ST 411/511 Homework 7

Due on March 4

Aashish Adhikari

Winter 2020

Instructions

This assignment is due by 11:59 PM, March 4, 2020 on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile (should you choose – recommended) using the provide .Rmd (R Markdown) template.** If you opt to not use R Markdown, please format your solutions in a similar manner as provided in this document. Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope. You should also use complete, gramatically correct sentences for your solutions.

Problems (25 points total)

Question 1

(a) (2 points) In comparing 10 groups, a researcher notices that the sample mean of group 7 is the largest and the sample mean of group 3 is the smallest. The researcher then decides to test the hypothesis that $\mu_7 - \mu_3 = 0$. Why should a multiple comparison procedure be used even though there is only one comparison being made?

It is a post-hoc comparison.i.e., the analysis is being done after the data was observed. As the question suggests, the researcher is comparing 10 groups and needs to adjust for the fact that the errors compound with the increasing number of statistical tests. (also called look-else where effect) Post-hoc methods like multiple comparison test allow us to account for this higher probability of error with increasing number of tests on the same data.

b) (2 points) When choosing coefficients for a contrast, does the choice of $\{C_1, C_2, \ldots, C_I\}$ give a different t-statistic than the choice of $\{4C_1, 4C_2, \ldots, 4C_I\}$? Explain why or why not.

If we scale the hypothesis appropriately as per the scaling of the coefficients, the t-statistic is still the same as the scaling happens on both sides of the equation. However, if the hypothesis equals zero, there is no requirement for scaling of the hypothesis in the first place. But I guess we might lose the property that "the coefficients sum to 1" if the multiplying constants are large enough.

Question 2 (Modified from *Sleuth* 6.17)

The relative head length (RHL) is measured for adders (a type of snake) on the Swedish mainland and on groups of islands in the Baltic Sea. Relative head length is adjusted for overall body length, determined separately for males and females. The data are below and additionally you know that the pooled estimate of standard deviation of the RHL measurements was 11.72 based on 230 degrees of freedom.

```
adder <- data.frame(Locality = c("Uppsala", "In-Fredeln", "Inre Hammnskar", "Norrpada",
                                 "Karringboskar", "Angskar", "SvenskaHagarna"),
                    SampleSize = c(21, 34, 20, 25, 7, 82, 48),
                    meanRHL = c(-6.98, -4.24, -2.79, 2.22, 1.27, 1.88, 4.98))
```

Consider the question: "Is the average of the mean relative head lengths for snakes on the Swedish mainland equal to the average of the mean relative head lengths for snakes on islands in the Baltic Sea?" Uppsala is the mainland, and the other six localities refer to islands in the Baltic Sea.

(a) (3 points) Give the coefficients for the linear combination you would use to test this question, and state the null hypothesis you would be testing using statistical notation.

For the mainland group, C(1) = 1 c(i) = -1/6 for i = each group except for the mainland

Null hypothesis: H0: Gamma = 1 * m(1) - (1/6) * (m(2)+m(3)+m(4)+m(5)+m(6)+m(7)) where m represents the true population mean

H0: Gamma = 0

In words, the null hypothesis : The average of the mean relative head lengths for snakes on Swedish mainland is equal to the average of mean relative head lengths for snakes on the islands in Baltic Sea.

(b) (4 points) What is the *t*-statistic for testing the hypothesis in part (a)? Please include in your answer your computed values of *g* and the standard error of *g*.

```
g <- 1 * (-6.98) - (1/6) * (-4.28-2.79+2.22+1.27+1.88+4.98)
g
```

```
## [1] -7.526667
```

g-value is -7.526.

```
sp2 = 11.72
c2sq = (-1/6)^2
c3sq = (-1/6)^2
c4sq = (-1/6)^2
c5sq = (-1/6)^2
c6sq = (-1/6)^2
c7sq = (-1/6)^2

se=sqrt(sp2*((1^2/21)+(c2sq)*((1/34)+(1/20)+(1/25)+(1/7)+(1/82)+(1/48))))
se
```

```
## [1] 0.8088454
```

```
tstat=(g-0)/(se)
tstat
```

```
## [1] -9.305445
```

The standard error is 0.8088. The t-statistic is -9.305.

