# IAF602: Statistical Methods for Data Analytics

## Assignment 3 Worksheet

1. Refer to Case Study 5.1.1. Verify the values obtained in the Statistical Conclusions on p.116.

> t.test(case0501$Lifetime[case0501$Diet == "N/R50"], case0501$Lifetime[case0501$Diet == "N/N85"])

Welch Two Sample t-test

data: case0501$Lifetime[case0501$Diet == "N/R50"] and case0501$Lifetime[case0501$Diet == "N/N85"]

t = 8, df = 122, p-value = 1e-13

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

7.34 11.87

sample estimates:

mean of x mean of y

42.3 32.7

For our t.test of Diet N/R50 against Diet N/N85 we find a one sided p-value that is similar to the one in the book (virtually zero), and a 95% confidence interval from 7.3 to 11.9. This is overwhelming evidence that the change in Diet increased the lifespan of the rats.

> t.test(case0501$Lifetime[case0501$Diet == "R/R50"], case0501$Lifetime[case0501$Diet == "N/R50"])

Welch Two Sample t-test

data: case0501$Lifetime[case0501$Diet == "R/R50"] and case0501$Lifetime[case0501$Diet == "N/R50"]

t = 0.5, df = 124, p-value = 0.6

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.95 3.13

sample estimates:

mean of x mean of y

42.9 42.3

There is no evidence that the change in diet from R/R50 to N/R50 increases or decreases lifetime for the rats. We find a one sided p value of 0.3 and a 95% confidence interval from -1.95 to 3.13.

> t.test(case0501$Lifetime[case0501$Diet == "N/R40"], case0501$Lifetime[case0501$Diet == "N/R50"])

Welch Two Sample t-test

data: case0501$Lifetime[case0501$Diet == "N/R40"] and case0501$Lifetime[case0501$Diet == "N/R50"]

t = 2, df = 129, p-value = 0.03

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.318 5.321

sample estimates:

mean of x mean of y

45.1 42.3

We find a one sided p value of 0.015 for our test of N/R40 against N/R50. This is moderate evidence that there is an increase in lifespan for the rats due to this change in diet. We also calculate a 95% confidence interval from 0.32 to 5.32.

> t.test(case0501$Lifetime[case0501$Diet == "N/R50"], case0501$Lifetime[case0501$Diet == "lopro"])

Welch Two Sample t-test

data: case0501$Lifetime[case0501$Diet == "N/R50"] and case0501$Lifetime[case0501$Diet == "lopro"]

t = 2, df = 123, p-value = 0.05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.0133 5.2097

sample estimates:

mean of x mean of y

42.3 39.7

We find a one sided p value for this test to be 0.025, which is moderately strong evidence that this change in diet for the rats increased lifetime. We find a 95% confidence interval from 0.01 to 5.21 for the decrease in lifetime comparing N/R50 to lopro.

> t.test(case0501$Lifetime[case0501$Diet == "N/N85"], case0501$Lifetime[case0501$Diet == "NP"])

Welch Two Sample t-test

data: case0501$Lifetime[case0501$Diet == "N/N85"] and case0501$Lifetime[case0501$Diet == "NP"]

t = 5, df = 94, p-value = 7e-06

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

3.09 7.49

sample estimates:

mean of x mean of y

32.7 27.4

We find a p value of virtually zero, which means there is convincing evidence that the NP diet mice live longer than the N/N85 diet mice. We also calculate a 95% confidence interval from 3.09 to 7.49.

2. Refer again to Case Study 5.1.1. Recompute the *p*-values and confidence intervals, but this time using a Bonferroni adjustment to adjust for the fact that there are five comparisons, ensuring that the familywise confidence level is at least 95%. Write a set of Statistical Conclusions for the new results.

