STAT 511A HW4

Kathleen Wendt

9/27/2019

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 USD using a 95% confidence interval. From past surveys, the monthly charge for rent-controlled apartments ranged from 1600-3200 USD.

Part 1A

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

I used the **Empirical Rule** (99%) to estimate the standard deviation:

$\hat{\sigma}$ = (max - min) / 6 = (3200 - 1600) / 6 = 266.6666667

Part 1B

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

```
rent_n <- seq(from = 41, to = 50, by = 1)
sd <- 267.67
alpha <- 0.05
me <- qt(1-(alpha/2), df = rent_n-1)*sd/sqrt(rent_n)
nme1b <- data.frame(rent_n, me)
nme1b
```

```
##    rent_n       me
## 1      41 84.48708
## 2      42 83.41184
## 3      43 82.37667
## 4      44 81.37912
## 5      45 80.41697
## 6      46 79.48818
## 7      47 78.59086
## 8      48 77.72327
## 9      49 76.88380
## 10     50 76.07097
```

A sample size of at least **46 renters** would allow for the construction of a 95% confidence interval with a margin of error less than 80 USD.

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 $\alpha = 0.05$.

Part 2A

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

```
power.t.test(n = 120, sd = 2, delta = 0.4,
             sig.level = 0.05,
             type = "one.sample", alternative = "one.sided")
```

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

Part 2B

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

Higher

Part 2C

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

Higher

Part 2D

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

Lower

Part 2E

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?

Higher

Part 2F

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, sd = 2, delta = 0.4,
             sig.level = 0.05,
             type = "one.sample", alternative = "one.sided")
```

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

A sample size of **156 men** would be needed to achieve power of 0.80.

Question 3

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

