This document contains two clinical demonstrations/examples of some of the statistical work I do.

2025-17-10

A. Pharmaceutical Example (Taken from a real project and modified to make it easier to understand).

If say, a pharmaceutical company is considering revising its production process for its 200 mg capsules of a certain medication used for treating a disease. Let me call the disease COVID900. Let's also assume that a quality control manager task me to present the data and resulting conclusions to the Board of Directors as a Statistician. Now, let's explore the process of getting accurate data and conclusion to the board of directors.

First

I have to get some data (Skipping the scientific methodology of identification of problem, stating the aim, ...etc.). My task will be to randomly sample about 20 capsules from the current production process and 20 capsules from the potential new process. I then store the data in the csv file called meds.csv which contains the mass of these sampled capsules in milligrams. As I analyze the data, the goal will be to check and verify all assumtions of methods and models I am using.

```r
#loading a few packages
library(lawstat)
library(DescTools)

meds <- read.csv("C:/Users/radje/Desktop/USB/WKU some/Statistical Methods/Statistical Methods. Everything/meds.csv")
```

Second,

I will have to provide a numerical and graphical summaries, a thorough description of the dataset that will be suitable for presentation to the Board of Directors.
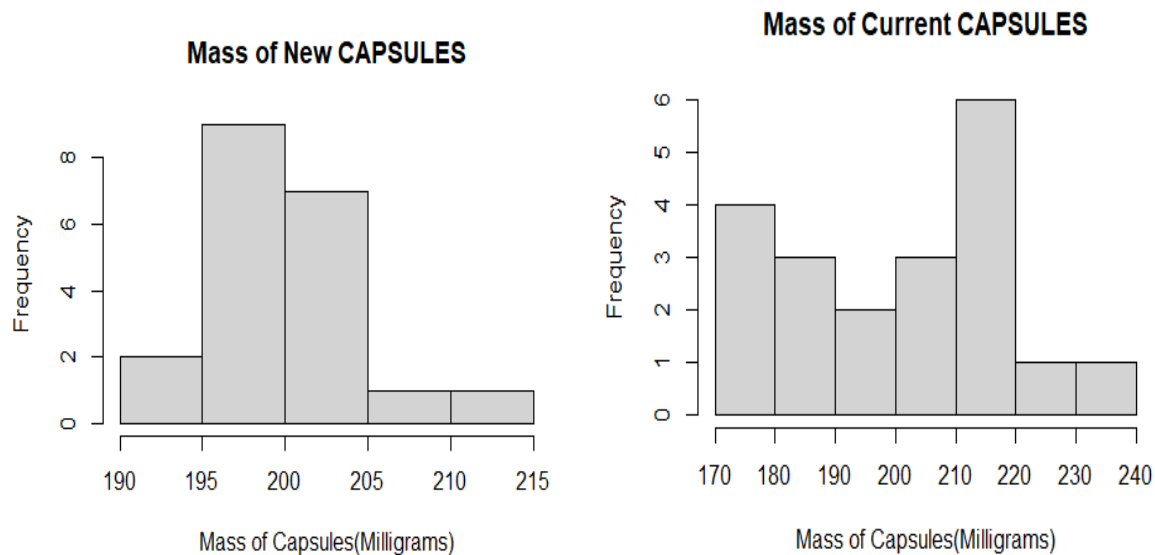
```r
attach(meds)
meds

##      New  Current
## 1  196.4    213.4
## 2  211.9    196.6
## 3  197.6    170.5
## 4  199.6    188.7
```

```
## 5   197.4    219.2
## 6   204.6    185.7
## 7   194.7    224.3
## 8   202.8    239.0
## 9   204.5    202.1
## 10 204.9    179.7
## 11 201.9    212.0
## 12 198.3    176.7
## 13 197.3    207.0
## 14 199.1    200.8
## 15 199.7    197.0
## 16 190.0    217.6
## 17 205.7    213.3
## 18 203.4    180.2
## 19 201.0    175.4
## 20 199.7    215.9
```

```
hist(New, main = "Mass of New CAPSULES", xlab ="Mass of
Capsules(Milligrams)", right=F)
```



```
hist(Current, main = "Mass of Current CAPSULES", xlab ="Mass of
Capsules(Milligrams)", right=F)
```

