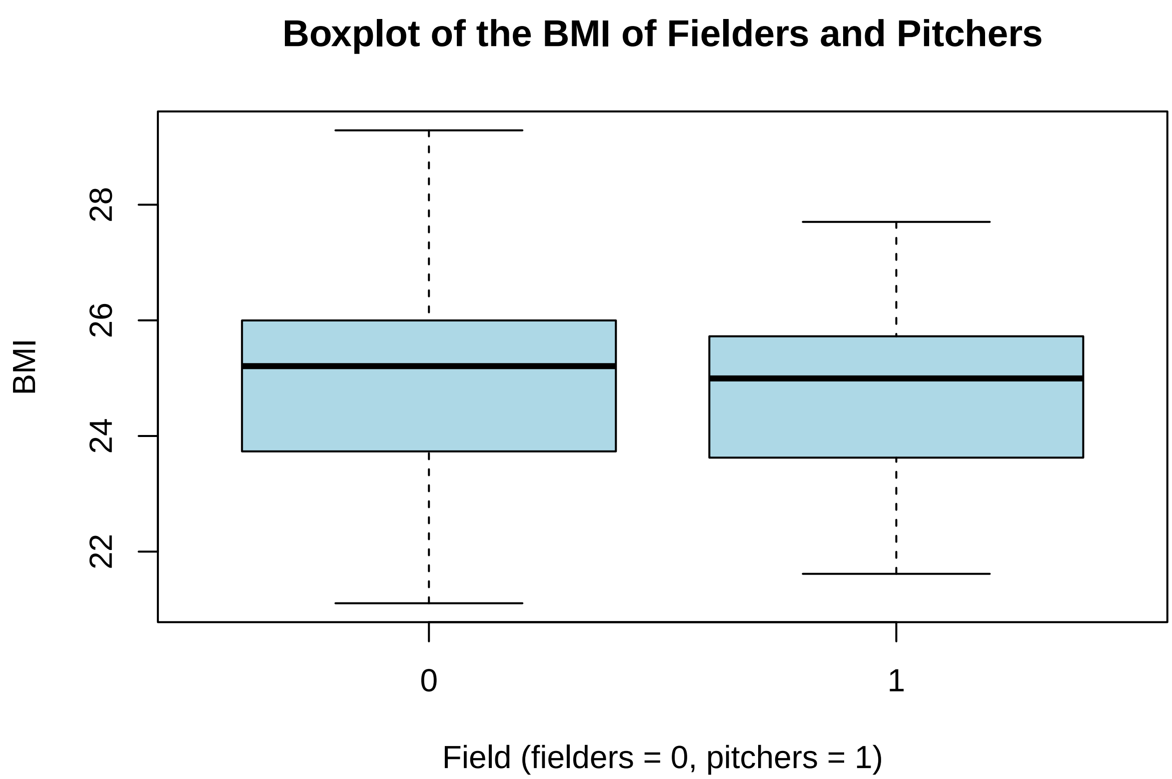
Take Home Exam 2

Matthew Bentz

4/4/2022

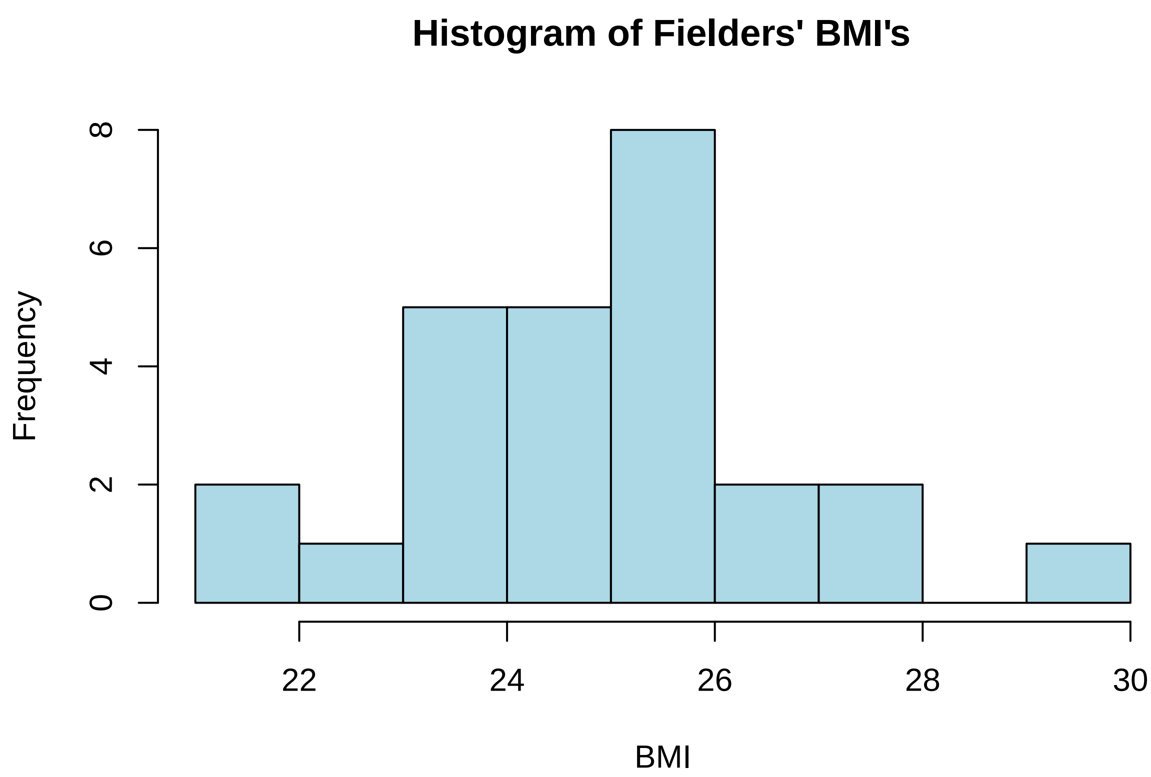
Part I

```
baseball <- read.csv("C:/Users/matth/Desktop/STATS/baseball.csv")

bmi <- (baseball$weight * 703) / baseball$height^2

baseball$bmi <- bmi

boxplot(baseball$bmi ~ baseball$field,
        main="Boxplot of the BMI of Fielders and Pitchers",
        xlab="Field (fielders = 0, pitchers = 1)", ylab="BMI",
        col="light blue")
```



The medians are relatively the same, around 25. The IQR's are also close together, with Q3 $\approx$ 