Explorations in statistics: the bootstrap

Douglas Curran-Everett

Division of Biostatistics and Bioinformatics, National Jewish Health, and Department of Biostatistics and Informatics 
and Department of Physiology and Biophysics, School of Medicine, University of Colorado Denver, Denver, Colorado 

Submitted 20 July 2009; accepted in final form 4 September 2009

Curran-Everett D. Explorations in statistics: the bootstrap. Adv Physiol Educ 33: 286–292, 2009; doi:10.1152/advan.00062.2009.— Learning about statistics is a lot like learning about science: the learning is more meaningful if you can actively explore. This fourth installment of Explorations in Statistics explores the bootstrap. The bootstrap gives us an empirical approach to estimate the theoretical variability among possible values of a sample statistic such as the sample mean. The appeal of the bootstrap is that we can use it to make an inference about some experimental result when the statistical theory is uncertain or even unknown. We can also use the bootstrap to assess how well the statistical theory holds: that is, whether an inference we make from a hypothesis test or confidence interval is justified.

Central Limit Theorem; R; sample mean; software; standard error

This fourth paper in Explorations in Statistics (see Refs. 3–5) explores the bootstrap, a recent development in statistics (8, 11–16) that evolved from the jackknife (25, 28, 36). Despite its brief history, the bootstrap is discussed in textbooks of statistics (26) and used in manuscripts published by the American Physiological Society (2, 10, 17–24, 35, 37).

The bootstrap gives us an empirical approach to estimate the theoretical variability among possible values of a sample statistic. In our previous explorations (3–5) we calculated the standard error of the mean, our estimate of the theoretical variability among possible values of a sample statistic. In our early explorations (3–5) we drew random samples of 9 observations from a standard normal distribution with mean μ = 0 and standard deviation σ = 1. These were the observations—the data—for samples 1, 2, and 1000:

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.422</td>
<td>1.006</td>
<td>0.371</td>
</tr>
<tr>
<td>0.147</td>
<td>1.138</td>
<td>0.797</td>
</tr>
<tr>
<td>1.715</td>
<td>0.517</td>
<td>0.838</td>
</tr>
<tr>
<td>1.725</td>
<td>0.517</td>
<td>1.715</td>
</tr>
<tr>
<td>1.103</td>
<td>1.560</td>
<td>0.702</td>
</tr>
<tr>
<td>1.006</td>
<td>0.485</td>
<td>0.702</td>
</tr>
<tr>
<td>0.587</td>
<td>-0.371</td>
<td>-0.371</td>
</tr>
</tbody>
</table>

Each time we drew a random sample we calculated some sample statistics. These were the statistics for samples 1, 2, and 1000:

<table>
<thead>
<tr>
<th>Sample</th>
<th>Mean</th>
<th>SE</th>
<th>t</th>
<th>LCI</th>
<th>UCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.147</td>
<td>0.070</td>
<td>3.364</td>
<td>0.325</td>
<td>1.232</td>
</tr>
<tr>
<td>2</td>
<td>0.797</td>
<td>0.070</td>
<td>3.364</td>
<td>0.325</td>
<td>1.232</td>
</tr>
</tbody>
</table>

In contrast to our early explorations that used 1000 samples from our standard normal population, our exploration of the bootstrap uses just the 9 observations from sample 1:

0.422, 1.006, 1.138, 0.147, 1.103, 0.371, 0.797, 0.702, 0.838

In our previous exploration (4) we used these observations to calculate a 90% confidence interval for the population mean:

[0.362, 1.232] = [0.36, 1.23]

Because this interval excluded 0, we declared, with 90% confidence, that 0 was not a plausible value of the population mean. This inference was consistent with the result of

1 The appendix reviews the origin of the name bootstrap.
2 For example, a sample mean, a sample standard deviation, or a sample correlation.
3 I developed the scripts for the early explorations (3–5) using R-2.6.2. I developed the script for this exploration using R-2.8.2 (deployed 22 Dec 2008), but it will run in R-2.6.2.
4 This file is available through the Supplemental Material for this article at the Advances in Physiology Education website.
5 The notation click A B means click A, then click B.
6 CRAN stands for the Comprehensive R Archive Network. A mirror is a duplicate server.
we do this? We draw at random—with replacement7—a sample.

