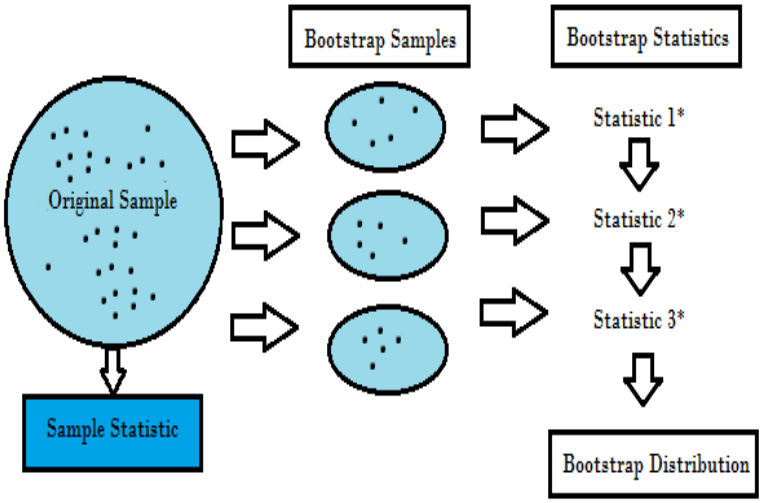
Bootstrapping



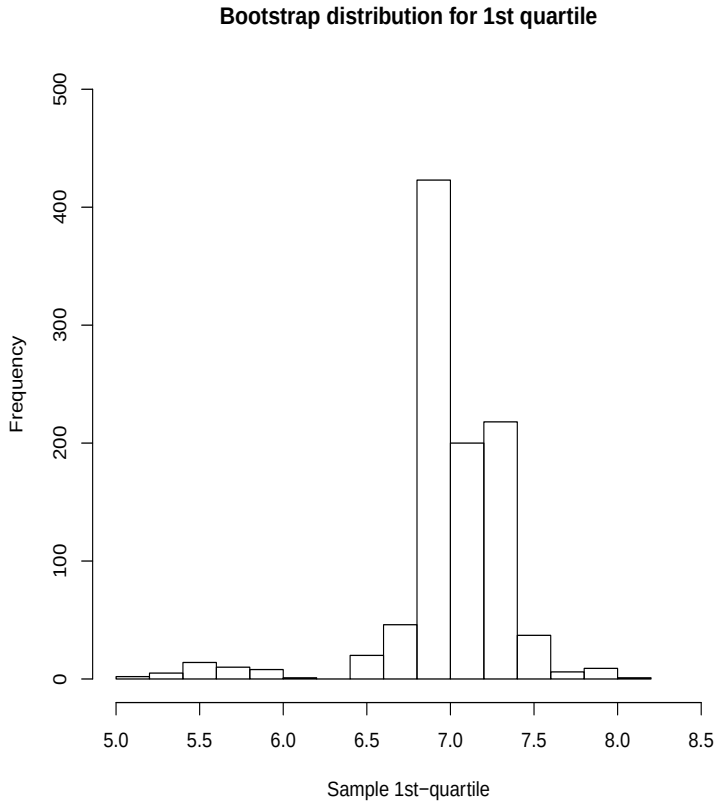
- Bootstrapping is a way of approximating the sampling distribution of estimators.
- It can be used to get standard errors, to construct confidence intervals, and to do hypothesis tests.
- It is particularly useful in situations where analytical results cannot be obtained.
- It is a very flexible technique which makes no assumptions about the form of the underlying distribution.
- Suppose we have the following values, assumed to be a random sample, and wanted to estimate the first quartile Q_1 of the population from which they were sampled.

```
2.83  4.60  5.13  5.54  6.51  6.90
6.91  6.93  7.01  7.22  7.27  7.29
7.41  7.52  7.93  7.96  7.97  8.09
8.20  8.29  8.43  8.63  8.77  9.34
9.36  9.59  9.63  9.93 10.75 12.66
```

- The sample quartile is $\hat{Q}_1 = 6.93$, and this will be our point estimate of the population quartile.
- To get a standard error for the estimate, we create a bootstrap distribution, and calculate the standard deviation of this distribution.
- The bootstrap distribution is formed by
    1. creating a large number (say 1000) new samples of the same size by sampling with replacement from the original sample
    2. calculating the statistic of interest, here the 1st quartile, for each new sample
- For example, one new sample obtained by sampling with replacement is

```
2.83  4.60  4.60  5.54  6.91  6.91
6.93  6.93  7.22  7.27  7.93  7.93
7.96  8.09  8.09  8.20  8.20  8.29
8.29  8.29  8.43  8.63  8.63  9.34
9.34  9.59  9.59  9.93  9.93 12.66
```

- You can see that in this sample, the value 4.60 was chosen twice but the value 6.51 was not selected.
- The value of the 1st sample quartile for this sample is again 6.93.
- If the process is repeated 1000 times, we obtain the bootstrap distribution of the 1st sample quartile, shown in the following histogram.

