Hw4

Amy Fox

9/24/2019

```
library(dplyr)
library(BSDA)
library(boot)
library(tidyr)
```

Question 1

The housing department in a large city wants to estimate the average rent for rent-controlled apartments in the city. They need to determine the number of renters to include in a survey in order to estimate the average rent to within $80 using a 95% confidence interval. From past surveys, the monthly charge for rent-controlled apartments ranged from $1600-$3200.

A. Suppose that based on the previous survey, almost all (>99%) apartment rents fell within $1600-$3200. Use this information to "estimate" the standard deviation.

The Empirical rule states that for normal distributions, >99% of the values will fall within the 3 standard deviations of the mean. Therefore, standard deviation = Range/6

```
(3200-1600)/6
```

```
## [1] 266.6667
```

The estimated standard devation is $266.67.

B. Using the standard deviation from above, find the (minimum) sample size required to achieve a 95% ME < $80.

One-sample t-test

```
n <- seq(25, 50, by = 1)
sd <- (3200-1600)/6
alpha <- 0.05
ME <- qt(1-alpha/2, df = n-1)*sd/sqrt(n)
data.frame(n, ME)
```

```
##     n        ME
## 1  25 110.07459
## 2  26 107.70900
## 3  27 105.48982
## 4  28 103.40254
## 5  29 101.43458
## 6  30  99.57497
## 7  31  97.81412
## 8  32  96.14359
## 9  33  94.55589
## 10 34  93.04440
## 11 35  91.60320
## 12 36  90.22702
## 13 37  88.91109
## 14 38  87.65115
## 15 39  86.44333
## 16 40  85.28414
## 17 41  84.17039
## 18 42  83.09918
## 19 43  82.06789
## 20 44  81.07408
## 21 45  80.11554
## 22 46  79.19022
## 23 47  78.29627
## 24 48  77.43193
## 25 49  76.59561
## 26 50  75.78583
```

Therefore, a sample size of at least 46 must be used.

Question 2

A national agency sets recommended daily allowances for many supplements. In particular, the allowance for zinc for adult men is 15 mg/day. The agency would like to determine if the average intake of zinc for adult men is greater than 15 mg/day. Suppose from a previous study they estimate the standard deviation to be 2 mg/day and they conjecture that the true population mean is 15.4 mg/day. The investigators plan to use a one-sample t-test with α =0.05.

** A. Find the power with n=120 for the scenario above.**

```
power.t.test(n = 120, delta = 15.4-15, sd = 2, sig.level = 0.05, type = "one.sample", alternative = "one
```

```
##
##      One-sample t test power calculation
##
##               n = 120
##           delta = 0.4
##              sd = 2
##       sig.level = 0.05
##           power = 0.703175
##     alternative = one.sided
```

The power is 0.703.

B. If the standard deviation was smaller (less than 2) would the power be higher or lower than that calculated in part A?

The power would be higher.

C. If the sample size was larger (more than 120) would the power be higher or lower than that calculated in part A?

The power would be higher.

D. If we used α =0.01 (instead of 0.05), would the power be higher or lower than that calculated in part A?

The power would be lower.

E. Using a conjectured mean of 16 mg/day (instead of 15.4), would the power be higher or lower than that calculated in part A?

The power would be higher.

F. Return to the original scenario and find the sample size required to achieve 80% power. Remember to “round” up to an integer value.

```
power.t.test(power = 0.8, delta = 15.4-15, sd = 2, sig.level = 0.05, type = "one.sample", alternative =
```

```
##
##      One-sample t test power calculation
##
##               n = 155.9257
##           delta = 0.4
##              sd = 2
##       sig.level = 0.05
##           power = 0.8
##     alternative = one.sided
```

The sample size for 80% power should be 156.