However, we do not use theory (7) to develop the notion of the standard error of the mean, our estimate of the theoretical variability among possible values of the sample mean (3). When we treated those 1000 sample means as observations, we calculated their standard deviation: 

\[ SE(\bar{y}) = 0.326. \]

The standard deviation of sample means is the standard error of the sample mean, SE(\(\bar{y}\)). The bootstrap estimates the standard error of a statistic using not a whole bunch of random samples from some theoretical population but actual sample observations.

The Bootstrap

In our previous exploration (4) we used the standard error of the mean, our estimate of the theoretical variability among possible values of the sample mean, to calculate a confidence interval for the mean of the underlying population. Rather than use theory (7) to develop the notion of the standard error of the mean, we used a simulation: we drew 1000 random samples from our population and—for each sample—we calculated the mean (3). When we treated those 1000 sample means as observations, we calculated their standard deviation:

\[ SE(\bar{y}) = 0.326. \]

The standard deviation of sample means is the standard error of the sample mean SE(\(\bar{y}\)). The bootstrap estimates the standard error of a statistic using not a whole bunch of random samples from some theoretical population but actual sample observations.

Table 1. Illustration of bootstrap samples

<table>
<thead>
<tr>
<th>(y)</th>
<th>(\bar{y}_1)</th>
<th>(\bar{y}_2)</th>
<th>(\bar{y}_3)</th>
<th>(\bar{y})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.422</td>
<td>0.285</td>
<td>...</td>
<td>1.034</td>
<td></td>
</tr>
<tr>
<td>1.103</td>
<td>1.235</td>
<td>1.235</td>
<td>0.285</td>
<td></td>
</tr>
<tr>
<td>1.006</td>
<td>0.912</td>
<td>1.825</td>
<td>1.103</td>
<td></td>
</tr>
<tr>
<td>1.034</td>
<td>0.912</td>
<td>1.034</td>
<td>0.912</td>
<td></td>
</tr>
<tr>
<td>0.285</td>
<td>1.825</td>
<td>1.006</td>
<td>1.103</td>
<td></td>
</tr>
<tr>
<td>-0.647</td>
<td>1.825</td>
<td>1.103</td>
<td>0.422</td>
<td></td>
</tr>
<tr>
<td>1.235</td>
<td>0.285</td>
<td>0.912</td>
<td>-0.647</td>
<td></td>
</tr>
<tr>
<td>0.912</td>
<td>1.006</td>
<td>1.103</td>
<td>0.912</td>
<td></td>
</tr>
<tr>
<td>1.825</td>
<td>-0.647</td>
<td>0.422</td>
<td>1.825</td>
<td></td>
</tr>
<tr>
<td>0.797</td>
<td>0.640</td>
<td>0.992</td>
<td>0.772</td>
<td></td>
</tr>
</tbody>
</table>

\(\bar{y}\), Actual observations from sample 1. Because the bootstrap procedure samples with replacement, an original observation can appear more than once in a bootstrap sample. The bootstrap sample means, \(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_b\) cluster around 0.797, the observed sample mean \(\bar{y}\).

The commands in lines 40–49 of Advances_Statistics_Code_Boot.R create three bootstrap samples from the observations in sample 1. To generate three bootstrap samples, we used a simulation: we drew 1000 random samples from Table 1: first line to Table 1: last line.

With this brief review of the observations from our first sample, we are ready to explore the bootstrap.

The Bootstrap

In our previous exploration (4) we used the standard error of the mean, our estimate of the theoretical variability among possible values of the sample mean, to calculate a confidence interval for the mean of the underlying population. Rather than use theory (7) to develop the notion of the standard error of the mean, we used a simulation: we drew 1000 random samples from our population and—for each sample—we calculated the mean (3). When we treated those 1000 sample means as observations, we calculated their standard deviation:

\[ SE(\bar{y}) = 0.326. \]

The standard deviation of sample means is the standard error of the sample mean SE(\(\bar{y}\)). The bootstrap estimates the standard error of a statistic using not a whole bunch of random samples from some theoretical population but actual sample observations.

Suppose we want to bootstrap the mean using the observations from the first sample: 0.422, 1.103, ..., 1.825. How do we do this? We draw at random—with replacement7—a sample of size 9 from these 9 actual observations (Table 1). We then repeat this process until we have drawn a total of b bootstrap samples.8 For each bootstrap sample, we calculate its mean.

