Q26

Rahul Atre

2023-11-06

Q26 [Resampling Methods]

We now consider the Boston housing data set.

1. Based on this data set, provide an estimate for the population mean of **medv**. Call this estimate $\hat{\mu}$.

```
bostonData = read.csv("Boston.csv") #Load Boston data

medv_mean = mean(bostonData$medv)
medv_mean
```

```
## [1] 22.53281
```

Ans: The estimate for the population mean of medv is approximately **22.5328**.

2. Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result.

Hint: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.

The standard error for a mean population is defined as follows:

$$SE(\hat{\mu}) = \frac{\sigma_{medv}}{\sqrt{n}}$$

So, using R to calculate this result:

```
medv_meanSE = sd(bostonData$medv) / sqrt(nrow(bostonData))
medv_meanSE
```

```
## [1] 0.4088611
```

Ans: The estimate of standard error of $\hat{\mu}$ is approximately **0.40886**.

3. Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from 2.?

```r
library(boot) #Load bootstrap and cross-validation functions into R

set.seed(50) #Set seed for reproducibility
boot.func <- function(data,index) {
  mean(data[index]) #Assign measure of accuracy for mean to estimate sample
}
boot(bostonData$medv,boot.func,1000) #Perform 1000 samples
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = bostonData$medv, statistic = boot.func, R = 1000)
##
##
## Bootstrap Statistics :
##     original       bias    std. error
## t1* 22.53281 -0.001728063   0.4100517
```

Using the bootstrap, the estimate for the mean standard error is **0.4100517**. This answer is extremely close to the one obtained in 2, at less than a 0.005 difference.

4. Based on your bootstrap estimate from 3., provide a 95% confidence interval for the mean of **medv** Compare it to the results obtained using **t.test(Boston$medv)**.

Hint: You can approximate a 95% confidence interval using the formula $[\hat{\mu} - 2SE(\hat{\mu}), \hat{\mu} + 2SE(\hat{\mu})]$.

We know from 3 that $SE(\hat{\mu}) = 0.4100517$. So:

```r
confidence_meanMedv = c(medv_mean - 2*0.4100517, medv_mean + 2*0.4100517)
confidence_meanMedv
```

```
## [1] 21.71270 23.35291
```

The 95% CI for the mean of medv is **(21.71270, 23.35291)**.

The t-test can be calculated as follows:

```r
t.test(bostonData$medv)
```

```
##
##  One Sample t-test
##
## data:  bostonData$medv
## t = 55.111, df = 505, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  21.72953 23.33608
## sample estimates:
## mean of x
##  22.53281
```

The 95% CI using the t-test is **(21.72953, 23.33608)**. Comparing the two values, we can see that both CI's are within one decimal point, thus showing that the bootstrap method provides a very good estimate.

5. Based on this dataset, provide an estimate, $\hat{\mu}_{med}$ for the median value of **medv** in the population.

```
median(bostonData$medv)
```

```
## [1] 21.2
```

From the above function call, the estimate for the median value of medv is **21.2**.

6. We now would like to estimate the standard error of $\hat{\mu}_{med}$. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.

The method of estimating the standard error of $\hat{\mu}_{med}$ using bootstrap will be similar to that of the mean:

```
set.seed(50) #Set seed for reproducibility
boot.func <- function(data,index) {
  median(data[index]) #Assign measure of accuracy for mean to estimate sample
}
boot(bostonData$medv, boot.func, 1000) #Perform 1000 samples
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = bostonData$medv, statistic = boot.func, R = 1000)
##
##
## Bootstrap Statistics :
##     original  bias    std. error
## t1*     21.2  -0.015   0.3997872
```

Using the bootstrap, the estimate for the median standard error is **0.3997872**. Similar to the mean, the standard error is quite small and relatively close to one another.

7. Based on this data set, provide an estimate for the tenth percentile of **medv** in Boston suburbs. Call this quantity $\hat{\mu}_{0.1}$. (You can use the **quantile()** function.)

```
quantile(bostonData$medv, 0.1)
```

```
##   10%
## 12.75
```

The estimate for the tenth percentile of medv in the Boston suburbs is **12.75**.

8 Use the bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings.

```r
set.seed(50) #Set seed for reproducibility
boot.func <- function(data,index) {
  quantile(data[index], 0.1) #Assign measure of accuracy to estimate quantile
}
boot(bostonData$medv, boot.func, 1000) #Perform 1000 samples
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = bostonData$medv, statistic = boot.func, R = 1000)
##
##
## Bootstrap Statistics :
##     original  bias    std. error
## t1*    12.75  0.0158   0.5028849
```

From the above bootstrap function call, the estimated standard error is given as **0.5028849**. This can be considered small value, though slightly higher than the median and mean values that were estimated using the bootstrap method.