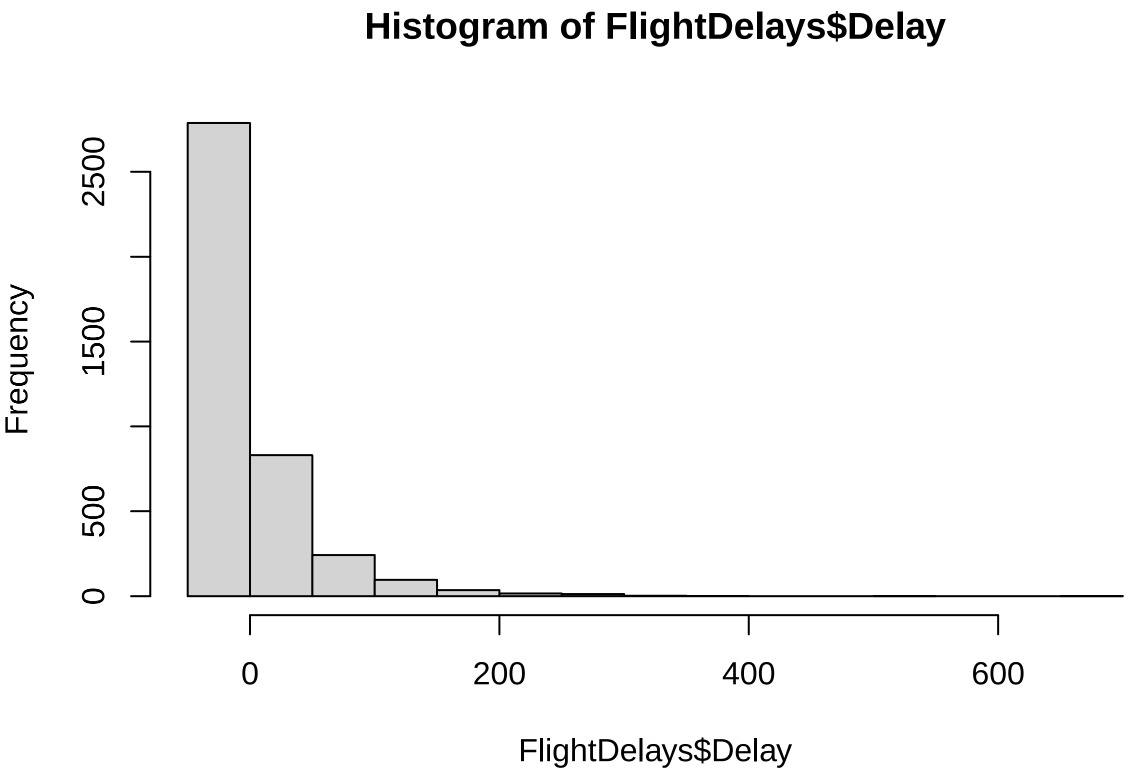
hw07

Victor Huang

2/23/2022

Q2. (Exercise 7.9) a).

```
FlightDelays <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/FlightDelays.csv")
hist(FlightDelays$Delay)
```



```
mean(FlightDelays$Delay)
```

```
## [1] 11.7379
```

```
sd(FlightDelays$Delay)
```

