Measuring Uncertainty With Bootstrap Methods

eCornell

7/26/2021

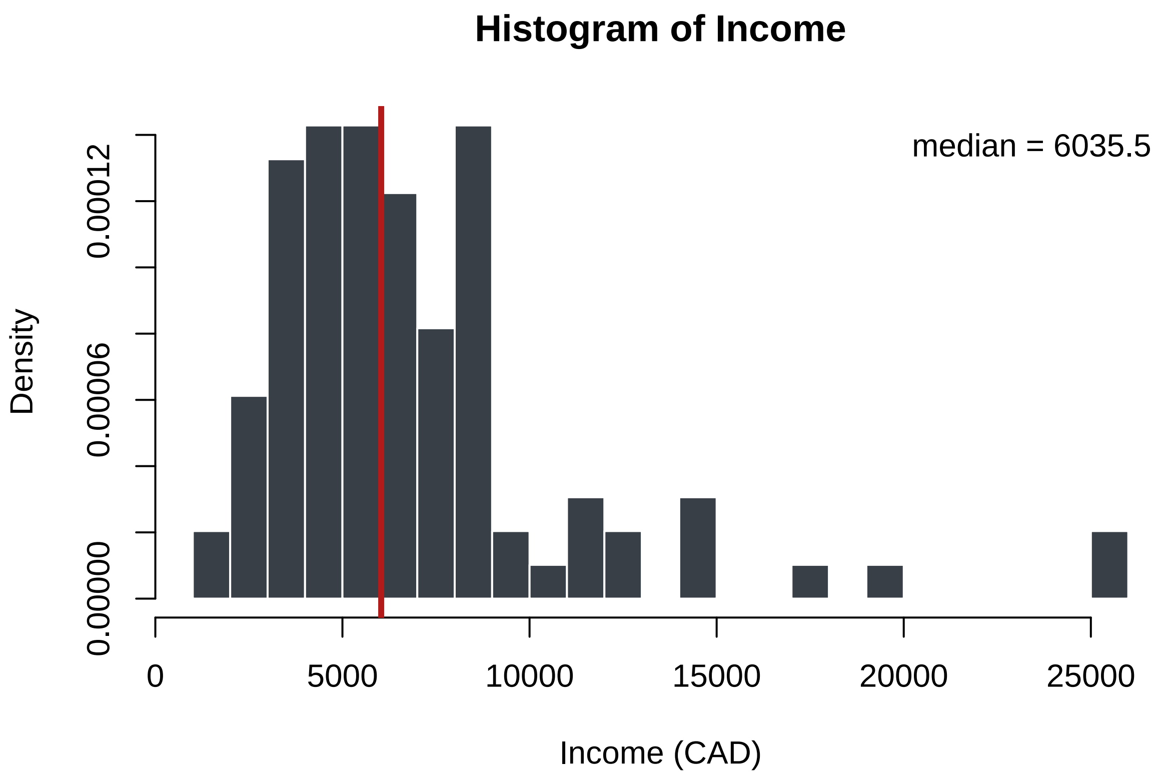
Scenario: Use this R Markdown file to help you create bootstrap samples of numerical data, and use them to create confidence intervals that you can use to test a hypothesis. The example in this file tests whether the median income in Canada in 1971 was 7000 CAD based on the Prestige data set in R.

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

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

# 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'
skyBlue = '#92b2c4' #skyblue
gold = '#fbb040' #gold
ecBlack = '#393f47' #ecBlack

# Create a histogram of the Prestige Data Set:
median(Prestige$income)
```

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

