Ole J. Forsberg

Practicum 3: October Customers

STAT 200: Introductory Statistics

November 19, 2017

The research question asks us to estimate the number of customers on November 20, 2017. The entire data set consists of the number of customers each day for the 628 days between July 1, 2015, and March 21, 2017. To answer the question, we must first select the appropriate sample of the data. Because Stillwater is a college town, one should pay attention to seasons — both semesters and sports. To perform a better analysis, I analyze the data using two different subsets.

The first analysis uses the 30 days of data around November 20. That is, I use the data between November 5 and December 5. This gives us three one-month windows on what we are to estimate. The second analysis uses all Mondays, since November 20 is a Monday. This gives us a much larger sample to work with, thus allowing us to create a more-precise confidence interval.

Comparing and contrasting the two confidence intervals will allow us to increase the quality of the final answers. This is how to test the representativeness of the sample. It is *imperative* that the sample be representative of the target population. I perform the same analysis on this different subset to see if there is a major difference in predictions.

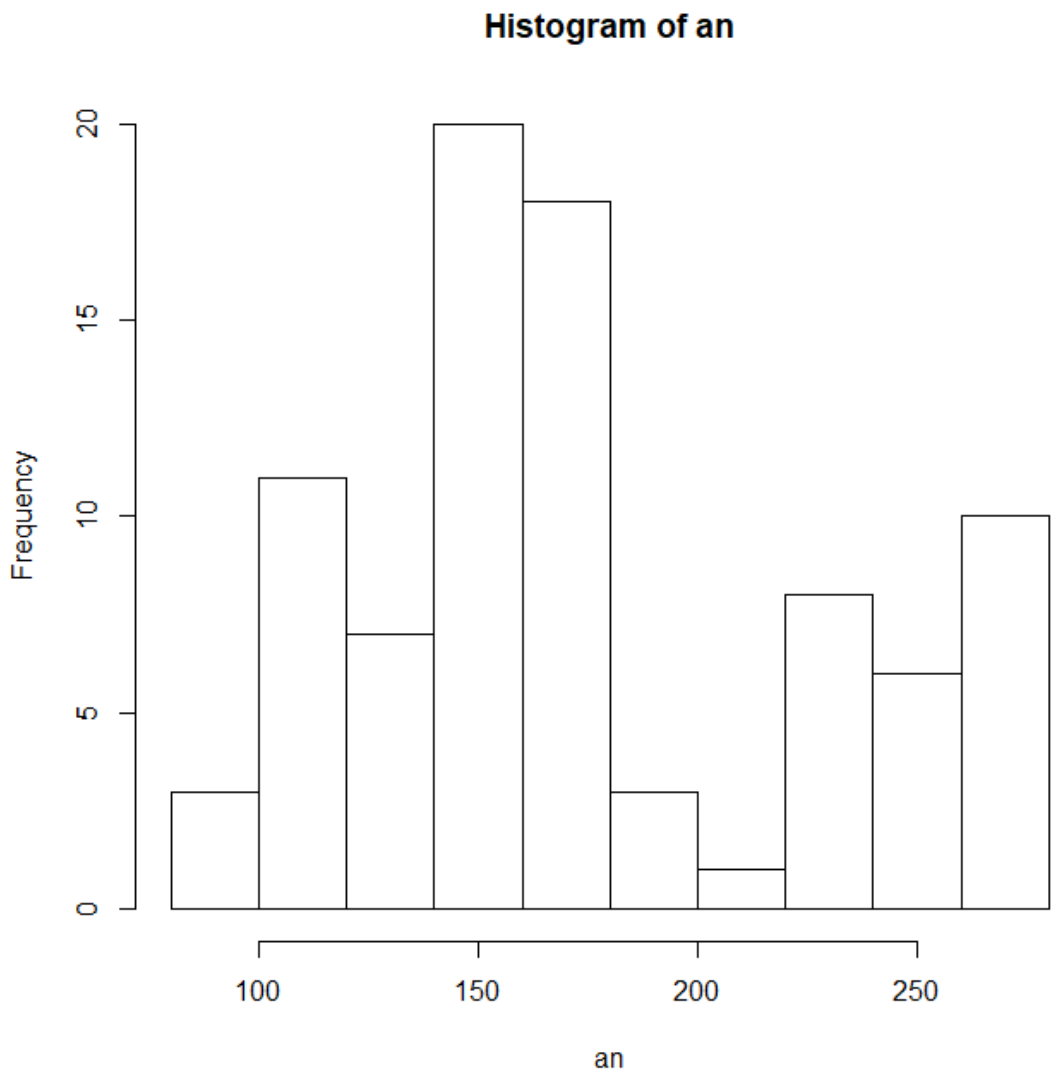
Analysis 1: Around November 20

This analysis examines all data within 15 days of November 20. The sample statistics for these n = 87 days are

Mean	176.3
Median	162.0
Standard deviation	52.91
Hildebrand ratio	0.271

According to the Shapiro-Wilk test, the data are not Normal (p-value < 0.0001). According to the Hildebrand rule, they are not symmetric (H = 0.271). With these results, one should use the non-parametric bootstrap. According to this procedure, we are 95% confident that the *expected* number of customers on November 20, 2017, to be between 165.4 and 187.8, with a best guess (point estimate) of 176.3.