(c) (2 points) Find the resulting *p*-value and state your conclusion.

```
p_val=2*(1-pt(9.313687,df=230))
p_val
```

```
## [1] 0
```

The p-value is 0. For any significance level alpha s.t. 0<alpha<=1, the obtained p-value is less than the significance level. Hence we reject the null hypothesis that the average relative head length of the means in the island groups is equal to the mean relative head length for the mainland group.

Question 3 (Modified from *Sleuth* 6.21)

Reconsider the education and future income data from your last homework (data: `ex0525`). Find *p*-values and 95% confidence intervals for the difference in means for all pairs of education groups in the following ways:

(a) (2 points) Using the Tukey-Kramer procedure.

```
data(ex0525)

#names(ex0525)
#head(ex0525)
#ggplot(data = ex0525, aes(x=Educ, y=Income2005))+geom_boxplot()

Education_group=lm(Income2005~Educ,data = ex0525)
Education_group
```

```
##
## Call:
## lm(formula = Income2005 ~ Educ, data = ex0525)
##
## Coefficients:
## (Intercept)    Educ13-15       Educ16      Educ<12      Educ>16
##       36865         8011        33132        -8563        39991
```

```
Education_group_Tukey_Test <- glht(Education_group, linfct = mcp(Educ = "Tukey"))
summary(Education_group_Tukey_Test)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
##
## Linear Hypotheses:
##                  Estimate Std. Error t value Pr(>|t|)
## 13-15 - 12 == 0      8011       2201   3.639   0.0024 **
## 16 - 12 == 0        33132       2571  12.885   <0.001 ***
## <12 - 12 == 0       -8563       4000  -2.141   0.1929
## >16 - 12 == 0       39991       2649  15.098   <0.001 ***
## 16 - 13-15 == 0     25121       2774   9.058   <0.001 ***
## <12 - 13-15 == 0   -16575       4133  -4.010   <0.001 ***
## >16 - 13-15 == 0    31980       2846  11.239   <0.001 ***
## <12 - 16 == 0      -41696       4341  -9.604   <0.001 ***
## >16 - 16 == 0        6858       3140   2.184   0.1764
## >16 - <12 == 0      48554       4388  11.066   <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```r
confint(Education_group_Tukey_Test)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
##
## Quantile = 2.7047
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                    Estimate    lwr         upr
## 13-15 - 12 == 0    8011.0607   2057.4471  13964.6743
## 16 - 12 == 0      33132.0768  26177.5614  40086.5922
## <12 - 12 == 0     -8563.4475 -19382.2599   2255.3648
## >16 - 12 == 0     39990.5665  32826.3993  47154.7337
## 16 - 13-15 == 0   25121.0161  17619.6563  32622.3759
## <12 - 13-15 == 0 -16574.5083 -27752.6931  -5396.3234
## >16 - 13-15 == 0  31979.5058  24283.3765  39675.6351
## <12 - 16 == 0    -41695.5244 -53437.3976 -29953.6512
## >16 - 16 == 0      6858.4897  -1635.6363  15352.6156
## >16 - <12 == 0    48554.0140  36686.7653  60421.2628
```

The requird values can be observed above. Note:James mentioned that it is okay to just show the result as it is.

(b) (2 points) Without adjusting for multiple comparisons.

```r
summary(Education_group_Tukey_Test, test=adjusted("none"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
##
## Linear Hypotheses:
##                    Estimate Std. Error t value Pr(>|t|)
## 13-15 - 12 == 0      8011       2201   3.639 0.000279 ***
## 16 - 12 == 0        33132       2571  12.885  < 2e-16 ***
## <12 - 12 == 0       -8563       4000  -2.141 0.032380 *
## >16 - 12 == 0       39991       2649  15.098  < 2e-16 ***
## 16 - 13-15 == 0     25121       2774   9.058  < 2e-16 ***
## <12 - 13-15 == 0   -16575       4133  -4.010 6.23e-05 ***
## >16 - 13-15 == 0    31980       2846  11.239  < 2e-16 ***
## <12 - 16 == 0      -41696       4341  -9.604  < 2e-16 ***
## >16 - 16 == 0        6858       3140   2.184 0.029062 *
## >16 - <12 == 0      48554       4388  11.066  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

