Resampling With Bootstrap Methods

eCornell

7/21/2021

Scenario: Use this R Markdown file to help you create bootstrap samples of numerical data. This file outlines two examples: - one in which you want to compare the exam scores of two different course sections, and - one in which you want to test a hypothesis about the median income in Canada based on the Prestige data set in R. You can start by creating bootstrap samples of the median income.

Step 1: Load the prestige data set and create colors.

```
library(carData) # The Prestige data set is available in the carData library

data(Prestige)  # Load the Prestige data set

# Exclude any observations that do not have an entry in the type column:
Prestige = Prestige[!is.na(Prestige$type),]

#eCornell Hex Codes:
crimson = '#b31b1b' #Crimson
lightGray = '#cecece' #lightGray
darkGray = '#606366' #darkGray
skyBlue = '#92b2c4' #skyblue
gold = '#fbb040' #gold
ecBlack = '#393f47' #ecBlack
```

Step 2: Bootstrap resampling on a small example.

Create the example data frame:

```
dat = data.frame(Name = c("Alice", "Bob", "Catie", "Dave", "Eve"), Score = c(10, 20, 30, 40, 50))
dat # View the data frame
```

```
##    Name Score
## 1 Alice    10
## 2   Bob    20
## 3 Catie    30
## 4  Dave    40
## 5   Eve    50
```

```
set.seed(1) #Set the seed for reproducibility
```

Then, to create bootstrap samples from this data set, you can use the `sample()` function with `replace = TRUE`.

```r
# This example data set will sample from the vector 1:5
1:5
```

```
## [1] 1 2 3 4 5
```

```r
# Sample with replacement to create the boot.id vector
boot.id = sample(1:5, size = 5, replace = TRUE)
boot.id
```

```
## [1] 1 4 1 2 5
```

```r
# Use the boot.id vector to create the bootstrap sample from your original data set
dat.boot = dat[boot.id,]
# View the bootstrapped data set dat.boot
dat.boot
```

```
##      Name Score
## 1   Alice    10
## 4    Dave    40
## 1.1 Alice    10
## 2     Bob    20
## 5     Eve    50
```

Step 3: Examine the prestige data set.

First, look at the profession types in the prestige data set.

```r
head(Prestige[Prestige$type == 'bc',])
```

```
##                           education income women prestige census type
## nursing.aides                  9.45   3485 76.14     34.9   3135   bc
## service.station.attendant      9.93   2370  3.69     23.3   5145   bc
## firefighters                   9.47   8895  0.00     43.5   6111   bc
## policemen                     10.93   8891  1.65     51.6   6112   bc
## cooks                          7.74   3116 52.00     29.7   6121   bc
## bartenders                     8.50   3930 15.51     20.2   6123   bc
```

```r
head(Prestige[Prestige$type == 'wc',])
```

```
##                     education income women prestige census type
## medical.technicians     12.79   5180 76.04     67.5   3156   wc
## radio.tv.announcers     12.71   7562 11.15     57.6   3337   wc
## secretaries             11.59   4036 97.51     46.0   4111   wc
## typists                 11.49   3148 95.97     41.9   4113   wc
## bookkeepers             11.32   4348 68.24     49.4   4131   wc
## tellers.cashiers        10.64   2448 91.76     42.3   4133   wc
```

```r
head(Prestige[Prestige$type == 'prof',])
```

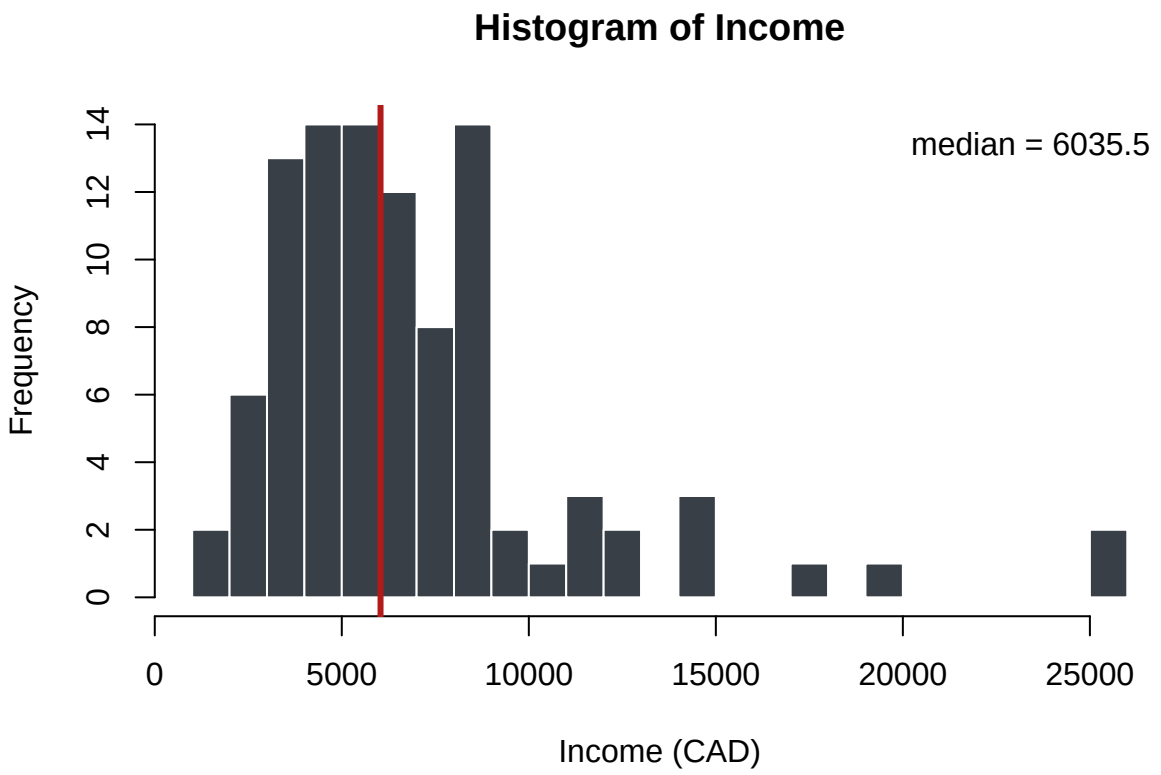
```
##                     education income women prestige census type
## gov.administrators      13.11  12351 11.16     68.8   1113 prof
## general.managers        12.26  25879  4.02     69.1   1130 prof
## accountants             12.77   9271 15.70     63.4   1171 prof
## purchasing.officers     11.42   8865  9.11     56.8   1175 prof
## chemists                14.62   8403 11.68     73.5   2111 prof
## physicists              15.64  11030  5.13     77.6   2113 prof
```