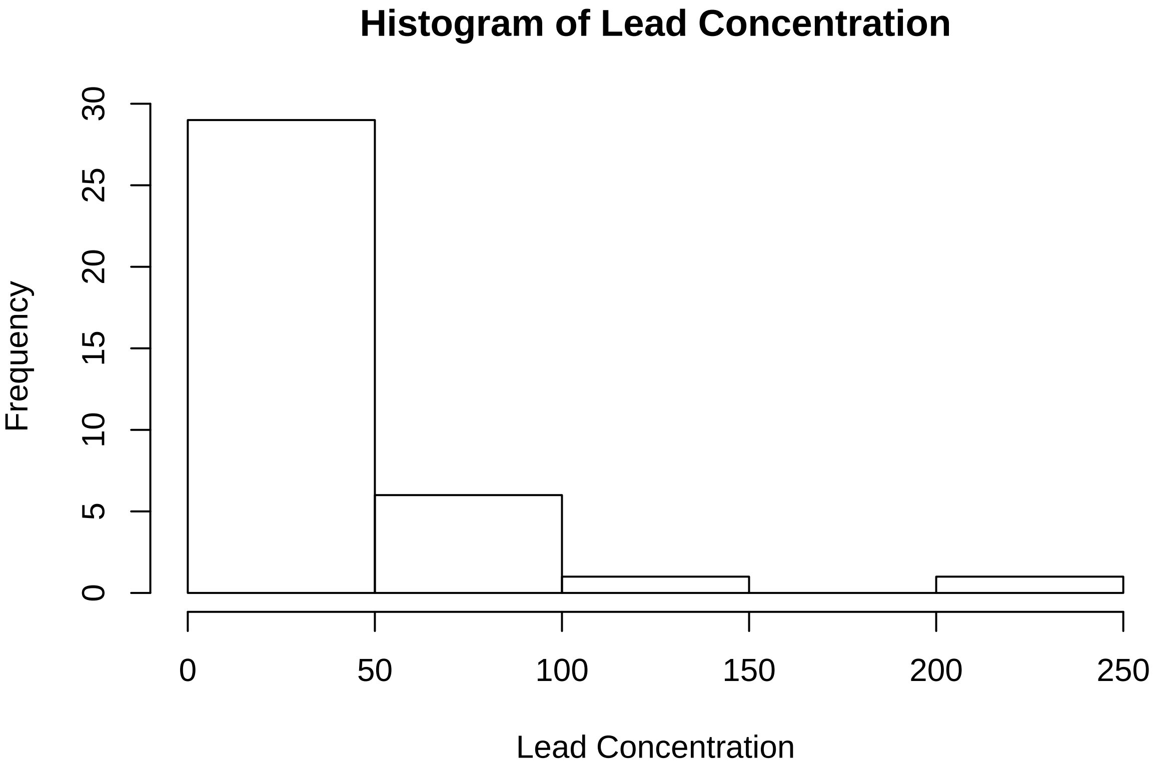
Question 3

Use the data from Problem 5.27 which deals with lead concentrations in estuarine creeks.

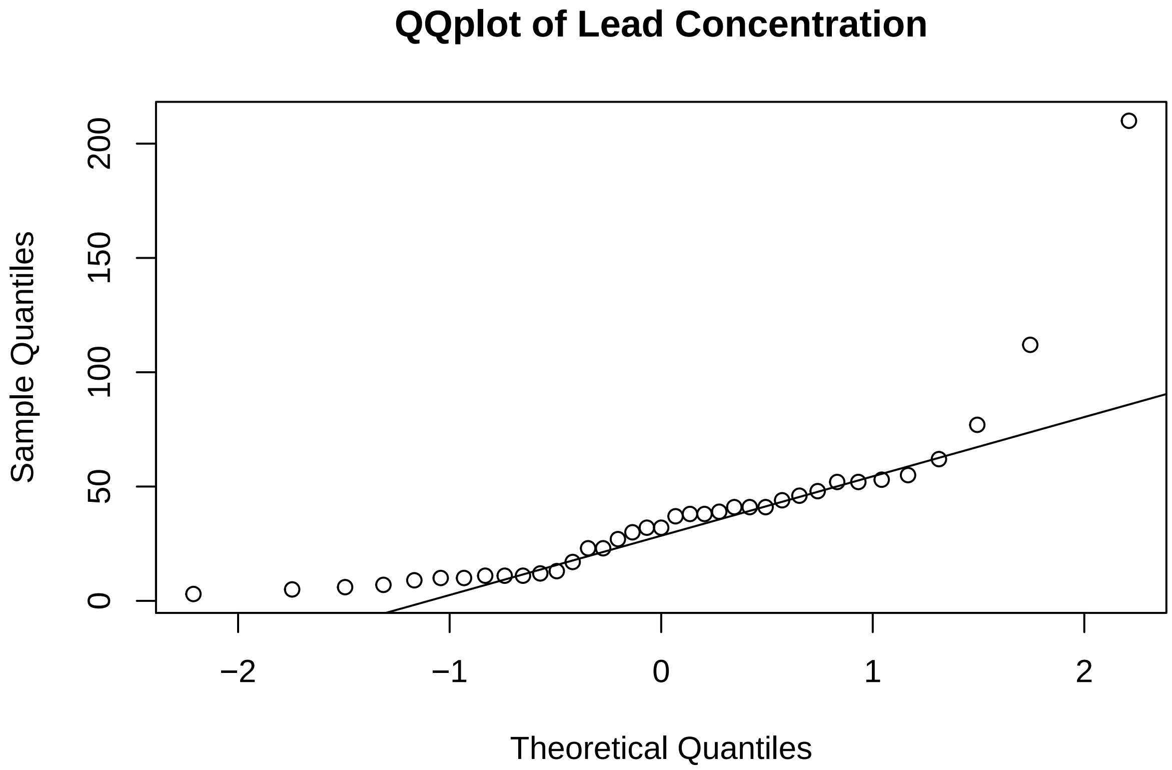
A. Construct a histogram, qqplot and run SW test of normality. What do you conclude about the normality of the data based on each of the criteria? Do the various plots and tests agree? (4 pts)

```
# read in data
lead_data <- read.csv("../Data/OTT_Final/ASCII-comma/CH05/ex5-27.txt") %>%
  rename(Lead = `X.Lead.`)

# construct histogram
hist(lead_data$Lead, main = "Histogram of Lead Concentration", xlab = "Lead Concentration")
```



```r
# qqplot
qqnorm(lead_data$Lead, main = "QQplot of Lead Concentration"); qqline(lead_data$Lead)
```



```r
# Shapiro-Wilk test of normality
shapiro.test(lead_data$Lead)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lead_data$Lead
## W = 0.69693, p-value = 1.928e-07
```

