Interpretation of Bootstrap Statistics

By Craig W. Slinkman

October 15, 2015

This document gives examples of bootstrap analysis to estimate the precision of using a sample statistic to estimate an unknown population parameter when the sample data is based on a *random sample*. We present the bootstrap method as an approach that will allows us to compute confidence intervals so that we can assess the quality of a point estimate. The quality of estimate is a function of both the level of confidence and the width of the confidence intervals.

# Quality of estimate

The quality of estimate is a function of both the level of confidence and the width of the confidence intervals.

If two confidence intervals have the same confidence level say 95%, then the confidence that is narrower has a small margin of error. The confidence with the smaller is a better estimate of the unknown population parameter than the the wider confidence interval because on average the sample mean will be closer to the unknown population mean.

If we have two confidence with the same margin of error but have the same width then the confidence interval with the larger confidence level is a superior interval because on average over the long run it is more likely to be correct.

# Parameters and statistics presented in this document

The table below gives the parameters and parameter expressions that we will present in this document.

|  |  |  |  |
| --- | --- | --- | --- |
| Estimated parameter | Parameter notation | Point estimate | Estimate notation |
| Population mean |  | Sample mean |  |
| Population proportion |  | Sample proportion |  |
| Difference in population means |  | Difference in sample means |  |
| Difference in population proportions |  | Difference in population proportions |  |

# The general bootstrap approach

In the analysis of the four cases presented above we will proceed as follows: 1. Draw a suitable graph of the sample data.

1. Compute the sample statistic. This is our point estimate.
2. Bootstrap the estimate.
3. Assess the bootstrap bias to see if the number of bootstrap replications is sufficient.
4. Assess the shape of the bootstrap sampling distribution.
5. Select the type of confidence interval. This depends upon the results of step 5.
6. Compute the confidence interval.
7. Provide and interpretation of the point estimate, confidence interval, interval interpretation, and assessment of the quality of the estimated interval.
8. Estimation of population mean

## Analysis of sample mean

We now present an analysis of estimating the mean commute time in Atlanta Georgia using the Compute Atlanta data in the R-package **Lock5Data**.  
We start with a plot of the data.

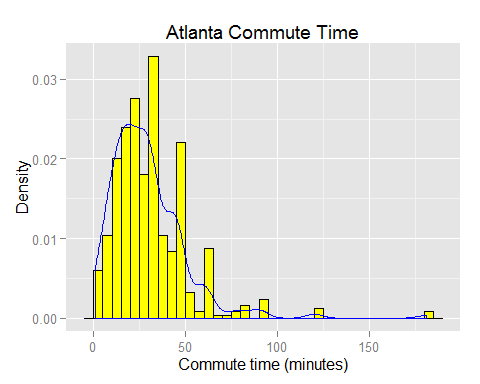
## Histogram and density plot of the sample data

### Graphical Analysis

Out variable of interest is the mean commute time between work and home. When we are analyzing a single quantitative variable we should use a histogram or a density plot or both. We choose to use booth.

## City Age Distance Time Sex  
## 1 Atlanta 19 10 15 M  
## 2 Atlanta 55 45 60 M  
## 3 Atlanta 48 12 45 M  
## 4 Atlanta 45 4 10 F  
## 5 Atlanta 48 15 30 F  
## 6 Atlanta 43 33 60 M

## City Age Distance Time Sex  
## 495 Atlanta 42 26 65 M  
## 496 Atlanta 28 3 5 F  
## 497 Atlanta 28 10 12 F  
## 498 Atlanta 32 18 60 F  
## 499 Atlanta 42 34 34 M  
## 500 Atlanta 40 13 25 F



We see that the Atlanta commute is positively skewed with one or two potential outliers. If we are only describing the data we would use the sample median as our measure of central tendency. However, we want to make inference about the commute times in Atlanta for all commuters.

## Sample statistics

In this section we present what R calls the six-number summary of the commute time variable. This is the same as the traditional five number summary but include the mean as well as the median.

summary( CommuteAtlanta$Time )

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 15.00 25.00 29.11 40.00 181.00

The sample mean is 29.11 minutes per commute. Symbolically we have

## Bootstrapping

In this document I do not show how to use R to compute the bootstrap statistics. If you wish to see how to do this look at the corresponding R-Markdown document **Bootstrap Analysis.RMD**.

## Loading required package: simpleboot  
## Loading required package: boot  
## Simple Bootstrap Routines (1.1-3 2008-04-30)

##   
## Bootstrap sampling distribution analysis   
## Statistic: Sample Mean   
## Replications: 10000   
## Sample size: 500   
## Estimated parameter value: 29.11   
## Bootstrap bais: -0.0060306   
## Standard error: 0.9314336

## Interpretation of the bootstrap output

### Statistic

We are using the statistic the sample mean to estimate the population mean .

### Replications

We created 10,000 sample means based on 10,000 bootstrap samples.

### Estimated parameter value

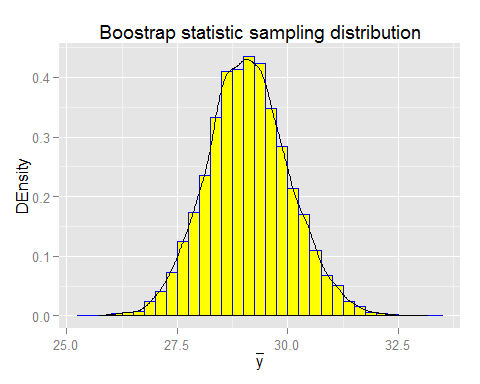
This should be equal to the original sample mean minutes per commute.

### Bias

This bootstrap bias is the difference between the sample mean the mean of all the bootstrap statistics. The magnitude of the bias should a relatively small value when compared to the sample mean. I our example the sample mean . The bootstrap bias is -0.0060306. The magnitude of the bias is very small when compared to the magnitude of the sample estimate.

If the magnitude of the bias is large when compared to the magnitude of sample statistic we would need to increase the number of replications until the magnitude of the bias is small when compared to the magnitude of the sample statistic.

## Bootstrap sampling distribution shape

We now investigate the shape of the bootstrap sampling distribution to determine the type of confidence interval we should use. A histogram of the sampling distribution is shown below.  


The bootstrap sampling distribution of the sample mean appears to be mound shaped.

## Selecting the appropriate confidence interval type

In this class we will only use either the normal confidence intervals or the percentile intervals. I will give you all possible intervals and you must select the appropriate interval. For our sample mean we have the following about.

Because the shape of the bootstrap statistic sampling distribution appears to be normal we will use the normal confidence intervals. Thus, if we wanted a 98% confidence interval we use the the 98% interval

We can compute the margin of error by dividing the with of the confidence interval by 2.

When we write up this analysis we should round these results.

## Interpretation

The point estimate of the mean commute in Atlanta Georgia is 29.1. We are 98% confident that the true but unknowable population mean is between 26.9 and round(bci$norm[3,3],1)`. This is an excellent estimated of mean commute tine because the interval width is very small and the confidence coefficient is high, that is 98%.

An alternative is to give the confidence interval in the form if the

.

The point estimate of the mean commute in Atlanta Georgia is 29.1. We are 98% confident that the true but population mean is in the interval minutes per commute. Because the margin is less than 10% of the value of the estimate of the parameter this is a relatively precise estimate.

# Thats all folks for now!