```
hist(Prestige$income, breaks = 20, freq = FALSE, col = ecBlack,
      border ='white', main = 'Histogram of Income', xlab = 'Income (CAD)')
abline(v=median(Prestige$income), col = crimson, lwd = 3)
legend('topright', legend = paste('median =', median(Prestige$income)), bty = 'n')
```



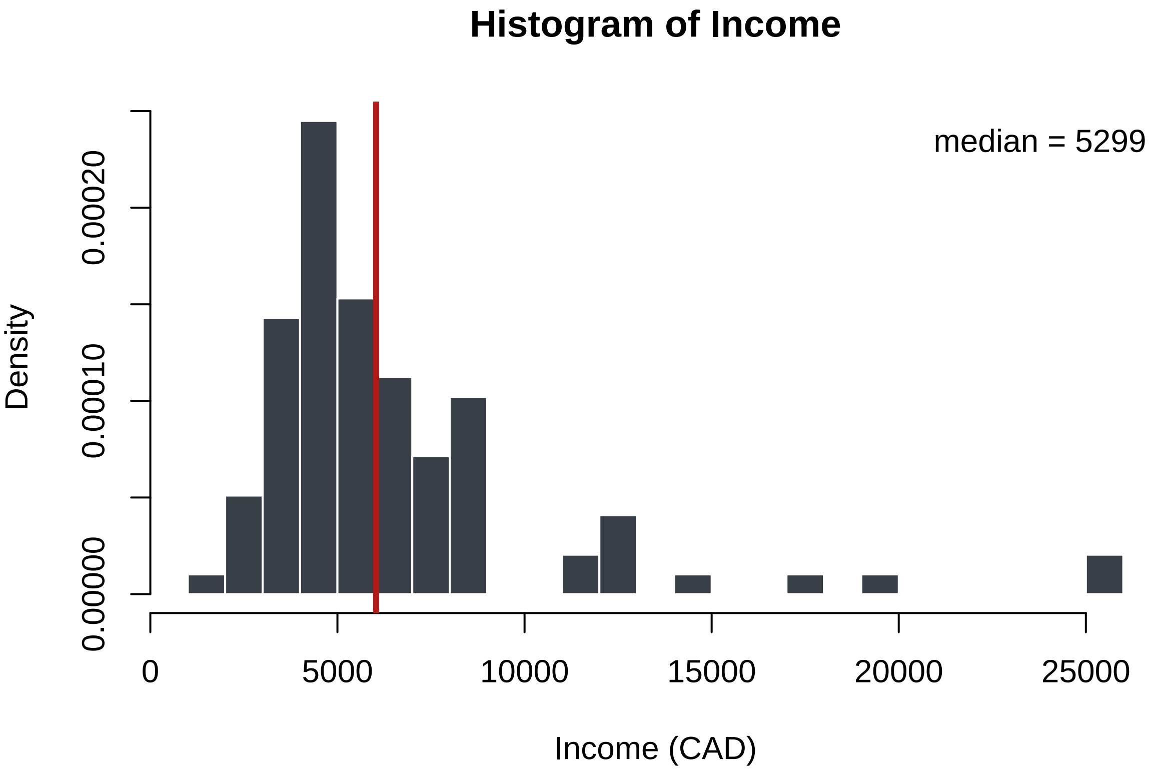
Step 2: View different bootstrap samples.

Run the following code chunk several times, examining how the histogram changes each time.

```r
# Set the seed for reproducibility
set.seed(1)
# Create bootstrap vector, boot.id
boot.id = sample(1:98, size = 98, replace = TRUE)

Prestige.boot = Prestige[boot.id,]

median(Prestige.boot$income)
```

```
## [1] 5299
```

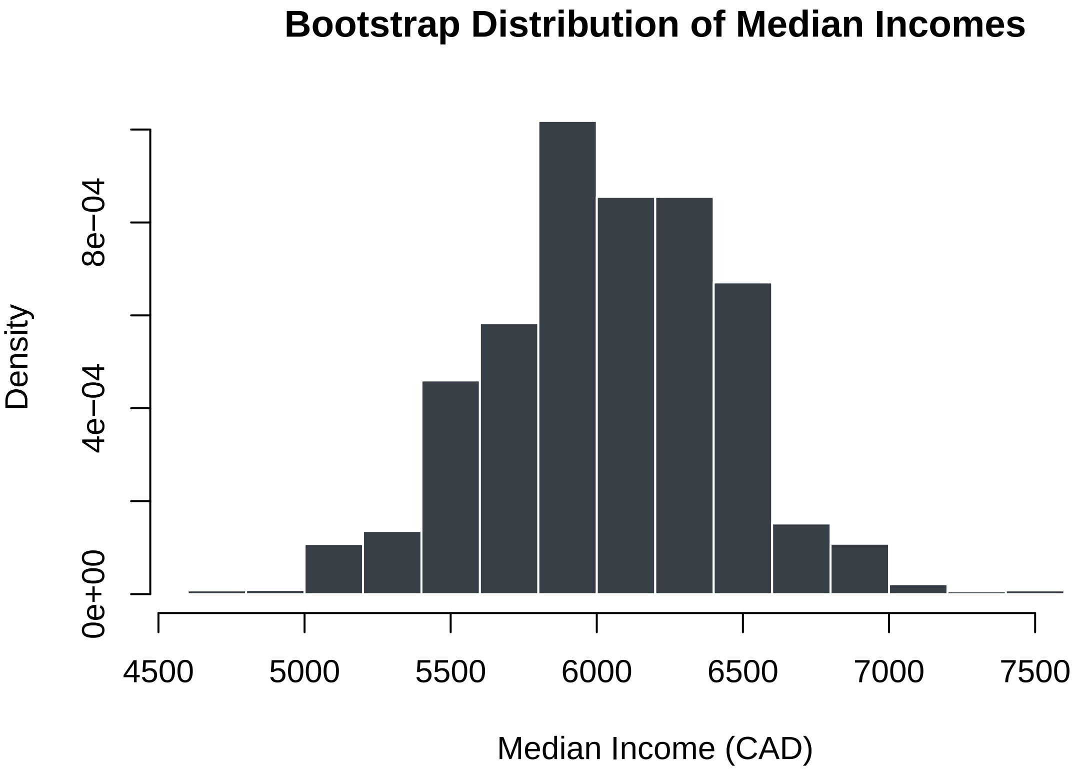
```r
hist(Prestige.boot$income, breaks = 20, freq = FALSE, col = ecBlack,
      border ='white', main = 'Histogram of Income', xlab = 'Income (CAD)')
abline(v=median(Prestige$income), col = crimson, lwd = 3)
legend('topright', legend = paste('median =', median(Prestige.boot$income)), bty = 'n')
```



Step 3: Bootstrap sample and create histogram of medians.

Draw B = 10,000 bootstrap samples, and make a histogram of median incomes. This histogram shows the variability of median incomes from the sample distributions.

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

B = 10000   # Number of bootstrap samples to draw
store_median = rep(0, B) # Vector to store B medians

# For loop to draw B bootstrap samples
for (n in 1:B){
  # Random sample of 98 values
  boot.id = sample(1:98, size = 98, replace = TRUE)
  # Create bootstrap sample from prestige data
  Prestige.boot = Prestige[boot.id,]
  # Store sample median of the boostrapped data in row n of the store_median vector
  store_median[n] = median(Prestige.boot$income)
} # End for loop

# Create a histogram of B median values:
hist(store_median, breaks = 20, freq = FALSE, col = ecBlack,
    border ='white', main = 'Bootstrap Distribution of Median Incomes',
    xlab = 'Median Income (CAD)')
```