All of the tests agree.

The histogram is skewed to the right, so it does not look normally distributed. In the qqplot, the data does not sit on the linear line; the curvature supports non-normality. Because the p-value for the Sharpiro-Wilk Normality test < 0.05, we can assume that it is NOT normally distributed.

B. Give the sample mean and median for this data.

```r
mean(lead_data$Lead)
```

```
## [1] 37.24324
```

```r
median(lead_data$Lead)
```

```
## [1] 32
```

The mean is 37.24 and the median is 32.

C. Use the sign test to test the null hypothesis that the median is equal to 30. Give the pvalue and make a conclusion. H0: $M = 30$ HA: $M \neq 30$

```
SIGN.test(lead_data$Lead, md = 30)
```

```
##
##  One-sample Sign-Test
##
## data:  lead_data$Lead
## s = 20, p-value = 0.6177
## alternative hypothesis: true median is not equal to 30
## 95 percent confidence interval:
##  17.34363 41.00000
## sample estimates:
## median of x
##          32
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level  L.E.pt U.E.pt
## Lower Achieved CI     0.9011 23.0000     41
## Interpolated CI       0.9500 17.3436     41
## Upper Achieved CI     0.9530 17.0000     41
```

The p-value is 0.6177. Because the p-value is greater than α =0.05, we fail to reject H0. Therefore, the true median could be equal to 30.

D. Give a 95% confidence interval for the median. Note: For consistency, please report the "Upper Achieved CI".

The 95% confidence interval for the median is (17, 41).

E. Give a (standard) 95% confidence interval for the mean.

```
t.test(lead_data$Lead)
```

```
##
##  One Sample t-test
##
## data:  lead_data$Lead
## t = 6.1023, df = 36, p-value = 5.074e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  24.86550 49.62099
## sample estimates:
## mean of x
##  37.24324
```

The normal 95% confidence interval for the mean is (24.86, 49.62).