Next, the boxplots

```
boxplot(meds,xlab="Type of production Process",ylab="Mass of
Capsules(Milligram)")
points(1,mean(New),col='red',cex=1.5)
points(1,mean(New),col='red',cex=1.5,pch=19)
points(2,mean(Current),col='red',cex=1.5,pch=19)
```
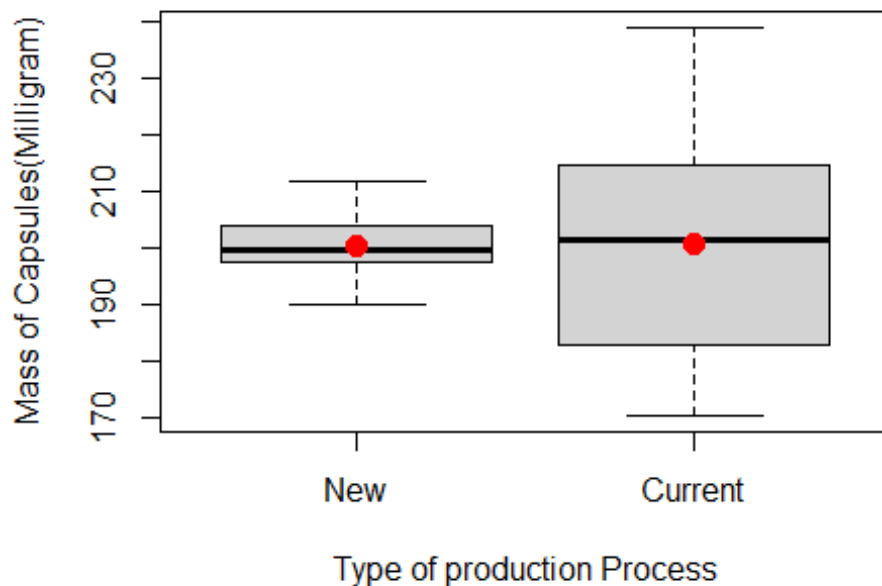
```
IQR(New)
## [1] 6.125
IQR(Current)
## [1] 29.7
sd(New)
## [1] 4.717925
sd(Current)
## [1] 18.86374
mean(Current)
## [1] 200.755
mean(New)
## [1] 200.525
```

Notice that the interquartile range for the New production process is 6.125 The interquartile range for the Current production process is 29.7 The standard deviation of the New production process is 4.717925. The standard deviation of the Current production process is 18.86374 The mean of the New production process is 200.525. The mean of the Current production process is 200.755
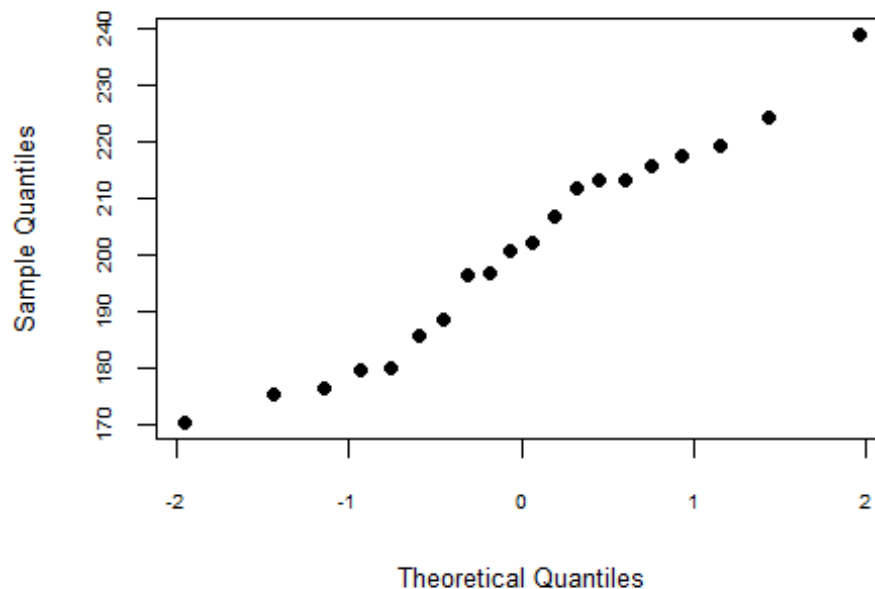
Third,

