Simonson_HW6

Martin A. Simonson

March 1, 2019

1.) Long Term Diet Study

Data are provided in *dietstudy.csv*. The response is the kg lost per subject 24 months after the start of the diet for the 272 people who completed the study.

a.) Test the null hypothesis that the mean weight loss is the same in the three diets. Write down null and alternate hypotheses using well-defined notations. Report your test-statistic (along with DFs), the p-value, and short conclusion (no statement of effeect).

- Let μ_1 represent the mean kg lost after 24 months for low-carbohydrate group, let μ_2 represent the mean kg lost after 24 months for the low-fat group, and let μ_3 represent the mean kg lost after 24 months in the Mediterranean group.

$H_0 : \mu_1 = \mu_2 = \mu_3$

vs.

$H_a : H_0$ is false

```
fit<-lm(WtLoss24 ~ Group, data=df)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: WtLoss24
##            Df Sum Sq Mean Sq F value  Pr(>F)
## Group       2  216.9 108.430  3.2358 0.04086 *
## Residuals 269 9013.9  33.509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The F-value is 3.2358 and the p-value is 0.04086. The numerator df is 2 and the denominator df is 269. We have some evidence to suggest that not all diets resulted in the same mean kg lost per subject after 24 months.

b.) Test for a difference between the low-fat and mediterranean diets. Write down the null and alternate hypotheses in terms of the contrast. Provide the estimate of teh contrast, its standard error, the t-ratio, degrees of freedom, and p-value. Wrie a short conclusion with direction of effect if applicable).

- Let μ_2 represent the mean kg lost after 24 months for the low-fat group, and let μ_3 represent the mean kg lost after 24 months in the Mediterranean group.

$H_0 : \mu_2 = \mu_3$

vs.

$H_a : \mu_2 \neq \mu_3$

```
diet.emm <- emmeans(fit, ~ Group)
diet.emm
```

```
##  Group            emmean    SE  df lower.CL upper.CL
##  Low-Carbohydrate   5.49 0.628 269     4.25     6.72
```

```
##   Low-Fat             3.30 0.597 269     2.13     4.48
##   Mediterranean       4.60 0.600 269     3.42     5.78
##
## Confidence level used: 0.95
```

```r
contrast(diet.emm,list("Low-Fat - Mediterranean"=c(0,1,-1)))
```

```
##   contrast                estimate    SE  df t.ratio p.value
##   Low-Fat - Mediterranean     -1.3 0.847 269 -1.533  0.1264
```

The estimate of the contrast is -1.3 with a standard error of 0.847. The t-ratio is -1.533 with 269 degrees of freedom. The p-value is 0.1264, therefore, we have no evidence against the null hypothesis and cannot conclude that the difference in mean kg lost per subject after 24 months differs between the low-fat and Mediterranean diets.

c.) Test for a difference between the low-fat and the Mediterranean diets. Write down the null and the alternative hypotheses in terms of the contrast. Provide an estimate of the contrast, it's standard error, the t-ratio, degrees of freedom, and the p-value. Write a short conclusion (include direction of effect if applicable).

- Let μ_3 represent the mean kg lost after 24 months in the Mediterranean group and let μ_4 represent the mean kg lost after 24 months for both low-carb and low-fat diets combined.

$H_0 : \mu_3 = \mu_4$

vs.

$H_a : \mu_3 \neq \mu_4$

```r
contrast(diet.emm, list("Mediterranean - Mean of other two"=c(0.5,0.5,-1)))
```