F. It should be clear from the diagnostics in part A that the assumption of normality is not met. Hence the CI from previous question is suspect. Give a 95% bootstrap studentized confidence interval for the mean. Hint: See "Boot Example2", but use a different value for set.seed.

```
#Define the function
mean.fun <- function (d, i)
{ m <- mean(d[i])
n <- length(i)
v <- (n-1)*var(d[i])/n^2
c(m, v)
}

set.seed(1200)

results2 <- boot(data = lead_data$Lead, mean.fun, R = 1000)
boot.ci(results2, type = "all")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results2, type = "all")
##
## Intervals :
## Level      Normal              Basic             Studentized
## 95%   (25.60, 49.17 )   (24.68, 47.27 )   (27.65, 57.63 )
##
## Level     Percentile            BCa
## 95%   (27.22, 49.81 )   (28.73, 55.92 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

The studentized bootstrap confidence interval is (27.65, 57.63).

G. Assuming that cumulative lead exposure is of interest, would the mean or the median be of more interest.

Because we're interested in the cumulative lead exposure, the mean would be of more interest.

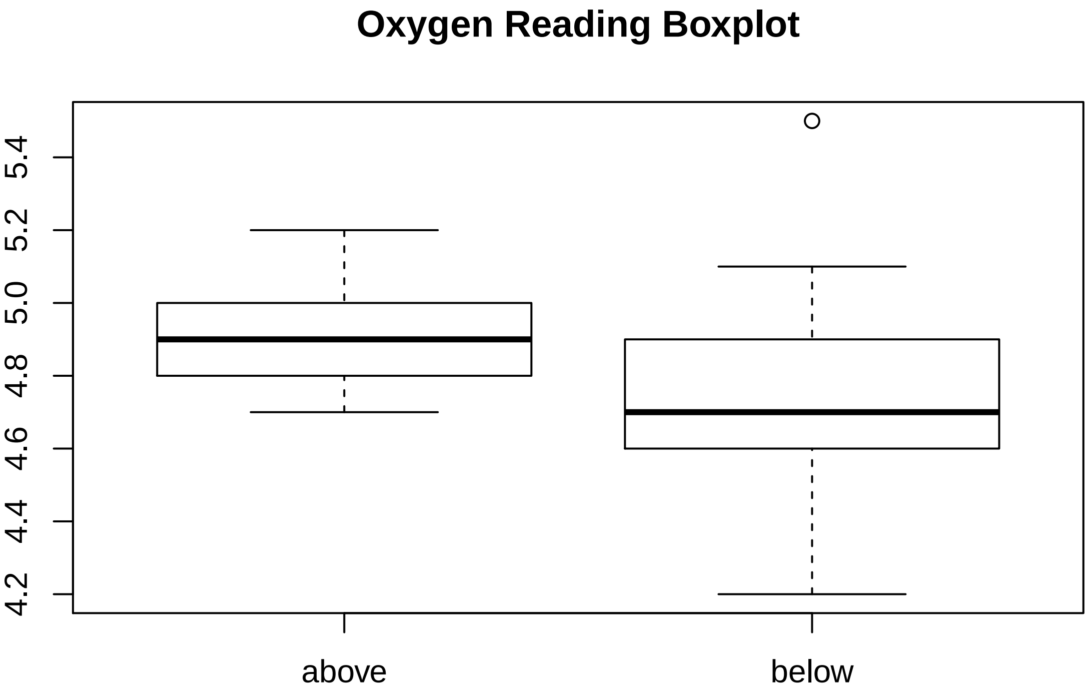
Question 4

Use the data from problem 6.6 which concerns dissolved oxygen readings for Above and Below town sites. Note: The values for the Below site do not match what is shown in the textbook.

A. Construct the side-by-side boxplots and include them in your assignment.

```
# read in data
oxygen_data <- read.csv("../Data/OTT_Final/ASCII-comma/CH06/ex6-6.txt") %>%
  rename(above = `X.Above.`,
         below = `X.Below.`)

# construct boxplot
boxplot(oxygen_data, main = "Oxygen Reading Boxplot")
```



B. Give the sample means and standard deviations for each site (Above and Below).

```
oxygen_data_tidy <- oxygen_data %>%
  gather(key = group, value = oxygen)

oxygen_data_tidy %>%
  group_by(group) %>%
  summarise(mean = mean(oxygen),
            sd = sd(oxygen))
```

```
## # A tibble: 2 x 3
##   group  mean    sd
##   <chr> <dbl> <dbl>
## 1 above  4.92 0.157
## 2 below  4.74 0.320
```

C. Considering the summary statistics from above, is the pooled variance t-test or Welch-Satterthwaite t-test preferred here? Justify your response using the rule of thumb from the notes.

To determine if the pooled variance or Welch-Satterthwaite test should be used, I first calculate the ratio of standard devations.

```
#smax - smin
0.320/0.157
```

```
## [1] 2.038217
```

Because the ratio is greater than 2, we do not assume equal variances, and the Welch-Satterthwaite t-test should be used.

D. Without assuming equal variances, give the 95% confidence interval for the difference between the means. Based on this interval, can we conclude that there is a difference between the population means? Explain.

```r
t.test(oxygen_data_tidy$oxygen ~ oxygen_data_tidy$group, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  oxygen_data_tidy$oxygen by oxygen_data_tidy$group
## t = 1.9551, df = 20.343, p-value = 0.06445
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01183912  0.37183912
## sample estimates:
## mean in group above mean in group below
##                4.92                4.74
```

The 95% confidence interval without assuming equal variances is (-0.012, 0.372). Because 0 is included in this confidence interval, we fail to reject H0. Therefore, we conclude that there is not a difference between the true population means.

E. Run the Welch-Satterthwaite t-test to test H0: H0:µ1- µ2=0 versus a two-sided alternative. Give the p-value and conclusion.

The p-value is 0.064 (based on above calculation). Because this value is greater than $\alpha = 0.05$, we fail to reject H0. Therefore, the true population means could be equal.