HW5_PH1700

Yue Zhang

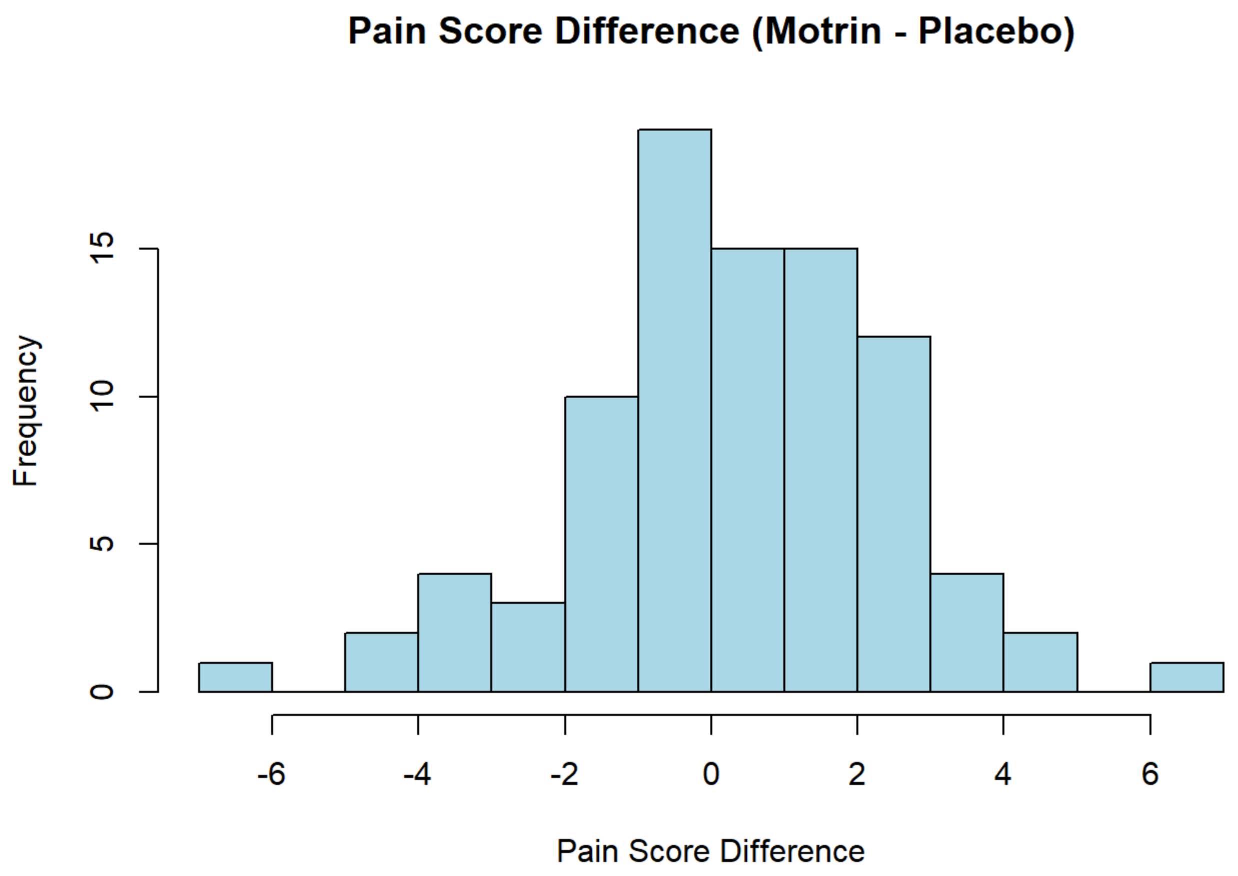
2024-10-02

8.81

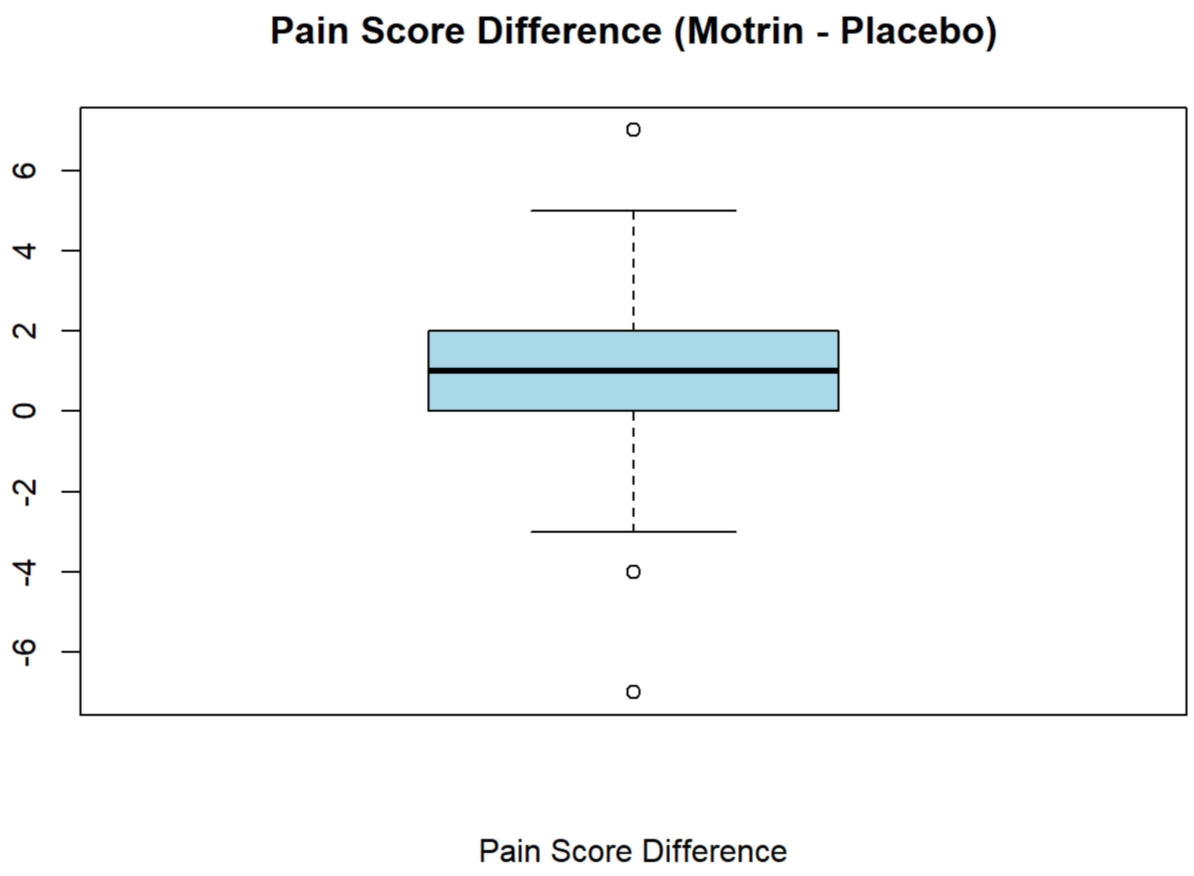
Research question: Is there a difference in degree of pain during maximal activity while on Motrin compared to placebo?