```
##   contrast                          estimate   SE  df t.ratio p.value
##   Mediterranean - Mean of other two   -0.206 0.74 269 -0.279  0.7805
```

The estimate of the contrast is -0.206 with a standard error of 0.74. The t-ratio is -0.279 with 269 degrees of freedom. The p-value is 0.7805 so there is no evidence against the null hypothesis that there is no difference in mean kg lost over a 24 month period between the Mediterranean diet and the average of the low-fat and low-carb diets.

d.) Compared to other long-term diet studies, dropout was low in this study. Even so, 50 of the initial 322 subjects did not complete the study. Why might dropout be a concern?

- Droupout is a potential concern because it would increase the likelihood of a false rejection of the null hypothesis based on p-value (Type I error). Individuals that were not losing weight during the course of the diet study may drop out and affect the mean change in weight for the whole group, leading to detection of an effect that does not exist.

2.) Cereal Leaf Beetles

The presence of harmful insects in farm fields is detected by erecting boards covered with a sticky material and then examining the insects trapped on the board. Some colors are more attractive than others. In a hypothetical experiment aimed at determining the best color for attracting cereal leaf beetles, ten randomly selected boards in each of four colors (yellow, gray, green, and blue) were randomly placed in a field of oats in July, and the number of beetles trapped on the boards were counted. For some reason, one of the blue boards' information was lost. Data are available in *beetles.csv*.

a) Test the null hypothesis of no difference in attractiveness (measured by the mean number of trapped beetles) between the four colors. List the hypotheses, give the F-ratio, p-value and a conclusion within the context of these data.

- Let μ_1, μ_2, μ_3, and μ_4 represent the mean number of trapped beetles on sticky boars colored Blue, Gray, Green, and Yellow, respectively. Therefore,

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

vs.

$H_a : H_0$ is false and at least one board color has a different mean number of trapped beetles than the other colors.

```
beetles<-read.csv("Data/beetles.csv")
summary(beetles)
```

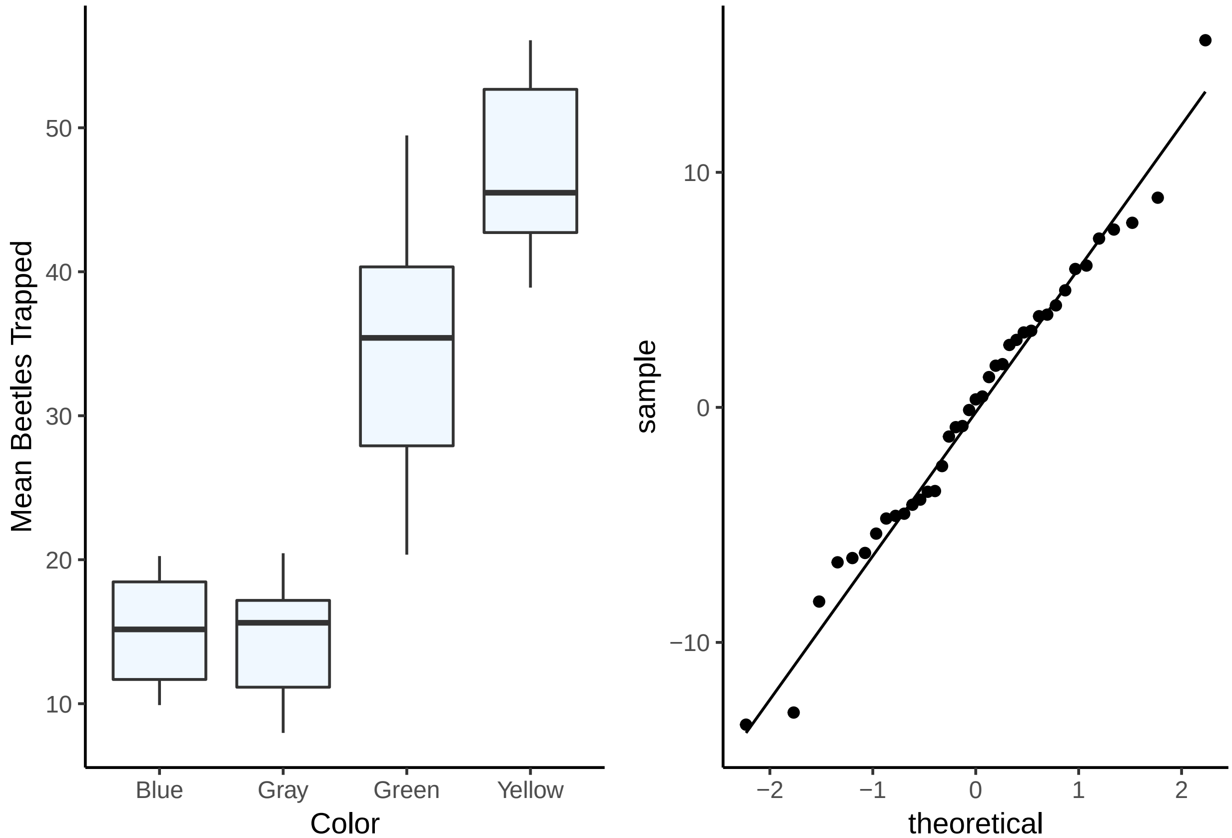
```
##    NumBeetles        Color
##  Min.   : 7.969   Blue  : 9
##  1st Qu.:15.747   Gray  :10
##  Median :20.449   Green :10
##  Mean   :28.029   Yellow:10
##  3rd Qu.:41.558
##  Max.   :56.079
```