26 and Q1 $\approx$ 24. It appears there are no outliers for the group, however, the spread of the fielders is larger than the pitchers. The overall shape for both groups is the same, with the caveat that the fielders' BMI's have a wider range.

```
hist(subset(baseball, field == 0)$bmi,
     main = "Histogram of Fielders' BMI's",
     xlab = "BMI", col = "light blue")
```




```
mean(subset(baseball, field == 0)$bmi)
```

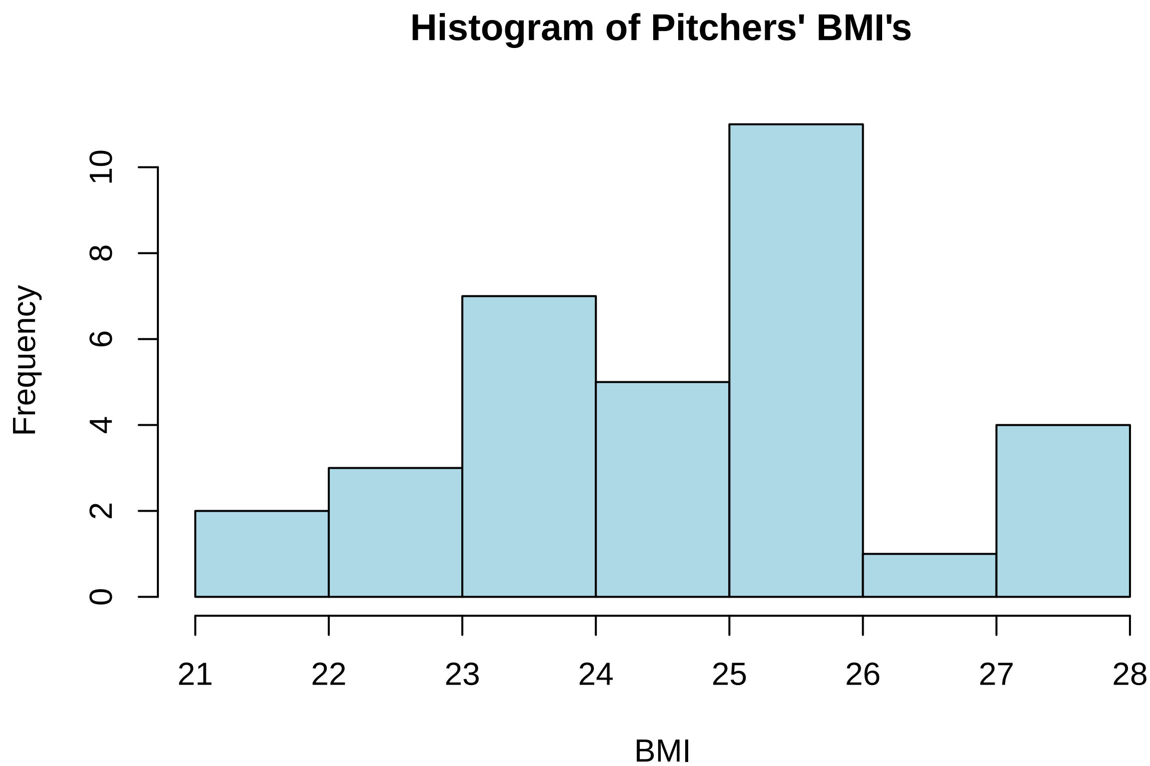
```
## [1] 24.98409
```

```
sd(subset(baseball, field == 0)$bmi)
```

```
## [1] 1.879712
```

