Code ▾

36-309 / 36-749 Homework 7: Multiple Testing

Due Wednesday, November 9, 11:59pm

Sanaz Saadatifar

Question 1: Beer Brewing (52pts)

Professor V. Staples (https://en.wikipedia.org/wiki/Vince_Staples) is conducting an experiment about beer brewing. In this experiment, several batches of beer are brewed for a fixed length of time and then clarified by isinglass flocculation, centrifugation, or filtration. The quantitative outcome is taste (no need to worry about how this is measured for this homework - assume that this has high construct validity).

Furthermore, Professor Staples has the following **planned comparisons**:

1. filtration vs. the average of isinglass and centrifugation
2. isinglass vs. centrifugation

Here is the dataset from this experiment:

Hide

```
beer = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/main/beer.cs
v")
#make sure the categorical variable is a factor
beer$method = factor(beer$method)
```

Here are the two variables in this dataset:

- `method` : “isinglass”, “filter”, or “centrifuge”
- `taste` : a quantitative outcome on a 0-to-100 scale (where higher denotes better taste)

a. (12pts) We’ll first consider running one-way ANOVA for this dataset. For this part, answer the following questions.

- (3pts) First, write code that runs the relevant one-way ANOVA for this dataset. Be sure to include `summary()` output in your code.

Hide

```
onewayModel = aov(taste ~ method, data = beer)
summary(onewayModel)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## method       2  696.9   348.5   12.49 7.16e-05 ***
## Residuals   37 1032.2    27.9                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

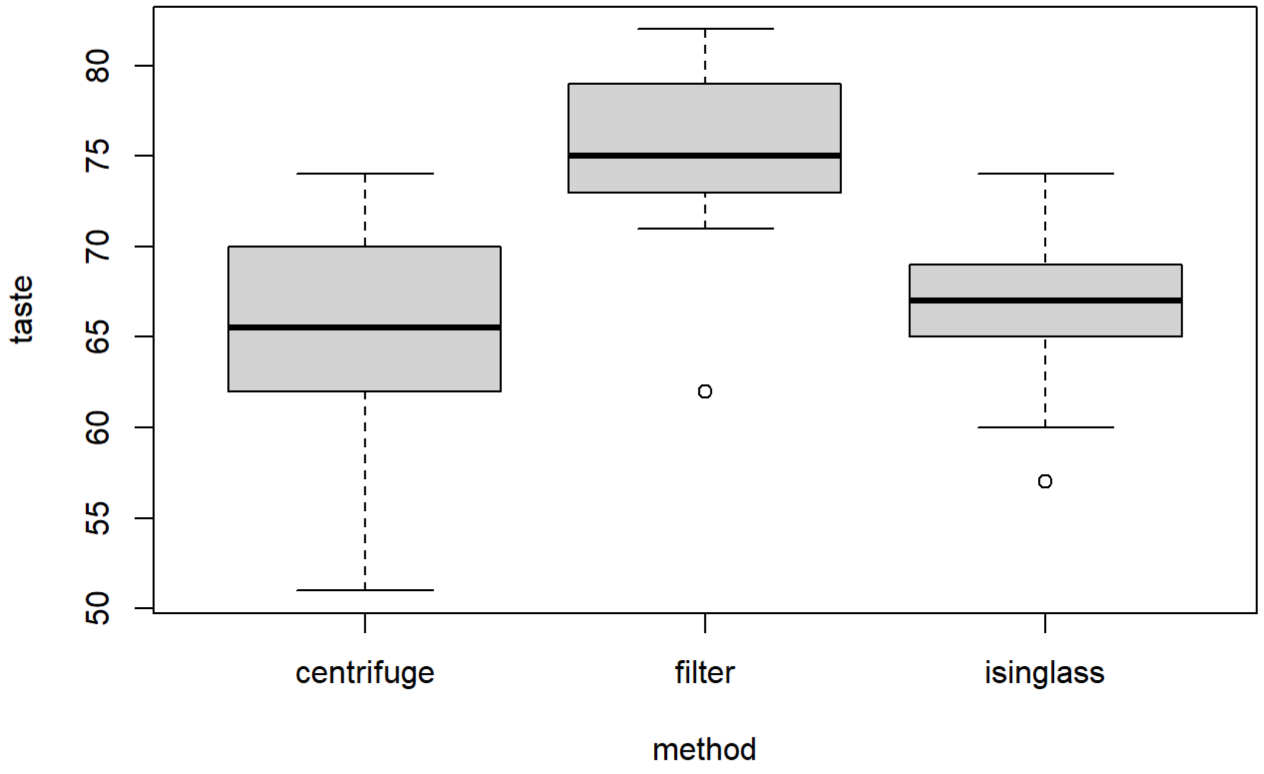
- (4pts) What is the null hypothesis tested by this one-way ANOVA? In your answer, use the symbols μ_I, μ_F, and μ_C, and explicitly state what these symbols mean. Then, based on your ANOVA output, do you reject or fail to reject this hypothesis? Explain in one sentence.

[The null hypothesis is that H0:μI=μF= μC, saying that mean is not different across different method categories. The alternative hypothesis HA is that mean is different across different method categories. Since the P-value is less than 0.05 (7.16e-05), then we reject the null.]

- (5pts) Finally, create a relevant side-by-side boxplot for these data that complements the one-way ANOVA. Then, interpret the boxplot in 1-3 sentences. In particular, in your interpretation, be sure to compare and contrast the *center* of the outcome variable across the treatment groups in this experiment.

Hide