Keeping the goal of this study in mind, I am going to use appropriate inferential tools to conduct a thorough statistical analysis of this dataset.

```
qqnorm(New, pch=19, main="New production Process", cex.axis=0.6, cex.lab=0.8)
```


New production Process
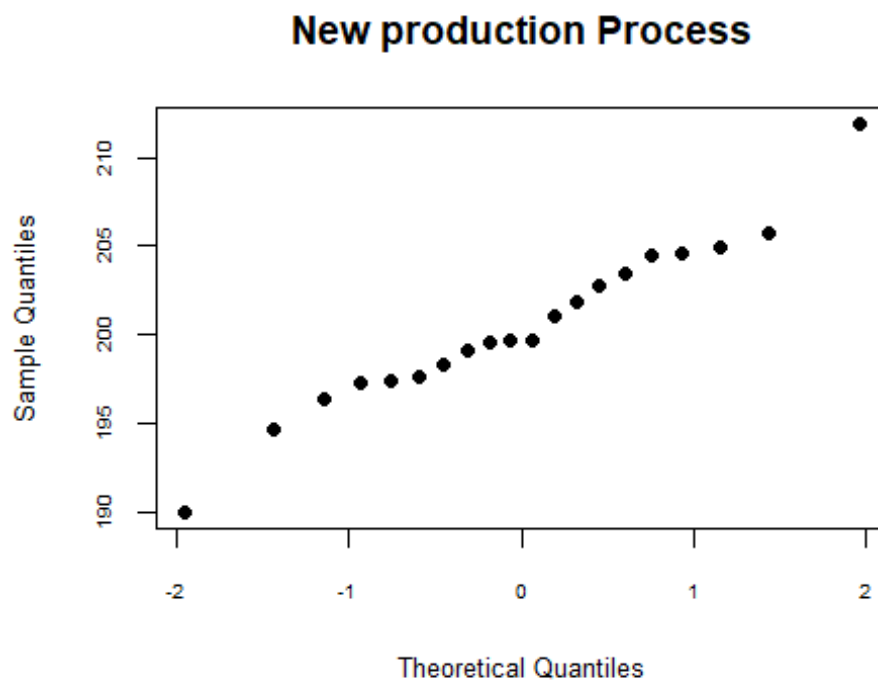
```
qqnorm(Current, pch=19, main="Current Productiion Process", cex.axis=0.6,
cex.lab=0.8)
```

Current Productiion Process

The normal probability plots are both pretty linear, and the boxplots are pretty symmetric. I do not see any major reasons to not operate under the assumption of normality. Additionally, we are told that random samples were selected from both types of production process. I have no reason to assume these samples weren't selected independently of one another.

```
var.test(New,Current, alternative="less")

##
##  F test to compare two variances
##
## data:  New and Current
## F = 0.062553, num df = 19, denom df = 19, p-value = 6.134e-08
## alternative hypothesis: true ratio of variances is less than 1
## 95 percent confidence interval:
##  0.0000000 0.1356302
## sample estimates:
## ratio of variances
##         0.06255279
```

Hypotheses

1. $H_0: \sigma_{\text{New}} \geq \sigma_{\text{Current}}$

$$H_a: \sigma_{\text{New}} < \sigma_{\text{Current}}$$

2. The assumptions of this test is described above.

3. F= 0.062553 (see output above)

4. p-value = 0(see output above)

5. Reject H0 (assuming alpha = 0.05)

At the 0.05 significance level (p-value = 0), there is sufficient evidence to conclude that the new production process for making the 200mg capsules for treating COVID900 has less variability in terms of its mass in milligrams compared to the current one.