> Bon.5.1 <- contrast(emm.fit5.1,method="pairwise",adjust="Bonferroni")

> confint(Bon.5.1)

contrast estimate SE df lower.CL upper.CL

(N/N85) - (N/R40) -12.43 1.24 343 -16.08 -8.77

(N/N85) - (N/R50) -9.61 1.19 343 -13.12 -6.10

(N/N85) - NP 5.29 1.30 343 1.44 9.13

(N/N85) - (R/R50) -10.19 1.26 343 -13.91 -6.48

(N/N85) - lopro -6.99 1.26 343 -10.71 -3.28

(N/R40) - (N/R50) 2.82 1.17 343 -0.64 6.28

(N/R40) - NP 17.71 1.29 343 13.91 21.52

(N/R40) - (R/R50) 2.23 1.24 343 -1.44 5.90

(N/R40) - lopro 5.43 1.24 343 1.76 9.10

(N/R50) - NP 14.90 1.24 343 11.23 18.56

(N/R50) - (R/R50) -0.59 1.19 343 -4.12 2.94

(N/R50) - lopro 2.61 1.19 343 -0.92 6.14

NP - (R/R50) -15.48 1.31 343 -19.35 -11.62

NP - lopro -12.28 1.31 343 -16.15 -8.42

(R/R50) - lopro 3.20 1.26 343 -0.53 6.93

Confidence level used: 0.95

Conf-level adjustment: bonferroni method for 15 estimates

> Bon.5.1

contrast estimate SE df t.ratio p.value

(N/N85) - (N/R40) -12.43 1.24 343 -10.060 <.0001

(N/N85) - (N/R50) -9.61 1.19 343 -8.090 <.0001

(N/N85) - NP 5.29 1.30 343 4.070 0.0010

(N/N85) - (R/R50) -10.19 1.26 343 -8.110 <.0001

(N/N85) - lopro -6.99 1.26 343 -5.570 <.0001

(N/R40) - (N/R50) 2.82 1.17 343 2.410 0.2490

(N/R40) - NP 17.71 1.29 343 13.780 <.0001

(N/R40) - (R/R50) 2.23 1.24 343 1.800 1.0000

(N/R40) - lopro 5.43 1.24 343 4.380 <.0001

(N/R50) - NP 14.90 1.24 343 12.010 <.0001

(N/R50) - (R/R50) -0.59 1.19 343 -0.490 1.0000

(N/R50) - lopro 2.61 1.19 343 2.190 0.4400

NP - (R/R50) -15.48 1.31 343 -11.850 <.0001

NP - lopro -12.28 1.31 343 -9.400 <.0001

(R/R50) - lopro 3.20 1.26 343 2.540 0.1750  
  
We are specifically looking at the 5 pairs,  
N/N85 and N/R50: confidence interval from 6.1 to 13.12, p value = <.0001  
R/R50 and N/R50: confidence interval from -4.12 to 2.94, p value = 1

N/R40 and N/R50: confidence interval from -0.64 to 6.28, p value = 0.2490

N/R50 and lopro: confidence interval from -0.92 to 6.14, p value = 0.44  
N/N85 and NP: confidence interval from 1.44 to 9.13, p value = 0.001

We now run a multiple comparison of means test, this time using the Bonferroni procedure. Previously we used the LSD procedure. This test varies from the LSD because the LSD is used for planned comparisons and controls an individual confidence interval. You can see our confidence intervals for the Bonferroni are wider than the LSD intervals. That is because the Bonferroni is controlling for a familywise confidence interval, so there are more intervals to be taken into account. Because we are controlling for a familywise and not individual intervals, our confidence intervals for the Bonferroni will be wider and our p values will be larger as well.

1. Refer one more time to Case Study 5.1.1. Use software to generate an analysis of variance table and add below.
   1. Identify each of the following values in/from the table:
      1. the pooled variance estimate
      2. the extra sum of squares for testing the separate-means model versus the equal-means model
      3. the sum of squares error, or SSE (sum of squares residuals from the full model)

> anova(lm(Lifetime~Diet, data = case0501))

Analysis of Variance Table

Response: Lifetime

Df Sum Sq Mean Sq F value Pr(>F)

Diet 5 12734 2546.8 57.104 < 2.2e-16 \*\*\*

Residuals 343 15297 44.6

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

1. The sqrt of the mean square for the full model will give us our pooled standard deviation.

> sqrt(44.6)

[1] 6.678323

ii. The extra sum of squares is the difference between the total sum of squares and the full model sum of squares

12734

iii. The SSE is the residuals from the full model

15297

b. Show how the ANOVA F-statistic is calculated using the values from (a) and their respective degrees of freedom.

