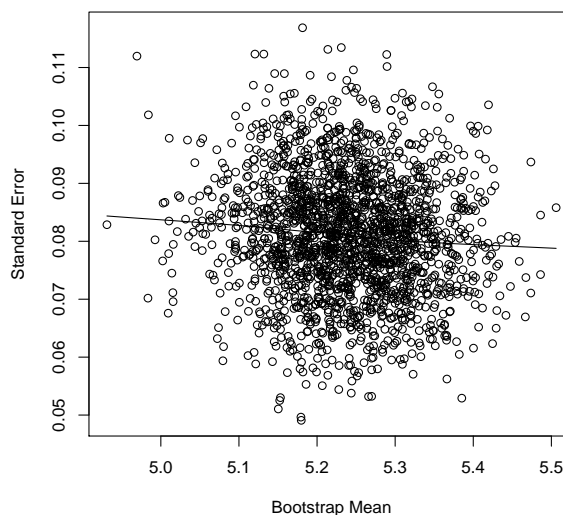


Figure 5: Scatterplot and lowess smooth of estimated standard errors against estimated means from each of 1999 bootstrap replicates taken from log-transformed data.

A call to `boot.ci`, with appropriate forward and backward transformations, produces all five intervals.

```
> boot.ci(pref.SRS.boot.2, h = log, hdot = function(x) 1/x, hinv = exp)
```

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 1999 bootstrap replicates

CALL :

```
boot.ci(boot.out = pref.SRS.boot.2, h = log, hdot = function(x) 1/x,
        hinv = exp)
```

Intervals :

Level	Normal	Basic	Studentized
95%	(181.6, 253.5)	(181.8, 253.0)	(181.9, 257.5)

Level	Percentile	BCa
95%	(179.6, 249.9)	(183.6, 256.7)

Calculations on Transformed Scale; Intervals on Original Scale

They are fairly similar, and do not differ wildly from the classical interval computed previously (175.6, 250.7).

It is unclear which of these intervals is the best to use. Davison and Hinkley (1997) recommend studentized intervals applied to data that have been transformed to stabilize variance. More details can be found in Davison and Hinkley (1997).

4 Other details

How many replicates? Efron and Tibshirani (1993) suggest up to 199 for standard errors and 1999 for quantiles.

References

- Casella, G., Berger, R. L., 1990. Statistical inference. Duxbury Press.
- Davison, A. C., Hinkley, D. V., 1997. Bootstrap methods and their application. Cambridge University Press.
- Efron, B., Tibshirani, R. J., 1993. An introduction to the bootstrap. Chapman and Hall.
- Hall, P., 1992. The bootstrap and Edgeworth expansion. Springer.