The bootstrap replications were drawn at random from the observations in sample 1 (see Table 1). The average of these sample means, Ave(\(\bar{y}\)), is 0.798, nearly identical to the observed sample mean \(\bar{y} = 0.797\). The standard deviation of these sample means, SD(\(\bar{y}\)), is 0.219, which estimates the standard deviation of the theoretical distribution of the sample mean, \(\sigma/\sqrt{n} = 0.333\). The commands in lines 65–78 of Advances_Statistics_Code_Boot.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figure 1: first line to Figure 1: last line.
Normal distribution

Nonnormal distribution

Empirical quantiles

Theoretical quantiles

Fig. 2. Normal quantile plots (right) of 100 observations drawn from a normal (top) or nonnormal (bottom) distribution. If observations are consistent with a normal distribution, then they will follow a rough straight line. If observations are inconsistent with a normal distribution, then they will deviate systematically from a straight line. The commands in lines 109–156 of Advances_Statistics_Code_Boot.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figure 2: first line to Figure 2: last line.

Empirical quantiles

Theoretical quantiles

Fig. 3. Normal quantile plot of 10,000 sample means, each with 9 observations. The bootstrap replications were drawn at random from the observations in sample 1 (see Table 1). The bootstrap sample means are inconsistent with a normal distribution. The commands in lines 165–171 of Advances_Statistics_Code_Boot.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figure 3: first line to Figure 3: last line.

Percentile confidence interval. In the bootstrap distribution of 10,000 sample means (see Fig. 1), 90% of the means are covered by the interval

\[ [0.426, 1.143] \]

The commands in lines 196–199 of Advances_Statistics_Code_Boot.R return these values.

Fig. 4. Normal quantile plot of 10,000 sample means, each with 9 observations. The bootstrap replications were drawn at random from the observations in sample 1000 (see Table 1). These bootstrap sample means are consistent with a normal distribution. The commands in lines 183–189 of Advances_Statistics_Code_Boot.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figure 4: first line to Figure 4: last line.

\[ \hat{y} - a, \hat{y} + a \] ,

where the allowance \( a \) was

\[ a = n^{1/2} \cdot \text{SD}(\bar{y}) \].

Recall that \( n^{1/2} \) is the 100\[1 - (\alpha/2)\]th percentile from the standard normal distribution and \( \text{SD}(\bar{y}) \) is the standard deviation of the sample means, \( \sigma \sqrt{n} \).

A normal-theory confidence interval based on bootstrap replications of the sample mean is similar. We just replace \( \hat{y} \) with \( \text{Ave}(\hat{y}) \), \( a \) with \( a^* \), and \( \text{SD}(\hat{y}) \) with \( \text{SD}(\text{Ave}(\hat{y})) \):

\[ [\text{Ave}(\hat{y}^*), a^*, \text{Ave}(\hat{y}) + a^*] \]

The bootstrap allowance \( a^* \) is

\[ a^* = n^{1/2} \cdot \text{SD}(\text{Ave}(\hat{y})) \].

Suppose we want to calculate a 90% bootstrap confidence interval for the population mean. In this situation, \( \alpha = 0.10 \) and \( n^{1/2} = 1.645 \). Therefore, the allowance \( a^* \) is

\[ a^* = n^{1/2} \cdot \text{SD}(\text{Ave}(\hat{y})) = 1.645 \cdot 0.219 = 0.360 \],

and the resulting 90% confidence interval is

\[ [0.438, 1.158] \approx [0.44, 1.16] \]

This bootstrap confidence interval resembles the confidence interval of \([0.36, 1.23]\) that we calculated before (4). The commands in lines 196–199 of Advances_Statistics_Code_Boot.R return these values. Your values will differ slightly.
Bias-corrected-and-accelerated confidence interval. If the number of observations in the actual sample is too small, if the average \( \bar{y}_n \) of the bootstrap sample means differs from the sample mean \( \bar{y} \), or if the bootstrap distribution of the sample mean is skewed, then normal-theory and percentile confidence intervals are likely to be inaccurate (16, 26). A bias-corrected-and-accelerated confidence interval adjusts percentiles of the bootstrap distribution to account for bias and skewness (16, 26). In this exploration, the 90\% bias-corrected-and-accelerated confidence interval is

\[
[0.376, 1.113] \div [0.38, 1.11] .
\]

The commands in lines 196–199 of Advances_Statistics_Code_Boot.R return these values. This confidence interval is shifted slightly to the left of the percentile interval of [0.43, 1.14]. In many situations, a bias-corrected-and-accelerated confidence interval will provide a more accurate estimate of our uncertainty about the true value of a population parameter such as the mean (16, 26).

