Assignment 7

Sean Leggett - BDA201 Winter 2020 March 21, 2020

Question

There is a cleaning company that offers cleaning services to other businesses. The company contracts to clean office space in increments of 100 square feet. The business determines its margins by determining how long it takes a crew to clean 100 square feet of office space and bases its rates on this information.

1. Because the company is relatively new, it has to estimate the time it takes to clean a 100 square feet of office space. The company estimates that it should take 5.5 hours to clean 100 square feet (m = 5.5). The company starts its business with this expectation and works for a week straight, collecting data as it proceeds in order to be certain that it is neither over- nor under-charging its clients. The data collected by the company can be seen in the table below. House per 100 square feet 5.5 6.9 5.8 5.5 4.5 4.6 7.8 6.4 5.6 7.1 4 4.5
2. If the cleaning company had a sample of seven rather than a sample of 12 upon which to base its conclusions, what would be the boundaries of the 98 percent confidence interval for the estimate of the number of hours? Assume that the sample mean and standard deviation are equal to those calculated above.
3. What happens to the confidence bounds if the company wants to increase its confidence interval to 99%. Explain the change.

Answer

We considered analysis outside the perspective of normal distribution to be outside the scope of this assignment.

Let's begin by assessing mean and standard deviation of the data collected:

```
collection <- data.frame("Hours" = c(5.5, 6.9, 5.8, 5.5, 4.5, 4.6, 7.8, 6.4, 5.6, 7.4, 4, 4.5))
collmean <- mean(collection$Hours)
collsd <- sd(collection$Hours)
collmean
```

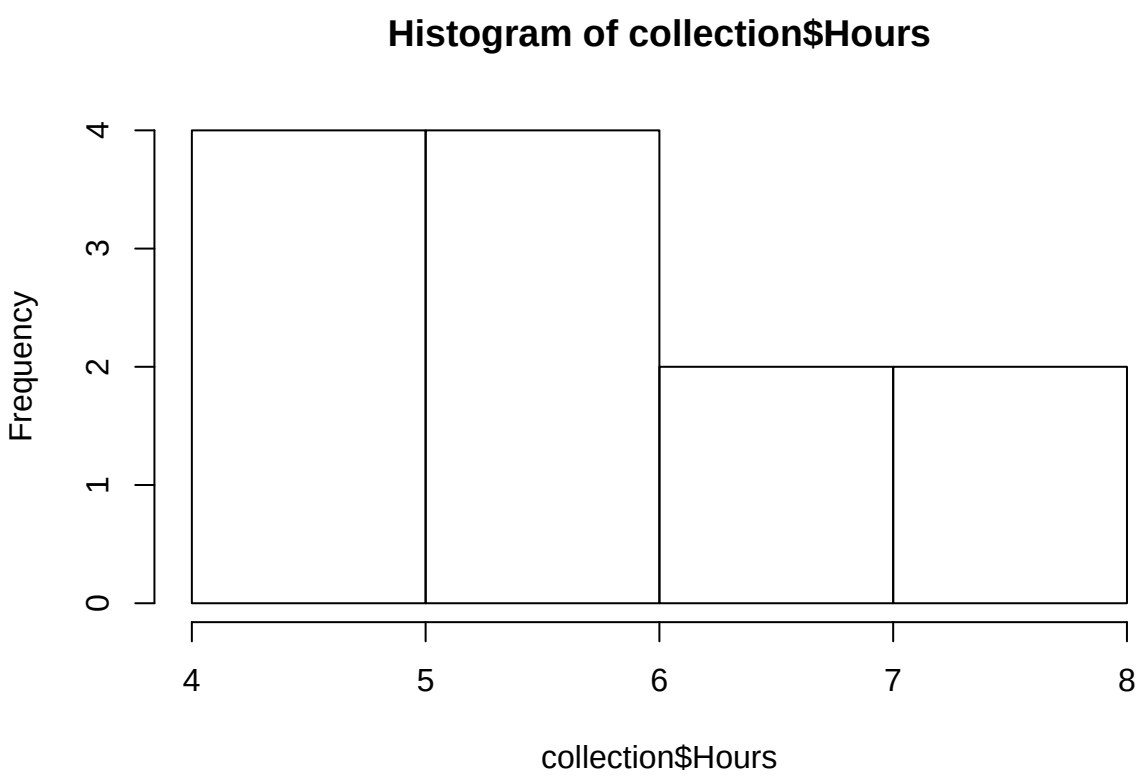
```
## [1] 5.708333
```

```
collsd
```

```
## [1] 1.217642
```

We would prefer to visual confirm standar deviation...

```
collhist <- hist(collection$Hours)
```




```
collhist
```

```
## $breaks
## [1] 4 5 6 7 8
##
## $counts
## [1] 4 4 2 2
##
## $density
## [1] 0.3333333 0.3333333 0.1666667 0.1666667
##
## $mids
## [1] 4.5 5.5 6.5 7.5
##
## $xname
## [1] "collection$Hours"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

We do not find this to be a comforting distribution. Outliers certainly affect the variable distribution. In this case, it could be down to different types of spaces. 100sqft of doctor's office is different to 100sqft of lawyer's office. However, we accept the data as reported.

Firstly, we would like to establish the boundaries around existing sample of 12 for comparison.

```
x <- 5.5
z_score <- qnorm(.990) ##qnorm of ..990 for 98% confidence
n <- 12 ## number of results in sample
s <- 1.2 ## standard deviation of results

lower_bound <- x - z_score * s/sqrt(n)
lower_bound
```

```
## [1] 4.694129
```

```
upper_bound <- x + z_score * s/sqrt(n)
upper_bound
```

```
## [1] 6.305871
```

Which means, with the 12 result sample, 98% of the time, the hours required to clean 100 sq ft of office space would be between 4.7 hours and 6.3 hours. An interval of 1.6 hours. Not particularly narrow.

Now, let's repeat the exercise using 7 results to establish the confidence interval, assuming that standard deviation and sample mean have not changed...

```
x <- 5.5
z_score <- qnorm(.990) ##qnorm of ..990 for 98% confidence
n <- 7 ## number of results in sample
s <- 1.2 ## standard deviation of results

lower_bound <- x - z_score * s/sqrt(n)
lower_bound
```

```
## [1] 4.444868
```

```
upper_bound <- x + z_score * s/sqrt(n)
upper_bound
```

```
## [1] 6.555132
```

Using 7 data points from the sample the results are 98% of the time, it will take between 4.4 and 6.6 hours to clean 100 square feet of office space. An interval even wider than with 12 points of data considered. Clearly, up to perhaps an unknown quantity, more data points in the sample yield a narrower confidence interval.

For Question 2)

Increasing the confidence interval to 99% we find the following:

```
x <- 5.5
z_score <- qnorm(.995) ##qnorm of ..995 for 99% confidence
n <- 7 ## number of results in sample
s <- 1.2 ## standard deviation of results

lower_bound <- x - z_score * s/sqrt(n)
lower_bound
```

```
## [1] 4.331714
```

```
upper_bound <- x + z_score * s/sqrt(n)
upper_bound
```

```
## [1] 6.668286
```

Therefore, 99% of the time, it will take between 4.3 hours and 6.7 hours to clean 100 squre feet of office space. A confidence interval of 2.4 hours which is a substanial range. With limited sample space and with such a high requirement for the confidence interval (99%) we are not surprised to see quite a broad range in the actual range in the confidence interval.