The following is a histogram of the number of customers who dined at the Lamplighter restaurant for the period of interest. Note that it appears to be rather skewed.



Analysis 2: Mondays

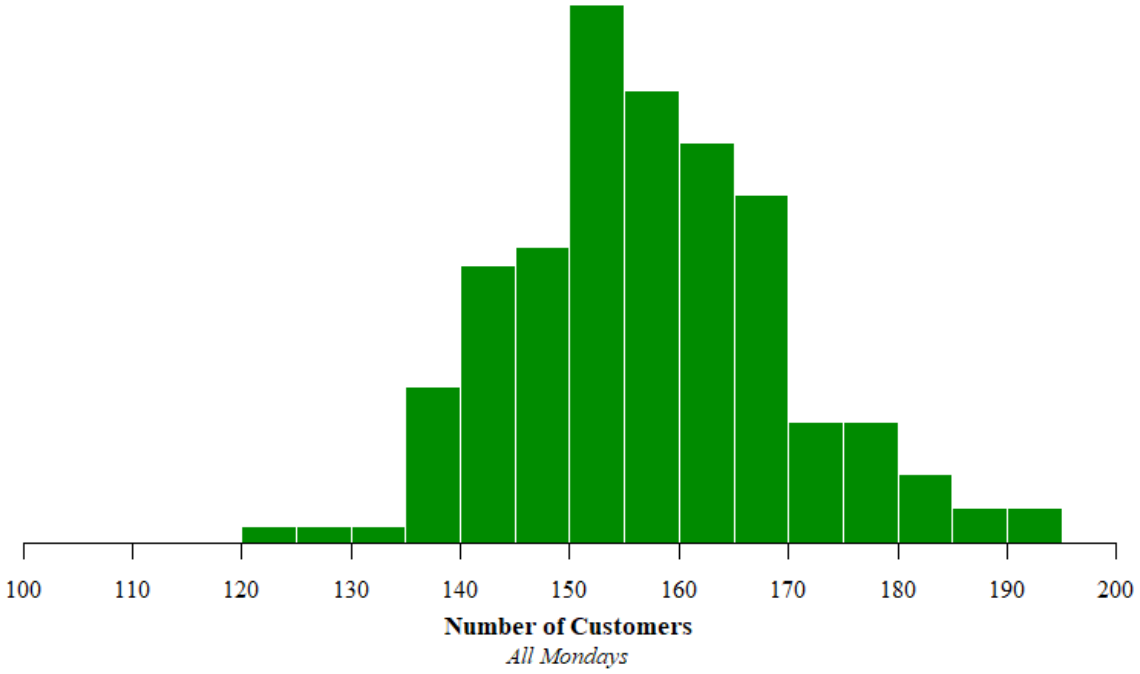
This analysis examines all Mondays. The sample statistics for these n = 167 days are

mean	157.7
median	157.0
standard deviation	12.57
Hildebrand ratio	0.052

According to the Shapiro-Wilk test, the data are sufficiently Normal (p-value = 0.4930). Thus, I shall use the one-sample t-procedure. According to this procedure, we are 95%

confident that the *expected* number of customers on November 20, 2017, to be between 155.7 and 159.6, with a best guess (point estimate) of 157.7.

The following is a histogram of the number of customers who dined at the Lamplighter restaurant on Mondays. Note that, compared to the histogram above, this is much narrower.



Conclusion

In this analysis, we calculated the expected number of customers on November 20, 2017, at the Lamplighter Restaurant. Focusing only on the 30 days around November 20, we are 95% confident that the expected number of customers will be between 165.4 and 187.8, with a best guess (point estimate) of 176.3. Using the Mondays, we are 95% confident that the expected number of customers will be between 155.7 and 159.6, with a best guess (point estimate) of 157.7.

As this first interval is based on a sample that is probably more representative than the second, it produces results that I find more trustworthy. In the future, I would suggest focusing only on Mondays during the Autumn term. This subset takes into consideration both the annual nature of restaurants, especially football season, and their weekly nature.

Appendix: R Script

```
##### Practicum 3 Example
#####

### Preamble
source("http://rfs.kvasaheim.com/stat200.R")

dt = read.csv("lamplighterSales1803.csv")
summary(dt)
attach(dt)

aroundNov20 = c( which(month==11 & day>=5), which(month==12 & day<=5) )
allMondays  = seq(2,1169,7)

an = dt[aroundNov20, ]
nm = dt[allMondays,  ]


### Dates around November 20
length(an$customers)
summary(an$customers)
sd(an$customers)

# Test assumptions
shapiro.test(an$customers)
hildebrand.rule(an$customers)

# Non-parametric bootstrap
mn = numeric()
for(i in 1:1e4) {
  x = sample(an$customers, replace=TRUE)
  mn[i] = mean(x)
}
mean(an$customers)
quantile(mn, c(0.025,0.975))

# Graphic
hist(an$customers, xlab="Customers")


### Mondays
length(nm$customers)
summary(nm$customers)
sd(nm$customers)

# Test assumptions
shapiro.test(nm$customers)
hildebrand.rule(nm$customers)
t.test(nm$customers)

# Graphic
hist(nm$customers, xlab="Customers")
title(sub="All Mondays", line=3.25)
```