```
boxplot(taste ~ method, data = beer)
```



[We see that the center of the outcome variable is different across method categories. For example this boxplot shows that filter category's mean taste or center of the outcome variable is higher compared to isinglass, and centrifuge. And no significant different is seen between isinglass and

centrifuge categories' center of outcome variable.]

b. (13pts) For this part, complete the following tasks:

- (3pts) Use the `levels()` function to figure out what the levels are for the `method` variable. What does R list as the first level, the second level, and the third level?

Hide

```
levels(beer$method)
```

```
## [1] "centrifuge" "filter"     "isinglass"
```

[centrifuge is the first level, filter is the second level, and isinglass is the third level]

- (6pts) Look back at the two **planned comparisons** presented at the beginning of this question. Write these planned comparisons in terms of a **contrast null hypothesis**, with some μs on the left-hand side and 0 on the right-hand side. **In your answer, write the μs in the same order as the levels of `method` are listed in R.** For example, if you said "isinglass" comes first, then you should write μ_I first here. (**Hint**: Your answer should ultimately be two null hypothesis equations, each of which has some linear combination of μs equal to zero.)

[H0: (1/2) μ1 - (1) μ2 + (1/2) μ3 = 0] [H0: (1) μ1 + (0) μ2 - (1) μ3 = 0]

- (4pts) Write the two sets of contrast coefficients that can be entered into a computer program to test these two planned null hypotheses. Again, the first coefficient should correspond to whatever level is listed first in R.

[The first contrast is (1/2, -1, 1/2), and the second contrast is (1, 0, -1).]

c. (12pts) Now we'll consider testing the two planned comparisons listed at the very beginning of this question. For reference, Professor Staples' planned comparisons were:

1. filtration vs. the average of isinglass and centrifugation
2. isinglass vs. centrifugation

For this part, answer the following questions.

- (6pts) First, explain why the **four conditions for planned comparisons** hold for these two comparisons. Be sure to provide an explanation for each condition.

[The are four conditions, 1) they should be chosen in advance which is hold in this case because it is planned before. 2) they are only used if the overall P-value H0 : μ1 = ⋯ = μk is ≤ α., this would also hold because we would say this only holds in case P-value is≤ α(0.05) 3) The contrast hypotheses are orthogonal. This means that the sum of products of coefficients are zero. It makes the hypotheses independent. This also holds because; (1/2)(1) – (1)(0)-(1/2)(1)=0 4) multiple comparisons is not greater than the degrees of freedom, which hold because df is k-1= 3-1 = 2 and we have 2 comparisons which is not greater than df.]

- (6pts) Now use the `glht()` function to test both of these planned comparisons. After doing this, state your scientific conclusion for each planned comparison. (**Hint**: In order to run the `glht()` function, you first need to write `library(multcomp)` in your .Rmd file. Furthermore, in order for `library(multcomp)` to work, you need to install the `multcomp` R package; if you haven't done this yet, see Task 2 of Lab7.)

Hide