(a)

```
Tennis = read.csv("/Biostat/Biostatistics/PHL_1700/Data/Raw/Tennis.csv")

# Since we have 88 observations which is larger than 30, we
# should be able to apply the CLT.

# Create variable to calculate the pain difference (Motrin
# - Placebo)
Tennis$pain_diff = ifelse(Tennis$drg_ord == 1, Tennis$painmx_2 -
    Tennis$painmx_4, Tennis$painmx_4 - Tennis$painmx_2)

# Plot histogram of the differences
hist(Tennis$pain_diff, main = "Pain Score Difference (Motrin - Placebo)",
    xlab = "Pain Score Difference", col = "lightblue", breaks = 10)
```



```
# Plot boxplot of the difference
boxplot(Tennis$pain_diff, main = "Pain Score Difference (Motrin - Placebo)",
    xlab = "Pain Score Difference", col = "lightblue")
```



The histogram shows a reasonably symmetric and bell-shaped distribution, indicating that the normalization of the data. Thus the CLT applies. The boxplot shows that there are very few outliers and the median of pain score difference is around zero, indicating the difference between Motrin and placebo is small.

(b)
A paired t-test should be used, since this is a cross-over design.

(c)
Null hypothesis: There is no difference in degree of pain during maximal activity while on Motrin compared to placebo Alternative hypothesis: There is difference in degree of pain during maximal activity while on Motrin compared to placebo

```
# Since we've calculated the pain difference, we'll use one
# sample t test
t.test(Tennis$pain_diff, mu = 0)
```

```
##
##  One Sample t-test
##
## data:  Tennis$pain_diff
## t = 3.4135, df = 87, p-value = 0.0009758
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.3370241 1.2766123
## sample estimates:
## mean of x
## 0.8068182
```

The p-value is 0.001 < 0.05, then we have enough evidence to reject the null hypothesis that there is difference in degree of pain during maximal activity while on Motrin compared to placebo.

(d)
The 95% CI is (0.34, 1.28) which means that we are 95% confident that the difference of the means will lies between 0.34 and 1.28. However, the null value is 0 which is not included in the CI. Thus, we can reject the null hypothesis.

(e)
Null hypothesis: There is no difference in degree of pain during maximal activity while on Motrin compared to placebo Alternative hypothesis: Motrin is associated with a lower degree of pain during maximal activity compared to placebo

```
t.test(Tennis$pain_diff, mu = 0, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  Tennis$pain_diff
## t = 3.4135, df = 87, p-value = 0.9995
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##      -Inf 1.199783
## sample estimates:
## mean of x
## 0.8068182
```

The 95% CI is (-inf, 1.20), which means that we are 95% confident that the difference of means will lies in the region. The null value (0) is included in the CI and our p-value (0.9995) is larger than 0.05, thus we don’t have enough evidence to reject the null hypothesis.

8.139

Research question: Do boys with better glycemic control have different growth patterns in weight than boys with poorer glycemic control?

(1)

```
Diabetes = read_dta("/Biostat/Biostatistics/PHL_1700/Data/Raw/DIABETES-1.DAT.dta")
write.csv(Diabetes, "diabetes.csv", row.names = FALSE)

# Calculate the average HbgAlc for each subject
Diabetes_clean = Diabetes %>%
    group_by(id) %>%
    summarize(avg_hbgAlc = mean(gly_a1c))
head(Diabetes_clean)
```

id <dbl>	avg_hbgAlc <dbl>
118130	9.044100
120882	9.347100
124129	10.045000
126139	10.256200
126180	8.603317
129511	9.366162

6 rows

(2)

```
# Calculate the median HgbAlc for all boys
Median_hbgAlc = median(Diabetes_clean$avg_hbgAlc)
```

(3)

```
# Categorized boys into two groups
Diabetes_clean = Diabetes_clean %>%
    mutate(group = ifelse(avg_hbgAlc < Median_hbgAlc, "Controlled",
        "Uncontrolled"))
head(Diabetes_clean)
```

id <dbl>	avg_hbgAlc <dbl>	group <chr>
118130	9.044100	Uncontrolled
120882	9.347100	Uncontrolled

id <dbl>	avg_hbgAlc <dbl>	group <chr>
124129	10.045000	Uncontrolled
126139	10.256200	Uncontrolled
126180	8.603317	Controlled
129511	9.366162	Uncontrolled