The F-statistic can be calculated by finding the ratio of the mean squares from the reduced model against the full model.

> 2546.8 / 44.6

[1] 57.10314

**3. Extra Credit**: Complete Exercise 15, p.171.

a. We can calculate the pooled standard deviation by taking the sum of the standard deviations and multiplying them by the sum of the samples, and then dividing by the sum of the samples

> pooledsd = sum(teaching$SD \* teaching$n)/sum(teaching$n)

> pooledsd

[1] 4.436

b. To find the coefficients to use for a linear combination, we can find the average of all of our scores and compare groups 1-5 to it, and then multiply by 2.

> mean(average)

[1] 29.3

Group 1: 2(30.2 – 29.3) = 0.33

Group 2: 2(28.8 - 29.3) = -0.5

Group 3: 2(26.2 - 29.3) = -0.5

Group 4: 2(31.1 - 29.3) = 0.33

Group 5: 2(30.2 - 29.3) = 0.33

c. Now that we have our coefficients, we can compare groups 2 and 3 to groups 1,4, and 5. We estimate the linear combination by finding the difference in means for 2 and 3 in comparison to 1,4, and 5

> ((28.8+26.2)/2) - ((30.2+31.1+30.2)/3)

[1] -3

We estimate that groups 1,4, and 5 have a group mean that is 3 points higher than the group mean of groups 2 and 3.

We can now estimate our (y) parameter with our coefficients by finding (g).

> g = 0.33\*(30.2) – 0.5\*(28.8) – 0.5\*(26.2) + 0.33\*(31.1) +0.33\*(30.2)

> g

[1] 2.695

We now calculate our standard error

> step1 = (((0.33)^2)/9)+(((-0.5)^2)/9)+(((-0.5)^2)/9)+(((0.33)^2)/9)+(((0.33)^2)/9)

> step1

[1] 0.09185556

> step2 = sqrt(step1)

> step2

[1] 0.3030768

> step3 = 4.436\*step2

> step3

[1] 1.344449  
  
We can now find our t-statistic by dividing (g) by our SE

> 2.695/1.344

[1] 2.005208

We now calculate our confidence interval

> 3 + (2.005)\*(1.344)

[1] 5.69472

> 3 - (2.005)\*(1.344)

[1] 0.30528

We calculate a 95% confidence interval between 0.305 and 5.695 for the increase in points between the group average of groups 2 and 3 and the group average of groups 1,4 and 5.

4. Complete Exercise 17, p. 142.

* 1. Reproduce the ANOVA table here, filling in the missing values.

Between groups DF: 7  
Between groups sum of squares: 35,819  
Within groups mean square: 1462  
Between groups mean square: 5117  
F-statistic: 3.5  
P value = 0.0099

* 1. How many groups were there?

There are 32 groups in this test

* 1. Assess the degree of evidence that the group means are different. State the hypotheses, compute an appropriate test statistic and *p*-value and interpret.

We can compute our test statistic by finding the ratio of the mean squares, which gives us an F-statistic of 3.5. We can then use an F-distribution table to find a p value of 0.0099, which is very convincing evidence that at least one of the means in these groups is different than the others.

5. Refer to the Cancer Survival (<https://math.tntech.edu/e-stat/DASL/page24.html>) study. Suppose that the researchers have specified no preplanned means comparisons, but wish to make all pairwise comparisons to help interpret the results. Analyze the data and write a brief statistical report including a summary of statistical conclusions, graphical display(s), and a section describing the details of the particular methods used.

This data set includes data from a study done about the location of the cancerous cells in the body and the survival time in days. The two variables that we will be taking a look at is the survival time and the organ in which the cancer affected in each patient. There are a total of 64 patients that were sampled for this dataset. There are a total of 5 different levels, or organs, included in the organs variable.