```
creek_lead <- readxl::read_xlsx("ex5-27.xlsx")
tibble::glimpse(creek_lead)
```

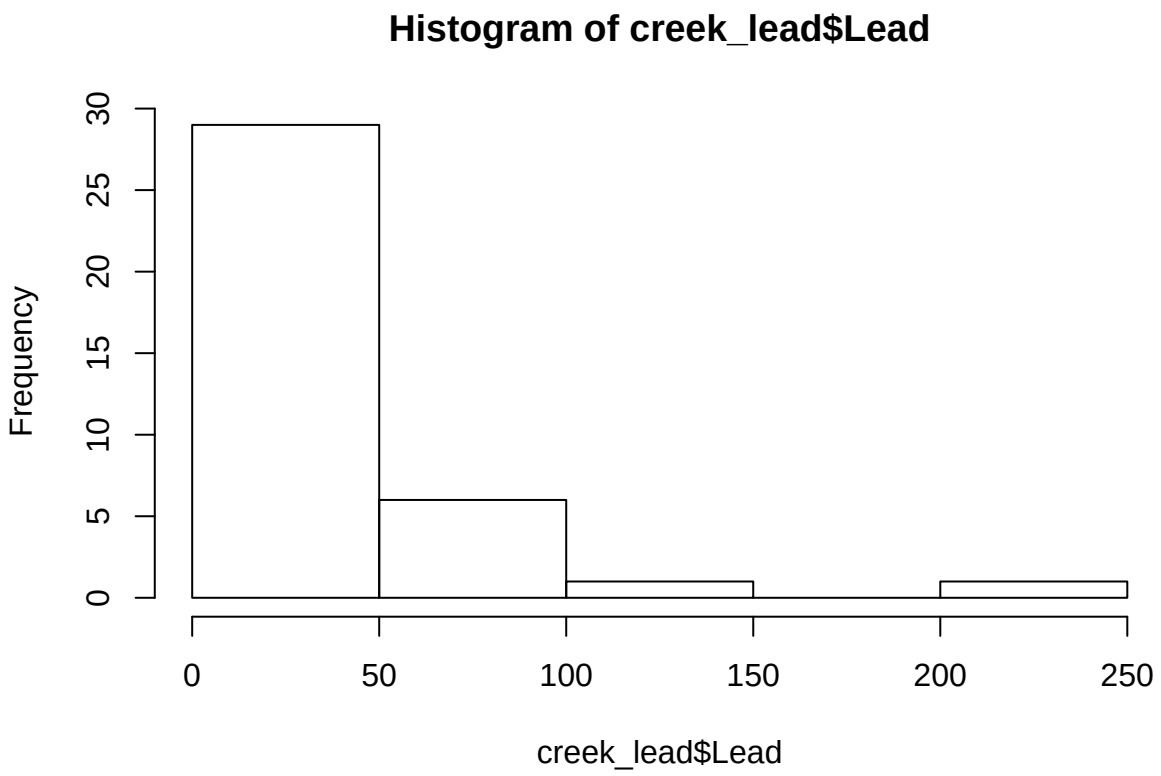
```
## Observations: 37
## Variables: 1
## $ Lead <dbl> 48, 41, 3, 77, 53, 37, 13, 210, 44, 41, 10, 38, 55, 46, 1...
```

Part 3A

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)

Histogram

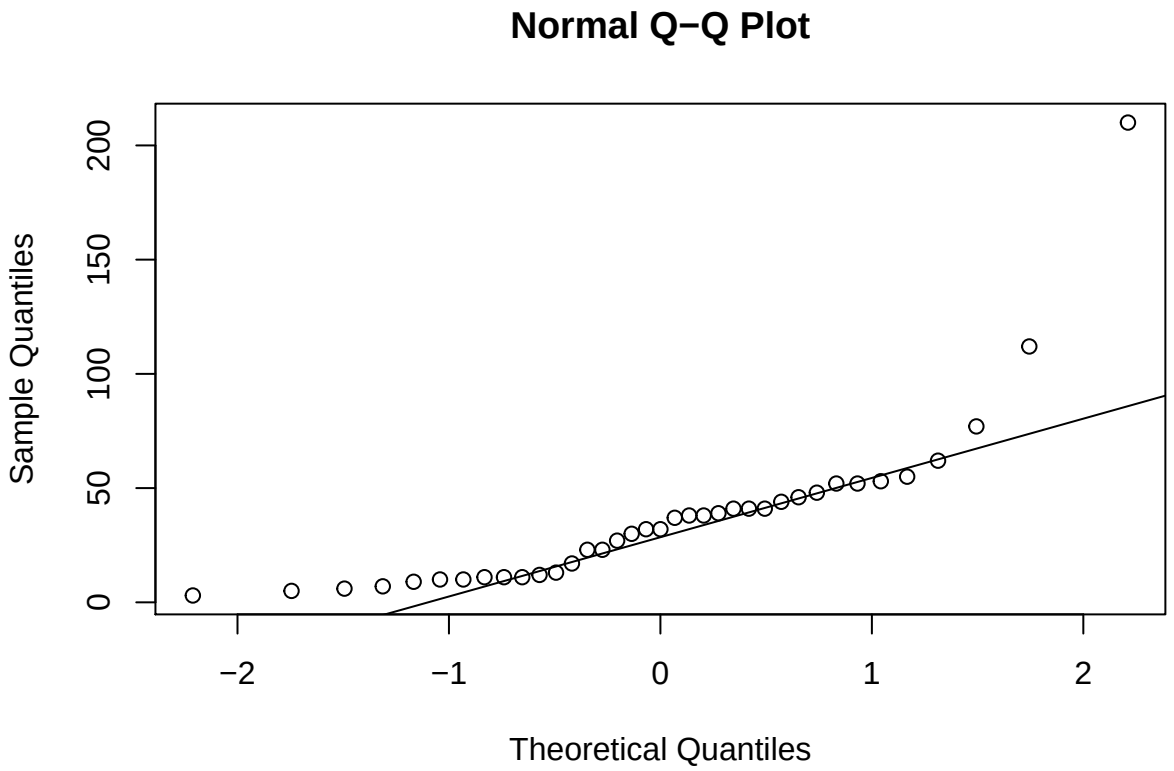
```
hist(creek_lead$Lead)
```



The above histogram indicates the lead levels in the estuarine creeks are not normally distributed and are positively skewed.

Q-Q Plot

```
qqnorm(creek_lead$Lead)
qqline(creek_lead$Lead)
```



The above Q-Q plot shows on-normal patterns consistent with those shown in the histogram. The data are not normal.

Shapiro-Wilk Test

```
shapiro.test(creek_lead$Lead)
```

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

Confirmed by the Shapiro-Wilk test, the data on lead levels in esturine creeks are not normal, $p < 0.05$.

Overall, visual inspection (histogram and Q-Q plot) and the Shapiro-Wilk test indicate the sample data are not normally distributed and are positively skewed.

Part 3B

Give the sample mean and median for this data.