6 rows

(4)

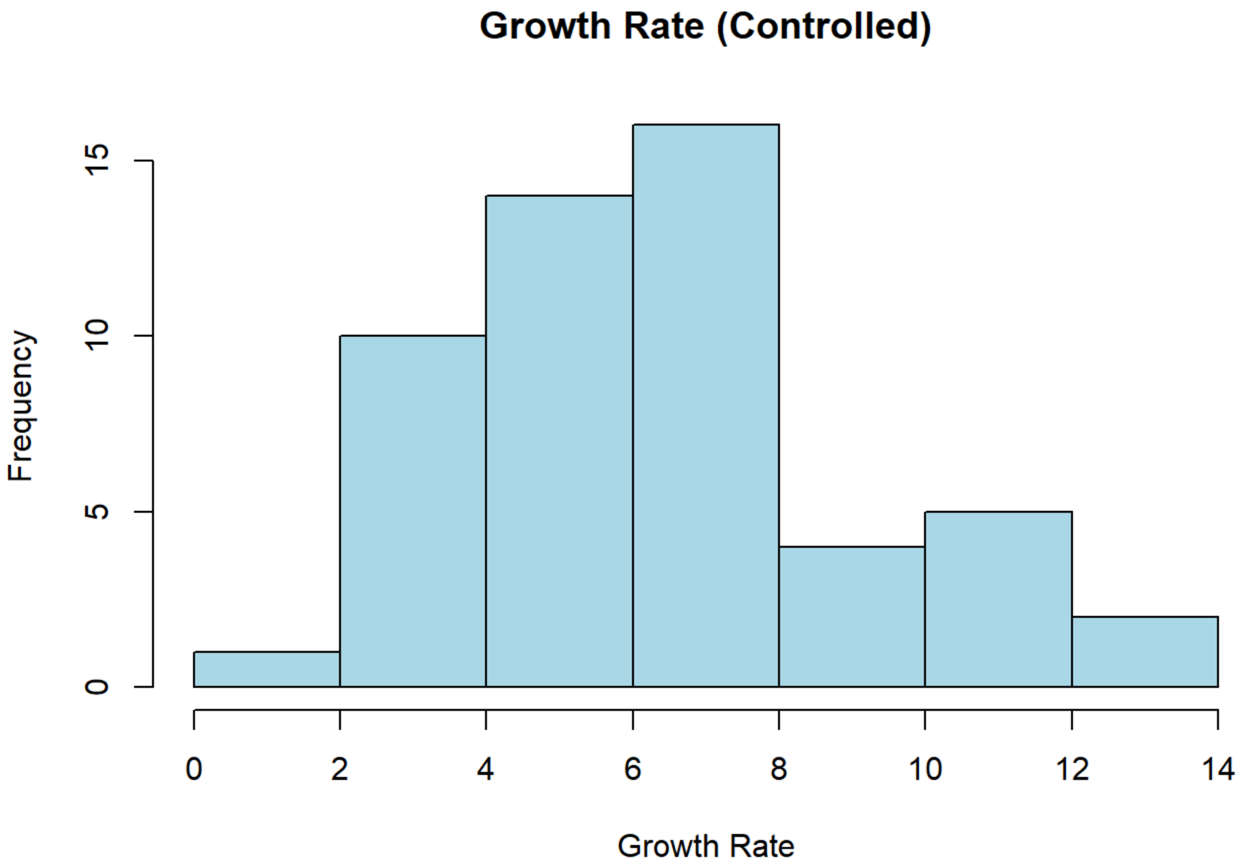
```
# Create growth variable
Growth = Diabetes %>%
    group_by(id) %>%
    summarize(first_weight = first(wt_kg), last_weight = last(wt_kg),
        first_age = first(age_yrs), last_age = last(age_yrs)) %>%
    mutate(growth_rate = (last_weight - first_weight)/(last_age -
        first_age))

Diabetes_clean = merge(Diabetes_clean, Growth, by = "id")
head(Diabetes_clean)
```

	id <dbl>	avg_hbgAlc <dbl>	group <chr>	first_weight <dbl>	last_weight <dbl>	first_age <dbl>	last_age <dbl>	growth_ra <db
1	118130	9.044100	Uncontrolled	54.9	61.5	14.2	15.1	7.3333
2	120882	9.347100	Uncontrolled	40.5	67.3	10.5	14.7	6.3809
3	124129	10.045000	Uncontrolled	43.2	70.9	10.7	15.0	6.4418
4	126139	10.256200	Uncontrolled	67.7	78.8	12.0	13.5	7.4000
5	126180	8.603317	Controlled	38.6	56.2	12.6	15.2	6.7692
6	129511	9.366162	Uncontrolled	36.3	79.9	10.3	15.2	8.8979