Fourth and finally,

Conclusion: Based on the results above, I will recommend to the Board of Directors that the company switch to the new production Process. Although both processes means are similar, I will prefer the production process that produces a more consistent, less variable results. Thus, the new production process is better.

B. Another clinical example taken from a real project and modified for easier comprehension.

Let assume scientists want to conduct an experiment to test the effects of five different treatment diets on some subjects, say turkeys. My task is to statistically evaluate the amount of weight gained at the end of the alloted time period. So, to desing this experiment and make high level informed decisions, I will begin by:

First

I will randomly assign at least six turkeys to each of five diet groups and treat/feed them for a fixed period of time. Then determine and store the amount of weight gained (in pounds) at the end of the time period. The five diets/treatment (and their designations) are then stored in a csv file called turkeys.csv. They are as follows:

- Control diet [Control]
- Control diet + level 1 of additive A [C1A]
- Control diet + level 2 of additive A [C2A]
- Control diet + level 1 of additive B [C1B]
- Control diet + level 2 of additive B [C2B]

```
turkeys <- read.csv("C:/Users/radje/Desktop/USB/WKU some/Statistical
Methods/Statistical Methods. Everything/turkeys.csv")
library(lawstat)
attach(turkeys)
```

Second,

Because this lends itself to an ANOVA, I will conduct an ANOVA to determine if there is a significant difference in the mean weight gain of turkeys/subjects for the five

diets/treatment. This requires the verification of all the necessary assumptions for an Anova.

```
boxplot(Gain~Group,horizontal=T,xlab="Type of Diet",ylab="Group")
```
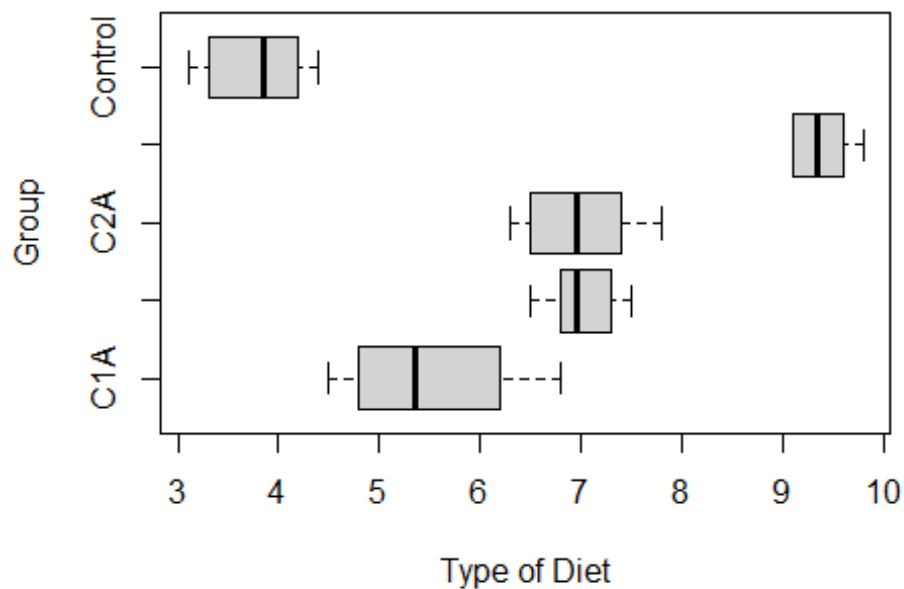


```
levene.test(Gain, Group)

##
##  Modified robust Brown-Forsythe Levene-type test based on the absolute
##  deviations from the median
##
## data:  Gain
## Test Statistic = 2.2257, p-value = 0.09506
```

The boxplots show fairly similar amounts of spread in the distributions. Furthermore, since the BFL test has a p value of 0.09506, we can assume equal population variances.