The shape of the graph is representative of normal, however, there are spikes on the left and right side. The graph is also right skewed. The standard deviation is $\approx$ 1.88 and the mean is $\approx$ 25. Our findings match the assumption that the data for the BMI's of fielders is distributed approximately normal.

```
hist(subset(baseball, field == 1)$bmi,
     main = "Histogram of Pitchers' BMI's",
     xlab = "BMI", col = "light blue")
```




```
mean(subset(baseball, field == 1)$bmi)
```

```
## [1] 24.65618
```

```
sd(subset(baseball, field == 1)$bmi)
```

```
## [1] 1.618991
```

The shape of this graph is representative of a normal distribution, however, there is a spike on the interval [27-28]. The graph is also somewhat bimodal at intervals [23-24] and [25-26] and left skewed. The standard deviation is ≈ 1.62 and the mean is ≈ 24.6. Our findings match the assumption that the data for the BMI's of pitchers is distributed approximately normal.

7.

```
var.test(subset(baseball, field == 0)$bmi, subset(baseball, field == 1)$bmi)
```

```
##
##  F test to compare two variances
##
## data:  subset(baseball, field == 0)$bmi and subset(baseball, field == 1)$bmi
## F = 1.348, num df = 25, denom df = 32, p-value = 0.422
```

```
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6446761 2.9198676
## sample estimates:
## ratio of variances
##           1.348011
```

The p-value for the null hypothesis, that the true ratio of variances is equal to 1, is greater than the alpha value .05. This means that we would fail to reject the hypothesis that the population standard deviations are equal. The test statistic (and equivalent ratio 1), 1.348, also falls within the 95% confidence interval of [0.645, 2.92]. We would reach the same conclusion from this, although the standard deviations are not technically exact.

Part II

8.

```r
t.test(subset(baseball, field == 0)$bmi, subset(baseball, field == 1)$bmi)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(baseball, field == 0)$bmi and subset(baseball, field == 1)$bmi
## t = 0.70666, df = 49.542, p-value = 0.4831
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6043368  1.2601590
## sample estimates:
## mean of x mean of y
##  24.98409  24.65618
```

Since $p \geq \alpha$, we would fail to reject the null hypothesis that the BMI of fielders is significantly different than the BMI of pitchers. p = .4831 and α = .05 at 95% level of confidence.

9.

```r
t.test(subset(baseball, field == 1)$bmi, subset(baseball, field == 0)$bmi)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(baseball, field == 1)$bmi and subset(baseball, field == 0)$bmi
## t = -0.70666, df = 49.542, p-value = 0.4831
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.2601590  0.6043368
## sample estimates:
## mean of x mean of y
##  24.65618  24.98409
```

Since $p \geq \alpha$, we would fail to reject the null hypothesis that the BMI of pitchers is significantly different than the BMI of fielders. This finding does not change our conclusion from question 7.

10.

```r
t.test(subset(baseball, field == 0)$bmi, subset(baseball, field == 1)$bmi, conf.level = .99)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(baseball, field == 0)$bmi and subset(baseball, field == 1)$bmi
## t = 0.70666, df = 49.542, p-value = 0.4831
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -0.9151224  1.5709447
## sample estimates:
## mean of x mean of y
##  24.98409  24.65618
```

The results do not change with a lower α value. This is because our interval grows larger, and we include more values in our confidence interval. Since our test statistic is included in the 95% confidence interval, we can already assume that it will be included in our 99% confidence interval.

11.

From question 8, our 95% confidence interval between fielders and pitchers BMI is:
[-0.6043368, 1.2601590]

12.

From question 9, our 95% confidence interval between pitchers and fielders BMI is:
[-1.2601590, 0.6043368]

This interval has been switched from including more positive values to including more negative ones, although their absolute values are equal. This is due to the way we compare their differences. When the pitchers' BMI's are being subtracted from the fielders' BMI's, we will get more positive values because on average the fielders' BMI's are greater and vice versa. Overall, this means that we can be 95% confident that the true population mean difference between pitchers and fielders lies between this interval. These means must be at the very least equal, and at the very most 1.26 apart.