Then, examine the histogram of incomes in this data set. Notice that the histogram is right skewed with some outliers, and has a median of 6035 CAD.

```
# Create a histogram with a vertical line at the observed median income
median(Prestige$income)
```

```
## [1] 6035.5
```

```
hist(Prestige$income, breaks = 20, freq = TRUE, col = ecBlack, border ='white',
     main = 'Histogram of Income', xlab = 'Income (CAD)')
# The abline() function creates a line on a plot. The argument v = x tells R to plot a vertical line at
abline(v=median(Prestige$income), col = crimson, lwd = 3)
# The legend command adds text to the plot.
legend('topright', legend = paste('median =', median(Prestige$income)), bty = 'n')
```



Step 4: Use bootstrap resampling on the prestige data set.

Set the seed for reproducibility.

```
set.seed(1)
```

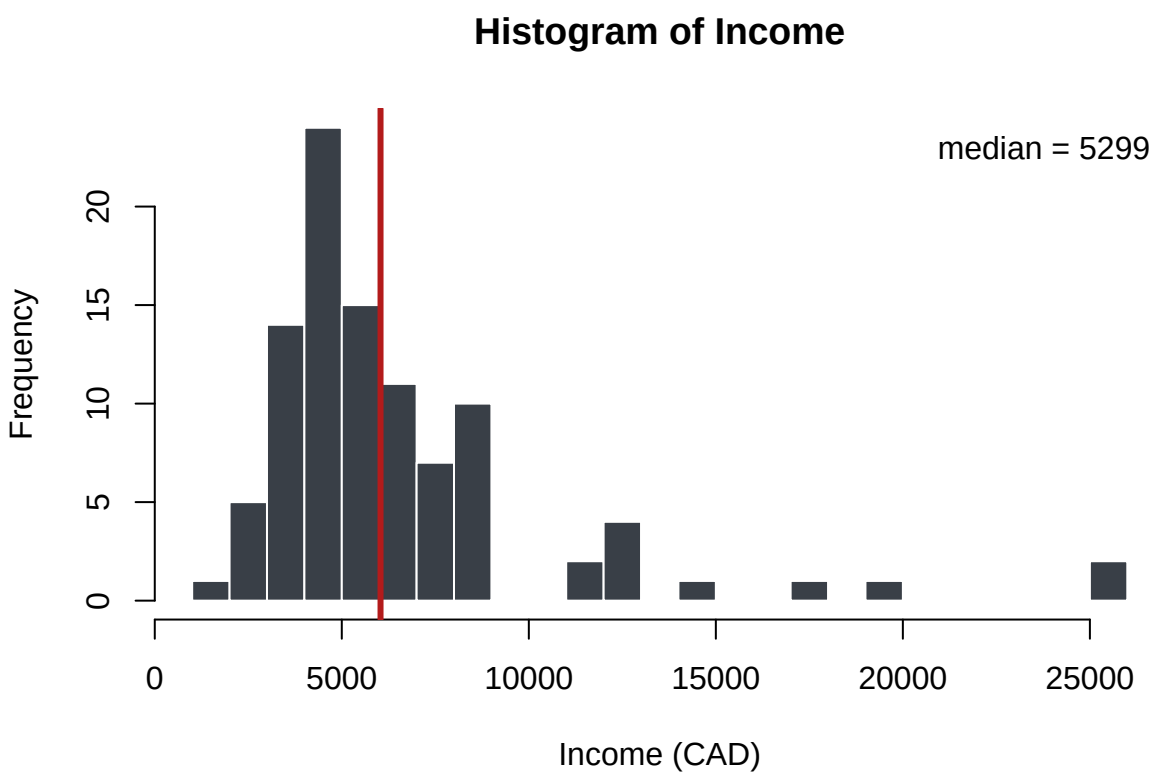
Use `sample()` function to create a bootstrap data set. Run this code chunk several times to see the variation.

```
# Create bootstrap vector, boot.id
boot.id = sample(1:98, size = 98, replace = TRUE)

# Create a new data frame, Prestige.boot, so that each row matches the rows sampled from the bootstrap
Prestige.boot = Prestige[boot.id,]

# Draw histogram of the bootstrapped data with the median plotted on it.
hist(Prestige.boot$income, breaks = 20, freq = TRUE, col = ecBlack,
     border = 'white', main = 'Histogram of Income', xlab = 'Income (CAD)')
```

```r
abline(v=median(Prestige$income), col = crimson, lwd = 3)
legend('topright', legend = paste('median =', median(Prestige.boot$income)), bty = 'n')
```



Notice that: - histograms of median income are changing across different bootstrap samples, and - median incomes change from sample to sample, but all of the bootstrap data sets maintain the overall shape of the histogram: right-skewed with some outliers.