```
fit2<-lm(NumBeetles ~ Color, data = beetles)
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: NumBeetles
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Color      3 7276.0 2425.34  61.237 5.259e-14 ***
## Residuals 35 1386.2   39.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-ratio is 61.237 with 3 and 35 degrees of freedom. The p-value is $<< 0.0001$ so we have very strong evidence to reject the null hypothesis and conclude that there is a true difference in the mean number of trapped beetles between blue, gray, green, and yellow boards.

b.) Do these data violate any assumptions of the test from (2a.)? Support your answers with the appropriate graphs, comment on the validity of each of the needed assumptions.



The observations within and between groups are independent and randomly assigned. The assumption of equal variance is not met; the sizes of the box-and-whisker pltos are unequal across all four groups. The data fall along a straight line in the QQ-plot so the normality assumption is visually confirmed.

Write down the null and alternative hypotheses in terms of the contrasts. Provide the estimates of the contrasts, their standard errors, the t-ratios, degrees of freedom and the p-values and write a short conclusion (including direction of effect as applicable).

c.) Is there any difference between the yellow and green color?

Let μ_1, μ_2 represent the mean number of trapped beetles on sticky boars colored Green and Yellow, respectively. Therefore,

$H_0 : \mu_1 = \mu_2$

vs.

$H_a : \mu_1 \neq \mu_2$

```
diet.emm2 <- emmeans(fit2,~Color)
contrast(diet.emm2, list("Green - Yellow" = c(0,0,1,-1)))
```

```
##  contrast       estimate   SE df t.ratio p.value
##  Green - Yellow    -13.3 2.81 35 -4.729  <.0001
```

The estimate of the contrast is -13.3 with a standard error of 2.81. The t-ratio is -4.729 with 35 degrees of freedom. The p-value is <0.001, therefore, we have very strong evidence against the null hypothesis. Data indicates that the green color boards captured fewer bugs than the yellow boards.

d.) Any difference between the gray and the average of the other colors?

Let μ_1 represent the mean number of trapped beetles on gray boards, and let μ_2 represent the mean number of trapped beetles on sticky boars colored blue, Green, and Yellow combined.

$H_0 : \mu_1 = \mu_2$

vs.

$H_a : \mu_1 \neq \mu_2$

```
contrast(diet.emm2, list("Gray - Mean of other Colors"=c(0.33,-1,0.33,0.33)))
```

```
##  contrast                    estimate  SE df t.ratio p.value
##  Gray - Mean of other Colors     17.2 2.3 35 7.474   <.0001
```

The estimate of the contrast is 17.2 with a standard error of 2.3. The t-ratio is 7.474 with 35 degrees of freedom. The p-value is < 0.0001 so we have very strong evidence to reject the null hypothesis that there is no difference in mean beetles trapped between gray boards and an average of the other three board colors. Data suggest that the gray boards capture more beetles, on average, than the average of the other three colored boards.