The normal probability plot of the residuals is not perfectly linear, but it is not so bad as to raise major cause for concern about the normality assumption. Since they randomly assigned the turkeys into the diet groups, we can assume independent random samples.

```
matrix(Gain, nrow=6,ncol=5,dimnames = list(c(),c("Control",
"C1A","C2A","C1B","C2B")))

##       Control C1A C2A C1B C2B
## [1,]      4.1 5.2 6.3 6.5 9.5
## [2,]      3.3 4.8 6.5 6.8 9.6
```

```
## [3,]      3.1 4.5 7.2 7.3 9.2
## [4,]      4.2 6.8 7.4 7.5 9.1
## [5,]      3.6 5.5 7.8 6.9 9.8
## [6,]      4.4 6.2 6.7 7.0 9.1

turkey.mat<-matrix(Gain, nrow=6,ncol=5,dimnames = list(c(),c("Control",
"C1A","C2A","C1B","C2B")))
turkey.df<-as.data.frame(turkey.mat)
attach(turkey.df)
turkey.df

##    Control C1A C2A C1B C2B
## 1      4.1 5.2 6.3 6.5 9.5
## 2      3.3 4.8 6.5 6.8 9.6
## 3      3.1 4.5 7.2 7.3 9.2
## 4      4.2 6.8 7.4 7.5 9.1
## 5      3.6 5.5 7.8 6.9 9.8
## 6      4.4 6.2 6.7 7.0 9.1

all.turkey.df<-c(Control,C1A,C2A,C1B,C2B)
means.turkey<-c(rep(mean(Control),6), rep(mean(C1A),6), rep(mean(C2A),6),
rep(mean(C1B),6), rep(mean(C2B),6))
res.turkeys<-all.turkey.df-means.turkey
qqnorm(res.turkeys, pch=19)
```


Normal Q-Q Plot
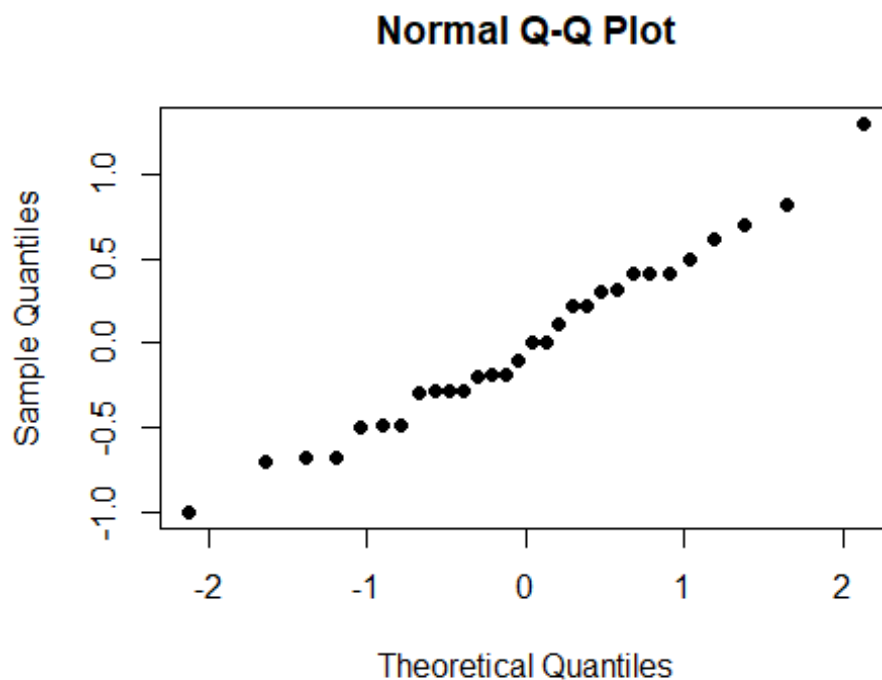
#Evaluation/Analyses 1.