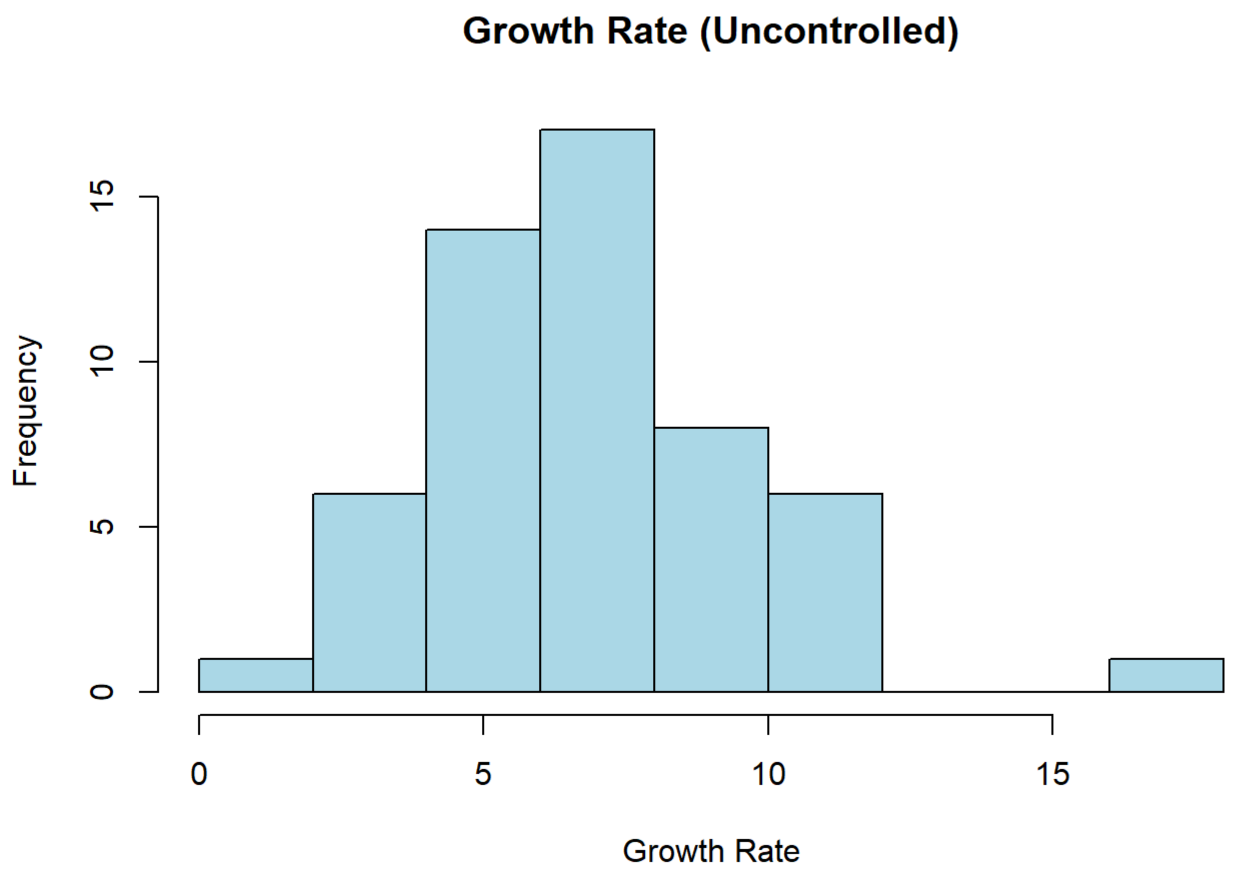
6 rows

(5)
First, we will check the normality of these two groups

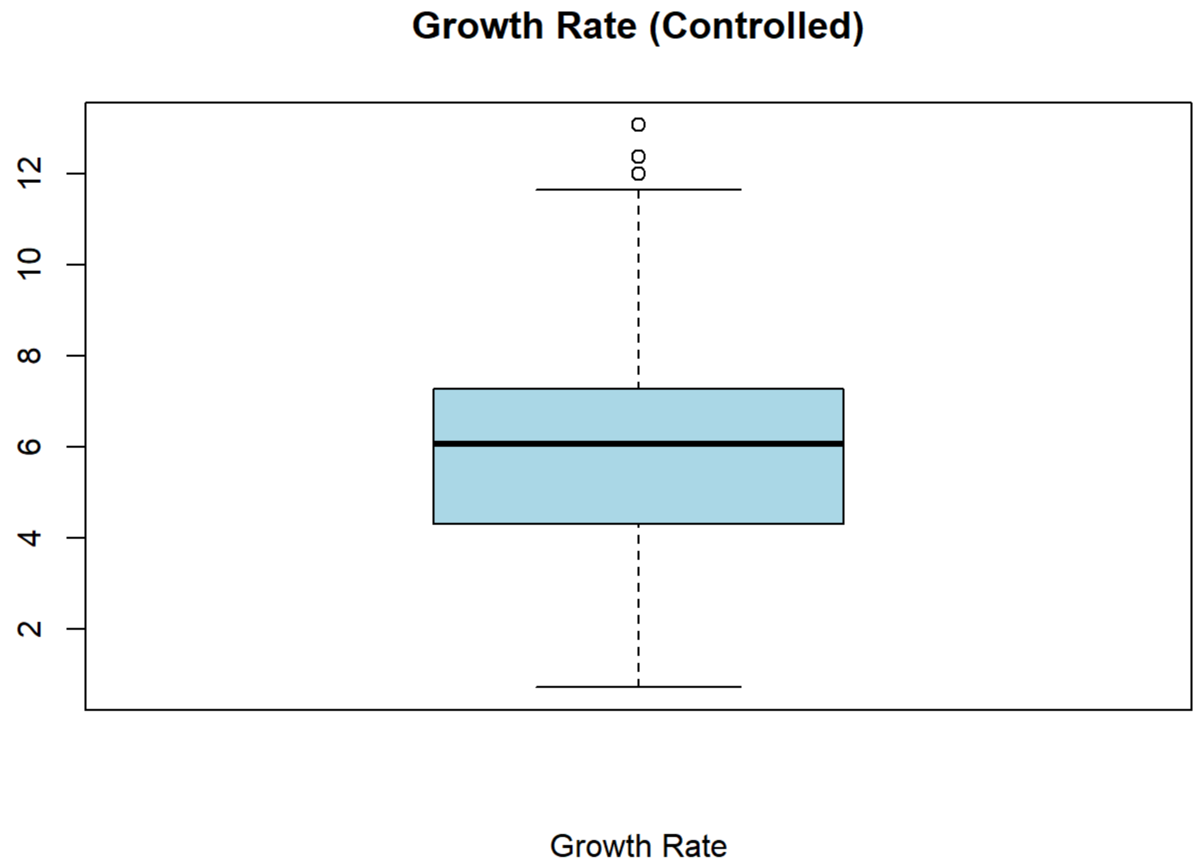
```
Controlled = Diabetes_clean$growth_rate[Diabetes_clean$group ==
    "Controlled"]
Uncontrolled = Diabetes_clean$growth_rate[Diabetes_clean$group ==
    "Uncontrolled"]
# Plot histogram of growth rate
hist(Controlled, main = "Growth Rate (Controlled)", xlab = "Growth Rate",
    col = "lightblue")
```