```r
library(multcomp)
```

```
## Warning: package 'multcomp' was built under R version 4.1.3
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Warning: package 'TH.data' was built under R version 4.1.3
```

```
## Loading required package: MASS
```

```
## 
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
## 
##     geyser
```

Hide

```r
glht.fit1 = glht(model = onewayModel,
                 linfct = mcp(method = c(1/2, -1, 1/2)))
summary(glht.fit1)
```

```
## 
##   Simultaneous Tests for General Linear Hypotheses
## 
## Multiple Comparisons of Means: User-defined Contrasts
## 
## 
## Fit: aov(formula = taste ~ method, data = beer)
## 
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0   -8.868      1.783  -4.973 1.54e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Hide

```
glht.fit2 = glht(model = onewayModel,
                 linfct = mcp(method = c(1, 0, -1)))
summary(glht.fit2)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = taste ~ method, data = beer)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(>|t|)
## 1 == 0  -0.8132     2.0344    -0.4    0.692
## (Adjusted p values reported -- single-step method)
```

[For the first one, since the p-value is less than 0.05 (1.54e-05), then we reject the null hypothesis, so filter category’s mean is different from the average of centrifuge and isinglass means. And since point estimate is -8.868, this means that average of isinglass and centrifuge mean population is less than filter’s mean population. For the second one, since the p-value is larger than 0.05 (0.692), then we fail to reject the null hypothesis, so centrifuge category’s mean is not different from the isinglass means.]

d. (15pts) Now we will consider using the Tukey procedure. Please answer the following:

- (5pts) Note that we’ve already tested the two planned comparisons in Part C. Which two **additional** null hypotheses can be tested using the Tukey procedure? Please write out your null hypotheses using the μ notation you’ve used in previous parts. Be sure to explain why these additional null hypotheses are appropriate for the Tukey procedure.

[Since Tukey can only test pairwise hypothesis, and we already tested centrifuge vs isinglass, so the only two left are filter vs centrifuge (H0: (1) μ1 - (1) μ2 - (0) μ3 = 0) and filter vs isinglass(H0: (0) μ1 + (1) μ2 - (1) μ3 = 0).]

- (5pts) Use the Tukey procedure to test the two additional null hypotheses you wrote in the previous bullet. What is your conclusion for both of these null hypotheses?

Hide

```
TukeyHSD(onewayModel)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = taste ~ method, data = beer)
##
## $method
##                             diff        lwr       upr     p adj
## filter-centrifuge      9.2747253   4.307877 14.241573 0.0001584
## isinglass-centrifuge   0.8131868  -4.153661  5.780035 0.9158962
## isinglass-filter      -8.4615385 -13.519529 -3.403548 0.0006516
```

[For the filter vs centrifuge (H0: (1) μ1 - (1) μ2 - (0) μ3 = 0), we can reject the null hypothesis since the p-value is less than 0.05 (0.0001584). centrifuge is greater than filter. For the filter vs isinglass(H0: (0) μ1 + (1) μ2 - (1) μ3 = 0), we can also reject the null hypothesis since the p-value is less than 0.05 (0.0006516). isinglass is greater than filter]

- (5pts) State one additional complex null hypothesis that **has not been tested in this homework** and for which we cannot test using the Tukey Procedure. (You do not need to test it.) Please write out your complex null hypothesis in words **as well as** mathematical symbols.