$$H_0: \mu_c = \mu_{CIA} = \mu_{C2A} = \mu_{C1B} = \mu_{C2B}$$

$$H_a: \text{Not all means are equal}$$

2. For the reasons explained above, I will conduct an ANOVA here.
3. F = 81.67(see output below)
4. p-value $\approx$ 0 (see output below)
5. Reject H0

At the 0.05 significance level (p-value $\approx$ 0), there is sufficient evidence to conclude that there is a difference in the mean amount of weight gained after the turkeys were fed with the five different diets.

```
summary(aov(Gain~Group))
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## Group          4 103.04  25.759   81.67 5.6e-14 ***
## Residuals     25   7.89   0.315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Third

Now, I use an appropriate multiple-comparisons procedure to determine which pairs of means are significantly different. I use Tukey because I am doing all comparison which leads to 5C2 (combination) number of tests.

```
TukeyHSD(aov(all.turkey.df~Group))
```

```
##    Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = all.turkey.df ~ Group)
##
## $Group
##                     diff        lwr        upr      p adj
## C1B-C1A       1.50000000  0.5477409  2.4522591 0.0008574
## C2A-C1A       1.48333333  0.5310743  2.4355924 0.0009765
## C2B-C1A       3.88333333  2.9310743  4.8355924 0.0000000
## Control-C1A  -1.71666667 -2.6689257 -0.7644076 0.0001572
## C2A-C1B      -0.01666667 -0.9689257  0.9355924 0.9999983
## C2B-C1B       2.38333333  1.4310743  3.3355924 0.0000010
## Control-C1B  -3.21666667 -4.1689257 -2.2644076 0.0000000
## C2B-C2A       2.40000000  1.4477409  3.3522591 0.0000009
## Control-C2A  -3.20000000 -4.1522591 -2.2477409 0.0000000
## Control-C2B  -5.60000000 -6.5522591 -4.6477409 0.0000000
```

The p-values indicate that there are significantly different pairs:

- C1B-C1A

- C2A-C1A
- C2B-C1A
- Control-C1A
- C2B-C1B
- Control-C1B
- C2B-C2A
- Control-C2A
- Control-C2B

Fourth,

I will also want to determine which of the additive diets result in a significantly different mean weight gain in comparison to the control diet. The appropriate test hes the Dunnett test.

```
all.turkey<-c(Control, C1A, C2A, C1B, C2B)
Group

##  [1] "Control" "Control" "Control" "Control" "Control" "Control" "C1A"
##  [8] "C1A"     "C1A"     "C1A"     "C1A"     "C1A"     "C2A"     "C2A"
## [15] "C2A"     "C2A"     "C2A"     "C2A"     "C1B"     "C1B"     "C1B"
## [22] "C1B"     "C1B"     "C1B"     "C2B"     "C2B"     "C2B"     "C2B"
## [29] "C2B"     "C2B"

DunnettTest(all.turkey, as.factor(Group), control="Control")

##
##   Dunnett's test for comparing several treatments with a control :
##     95% family-wise confidence level
##
## $Control
##                diff    lwr.ci   upr.ci     pval
## C1A-Control 1.716667 0.8710423 2.562291 5.3e-05 ***
## C1B-Control 3.216667 2.3710423 4.062291 4.8e-10 ***
## C2A-Control 3.200000 2.3543756 4.045624 1.5e-11 ***
## C2B-Control 5.600000 4.7543756 6.445624 < 2e-16 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The additive diets/treatments that differ significantly in mean weight compared to the control are CA1, C1B, C2A and C2B. Since each of those corresponding intervals are entirely positive, I conclude that those additive diets produced a significantly larger mean weight gain than the control diet. Even though these are two-sided intervals (rather than one-sided).