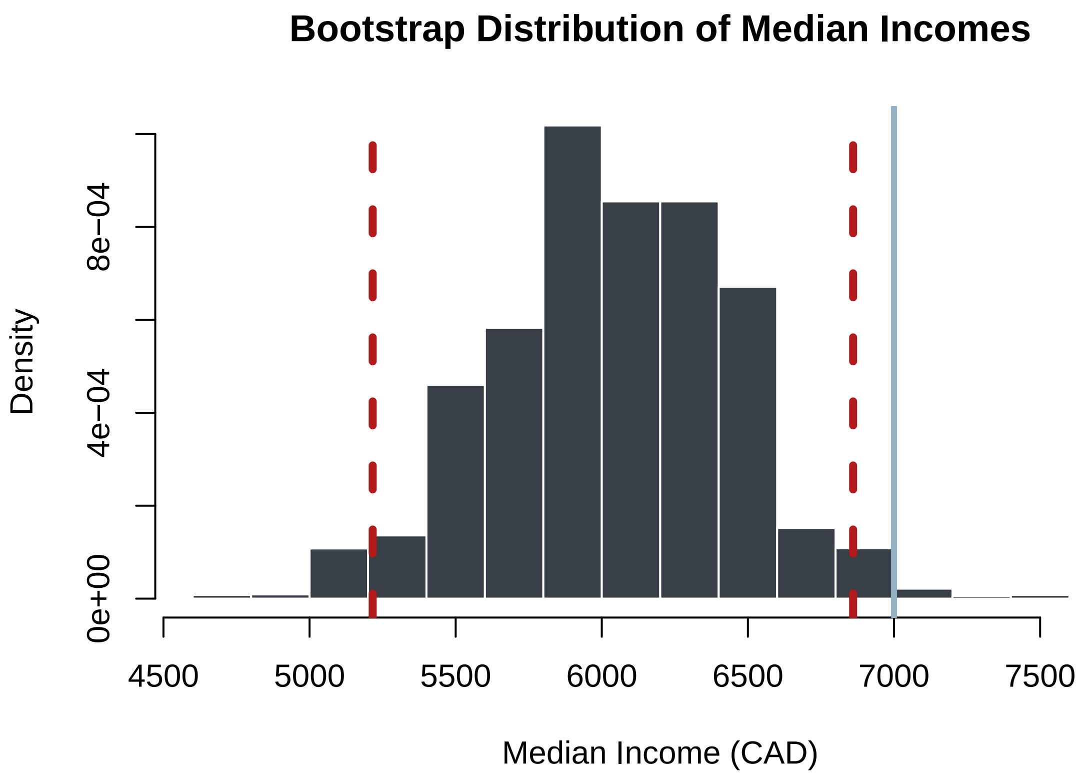
Step 4: Calculate the confidence intervals.

Take the 2.5th and 97.5th percentile of this distribution to form a 95% bootstrap confidence interval. `95% confidence` reflects the fact that when we perturbed the data many, many times and re-calculated the median income, it varied quite a bit but 95% of the time it remained within these two dashed bars.

```
# Calculate the confidence interval:
ci.95 = quantile(store_median, probs = c(0.025, 0.975))
ci.95
```

```
##     2.5%    97.5%
## 5215.975 6860.000
```

```
# Plot the confidence interval on the histogram
hist(store_median, breaks = 20, freq = FALSE, col = ecBlack,
    border ='white', main = 'Bootstrap Distribution of Median Incomes',
    xlab = 'Median Income (CAD)')
abline(v = ci.95, col = crimson, lwd = 4, lty = 2)

# Test if the population median income is 7000 CAD or not.
# Plot a vertical line over the histogram at 7000
abline(v=7000, col = skyBlue, lwd = 3)
```



Step 4: Use the confidence intervals to test a hypothesis.

Test if the population median income is 7000 CAD or not.

Since 7000 CAD is not included in 95% CI, we will conclude that the population median income is NOT 7000 CAD, and we are 95% confident about our conclusion.