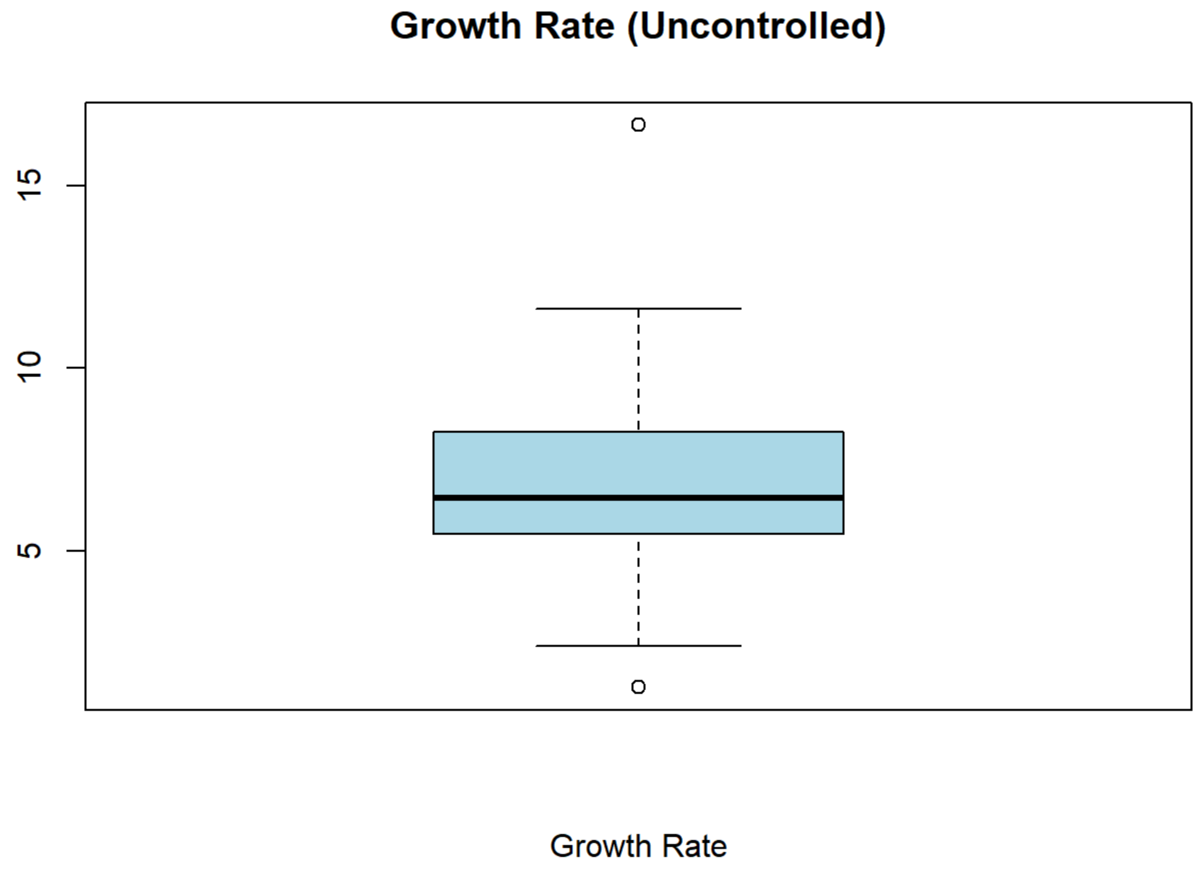
```
hist(Uncontrolled, main = "Growth Rate (Uncontrolled)", xlab = "Growth Rate",
    col = "lightblue")
```