3.) Calf Leanness

Consider an experiment to test the effect of a drug on lean percentage in calves. 25 calves are to be used in the experiment. Five treatments are randomly assigned to the 25 calves in a completely randomized manner so that there are 5 calves for each treatment, which consists of an injection of a solution containing the different concentrations of the drug. Treatment 1 contains 1 mg/L; treatment 2 contains 2 mg/L; treatment 3 contains 4 mg/L; treatment 4 contains 8 mg/L; and treatment 5 contains 16 mg/L of the drug. The data are provided in *LeanPerc.csv*

a.) Could the mean lean percentage for the five groups be the same? perform the one-way ANOVA F-test of the lean percentages. List the hypotheses, give the F-ratio, p-value, and a conclusion within the context of the data.

- Let μ_1, μ_2, μ_3, μ_4, and μ_5 represent the mean lean percentage for treatments 1:5, respectively, where drug concentration varies (increasing exponentially) by treatment.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

vs.

$H_a : H_0$ is false and at the mean lean percentage for at least one treatment is different than the others.

```
df<-read.csv("Data/LeanPerc.csv", header=T)
df$Treatment<-as.factor(df$Treatment)
str(df)
```

```
## 'data.frame':    25 obs. of  2 variables:
##  $ Treatment: Factor w/ 5 levels "1","2","4","8",..: 1 1 1 1 1 2 2 2 2 2 ...
##  $ LeanPerc : int  40 32 35 42 29 50 37 34 33 46 ...
```

```
fit3<- lm(LeanPerc ~ Treatment, data = df)
anova(fit3)
```

```
## Analysis of Variance Table
##
## Response: LeanPerc
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Treatment  4 4539.8 1134.96  12.546 2.877e-05 ***
## Residuals 20 1809.2   90.46
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-ratio is 12.546 with 4 and 20 degrees of freedom. The p-value is > 0.001 so we have very strong evidence to reject the null hypothesis that no difference in mean lean percentage exists between the five treatments.

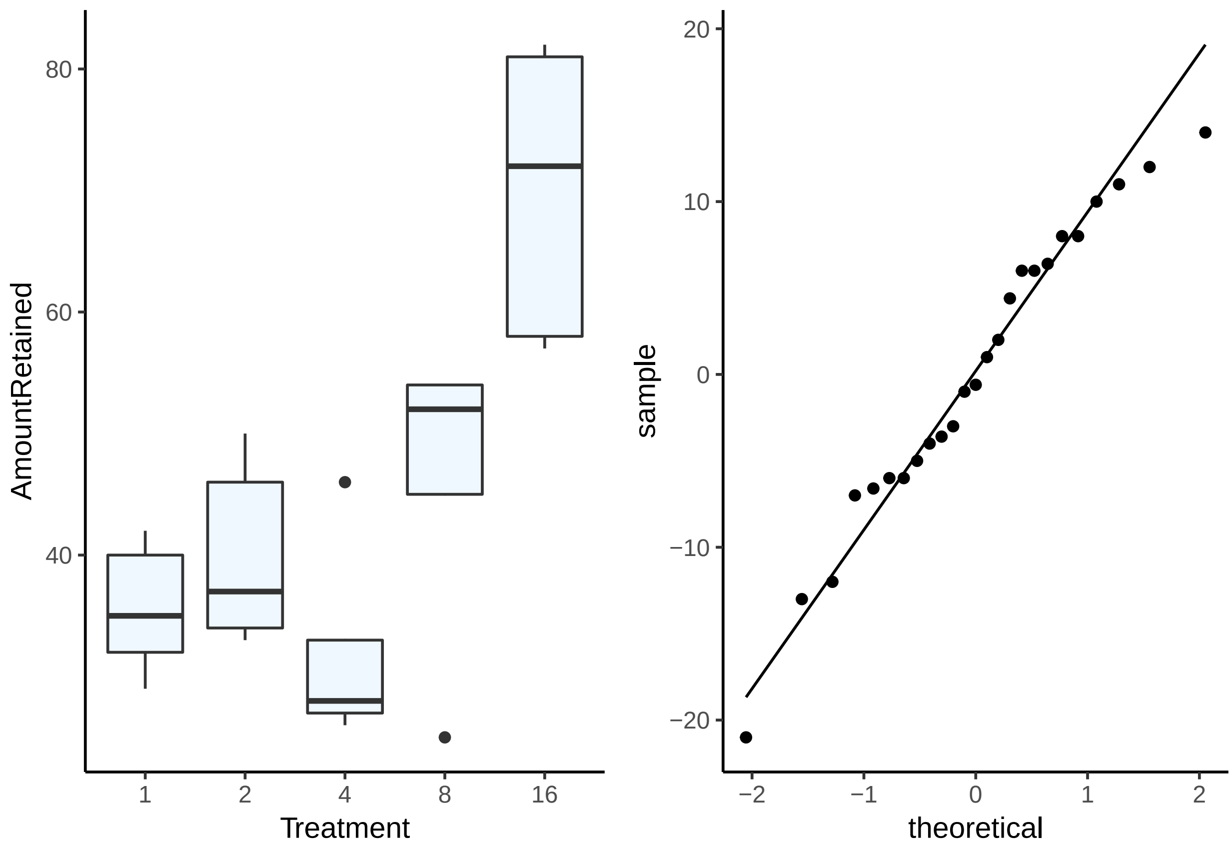
b.) show a table of group means and group standard deviations. Similar to the two-sample problems, the standard deviations of the observations in each group can be compared to assess whether the constant standard deviation assumption is approximately satisfied. The ratio of the largest and smallest standard deviation should be no more than 3. Is that the case here?