Fifth,

Next will be to construct identify contrasts and use the most appropriate multiple-comparison procedure to address the following comparisons. Which comparisons yield significant results (i.e., show a significant difference)? - i. Compare the mean for the control diet to the average of the means for the additive diets. - ii. Compare the mean gain of the additive A diets to the control diet. - iii. Compare the mean gain of the additive B diets to the control diet. - iv. Compare the mean gain of the additive A diets to the additive B diets.

NB: Using R defined levels in R

Let a_1, a_2, a_3, a_4, and a_5 be the constants for the Control, C1A, C2A, C1B, and C2B means of the weights of turkeys.

a.

$$4\mu_{Control} + (-1)\mu_{C1A} + (-1)\mu_{C2A} + (-1)\mu_{C1B} + (-1)\mu_{C2B} = 0$$

The coefficients are:

$$a_1 = 4, \quad a_2 = -1, \quad a_3 = -1, \quad a_4 = -1, \quad a_5 = -1.$$

b.

$$-2\mu_{Control} + 1\mu_{C1A} + 1\mu_{C2A} + 0\mu_{C1B} + 0\mu_{C2B} = 0$$

The coefficients are:

$$a_1 = -2, \quad a_2 = 1, \quad a_3 = 1, \quad a_4 = 0, \quad a_5 = 0.$$

c.

$$-2\mu_{Control} + 0\mu_{C1A} + 0\mu_{C2A} + 1\mu_{C1B} + 1\mu_{C2B} = 0$$

The coefficients are:

$$a_1 = -2, \quad a_2 = 0, \quad a_3 = 0, \quad a_4 = 1, \quad a_5 = 1.$$

d.

$$0\mu_{Control} + 1\mu_{C1A} + 1\mu_{C2A} + (-1)\mu_{C1B} + (-1)\mu_{C2B} = 0$$

The coefficients are:

$$a_1 = 0, \quad a_2 = 1, \quad a_3 = 1, \quad a_4 = -1, \quad a_5 = -1.$$

Thus,

```
library(DescTools)

a.vec<-c(1/4, 1/4, 1/4, 1/4, -1)
```

```
b.vec<-c(1/2, 0, 1/2, 0, -1)
c.vec<-c(0, 1/2, 0, 1/2, -1)
d.vec<-c(1/2, -1/2, 1/2, -1/2, 0)

cont<-cbind(a.vec,b.vec,c.vec,d.vec)

ScheffeTest(aov(all.turkey~Group), which = "Group", contrasts = cont)

##
##   Posthoc multiple comparisons of means: Scheffe Test
##      95% family-wise confidence level
##
## $Group
##                               diff     lwr.ci     upr.ci     pval
## C1A,C1B,C2A,C2B-Control   3.433333   2.581816   4.284850 4.7e-11 ***
## C1A,C2A-Control           2.458333   1.525543   3.391123 2.5e-07 ***
## C1B,C2B-Control           4.408333   3.475543   5.341123 1.4e-12 ***
## C1A,C2A-C1B,C2B          -1.950000  -2.711620  -1.188380 4.3e-07 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: Significant contrasts are indicated by highlighted p-values and output in the above output. There is significant evidence that the mean of the control diet is different from the average of the means of the additive diets in terms of weight gain. Furthermore, the entirely positive interval suggests that the average of the means of the additive diets is significantly higher than the control. There is also significant evidence that the mean weight gain for the additive A diets is significantly different than the Control. Again, the entirely positive interval suggests that additive A are more effective than the Control in terms of weight gain. Similarly, there is also significant evidence that the mean weight gain for the additive B diets is significantly different than the Control. Again, the entirely positive interval suggests that additive B are more effective than the Control in terms of weight gain. Finally, there is also significant evidence that the mean weight gain for the additive A diets is significantly different than the additive B diets. Again, the entirely negative interval suggests that additive B are more effective than the additive A in terms of weight gain.