```
mean(creek_lead$Lead)
```

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

The mean of the lead levels is 37.2432432.

```
median(creek_lead$Lead)
```

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

The median of the lead levels is 32.

Part 3C

Use the sign test to test the null hypothesis that the median is equal to 30. Give the p-value and make a conclusion.

```
# install.packages("BSDA")
BSDA::SIGN.test(x = creek_lead$Lead, md = 30)
```

```
##
##  One-sample Sign-Test
##
## data:  creek_lead$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 null hypothesis is that median (M) of lead levels in the breaks is 30. The alternative hypothesis is that the true M is not 30. Using $\alpha = 0.05$, I fail to reject the null hypothesis that the true median lead level is 30, $p = 0.6177$.

Part 3D

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

(17, 41)

Part 3E

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

```
creek_mean_ttest <- broom::tidy(t.test(creek_lead$Lead))
tibble::glimpse(creek_mean_ttest)
```

```
## Observations: 1
## Variables: 8
## $ estimate    <dbl> 37.24324
## $ statistic   <dbl> 6.102305
## $ p.value     <dbl> 5.074225e-07
## $ parameter   <dbl> 36
## $ conf.low    <dbl> 24.8655
## $ conf.high   <dbl> 49.62099
## $ method      <chr> "One Sample t-test"
## $ alternative <chr> "two.sided"
```

The 95% confidence interval for μ lead level is 24.8654952 to 49.6209913.

Part 3F

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.

```
creek_n <- length(creek_lead$Lead) # set n
set.seed(20190926) # set seed with today's date
resamples <- lapply(1:10000, function(i) # create df with 10000 draws
sample(creek_lead$Lead, size = creek_n, replace = TRUE)) # sample into df
dim(resamples) # check dimension of df
```

```
## NULL
```

```
length(resamples) # check if 10000 samples were drawn
```

```
## [1] 10000
```