```
df.emm<-emmeans(fit3, ~ Treatment)
df.emm
```

```
##  Treatment emmean   SE df lower.CL upper.CL
##  1           35.6 4.25 20     26.7     44.5
##  2           40.0 4.25 20     31.1     48.9
##  4           32.0 4.25 20     23.1     40.9
##  8           46.0 4.25 20     37.1     54.9
##  16          70.0 4.25 20     61.1     78.9
##
## Confidence level used: 0.95
```

There are no instances wehere the ratio of the standard deviations is greater than 3, indeed, the standard deviations all appear to be equal because the sample size is the same accross all five treatments.

c.) Graphically check the assumptions.



The widths of the boxes and lengths of the whiskers are uneven, so the equal variance assumption is violated. There is also some curvature at the extremes of the QQ plots so the normality assumption also appears to be violated.

4.) Balloon Angioplasty

In a study of balloon angioplasty, patients with coronary artery disease were randomly assigned to one of four treatment groups: placebo, probucol (an experimental drug), multivitamins (a combination of beta carotene, vitamin E, and vitamin C), or probucol combined with multivitamins. Balloon angioplasty was performed on each of the patients. Later, "minimal luminal diameter" (a measurement of how well the angioplasty did in dilating the artery) was recorded for each of the patients. Summary statistics are given in the following table:

value	Placebo	Probucol	Multivitamins	Probucol and Multivitamins
n	62	58	54	56
Mean	1.43	1.79	1.40	1.54
SD	0.58	0.45	0.55	0.61

a) Complete the ANOVA table below, showing all the details of how the corresponding values were obtained. p-values can be obtained using the "P from F" item at https://www.graphpad.com/quickcalcs/pvalue1.cfm

- *Degrees of Freedom*: 229 = N-1 = 230. g = 4. Between DF = g-1 = 4-1 = 3. Within DF = N-g = 230-4 = 226.
- *SumSquares*: $S_p = [(n1-1)S1^2 + (n2-1)S2^2 + (n3-1)S3^2 + (n4-1)S4^2]/[(n1-1) + (n2-1) + (n3-1) + (n4-1)]$ Therefore, $S_p = 0.3034$ = MSE.
- *SSres* = MSE(N-g) = 68.57.
- *SSBet* = SSTot-SSRes = 5.42.
- *MSBet* = SSBet / (g-1) = 1.81.
- *F* = MSBet / MSE = 5.97

Source	DF	SS	MS	F	p-value
Between	3	5.42	1.81	5.97	0.0006
Within	226	68.57	0.3034	-	-
Total	229	73.99	0.3231	-	-

b.) Conduct an F-test to determine if there is evidence tha the means are not all the same among the treatments. Write down the null and alternate hypotheses, and report the test statistic, p-value and degrees of freedom. Give a conclusion within the context of the study.

- Let μ_1, μ_2, μ_3, and μ_4 represent the mean minimal luminal diameter for Placebo, Probucol, Multivitamins, and Probucol and Multivitamins combined (respectively).

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

vs.