[H0: (1) μ1 - (1/2) μ2 + (1/2) μ3 = 0, centrifuge vs. the average of isinglass and filter]

Question 2: Revisiting Spaghetti and Marshmallow Buildings (48 points)

In Homework5 (Question 2), we analyzed this dataset:

Hide

```
build = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/main/build.c
sv")
#make sure the categorical variables are factors
build$age = factor(build$age)
build$gender = factor(build$gender)
```

As a reminder, in this dataset, the experimental units are groups of 4 subjects who are given a box of spaghetti and a bag of marshmallows and told to build the highest structure possible. The groups are either all kindergarteners, all 5th graders, or all graduate students. Furthermore, the groups are either all male, all female, or 2 of each (mixed gender). Thus, there are two factors (`age` and `gender`) each with three levels, implying that two-way ANOVA is the appropriate analysis for this dataset.

In Homework5, we found that the “no interaction” (additive) two-way ANOVA model is most appropriate for this dataset (because we found that the interaction p-value was substantially greater than 0.05). Thus, we will focus on the no interaction two-way ANOVA model in this homework:

Hide

```
summary(aov(cmHigh ~ age + gender, data = build))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)    
## age          2   4125  2062.6  58.264 2.6e-14 ***
## gender       2     30    15.1   0.427   0.655    
## Residuals   55   1947    35.4                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a. (7pts) For this part, answer the following:

- (3pts) Look at the above two-way ANOVA table. According to the definition of planned comparisons, the number of comparisons we test cannot be greater than the degrees of freedom. From the two-way ANOVA table, what is the maximum number of planned comparisons we could test for this dataset? Explain your reasoning in 1 sentence. (**Hint**: It may be helpful to refer back to Lecture9.)

[Since the p-value for the gender is greater than 0.05, no follow-up tests are required in terms of gender. However for the age, the p-value is less than 0.05, so follow-up comparisons would be beneficial to do. The df is 2, and for planned comparison, the number of planned comparisons should not be greater than df, so the maximum number would be 2.]

- (4pts) Because we've already analyzed this dataset previously, it will be impossible to conduct planned comparisons for this dataset. According to the p-values in the two-way ANOVA table, which sets of group means should we conduct post-hoc follow-up tests for? Explain your reasoning in 1-3 sentences.

[Since the p-value for the gender is greater than 0.05, no follow-up tests are required in terms of gender. However for the age, the p-value is less than 0.05, so follow-up comparisons would be beneficial to do for age.so the μ(5th graders), μ(graduate students), and μ(kindergarteners) means can be tested. For example to check if H0: μ(5th graders)= μ(graduate students), H0: μ(5th graders)= μ(kindergarteners), and H0: μ(graduate students)= μ(kindergarteners).]

b. (14pts) For this part, answer the following:

- (6pts) For the group means you identified in the second bullet of Part A, write out every possible paired comparison in null hypothesis form using notation that looks like this: H_0: $\mu_1 = \mu_2$. For your μs, please use intuitive notation, and explicitly state what your notation means (i.e., what each of your μs represents).

Hide