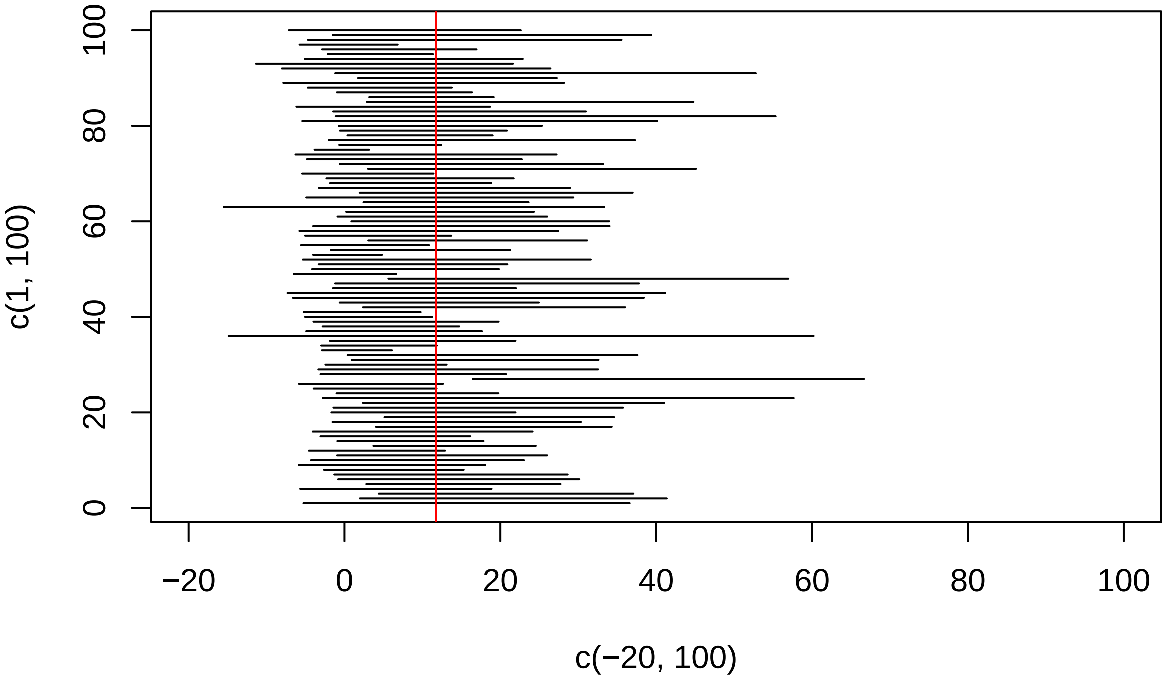
```
## [1] 41.6305
```

The histogram shows a uni-modal, extremely right-skewed graph with mean 11.7379 and SD 41.6305.

b).

```
mu <- mean(FlightDelays$Delay)
counter <- 0
plot(c(-20, 100), c(1, 100), type = "n")

for(i in 1:1000){
  x <- sample(FlightDelays$Delay, 30, replace = FALSE)
  L <- t.test(x)$conf.int[1]
  U <- t.test(x)$conf.int[2]
  if(L < mu && mu < U)
    counter <- counter + 1
  if(i <= 100)
    segments(L, i, U, i)
}
abline(v = mu, col = "red")
```



```
counter/1000
```

```
## [1] 0.866
```

Using random samples of size 30, we were only able to capture the true mean 86.3% of the times. As such, it did not capture the true mean 95% of the times.

c) The confidence interval measures the range where the the “95% confident” lies. As such, with a larger/smaller range of numbers, to incorporate that, the CI would also have to be correspondingly larger/smaller

Q3. (Exercise 7.10)