Limitations. As useful as the bootstrap is, it cannot always salvage the statistical analysis of a small sample. Why not? If the sample is too small, then it may be atypical of the underlying population. When this happens, the bootstrap distribution will not mirror the theoretical distribution of the sample statistic. The trouble is, it may not be obvious how small too small is. A statistician (see Ref. 6, guideline 1) and an estimate of power can help.

In our second and third explorations (4, 5) we concluded that the observations from sample 1 were consistent with having come from a population that had a mean other than 0. Only because we had defined the underlying population (see Ref. 3) could we have known that we had erred. When we do a single experiment, we can have enough unusual observations so that it just appears the observations came from a different population (5). A small sample size exacerbates the potential for this phenomenon. This is what happened when we bootstrapped the sample mean using observations from the first sample.

We know the theoretical distribution of the sample mean is exactly normal (3), but the distribution of those bootstrap means is clearly not normal (see Fig. 3). This happens because the observations from the first sample are atypical of the underlying population. The message? All bets are off if you bootstrap a sample statistic using observations from a sample that is too small.

The Bootstrap in Data Transformation

In our previous exploration (4) we delved into confidence intervals by drawing random samples from a normal distribution. That we chose a normal distribution for our population was no accident. For normal-theory confidence intervals to be meaningful, one thing must be approximately true: if the random variable \( Y \) represents the physiological thing we care about, then the theoretical distribution of the sample mean \( \bar{Y} \) with \( n \) observations must be distributed normally with mean \( \mu \) and standard deviation \( \sigma/\sqrt{n} \). In our simulations (3–5) this assumption was satisfied exactly (7). In a real experiment we never know if this assumption is satisfied, but we know it will be satisfied at least roughly—regardless of the population distribution from which the sample observations came—as long as the sample size \( n \) is big enough (27).

What happens, however, if we doubt that the theoretical distribution of the sample mean is consistent with a normal distribution? Our suspicion would be natural: this distribution is, after all, a theoretical one. One approach is to transform the sample observations. Common transformations include the logarithm and the inverse. Box and Cox (1) described a family of power transformations in which an observed variable \( y \) is transformed into the variable \( w \) using the parameter \( \lambda \):

\[
w = \begin{cases} 
(y^\lambda - 1)/\lambda & \text{for } \lambda \neq 0 \\
\ln y & \text{for } \lambda = 0.
\end{cases}
\]

The logarithmic, \( \lambda = 0 \), and inverse, \( \lambda = -1 \), transformations belong to this family.

Draper and Smith (9) summarized the steps needed to estimate \( \lambda \) and its approximate 100(1 – \( \alpha \))% confidence interval. These steps include, for each trial value of \( \lambda \), the calculation of the maximum likelihood \( \ell_{\max} \):

\[
\ell_{\max} = -\frac{1}{2} \ln (SS_{\text{residual}}/n) + (\lambda - 1) \sum \ln y_i , \quad (1)
\]

where \( SS_{\text{residual}} \) is the residual sum of squares in the fit of a general linear model to the observations. The optimum estimate of \( \lambda \) maximizes \( \ell_{\max} \) (Fig. 5).

Without question, transformation can be useful (1, 9). But really, who wants to identify the optimum transformation by solving Eq. 1? The bootstrap provides another way.

The textbook I use in my statistics course provides the data: measurements of C-reactive protein (Table 2). Problem 7.26 (Ref. 26, p. 442–443) asks students to study the distribution of these 40 observations and to calculate a 95\% confidence interval for the true C-reactive protein mean. The real value of problem 7.26 is that it then asks students if a confidence interval is even appropriate for these data.