H_a : H_0 is false and at the mean minimal luminal diameter is not equal between treatment groups.

- **Answer**: The F-statistic is 5.97 with 3 and 226 degrees of freedom. The p-value is 0.0006, indicating we have strong evidence to reject the null hypothesis and we can conclude that the mean minimal luminal diameters are not all equal between Placebo, Probucol, Multivitamin, and Probucul+Multivitamin groups.

5.) Conceptual ANOVA Questions

Use a dataset (e.g. the one in Problem #2) and a software to check.

a.) What would happen to the p-value in an ANOVA if a constant (say, 29) was added to each data point?

- **Answer**: The p-value would stay the same. Although the magnitude of each observation has changed, it has changed proportionally for all observations. Therefore, the *difference* between treatments remains unchanged. An ANOVA table is provided below, which is identical to the ANOVA table from question 2.

```
beetles<-read.csv("Data/beetles.csv")
beetles$NumBeetles2<-beetles$NumBeetles+29
summary(beetles)
```

```
##    NumBeetles        Color     NumBeetles2
##  Min.   : 7.969   Blue  : 9   Min.   :36.97
##  1st Qu.:15.747   Gray  :10   1st Qu.:44.75
##  Median :20.449   Green :10   Median :49.45
##  Mean   :28.029   Yellow:10   Mean   :57.03
##  3rd Qu.:41.558               3rd Qu.:70.56
##  Max.   :56.079               Max.   :85.08
```

```
new.fit<-lm(NumBeetles2~Color,data=beetles)
anova(new.fit)
```

```
## Analysis of Variance Table
##
## Response: NumBeetles2
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Color      3 7276.0 2425.34  61.237 5.259e-14 ***
## Residuals 35 1386.2   39.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b.) What would happen to the p-value in an ANOVA if the dataset was doubled by duplicating the data? (That is, you create a new dataset made of two copies of the data, thus doubling the number of observations, but keeping the same number of treatment groups).

- **Answer**: The p-value should decrease. Although the data is "doubled" the main effect of this change is the increase of sample size per group, which reduces estimates of error and leads to lower p-values.

```
beetles<-read.csv("Data/beetles.csv")
# rbind to attach data frame to bottom of itself
beetles2<-rbind(beetles,beetles)
# testing length of new data frame compared to old
length(beetles2$NumBeetles)-length(beetles$NumBeetles)
```

```
## [1] 39
```

```
# New frame is 39 observations longer than old, which had 39 observations.
# we're good.
```

```r
new.fit.2<-lm(NumBeetles~Color, data = beetles2)
anova(new.fit.2)
```

```
## Analysis of Variance Table
##
## Response: NumBeetles
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Color      3 14552.0  4850.7  129.47 < 2.2e-16 ***
## Residuals 74  2772.4    37.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```