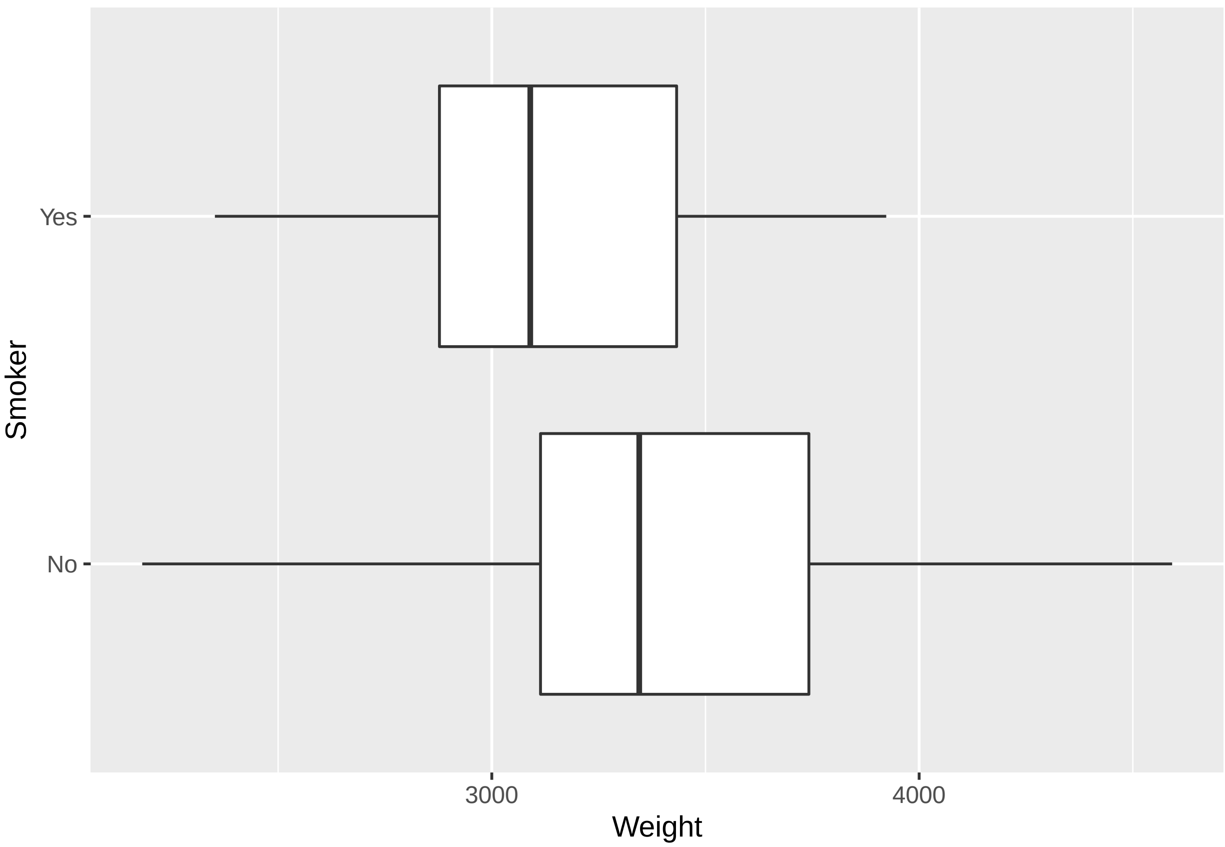
```r
Olympics2012 <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/Olympics2012.csv")
t.test(Olympics2012$Age)
```

```
##
##  One Sample t-test
##
## data:  Olympics2012$Age
## t = 21.195, df = 41, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  24.70732 29.91173
## sample estimates:
## mean of x
##  27.30952
```

The distribution is shown above and we get a 95% CI between 24.70732 and 29.91173

Q4. (Exercise 7.14) a).

```r
Girls2004 <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/Girls2004.csv")
ggplot(Girls2004, aes(x = Weight, y = Smoker)) + geom_boxplot()
```



Looking at the boxplots above, we can see that the average non-smoker mother baby is on average heavier than their smoker mother baby counterparts with

b).

```r
t.test(Weight ~ Smoker, data = Girls2004, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Weight by Smoker
## t = 1.7028, df = 78, p-value = 0.09258
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -48.5319 622.4186
## sample estimates:
##  mean in group No mean in group Yes
##          3401.580          3114.636
```

We get a 95% confidence interval between -48.5319 and 622.4186. What this means is that we are 95% confident that the average weight difference between smoking and non-smoking mothers baby weight difference will fall between -48.5319 and 622.4186.

Q5. (Exercise 7.19)

```r
Groceries <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/Groceries.csv")
t.test(Groceries$Walmart, Groceries$Target, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Groceries$Walmart and Groceries$Target
## t = -0.47046, df = 29, p-value = 0.6415
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3030159  0.1896825
## sample estimates:
## mean of the differences
##             -0.05666667
```

```r
GroceriesFiltered<- Groceries[-c(2, 26, 28), ]
t.test(GroceriesFiltered$Walmart, GroceriesFiltered$Target, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  GroceriesFiltered$Walmart and GroceriesFiltered$Target
## t = -1.783, df = 26, p-value = 0.08627
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.26312656  0.01868211
## sample estimates:
## mean of the differences
##              -0.1222222
```

We have a 95% CI between -0.3030159 and 0.1896825, meaning we are 95% confident that the average difference in price between the same product being sold at Target and Walmart will fall between -0.3030159 and 0.1896825. After removing the outliers, we see huge change in p-value from 0.6415 to 0.08627. While it still isn't discerniblely different (p-val < 0.05), we can still say the outlier data was influential by seeing the difference in p-val and 95% CI range.