Cursory examinations of the histogram and normal quantile plot reveal that the C-reactive protein values are skewed and inconsistent with a normal distribution (Fig. 6, top). Still, if the theoretical distribution of the sample mean is roughly normal—if

\[
\bar{Y} \sim N(\mu, \sigma/\sqrt{n}) .
\]

In our second exploration (5) we used the test statistic \( t \) to investigate hypothesis tests, test statistics, and \( P \) values. A \( t \) statistic shares this assumption.

By using the hypothesis test to determine if the sample mean is different from 0, we are using the central limit theorem. The C-reactive protein distribution is not normal, but it is close enough for the central limit theorem to apply. We conclude that the C-reactive protein values are likely to be normally distributed (Fig. 6, bottom). Thus, the 95\% confidence interval calculated using \( t \) is likely to be valid. We can use the bootstrap to confirm this conclusion.

The commands in lines 180–189 of Advances_Statistics_Code_Boot.R return these values. This confidence interval is

\[
[0.376, 1.113] \div [0.38, 1.11] .
\]

The logarithmic, \( \lambda = 0 \), and inverse, \( \lambda = -1 \), transformations belong to this family.

Draper and Smith (9) summarized the steps needed to estimate \( \lambda \) and its approximate 100(1 – \( \alpha \))% confidence interval. These steps include, for each trial value of \( \lambda \), the calculation of the maximum likelihood \( \ell_{\max} \):

\[
\ell_{\max} = -\frac{1}{2} \ln (SS_{\text{residual}}/n) + (\lambda - 1) \sum \ln y_i , \quad (1)
\]

where \( SS_{\text{residual}} \) is the residual sum of squares in the fit of a general linear model to the observations. The optimum estimate of \( \lambda \) maximizes \( \ell_{\max} \) (Fig. 5).

Without question, transformation can be useful (1, 9). But really, who wants to identify the optimum transformation by solving Eq. 1? The bootstrap provides another way.

The textbook I use in my statistics course provides the data: measurements of C-reactive protein (Table 2). Problem 7.26 (Ref. 26, p. 442–443) asks students to study the distribution of these 40 observations and to calculate a 95\% confidence interval for the true C-reactive protein mean. The real value of problem 7.26 is that it then asks students if a confidence interval is even appropriate for these data.

Cursory examinations of the histogram and normal quantile plot reveal that the C-reactive protein values are skewed and inconsistent with a normal distribution (Fig. 6, top). Still, if the theoretical distribution of the sample mean is roughly normal—\( \bar{Y} \sim N(\mu, \sigma/\sqrt{n}) \)—the 95\% confidence interval calculated using \( t \) is likely to be valid. We can use the bootstrap to confirm this conclusion.
**Table 2. C-reactive protein data**

<table>
<thead>
<tr>
<th>y</th>
<th>w</th>
<th>y</th>
<th>w</th>
<th>y</th>
<th>w</th>
<th>y</th>
<th>w</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>73.20</td>
<td>4.307</td>
<td>0</td>
<td>0</td>
<td>15.74</td>
<td>2.818</td>
</tr>
<tr>
<td>3.90</td>
<td>1.589</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5.64</td>
<td>1.893</td>
<td>46.70</td>
<td>3.865</td>
<td>4.81</td>
<td>1.760</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8.22</td>
<td>2.221</td>
<td>0</td>
<td>0</td>
<td>9.57</td>
<td>2.358</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.36</td>
<td>1.850</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5.62</td>
<td>1.890</td>
<td>26.41</td>
<td>3.111</td>
<td>0</td>
<td>0</td>
<td>9.37</td>
<td>2.339</td>
</tr>
<tr>
<td>3.92</td>
<td>1.593</td>
<td>22.82</td>
<td>3.171</td>
<td>5.66</td>
<td>1.896</td>
<td>20.78</td>
<td>3.081</td>
</tr>
<tr>
<td>6.81</td>
<td>2.055</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7.10</td>
<td>2.092</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>30.61</td>
<td>3.453</td>
<td>59.76</td>
<td>4.107</td>
<td>7.89</td>
<td>2.185</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>3.49</td>
<td>1.502</td>
<td>12.38</td>
<td>2.594</td>
<td>5.53</td>
<td>1.876</td>
</tr>
</tbody>
</table>

Where: y, Actual C-reactive protein observations (in mg/l) (26). The average, standard deviation, and standard error of these observations are Ave(y) = 10.03, SD(y) = 16.56, and SE(y) = 2.62. The transformed observations, w, result from ln (y + 1). The average, standard deviation, and standard error of the transformed observations are Ave(w) = 1.495, SD(w) = 1.391, and SE(w) = 0.220. For each set of observations, if we want to calculate a 95% confidence interval for the true mean, then α = 0.05, the degrees of freedom ν = 39, and tν,α/2 = 2.023 (see Ref. 4).