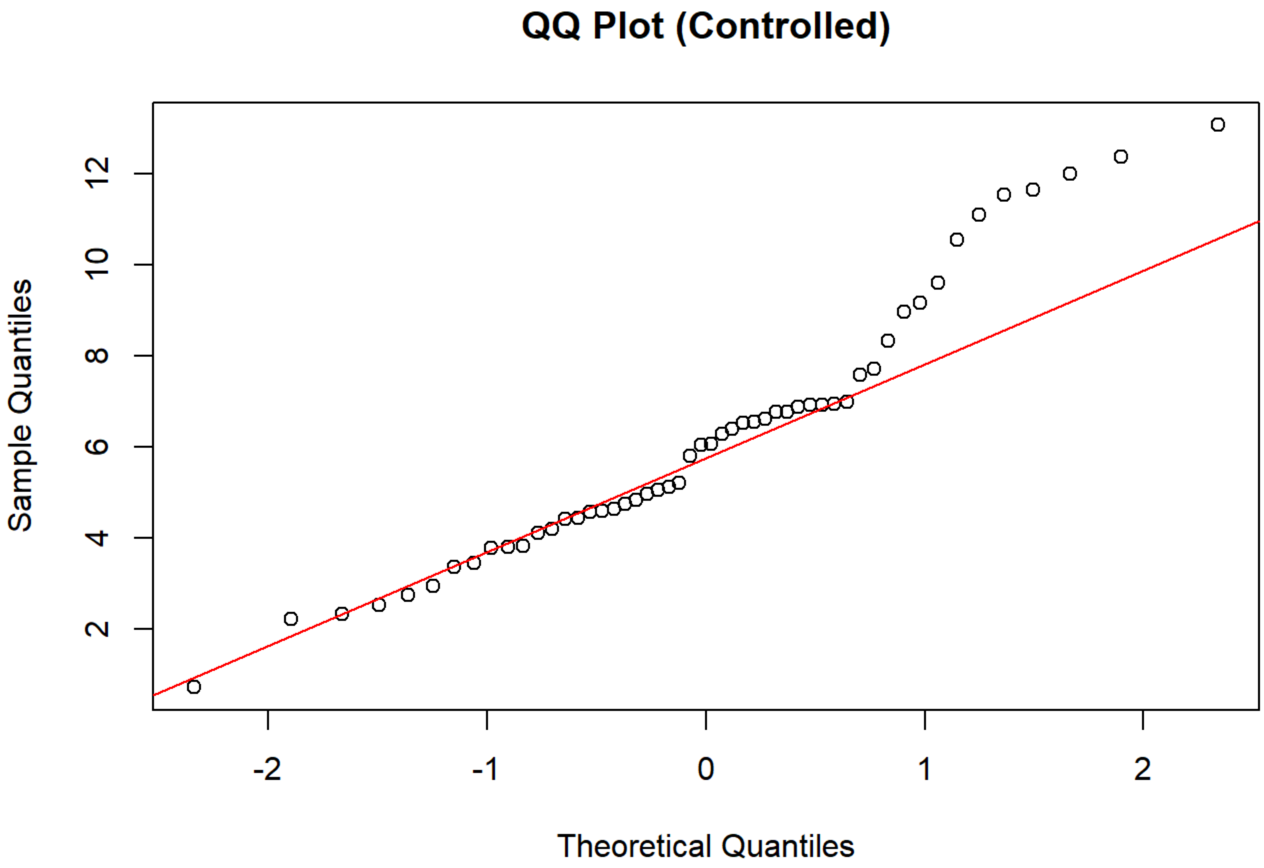
```
boxplot(Controlled, main = "Growth Rate (Controlled)", xlab = "Growth Rate",
    col = "lightblue")
```



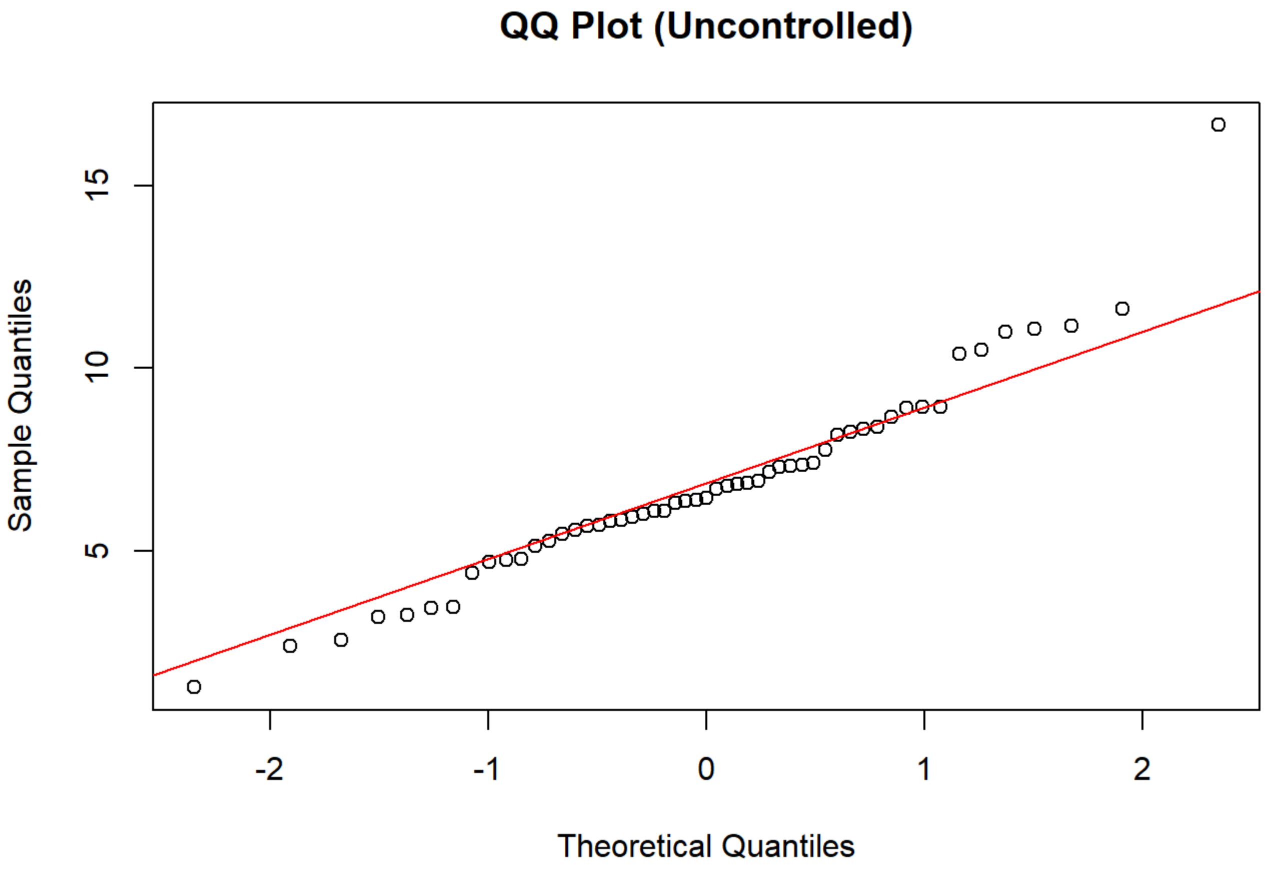
```
boxplot(Uncontrolled, main = "Growth Rate (Uncontrolled)", xlab = "Growth Rate",
    col = "lightblue")
```



```
qqnorm(Controlled, main = "QQ Plot (Controlled)")
qqline(Controlled, col = "red")
```



```
qqnorm(Uncontrolled, main = "QQ Plot (Uncontrolled)")
qqline(Uncontrolled, col = "red")
```



```
shapiro.test(Controlled)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Controlled
## W = 0.94945, p-value = 0.0276
```

```
shapiro.test(Uncontrolled)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Uncontrolled
## W = 0.95108, p-value = 0.02996
```