```
levels(build$age)
```

```
## [1] "5th grade"         "graduate students" "kindergarten"
```

[Based on the results of level, "5th grade" represents level 1, "graduate students" represents level 2, and "kindergarten" represents level 3. Hence to compare "5th grade" vs "graduate students", the null hypothesis would be H0: μ1= μ2. to compare "5th grade" vs " kindergarten ", the null hypothesis would be H0: μ1= μ3. to compare" graduate students " vs " kindergarten ", the null hypothesis would be H0: μ2= μ3]

- (4pts) What type of "Corrections for Unplanned Comparisons" procedure would be most appropriate for the paired comparisons you identified in the previous bullet point? Explain your reasoning in 1 sentence.

[for this case, since we don't have a control group dunnet would not be appropriate, so for just pairwise comparison Tukey and Bonferroni can be considered and since Bonferroni is very conservative and inflates the P-value, Tukey would be most appropriate one.]

- (4pts) Let's say that, within the age factor, "graduate students" would be considered the "control group", and within the gender factor, "mixed" would be considered the "control group". If we were only interested in comparing each treatment group (kindergarten or 5th grade for `age` , male or female for `gender`) to its respective control group, what type of "Corrections for Unplanned Comparisons" procedure would be most appropriate? Explain your reasoning in 1 sentence.

[Dunnet would be the most appropriate one because it is used when we have a control group an dwe want to comapre the different treatment groups with control group only in pairwise tests.]

c. (15pts) In the first bullet of Part B, you should have wrote down several paired comparison hypotheses. In this part, we'll consider several ways to conduct these (unplanned) paired comparisons. For this part, answer the following questions.

- (5pts) First, use the `pairwise.t.test()` function to obtain "t-test like" p-values for each of the paired comparison hypotheses you wrote in Part B. To do this, within the `pairwise.t.test()` function, set `p.adjust.method = "none"` . For this part, all you need to do is write code that displays p-values for each paired comparison using `pairwise.t.test()` . (**Hint**: To remember how to use `pairwise.t.test()` , revisit the end of Task1 of Lab7.)

Hide

```
pairwise.t.test(x = build$cmHigh, g = build$age, p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  build$cmHigh and build$age
##
##                    5th grade graduate students
## graduate students 1.5e-05   -
## kindergarten      8.6e-08   1.6e-15
##
## P value adjustment method: none
```

- (5pts) Now we will obtain p-values that correct for multiple hypothesis testing. First, use the `pairwise.t.test()` function to obtain Bonferroni-based p-values for each of the paired comparison hypotheses you identified in Part B. All you need to do is write code that displays the appropriate p-values for this problem.

Hide

```
pairwise.t.test(x = build$cmHigh, g = build$age, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  build$cmHigh and build$age
##
##                   5th grade graduate students
## graduate students 4.4e-05   -
## kindergarten      2.6e-07   4.8e-15
##
## P value adjustment method: bonferroni
```

- (5pts) Now, write code that produces Tukey-based p-values for each of the paired comparison hypotheses. To do this, appropriately use the `TukeyHSD()` function. All you need to do is write code that displays the appropriate p-values for this problem. (**Hint**: You'll have to input a one-way ANOVA model int the `TukeyHSD()` function.)

Hide