Staying Current

Table 2. C-reactive protein data

<table>
<thead>
<tr>
<th>y</th>
<th>w</th>
<th>y</th>
<th>w</th>
<th>y</th>
<th>w</th>
<th>y</th>
<th>w</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>73.20</td>
<td>4.307</td>
<td>0</td>
<td>0</td>
<td>15.74</td>
<td>2.818</td>
</tr>
<tr>
<td>3.90</td>
<td>1.589</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5.64</td>
<td>1.893</td>
<td>46.70</td>
<td>3.865</td>
<td>4.81</td>
<td>1.760</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8.22</td>
<td>2.221</td>
<td>0</td>
<td>0</td>
<td>9.57</td>
<td>2.358</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.36</td>
<td>1.850</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5.62</td>
<td>1.890</td>
<td>26.41</td>
<td>3.111</td>
<td>0</td>
<td>0</td>
<td>9.37</td>
<td>2.339</td>
</tr>
<tr>
<td>3.92</td>
<td>1.593</td>
<td>22.82</td>
<td>3.171</td>
<td>5.66</td>
<td>1.896</td>
<td>20.78</td>
<td>3.081</td>
</tr>
<tr>
<td>6.81</td>
<td>2.055</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7.10</td>
<td>2.092</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>30.61</td>
<td>3.453</td>
<td>59.76</td>
<td>4.107</td>
<td>7.89</td>
<td>2.185</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>3.49</td>
<td>1.502</td>
<td>12.38</td>
<td>2.594</td>
<td>5.53</td>
<td>1.876</td>
</tr>
</tbody>
</table>

The Central Limit Theorem holds—then the 95% confidence interval [4.74, 15.33] mg/l will be meaningful. But we have an intractable problem: we have no way of knowing if the sample size of 40 is big enough for the theoretical distribution of the sample mean to be roughly normal.

**Problem 7.27** (Ref. 26, p. 443) tells students a log transformation10 decreases skewness and asks them to transform the actual C-reactive protein values as the optimal transformation (see Fig. 5).

![Fig. 6. Distributions (left) and normal quantile plots (right) of the actual and transformed C-reactive protein observations. The transformation changed the distribution of the observations, but the transformed values remain inconsistent with a normal distribution. The commands in lines 232–252 of Advances_Statistics_Code_Boot.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figure 6: first line to Figure 6: last line.](http://advan.physiology.org/)

C-reactive protein observations by adding 1 to each observation and taking the logarithm of that number:

$$w = \ln(y + 1)$$

Problem 7.27 then asks students to study the distribution of the 40 transformed observations and to calculate a 95% confidence interval for the true transformed C-reactive protein mean.

A histogram and normal quantile plot show that the transformed values are less skewed but still inconsistent with a normal distribution (Fig. 6, bottom). As my students tell me, “Better, but not great.” The 95% confidence interval [1.05, 1.94] reverts to

$$[e^{1.05} - 1, e^{1.94} - 1] = [1.86, 5.96] \text{ mg/l}$$

At this point, we have another problem: we have assumed the log transformation $\ln (y + 1)$ is useful— that it gives us a meaningful confidence interval— but what evidence do we have that it really is? You guessed it: the bootstrap.

A bootstrap distribution estimates the theoretical distribution of some sample statistic. In this problem, the bootstrap sample means from the actual observations are inconsistent with a normal distribution (Fig. 7, top). This means the sample size of 40 is not big enough for the theoretical distribution of the sample mean to be roughly normal; as a result, the confidence interval [4.74, 15.33] mg/l is misleading. On the other hand, the bootstrap sample means from the transformed C-reactive protein observations are consistent with a normal distribution (Fig. 7, bottom): the confidence interval [1.05, 1.94] = [1.86, 5.96] mg/l is a useful tool for inference.

