Introduction to the Bootstrap

Math 445, Spring 2016

One Sample Bootstrap

Motivating example

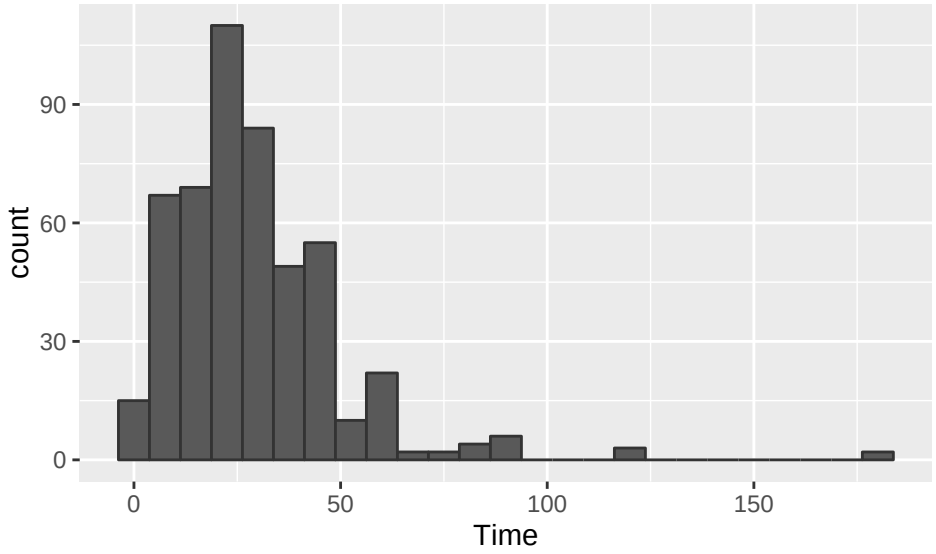
The file `CommuteAtlanta.csv` contains consists of a random sample of 500 people living in Atlanta, and provides information on their commute. The data were obtained from the U.S. Census Bureau's American Housing Survey (AHS), which contains information about housing and living conditions for samples from certain metropolitan areas. This sample includes only individuals that worked somewhere outside of their home.

```
commute <- read.csv("data/CommuteAtlanta.csv")
str(commute)
```

```
## 'data.frame':    500 obs. of  5 variables:
##  $ City    : Factor w/ 1 level "Atlanta": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age     : int  19 55 48 45 48 43 48 41 47 39 ...
##  $ Distance: int  10 45 12 4 15 33 15 4 25 1 ...
##  $ Time    : int  15 60 45 10 30 60 45 10 25 15 ...
##  $ Sex     : Factor w/ 2 levels "F","M": 2 2 2 1 1 2 2 1 2 1 ...
```

Below is a histogram of the commute times:

```
ggplot(data = commute) +
  geom_histogram(mapping = aes(x = Time), colour = "grey20", binwidth = 7.5)
```



Suppose that you are interested in moving to Atlanta and want to better understand the average commute time. It is easy to compute an average commute time from this sample, but what does this tell you about all commuters in Atlanta (i.e. the population)?

```r
# sample mean
mean(commute$Time)
```

```
## [1] 29.11
```

```r
# sample standard deviation
sd(commute$Time)
```

```
## [1] 20.71831
```

Target: a plausible range of values for the average commute time of all commuters in Atlanta, GA.

Algorithm

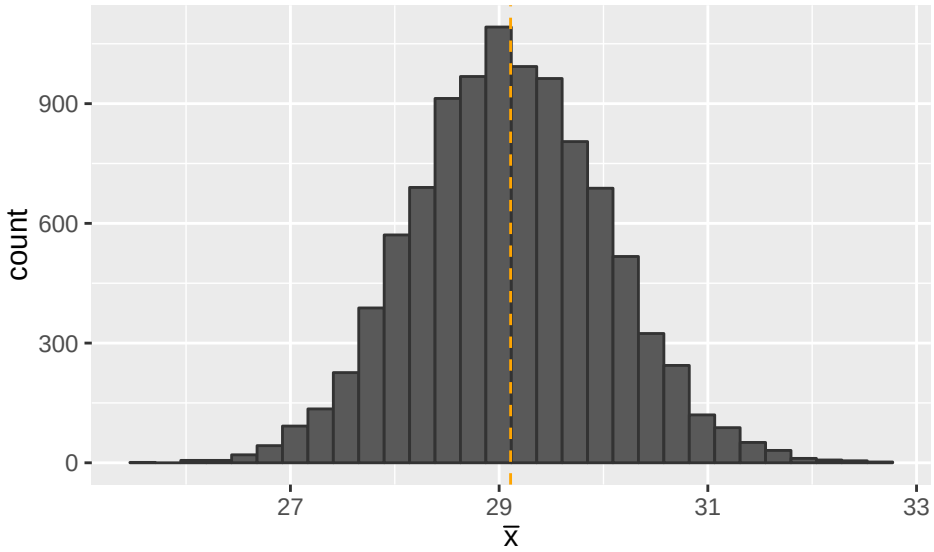
Given a sample of size n from a population,

1. Draw a resample of size n with replacement from the sample. Compute a statistic that describes the sample, such as the sample mean.
2. Repeat this resampling process many times, say 10,000.
3. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape.

Example implementation

The below code snippet boostraps our sample of 500 commuters in Atlanta in an effort to better understand the average commute time.

```r
n <- nrow(commute)
N <- 10^4
time_mean <- numeric(N)
for (i in 1:N) {
  x <- sample(commute$Time, size = n, replace = TRUE)
  time_mean[i] <- mean(x)
}
```



Bootstrap distribution

```r
# bootstrap mean
mean(time_mean)
```

```
## [1] 29.10232
```

```r
# bias
mean(time_mean) - mean(commute$Time)
```

```
## [1] -0.0076842
```

```r
# standard error
sd(time_mean)
```

```
## [1] 0.9198481
```

```r
quantile(time_mean, probs = c(.025, .975))
```

Bootstrap percentile confidence intervals

```
##     2.5%    97.5%
## 27.31995 30.94400
```