```
onewayModel_Q2 = aov(cmHigh ~ age, data = build)
TukeyHSD(onewayModel_Q2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = cmHigh ~ age, data = build)
##
## $age
##                                   diff        lwr       upr    p adj
## graduate students-5th grade     -8.825 -13.307002 -4.342998 4.35e-05
## kindergarten-5th grade          11.430   6.947998 15.912002 3.00e-07
## kindergarten-graduate students  20.255  15.772998 24.737002 0.00e+00
```

d. (12pts) For this part, answer the following questions:

- (3pts) In Part C, you should have found three different p-values for the *graduate-vs-5th-grade* comparison: One from a typical t-test, one from Bonferroni, and one from Tukey. Which one is the largest (the one from t-test, Bonferroni, or Tukey)? Which one is the smallest (the one from t-test Bonferroni, or Tukey)?

[The largest is the Bonferroni = 4.4e-05, and the smallest is the t-test = 1.5e-05]

- (3pts) When considering **only the Type 1 error rate**, which of the p-values (t-test, Bonferroni, or Tukey) would you most prefer? Explain your reasoning in 1-2 sentences.

[In terms pf type 1 error rate, I would select the Bonferroni, because its p-value is inflated so that it controls the type error rate. If I select the t-test p-value, we will falsely reject the null hypothesis at a rate greater than 5% if we use these p-values (i.e., our Type 1 error rate will be higher than 5%).]

- (3pts) When considering **only power**, which of the p-values (t-test, Bonferroni, or Tukey) would you most prefer? Explain your reasoning in 1-2 sentences.

[In terms of only power, I would select the t-test, because it has the smallest p-value. If I selected the Bonferroni, we will correctly reject the null hypothesis less often (because the p-values are higher), meaning that we have less power, so the smallest p-value is better in terms of power.]

- (3pts) When considering **Type 1 error rate and power together**, which of the p-values (t-test, Bonferroni, or Tukey) would you most prefer? Explain your reasoning in 1-2 sentences.

[I would select tukey, because, Bonferroni correction is often considered to be overly conservative although it guarantees to control the Type 1 error rate, it also sacrifices a lot of statistical power; more statistical power can be gained using more nuanced procedures like the Tukey procedure without sacrificing validity.]

Question 3 (ONLY REQUIRED FOR 36-749 STUDENTS; BONUS QUESTION FOR 36-309 STUDENTS; 5pts)

Say that Professor T. Yorke (https://en.wikipedia.org/wiki/Thom_Yorke) ran a randomized experiment with **four treatment groups**. He decided to run a t-test for **each pairwise comparison** among the treatment groups (e.g., treatment 1 vs treatment 2, treatment 1 vs treatment 3, etc.) Thus, Professor Yorke conducted multiple t-tests, but he did not do any corrections for multiple hypothesis testing. In this question, Professor Yorke will give you information about his (uncorrected) tests, and we'll incorporate corrections for multiple hypothesis testing. Answer the following two questions.

- (2.5pts) Professor Yorke says: "For the t-test where I compared the first treatment group to the second treatment group, I got a p-value of 0.01." What would be the "adjusted p-value" after the Bonferroni correction? Explain how you arrived at your answer in 1-2 sentences. (**Hint**: In previous parts of the homework, you should have implemented Bonferroni-corrected p-values. This question demonstrates that you don't even need the actual data to figure out the Bonferroni-corrected p-value.)

[Since there are 4 treatment groups, the size of the family of comparisons is m=6. Considering that p ≤α/m. to keep the α as 0.05 we need to correct the p-value as p multiplied by m, then corrected p-value is 0.01*6 = 0.06]

- (2.5pts) Now Professor Yorke says: "Now I want to compute the 95% confidence interval for the first-versus-second treatment group t-test. I assigned 20 subjects to each treatment group, the mean difference between the first and second treatment groups (first minus second) was 3.25, and the standard error was 1.25. Thus, the 95% confidence interval for this mean difference is:"

Hide

```
alpha = 0.05
meanDiff = 3.25
se = 1.25
n = 40
t.quant = qt(p = 1-alpha/2, df = n - 2)

#Lower and upper bound of 95% CI
c(meanDiff - t.quant*se, meanDiff + t.quant*se)
```

```
## [1] 0.7195073 5.7804927
```

Professor Yorke is telling the truth: This is indeed the (uncorrected) 95% confidence interval for the two-sample t-test. If you're unsure why, it may be helpful to review the 36-749 question on Homework1.

Your goal for this problem is to compute the **Bonferroni-corrected 95% confidence interval**, i.e., the 95% confidence interval that incorporates the Bonferroni correction for multiple testing. For this part, write code that produces the desired confidence interval, and then state what the confidence interval is and how you arrived at your answer. (**Hint**: Note that, for a single hypothesis test, we reject the null hypothesis if the p-value is less than alpha, which is typically 0.05. This is indeed why the above code produces the 95% confidence interval for a single hypothesis test. Given this, you should write code similar to the above that incorporates the Bonferroni correction, which recognizes that Professor Yorke didn't conduct just one hypothesis test.)

Hide

```
alpha = 0.05/6
meanDiff = 3.25
se = 1.25
n = 40
t.quant = qt(p = 1-alpha/2, df = n - 2)

#lower and upper bound of 95% CI
c(meanDiff - t.quant*se, meanDiff + t.quant*se)
```

```
## [1] -0.2293429  6.7293429
```

[First we needed to correct the alpha level, as it is not 0.05 in this case. The correct alpha level is α/m = 0.05/6 . after running the above code, the correct CI is (-0.2293429 , 6.7293429)]