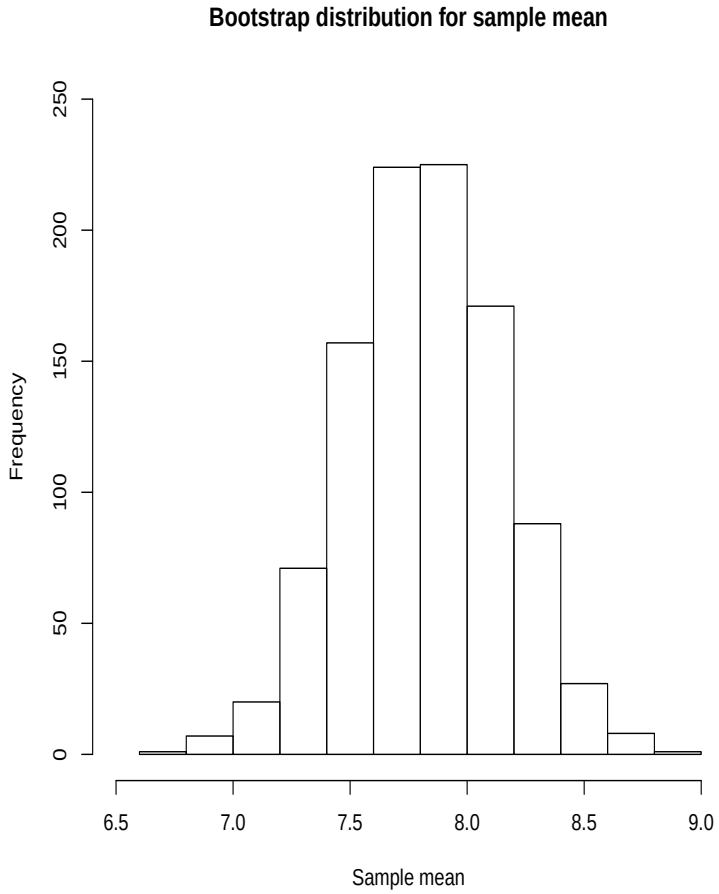


- Each of the 1000 points depicted is the value of the 1st quartile from a bootstrap sample.
- Note that this distribution has a peak near the sample value of 6.93.
- The distribution does not look very normally distributed.
- The standard deviation of this distribution is .3338, and we can use this to quantify the uncertainty in the point estimate, ie as the standard error.
- A 95% confidence interval for the 1st quartile is given by the 2.5th and 97.5th percentiles of this distribution, and is (5.88, 7.44)
- The example above uses an unusual statistic, the first quartile, for which there are no simple analytic results.

- Suppose instead we were interested in estimating the mean μ of the population using the sample mean.
- For the data set on page 1, the sample mean is $\bar{x} = 7.82$.
- The sample variance is $s^2 = 3.5733$, which gives the standard error $s_{\bar{x}} = s/\sqrt{n} = 1.890/5.48 = 0.345$.
- A 95% confidence interval for the mean, using the standard theory is

$$\begin{aligned} 7.82 &\pm 1.96(.345) \\ 7.82 &\pm .676 \\ (7.144 &\quad , \quad 8.496) \end{aligned}$$

- Using the bootstrap, we select 1000 samples with replacement (or we may use the same 1000 samples for getting the 1st-quartile distribution), and obtain the following histogram of sample means (see page 5).
- The standard deviation of this distribution is $s_B = .3382$, which is very close to the standard error obtained above.
- The 2.5th and 97.5th percentiles of the bootstrap distribution give the 95% bootstrap confidence interval (7.133, 8.483).
- Note that the bootstrap distribution for the sample mean is approximately normal, and that the confidence intervals are nearly the same.


Bootstrap distribution for sample mean
Frequency
0
50
100
150
200
250
6.5
7.0
7.5
8.0
8.5
9.0
Sample mean