![Fig. 7. Bootstrap distributions (left) and normal quantile plots (right) of 10,000 sample means, each with 40 observations. The bootstrap replications were drawn at random from the actual or transformed C-reactive protein observations in Table 2. The bootstrap sample means from the actual observations are inconsistent with a normal distribution. In contrast, the bootstrap sample means from the transformed observations are consistent with a normal distribution. The commands in lines 259–279 of Advances_Statistics_Code_Boot.R create this data graphic. To generate this data graphic, highlight and submit the lines of code from Figure 7: first line to Figure 7: last line.](http://advan.physiology.org/)
Summary

As this exploration has demonstrated, the bootstrap gives us an approach we can use to assess whether an inference we make from a normal-theory hypothesis test or confidence interval is justified: if the bootstrap distribution of a statistic such as the sample mean is roughly normal, then the inference is justified. But the bootstrap gives us more than that. We can use a bootstrap confidence interval to make an inference about some experimental result when the statistical theory is uncertain or even unknown (14, 26). Although we have explored the bootstrap using the sample mean, we can use the bootstrap to make an inference about other sample statistics such as the standard deviation or correlation coefficient.

In the next installment of this series, we will explore power, a concept we mentioned in our exploration of hypothesis tests (5). Power is the probability that we reject the null hypothesis given that the null hypothesis is false. The notion of power is integral to hypothesis testing, confidence intervals, and grant applications.

APPENDIX

In 1993, Efron and Tibshirani (16) wrote that bootstrap, a term coined originally by Efron in 1979 (11), was inspired by the colloquialism pull yourself up by your bootstraps, a nonsensical notion generally attributed to an escape of the fictitious Baron Munchausen (30). When I learned this, I embarked on a search for the escape. With the generous assistance of Dr. Sarah Kareem (Department of English, University of California, Los Angeles, CA) and Bernhard Wiebel (Munchausen Library, Zurich, Switzerland), this is what I discovered.

The complete title of the Baron’s adventures (30), written by Rudolf Erich Raspe, was Baron Munchausen’s Narrative of His Marvellous Travels and Campaigns in Russia. Humbly Dedicated and Recommended to Country Gentlemen; and, If They Please, to be Repeated as Their Own, After a Hunt at Horse Races, in Watering-Places, and Other Such Polite Assemblies; Round the Bottle and Fire-Side.

In chapter VI, the Baron throws a silver hatchet at two bears in hopes of rescuing a bee.11 The hatchet misses both bears and ends up on the moon. To retrieve the hatchet, the Baron grows and promptly climbs a Turkey-bean after it had attached itself to the moon. He finds his hatchet but then discovers the bean has dried up, so he braids a rope of straw to use for the descent. The Baron is partway down the straw rope when the next catastrophe hits. The details of the Baron’s escape differ according to the account you happen to read:

I was four or five miles from the earth at least, when it broke; I fell to the ground with such amazing violence, that I found myself stunned, and in a hole nine fathoms deep at least, made by the weight of my body falling from so great a height: I recovered, but knew not how to get out again; however, I dug slopes or steps or with my [finger] nails (the Baron’s nails were then of forty years’ growth), and easily accomplished it.

It was from this 9-fathom-deep hole that the Baron is rumored to have extricated himself by his bootstraps. Although the notion of pulling yourself up by your bootstraps is entirely consistent with Baron Munchausen’s flair for the dramatic, I failed to find any evidence that the Baron availed himself of this life-saving technique.

ACKNOWLEDGMENTS

I thank John Ludbrook (Department of Surgery, The University of Melbourne, Victoria, Australia) and Matthew Strand (National Jewish Health, Denver, CO) for giving helpful comments and suggestions, Sarah Kareem (Department of English, University of California, Los Angeles, CA) for humorizing my questions and for graciously providing a lot of information about Baron Munchausen, and Bernhard Wiebel (Munchausen Library, Zurich, Switzerland) for searching more than 80 English editions of the Baron’s adventures for mention of bootstraps.

REFERENCES


11 It is a long story.


30. Raspe RE. *Baron Munchausen’s Narrative of his Marvellous Travels and Campaigns in Russia*. Oxford: Smith, 1785, p. 45.


