Math notes 2: Bootstrapping

The bootstrap is a method for estimating parameters (and assessing the accuracy of such estimates) without making assumptions about the distribution of the random quantities involved. The basic idea is as follows:

Suppose we want to estimate some parameter (a mean, regression coefficient, or correlation for example) using a data sample $x_1, \ldots, x_n$. We assume that the sample is a random sample of values drawn from some distribution with density f say (or equivalently with distribution function F[1]). We can think of the parameter, which we denote by θ, as being some function of f or F, so we write it as $\theta(F)$ to emphasize this. For example, the sample mean is the function

$$\theta(F) = \int x f(x) dx$$

and the sample median is the value $\theta(F)$ defined by the equation

$$\int_{-\infty}^{\theta(F)} f(u) du = 1/2.$$

We don't know F as we only have the data. However, we can make a good guess about F by using the empirical distribution function (EDF). This is defined as the function that jumps up 1/n at each data point. Suppose the data are labeled in ascending order: $x_1 \le x_2 \cdots x_{n-1} \le x_n$. The formal mathematical definition of the EDF is

$$\hat{F}(x) = \begin{cases} 0, \text{ if } x < x_1, \\ i/n, \text{ if } x_i \le x < x_{i+1},\ i = 1, 2, \ldots, n\text{-}1 \\ 1, \text{ if } x_n \le x. \end{cases}$$

Figure 1 overleaf shows the distribution function for the normal distribution with mean 0 and variance 1, and also the EDF based on a sample of 50 observations.

The basic idea of the bootstrap is simple: we estimate the parameter $\theta(F)$ with $\theta(\hat{F})$. This is the same as estimating the population mean with the sample mean, and the population median with the sample median. Note that $\theta(\hat{F})$ depends on F through a sample drawn from F. An alternative notation that emphasizes this might be $\theta(\hat{F})$ = θ(sample, F).

[1] The distribution function F is related to the density f by the equation $F(x) = \int_{-\infty}^{x} f(u) du$.

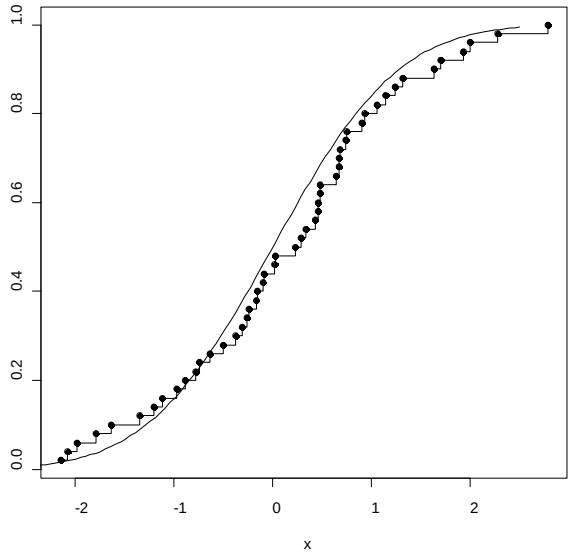


Figure 1. Distribution function and EDF based on 50 observations for the normal distribution.

We can get an idea of the accuracy of the estimate by considering the errors θ(sample, F) - $\theta(F)$. The only problem is that we don't know F. But we can look at the errors $\theta(sample, \hat{F}) - \theta(\hat{F})$, which should be similar since $\hat{F}$ is similar to F. We can generate as many errors $\theta(sample, \hat{F}) - \theta(\hat{F})$ as we like using the following procedure:

1. Draw a sample of size n from the distribution $\hat{F}$. Since this distribution puts probability 1/n at each data point, this is equivalent to sampling n values from $x_1, \ldots, x_n$ **with replacement**. This is called a *bootstrap sample*.
2. Calculate the parameter estimate using the bootstrap sample, and subtract the estimate calculated using the original sample. This produces a typical error $\theta(sample, \hat{F}) - \theta(\hat{F})$.

We can repeat this as many times as we like. A histogram of the resulting values will give us a picture of the error distribution. We can calculate the standard deviation of the values to obtain the standard error of the estimate. We can obtain a confidence interval from quantiles of the values.

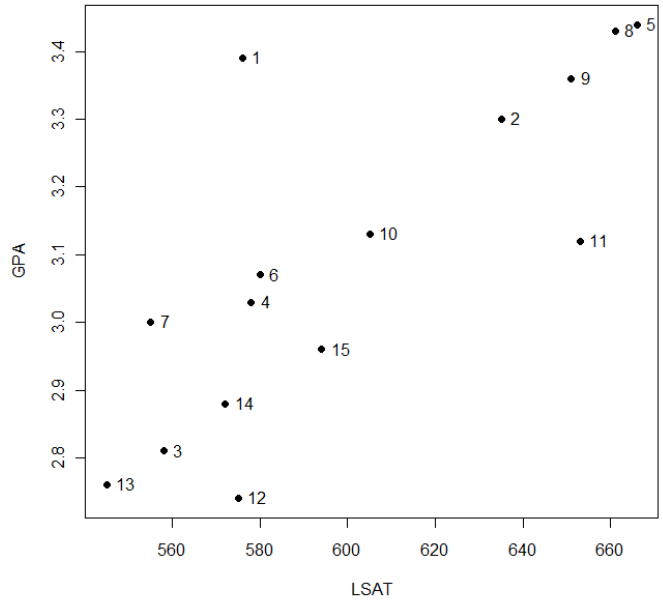
Example

The "law school data"