First, we can explore the data by creating a boxplot of the average survival time for each of the organs in our dataset.

![Chart, box and whisker chart

Description automatically generated]()

If we take a look at this boxplot, it’s pretty clear to see which organ has the highest survival time of the five groups. But, its equally if not more important for us to know which organ has the lowest survival time. Between Bronchus, Colon, and Stomach, it’s not entirely clear how those organs compare to each other. Also, it is not great analytical practice to compare these plots that all have a much different skewness to them. While this boxplot is helpful, it could be improved if we were to take a look at the log transformed values of these plots.

![Chart, box and whisker chart

Description automatically generated]()

Now, if we take a look at the log transformed plots, they are much more normal, have fewer outliers, and are easier for us to compare the qualities of these plots such as the median, min, max, and quartiles of each variable. This overall helps to make our data more interpretable.

We now look to make all pairwise comparisons and find statistical conclusions about any differences between these groups. The researchers want to take all pairwise comparisons into account, and there are no planned comparisons. We can run a multiple comparison procedure called the Tukey-Kramer to make all unplanned pairwise comparisons within our groups. The difference between the TK procedure and the more common LSD procedure is that the TK is used when there are no planned comparisons going into the analysis (unlike the LSD), and it controls the familywise confidence interval, whereas the LSD controls individual confidence intervals. The TK procedure, while creating a larger level of confidence in our test, does increase the margin of error in our test. The following procedure can be seen below.

> TK.cancer

contrast estimate SE df t.ratio p.value

Breast - Bronchus 1184.3 259 59 4.571 0.0002

Breast - Colon 938.5 259 59 3.622 0.0053

Breast - Ovary 511.6 340 59 1.506 0.5631

Breast - Stomach 1109.9 274 59 4.046 0.0014

Bronchus - Colon -245.8 230 59 -1.070 0.8208

Bronchus - Ovary -672.7 318 59 -2.116 0.2271

Bronchus - Stomach -74.4 247 59 -0.302 0.9981

Colon - Ovary -426.9 318 59 -1.343 0.6659

Colon - Stomach 171.4 247 59 0.695 0.9568

Ovary - Stomach 598.3 330 59 1.811 0.3773

> confint(TK.cancer)

contrast estimate SE df lower.CL upper.CL

Breast - Bronchus 1184.3 259 59 455 1913

Breast - Colon 938.5 259 59 209 1668

Breast - Ovary 511.6 340 59 -445 1468

Breast - Stomach 1109.9 274 59 338 1882

Bronchus - Colon -245.8 230 59 -892 400

Bronchus - Ovary -672.7 318 59 -1567 222

Bronchus - Stomach -74.4 247 59 -769 620

Colon - Ovary -426.9 318 59 -1322 468

Colon - Stomach 171.4 247 59 -523 866

Ovary - Stomach 598.3 330 59 -332 1528

Above are all of our computed t ratios, p values, and confidence intervals for the unplanned pairwise functions of the TK procedure. There are only 3 of the 10 pairs that show strong statistical evidence of a difference in means between the two groups. They will be listed below.

The p value for Breast-Bronchus is 0.0002, which is convincing statistical evidence that there is a difference in means between the survival time of the two groups. We calculate a 95% confidence interval for the difference in survival time of the groups between 455 days and 1913 days.

The p value for Breast-Colon is 0.0053, which is very strong statistical evidence that there is a difference in means between the survival time of the two groups. We calculate a 95% confidence interval for the difference in survival time of the groups between 209 days and 1668 days.

The p value for Breast-Stomach is 0.0014, which is convincing statistical evidence that there is a difference in means between the survival time of the two groups. We calculate a 95% confidence interval for the difference in survival time of the groups between 338 days and 1882 days.

The p values for the remaining 7 comparisons are too large to even suggest that there is any difference in the means of the survival rate in days for those comparisons.

6. Submit solution to assigned practice exercise(s) as requested in Discussions on Canvas. (Do not submit again here)