Q6. (Exercise 7.23)

```r
upper <- 5.29 - 1.15*(3.52/sqrt(500))
upper
```

```
## [1] 5.108968
```

Given the data, we get a one-sided upper t confidence interval from 5.109 to positive infinity for our true mean tax error

Q7. (Exercise 7.42)

Since we know that $X_1, X_2, ..., X_n$ are random sample from $N(\mu, \sigma^2)$ with sample mean $\bar{X}$ and variance S^2, we can conclude that $(n-1)S^2/\sigma^2$ has a χ^2 distribution with $n-1$) degrees of freedom based on Theorem B.10.5. By the definition of q_1 and q_2 being the $\alpha/2$ and $1-\alpha/2$ quantiles for a χ^2 distribution, we can conclude that the confidence interval for any σ^2 can be given by $((n-1)S^2/q_2),((n-1)S^2/q_1)$ by the definition of what a confidence interval is.

Q8. (Exercise 7.43)

```r
weights <- c(560, 568, 580, 550, 581, 581, 562, 550)
v <- var(weights)
n <- length(x)
(n-1)*v/qchisq(c(.95,.05), n-1)
```

```
## [1] 117.9864 283.5464
```

We get a 90% confidence interval between 117.9864 and 283.5464 for the variance σ^2

Q9. (Exercise 7.44)

Q10. (Exercise 7.27) a).

```r
prop.test(34, 350, conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  34 out of 350
## X-squared = 225.6, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.06913692 0.13429260
## sample estimates:
##          p
## 0.09714286
```

We get a 95% CI between 0.06913692 and 0.13429260 for the drug-taking group

b).

```
prop.test(56, 350, conf.level = 0.95)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  56 out of 350
## X-squared = 160.48, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1240384 0.2036158
## sample estimates:
##    p
## 0.16
```

We get a 95% CI between 0.1240384 and 0.2036158 for the placebo group

c). Yes, they do. We can conclude that the drug is effective as the range for the drug group is smaller and because the p-value is smaller as well.

d).

```
a <- c(34,56)
b <- c(350,350)
prop.test(a,b,conf.level = 0.95)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  a out of b
## X-squared = 5.623, df = 1, p-value = 0.01773
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.11508783 -0.01062645
## sample estimates:
##     prop 1     prop 2
## 0.09714286 0.16000000
```

We get a 95% CI between -0.11508783 and -0.01062645

e). Since we are trying to compare the drug with the placebo, the relevant parameter would be the difference as that would mark the difference the drug would make if it was administered to if it wasn't. As such, we should the difference (d) to compare the effectiveness of the drug against the placebo.

Q11. (Exercise 7.29)