```
confint(Education_group_Tukey_Test, calpha = univariate_calpha())
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
##
## Quantile = 1.9609
## 95% confidence level
##
##
## Linear Hypotheses:
##                   Estimate    lwr         upr
## 13-15 - 12 == 0    8011.0607   3694.7196  12327.4018
## 16 - 12 == 0      33132.0768  28090.0868  38174.0669
## <12 - 12 == 0     -8563.4475 -16407.0342   -719.8609
## >16 - 12 == 0     39990.5665  34796.5799  45184.5531
## 16 - 13-15 == 0   25121.0161  19682.5665  30559.4657
## <12 - 13-15 == 0 -16574.5083 -24678.6382  -8470.3783
## >16 - 13-15 == 0  31979.5058  26399.8492  37559.1623
## <12 - 16 == 0    -41695.5244 -50208.3257 -33182.7231
## >16 - 16 == 0      6858.4897    700.2894  13016.6899
## >16 - <12 == 0    48554.0140  39950.3161  57157.7120
```

The required values can be observed above. Note:James mentioned that it is okay to just show the result as it is.

(c) (3 points) What do you notice by comparing these two methods? Discuss differences in which tests are significant, how wide the confidence intervals are, and which confidence intervals contain 0.

While adjusting (Tukey-Kramer method): There are two confidence intervals for groups <12 - 12 and >16 - 16 which contain the value 0. Thus, we fail to reject the hypothesis that their means are equal for these groups. The p-values for these groups are greater than any significance level alpha, 0< alpha <= 1. For other groups' pairs, the p-values are less than alpha. Thus, we reject those hypotheses.

Without adjusting: There are no pairs of groups that contain 0 in the interval. All p-values are less than the significance level.

The confidence intervals while adjusting are wider than the intervals without adjusting. From above observations, we can see that "without adjusting" leads to more type-I errors.

(d) (3 points) Use the Dunnett procedure to compare every other group to the group with 12 years of education. Look at both the p-values and confidence intervals. Which group means apparently differ from the mean for those with 12 years of education?

```
levels(ex0525$Educ)
```

```
## [1] "12"    "13-15" "16"    "<12"   ">16"
```

```
ex0525$Educ<-relevel(ex0525$Educ,ref = "12")
levels(ex0525$Educ)
```

```
## [1] "12"    "13-15" "16"    "<12"   ">16"
```

```
Education_group_mod=lm(Income2005~Educ,data = ex0525)

Education_group_Dunnet=glht(Education_group_mod,linfct=mcp(Educ="Dunnet"))
summary(Education_group_Dunnet)
```

```
## 
##   Simultaneous Tests for General Linear Hypotheses
## 
## Multiple Comparisons of Means: Dunnett Contrasts
## 
## 
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
## 
## Linear Hypotheses:
##                  Estimate Std. Error t value Pr(>|t|)    
## 13-15 - 12 == 0     8011       2201   3.639  0.00112 ** 
## 16 - 12 == 0       33132       2571  12.885  < 1e-04 ***
## <12 - 12 == 0      -8563       4000  -2.141  0.11804    
## >16 - 12 == 0      39991       2649  15.098  < 1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
confint(Education_group_Dunnet)
```

```
## 
##   Simultaneous Confidence Intervals
## 
## Multiple Comparisons of Means: Dunnett Contrasts
## 
## 
## Fit: lm(formula = Income2005 ~ Educ, data = ex0525)
## 
## Quantile = 2.4804
## 95% family-wise confidence level
## 
## 
## Linear Hypotheses:
##                  Estimate    lwr         upr        
## 13-15 - 12 == 0   8011.0607   2551.0609  13471.0605
## 16 - 12 == 0     33132.0768  26754.1600  39509.9937
## <12 - 12 == 0    -8563.4475 -18485.2727   1358.3776
## >16 - 12 == 0    39990.5665  33420.3800  46560.7529
```

The groups for which the means are different from the educational group with 12 years of education are:

13-15 and 12 pvalue:0.00111 Conf interval:2550.7819 13471.3396

16 and 12 pvalue:< 1e-04 Conf interval:26753.8340 39510.3196

greater than 16 and 12.pvalue: < 1e-04 Conf interval:33420.0443 46561.0887

Only the groups less than 12 years of education and 12 years of eductaion have 0 in their confidence intervals.

(e) (2 points) Taking all of these tests together, what general statements would you make about the relationship between Education and Income?

We can claim that the pair of control group less than 12 and 12 years of education can have equal means of income as in both Tukey-Kramer with multiple comparisons and Dunnett Constrasts methods, the p-values are greater than the significance level alpha. Note that the difference of their means of incomes, 0, lies in the confidence intervals.