This famous data set has been used to illustrate the basic ideas of bootstrapping, and consists of 15 observations on US law schools. We have two variables: LSAT, the average score of applicants for admission to the law school on the LAW School Admissions Test, and GPA, the average grade point average of the applicants. The data are

```
   LSAT  GPA
1   576 3.39
2   635 3.30
3   558 2.81
4   578 3.03
5   666 3.44
6   580 3.07
7   555 3.00
8   661 3.43
9   651 3.36
10  605 3.13
11  653 3.12
12  575 2.74
13  545 2.76
14  572 2.88
15  594 2.96
```

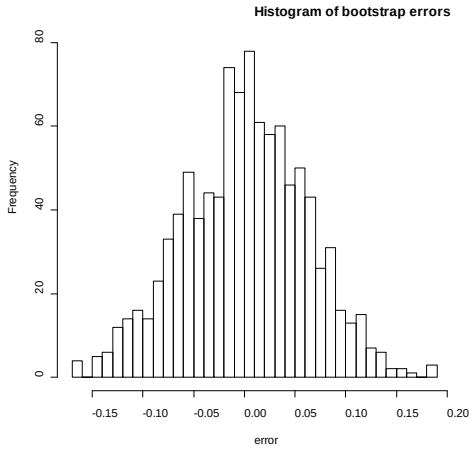
A scatterplot of the data is shown below.



For the moment we concentrate only on the variable GPA and estimate its mean. The mean value of the 15 GPA values is 3. 095. How accurate is this as an estimate of the mean of all law schools? To simulate say N=1000 values of the errors $\theta(sample, \hat{F}) - \theta(\hat{F})$, we can use the following R code:

```
GPA = c(3.39, 3.30, 2.81, 3.03, 3.44, 3.07, 3.00, 3.43,
            3.36, 3.13, 3.12, 2.74, 2.76, 2.88, 2.96)
N = 1000
err = numeric(N)
for(i in 1:N){
 use = sample(15,15, replace=TRUE)
 err[i] = mean(GPA[use]) - mean(GPA)  # GPA[use] is the bootstrap sample
}
```

We can examine the distribution of the errors by means of a histogram:



The standard error of the estimate is the standard deviation of the errors:

```
> sd(err)
[1] 0.06076417
```

We can get a 90% confidence interval for the true mean by noting that 90% of the bootstrap errors lie between the 5th and 95th quantiles of the errors. This leads to the following confidence interval

```
> c( mean(GPA) - quantile(err, 0.95), mean(GPA) - quantile(err, 0.05) )
     95%      5%
2.994667 3.198000
```

This works because we are assuming that the quantiles of the bootstrap errors $\theta(sample, \hat{F}) - \theta(\hat{F})$ are similar to the quantiles of the true errors θ(sample, F) - $\theta(F)$. Thus, we have approximately with 90% probability

quantile(err, 0.05) $\leq \theta$(sample, F) - $\theta(F)$) $\leq$ quantile(err, 0.95)

which implies the confidence interval above.

An alternative way of getting a confidence interval would be to appeal to the central limit theorem and assume the sampling distribution of the mean was normal with mean the true mean and variance σ^2/n. This leads to the usual confidence interval for the mean

$$\overline{x} - \frac{s}{\sqrt{n}} q_{0.95} \leq \mu \leq \overline{x} + \frac{s}{\sqrt{n}} q_{0.95}$$

For the GPA data this gives

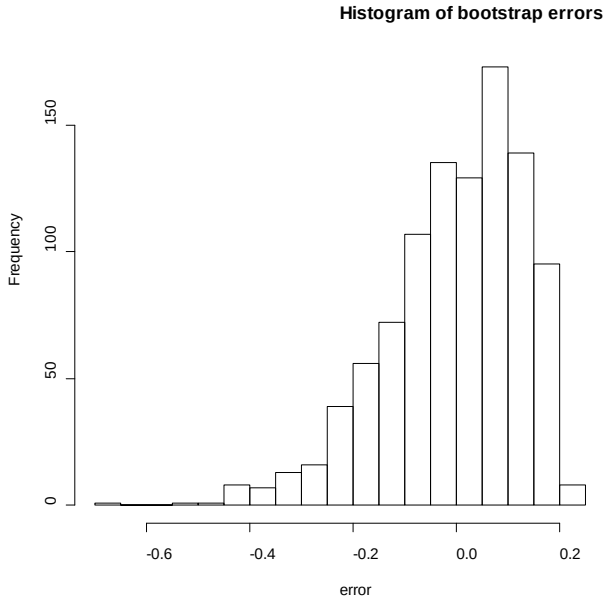
```
> c( mean(GPA) - sd(GPA)*qnorm(0.95)/sqrt(15), mean(GPA) +
sd(GPA)*qnorm(0.95)/sqrt(15) )
[1] 2.991247 3.198086
```

which is very close to the bootstrap version.

In this simple case we have some theory (the central limit theorem) that tells us what the sampling distribution is. The bootstrap really comes into its own when such theory is lacking.

We can use the bootstrap for bivariate data as well. Suppose we want to estimate the correlation between the LSAT scores and the GPA. The sample correlation is 0.776. Once again we can generate bootstrap errors: all we need to do is make a small adjustment to the code:

```
LSAT = c(576, 635, 558, 578, 666, 580, 555, 661, 651, 605,
          653, 575, 545, 572, 594)
N = 1000
err = numeric(N)
for(i in 1:N){
 use = sample(15,15, replace=TRUE)
 err[i] = cor(LSAT[use], GPA[use]) - cor(LSAT,GPA)
}
hist(err, nclass=30, xlab = "error", main = "Histogram of bootstrap errors")
```



Note that the sampling distribution is no longer symmetrical. We calculate a confidence interval for the true correlation in the same way:

```
> c( cor(LSAT,GPA) - quantile(err, 0.95),
          cor(LSAT,GPA) - quantile(err, 0.05) )
     95%       5%
0.604940 1.021548
```

Since the correlation can't be bigger than 1 we would use (0.605 1.000).