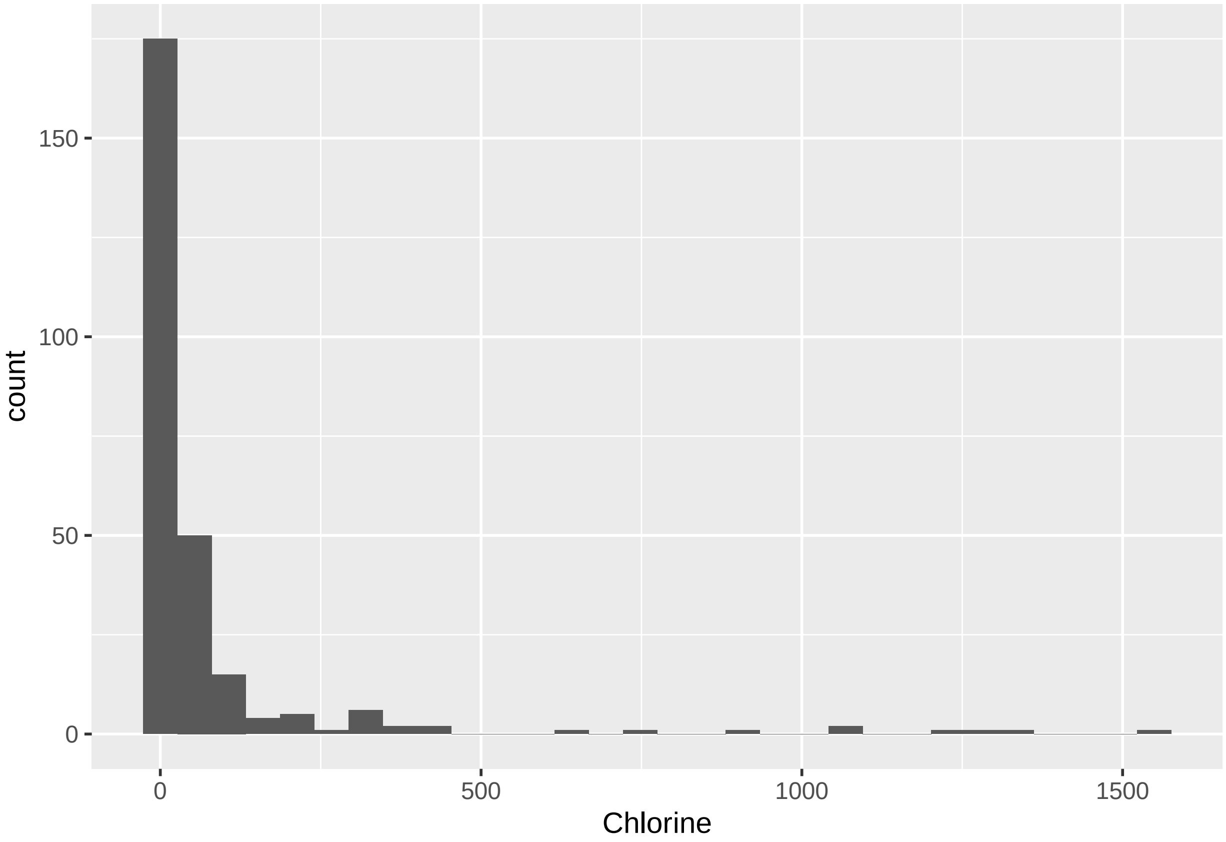
```
Bangladesh <- read.csv("https://sites.google.com/site/chiharahesterberg/data2/Bangladesh.csv")
chlorine <- with(Bangladesh, Chlorine[!is.na(Chlorine)])
summary(chlorine)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    5.00   14.20   78.08   55.50 1550.00
```

```
ggplot(Bangladesh, aes(x = Chlorine)) + geom_histogram()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



From the summary and histogram above we can see that the distribution of chlorine levels is uni-modal and extremely right skewed with Minimum = 1.00, Q1 = 5.00, Median = 14.20, Mean = 78.8, Q3 = 55.50, and Maximum = 1550.00.

b).

```
t.test(Bangladesh$Chlorine)
```

```
##
##  One Sample t-test
##
## data:  Bangladesh$Chlorine
## t = 6.0979, df = 268, p-value = 3.736e-09
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   52.87263 103.29539
## sample estimates:
## mean of x
##  78.08401
```

We get a 95% confidence interval between 52.87263 and 103.29539

c).

```
xbar <- mean(chlorine)
n <- length(chlorine)
B <- 10^4
Tstar <- numeric(B)
for (i in 1:B) {
boot_chlorine <-sample(chlorine, n, replace = TRUE)
Tstar[i] <- (mean(boot_chlorine) - xbar) / (sd(boot_chlorine)/sqrt(n))
}
quantile(Tstar, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## -2.708123  1.625365
```

```
q11Lower <- xbar - 1.66*(sd(chlorine)/sqrt(n))
q11Upper <- xbar + 2.69*(sd(chlorine)/sqrt(n))
q11Lower
```

```
## [1] 56.82756
```

```
q11Upper
```

```
## [1] 112.5297
```

Using bootstrapping we get a 95% confidence interval between 56.82756 and 112.5297. Since we are looking for mean, the bootstrap distribution would be used.