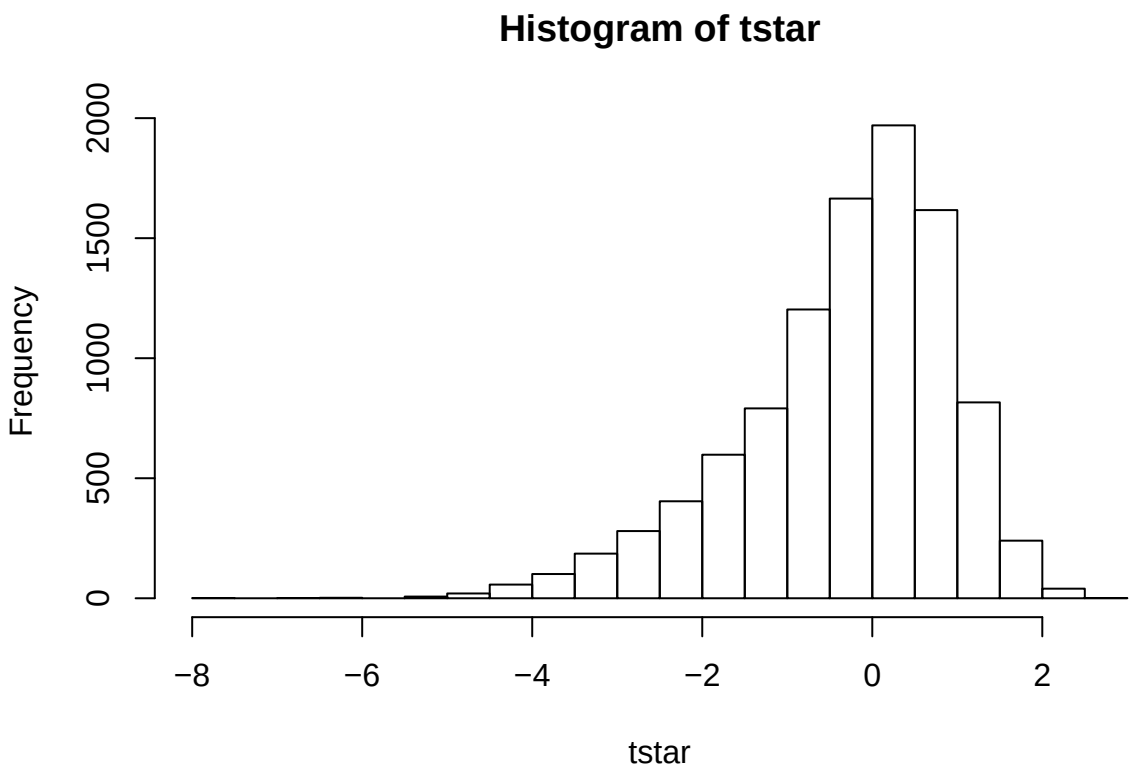
```
sort(resamples[[1]])
```

```
##  [1]   3   3   6   7   7   7  10  10  11  11  11  11  11  11  12  12  13
## [18]  17  17  30  32  32  37  37  38  39  39  41  41  41  41  46  52  52
## [35]  62  77 112
```

```
resamples <- simplify2array(resamples)
dim(resamples)
```

```
## [1]    37 10000
```

```
colmeans <- apply(resamples, 2, mean)
colsd <- apply(resamples, 2, sd)
tstar <- (colmeans - mean(creek_lead$Lead))/(colsd/sqrt(creek_n))
hist(tstar)
```




```
t025 <- quantile(tstar, prob = 0.975)
t975 <- quantile(tstar, prob = 0.025)
t025
```

```
##    97.5%
## 1.525357
```

```
t975
```

```
##      2.5%
## -3.303334
```

```
LB <- mean(creek_lead$Lead) - t025*sd(creek_lead$Lead)/sqrt(creek_n)
UB <- mean(creek_lead$Lead) - t975*sd(creek_lead$Lead)/sqrt(creek_n)
CI <- c(LB,UB)
names(CI) <- c()
CI
```

```
## [1] 27.93377 57.40396
```

Part 3G

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

Mean

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.

Note: The data is in "wide" format. All questions can be answered using the current format. An alternative is to "transpose" the data to "long format". This is NOT required, but may be handy. For example:

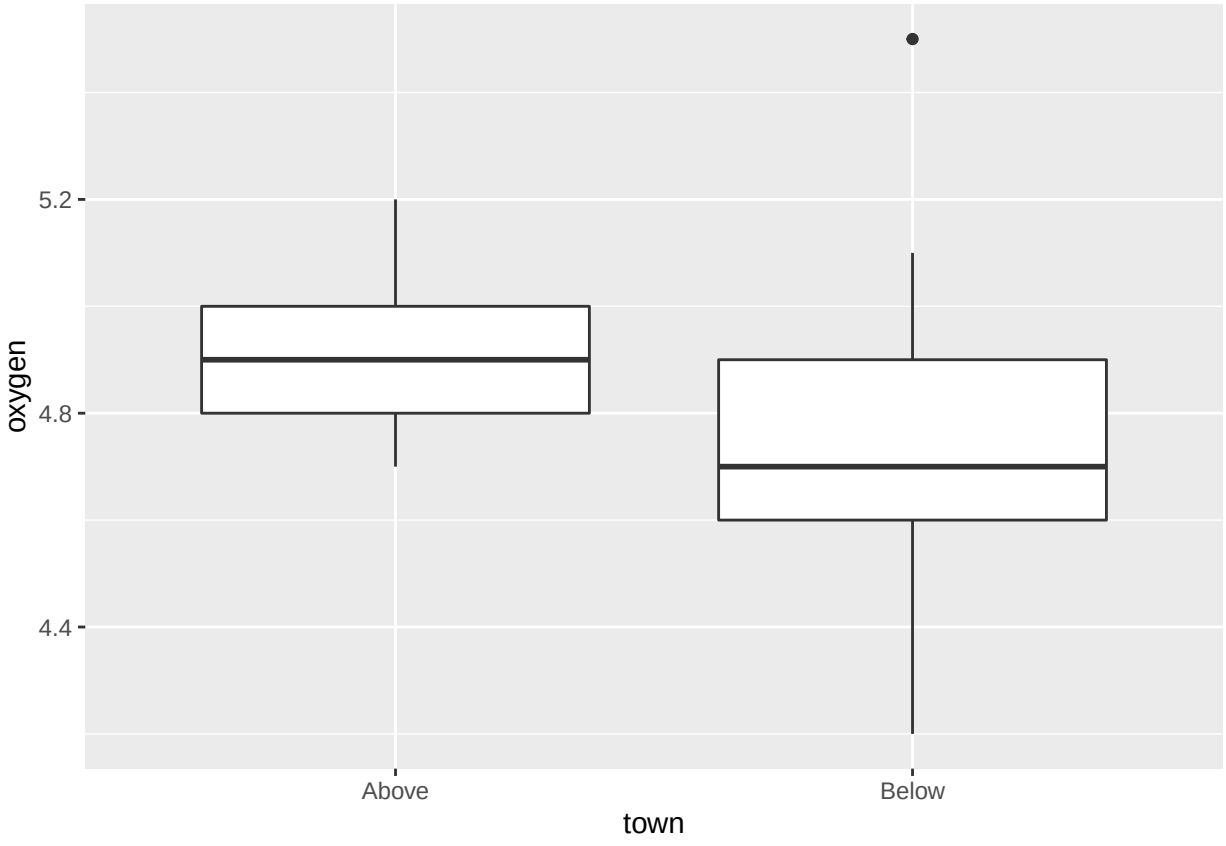
```
library(magrittr)
oxygen_data <- readxl::read_xlsx("ex6-6.xlsx") %>%
  tidyr::gather(key = town, value = oxygen)
tibble::glimpse(oxygen_data)
```

```
## Observations: 30
## Variables: 2
## $ town   <chr> "Above", "Above", "Above", "Above", "Above", "Above", "...
## $ oxygen <dbl> 5.2, 4.8, 5.1, 5.0, 4.9, 4.8, 5.0, 4.7, 4.7, 5.0, 4.7, ...
```

Part 4A

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

```
library(tidyverse)
library(ggplot2)
oxygen_data %>%
  ggplot(aes(x = town, y = oxygen)) +
  geom_boxplot()
```



Part 4B

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

```
oxygen_data %>%
  group_by(town) %>%
  summarize(n = n(),
            mean = mean(oxygen),
            sd = sd(oxygen),
            se = sd/sqrt(n)) %>%
  ungroup()
```

```
## # A tibble: 2 x 5
##   town      n  mean    sd     se
##   <chr> <int> <dbl> <dbl>  <dbl>
## 1 Above    15  4.92 0.157 0.0405
## 2 Below    15  4.74 0.320 0.0827
```

Part 4C

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.

Using the rule of thumb (max sd / min sd < 2 to assume equal variances), I estimate the value to be 2.0431386, which is slightly above 2, indicating that equal variances cannot be assumed and a Welch-Satterthwaite t-test is preferred.

Part 4D

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.

```
oxygen_welch <- broom::tidy(t.test(oxygen ~ town,
                                   data = oxygen_data,
                                   var.equal = FALSE))
oxygen_welch
```

```
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1    0.180      4.92      4.74      1.96  0.0644      20.3  -0.0118
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
```

The 95% confidence interval for the difference between the mean levels of oxygen by town is (-0.0118391, 0.3718391). Based on this interval, we can conclude that there is no statistically significant difference between means because 0 is included in the interval.

Part 4E

Run the Welch-Satterthwaite t-test to test the null hypothesis (μ 1 - μ 2 = 0) versus a two-sided alternative. Give the p-value and conclusion.

```
oxygen_welch
```

```
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1    0.180      4.92      4.74      1.96  0.0644      20.3  -0.0118
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
```

We fail to reject the null hypothesis that there is no difference between the mean oxygen levels in Above and Below, $p = 0.0644481$. These results suggest that there is no statistically significant difference between towns in oxygen levels.