The Shapiro-Wilk test shows that both groups’ p-values are less than 0.05, meaning the data significantly deviates from normality. So we cannot test for equal variances. Thus, we’ll conduct the two samples t-test with unequal variances

Null hypothesis: There is no difference between the growth rate in controlled group and uncontrolled group
Alternative hypothesis: There is difference between the growth rate in controlled and uncontrolled groups
Alternative hypothesis: The two groups have unequal variances

```
t.test(Controlled, Uncontrolled, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  Controlled and Uncontrolled
## t = -1.0501, df = 102.38, p-value = 0.2961
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.6542241  0.5089172
## sample estimates:
## mean of x mean of y
##  6.246733  6.819387
```

Since the p-value(0.2961) is larger than 0.05, we don’t have enough evidence to reject the null hypothesis. Therefore, there is no difference between the growth rate of boys with better glycemic control and boys with pooer glycemic control.

Additional problems

(a)

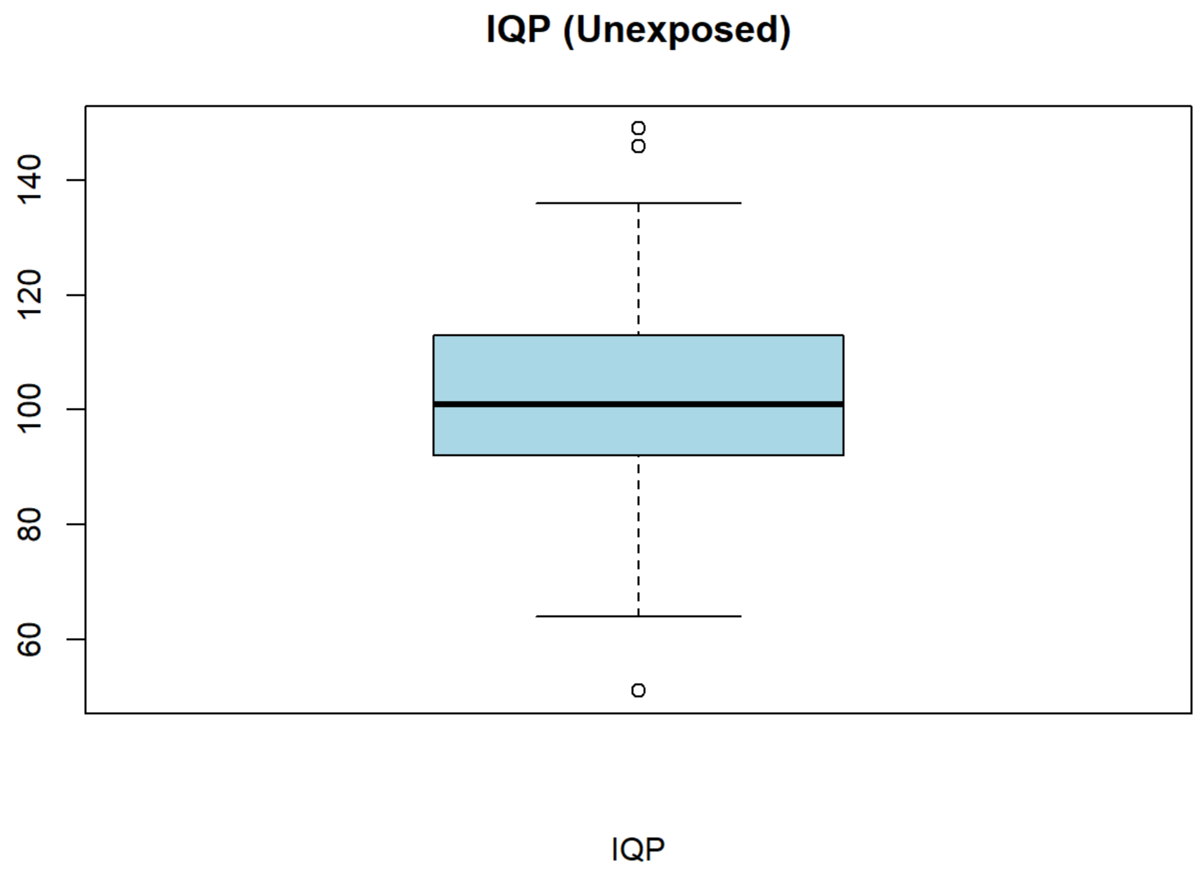
```
Lead = read_dta("/Biostat/Biostatistics/PHL_1700/Data/Raw/LEAD-1.DAT.dta")
write.csv(Lead, "Lead.csv", row.names = FALSE)

Lead_clean = Lead %>%
    mutate(Group = ifelse(lead_grp == 1, "unexposed", "exposed"))
head(Lead_clean)
```

id	area	ageyrs	sex	iqv_inf	iqv_comp	iqv_ar	iqv_ds	iqv_raw	iqp_pc ▶
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
101	3	11.08	1	3	4	3	5	15	10
102	3	9.42	1	7	9	7	6	29	8
103	3	11.08	1	4	9	5	3	21	10
104	2	6.92	1	4	6	6	6	22	5
105	1	11.25	1	5	4	8	5	22	5
106	2	6.50	1	5	12	11	9	37	14

6 rows | 1-10 of 40 columns

(b)

```
Unexposed = Lead_clean %>%
    filter(Group == "unexposed") %>%
    select(id, iqp)
boxplot(Unexposed$iqp, main = "IQP (Unexposed)", xlab = "IQP",
    col = "lightblue")
```



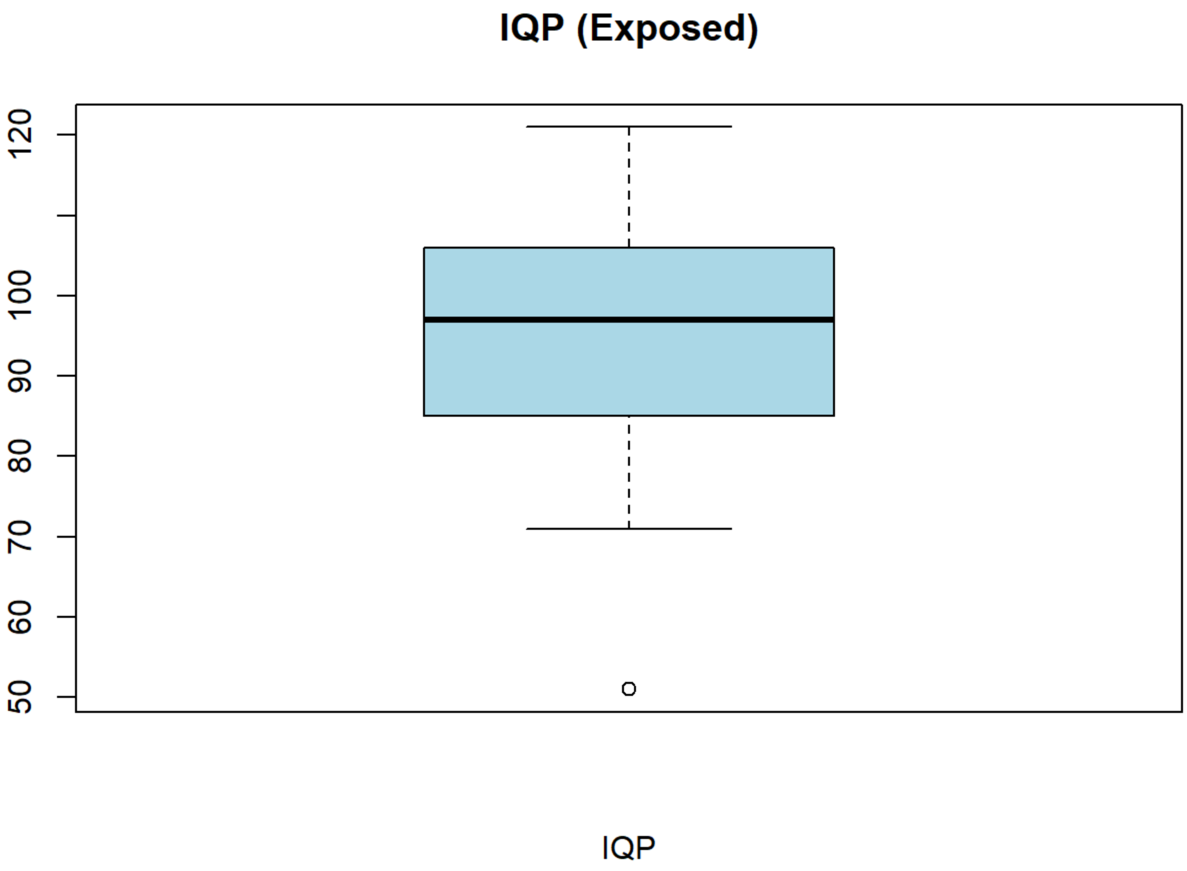
```
Q1_un = quantile(Unexposed$iqp, 0.25)
Q3_un = quantile(Unexposed$iqp, 0.75)
IQR_value_un = Q3_un - Q1_un

lower_bound_un = Q1_un - 1.5 * IQR_value_un
upper_bound_un = Q3_un + 1.5 * IQR_value_un

outliers_un = Unexposed %>%
    filter(iqp < lower_bound_un | iqp > upper_bound_un)

print(as.numeric(outliers_un$id))
```

```
## [1] 117 135 139
```

The outliers ids are: 117,135,139

(c)

```
Exposed = Lead_clean %>%
    filter(Group == "exposed") %>%
    select(id, iqp)
boxplot(Exposed$iqp, main = "IQP (Exposed)", xlab = "IQP", col = "lightblue")
```



```
Q1_e = quantile(Exposed$iqp, 0.25)
Q3_e = quantile(Exposed$iqp, 0.75)
IQR_value_e = Q3_e - Q1_e

lower_bound_e = Q1_e - 1.5 * IQR_value_e
upper_bound_e = Q3_un + 1.5 * IQR_value_e

outliers_e = Exposed %>%
    filter(iqp < lower_bound_e | iqp > upper_bound_e)

print(as.numeric(outliers_e$id))
```

```
## [1] 314
```

The outlier id is 314.

For those outliers, we’ll perform the test both with and without the outliers. If the results are similiar, then the outliers can be included. If the results varies, then the outliers should be omitted.