Code ▾

36-309 / 36-749 Homework 1: Honey, I Shrunk the Kid’s Cough

Due Wednesday, September 14, 11:59pm

Sanaz Saadatifar

Read all instructions carefully. Answer only what is asked and do not turn in extraneous material!

This homework is based on the paper “Effect of Honey, Dextromethorphan, and No Treatment on Nocturnal Cough and Sleep Quality for Coughing Children and Their Parents” (https://jamanetwork.com/journals/jamapediatrics/article-abstract/571638). You do not need to read the article.

Professor J. Seba (https://en.wikipedia.org/wiki/Nujabes) has conducted a study to assess how different treatments affected children’s nighttime cough and sleep quality. Professor Seba’s study compared three treatments:

1. A single dose of buckwheat honey
2. A single dose of honey-flavored dextromethorphan (an over-the-counter drug)
3. No treatment (i.e., a “control”)

The subjects of the study were children with respiratory infections, who thus had trouble coughing during the night. The study was run over two nights. On the first night, no treatment was given to any child; instead, their baseline cough frequency was measured by their parents. Then, for the second night, each child was randomized to one of the three above treatments, and then cough frequency was again measured. In the study “cough frequency” was rated on a 0-to-6 scale (ranging from 0 = “Not at all frequent” to 6 = “Extremely frequent”) by the children’s parents. As a result, Professor Seba and his students were able to compute the change in cough frequency between the first night (without treatment) and the second night (with treatment). In addition to cough frequency for the two nights, other characteristics were measured for each child.

In Problem 1, you will compare the honey group and the control group (i.e., the first and third groups above). In Problem 2, you will compare all three groups. By the end of this homework, you will have conducted more-or-less the main analyses in this paper, which is well-regarded and widely cited.

Problem 1: Comparing the Honey and Control Groups (49 points)

To get you started, Professor Seba provides you with this dataset:

Hide

```
honeyDataSubset =
  read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/master/honeyDataSu
bset.csv")
```

This dataset contains information on all subjects assigned to the honey group or control group.

a. (11pts) For this part, answer the following questions:

- (8pts) The above dataset contains the following variables: `age` (in years), `gender` ("male" or "female"), `cough1` and `cough2` (cough frequency on the first and second nights, respectively), and `group` ("honey" or "none"). First, using the `head()` function, display the first few rows of the `honeyDataSubset` dataset. Then, for each of these five variables, state its **role** (i.e., outcome or explanatory) and state its **type** (i.e., quantitative or categorical, discrete or continuous, nominal or ordinal). For each answer, explain your rationale.

Hide

```
head(honeyDataSubset)
```

```
##         age gender cough1 cough2 group
## 1 11.583333   male      6      3 honey
## 2  3.666667   male      4      1 honey
## 3  6.500000   male      4      2 honey
## 4 12.750000   male      4      5 honey
## 5 13.416667 female      4      3 honey
## 6  4.916667 female      4      3 honey
```

[Age: explanatory, quantitative, continuous because it has any number and we do not see just counts. Gender: explanatory, categorical, nominal because there is no real data and often coded by numbers. Cough1: outcome, categorical, ordinal because it shows levels with a meaningful order. Cough2: outcome, categorical, ordinal because it shows levels with a meaningful order. Group: explanatory, categorical, nominal because there is no real data and often coded by numbers]

- (3pts) Professor Seba is particularly interested in the difference in cough frequency before treatment (`cough1`) and after treatment (`cough2`). However, this outcome is technically not in this dataset. Using the dollar-sign notation, you can create a new variable in a dataset (i.e., a new column) as such:

Hide

```
#define the outcome, coughChange, as cough1 - cough2:
honeyDataSubset$coughChange = honeyDataSubset$cough1 - honeyDataSubset$cough2
```

[I thought about asking you to code this up yourself, but I wanted to save you some time, especially for the first homework.] That said, Professor Seba asks the following: How many possible values could `coughChange` take on? Answer and explain in 1-2 sentences. (**Hint**: This question isn't asking about the different values of `coughChange` that are observed in the actual data. Rather, it's asking how many *possible* values `coughChange` could yield.)

[Both Cough 2 and Cough1 are taking values 0 to 6. The minimum CoughChange value would be the 0-6= -6, and the maximum value of CoughChange would be 6-0 = 6, therefore, CoughChange can possibly take all 13 values from -6 to 6.]

b. (9pts) Now we'll conduct some **non-graphical EDA** for this dataset. For this part, answer the following questions.

- (4pts) What is the mean `age` in this dataset? Also, what is the youngest and oldest age in this dataset? Be sure to include the code you used to answer this question.

Hide

```
mean(honeyDataSubset$age)
```

```
## [1] 8.68287
```

Hide

```
min(honeyDataSubset$age)
```

```
## [1] 2.333333
```

Hide

```
max(honeyDataSubset$age)
```

```
## [1] 17.66667
```

[The mean age is 8.68287. the youngest kid is 2.333333 years old, and the oldest is 17.66667 years old.]

- (5pts) Now produce a contingency table involving the `gender` and `group` variables. After producing your table, use your table to answer the following questions. How many subjects were assigned to the honey group, and how many subjects were assigned to the control group? Furthermore, using your common-sense judgment (rather than statistical tests), would you say that about an equal proportion of males and females were assigned to the honey and control groups, or are the honey and control groups highly imbalanced in terms of the `gender` variable?

Hide

```
table(honeyDataSubset$group)
```

```
## 
## honey  none 
##    35    37
```

Hide

```
table(honeyDataSubset$group, honeyDataSubset$gender)
```

```
## 
##          female male
##   honey      18   17
##   none       17   20
```

[35 subjects were assigned to the honey group, and 37 subjects were assigned to the control group. In terms of proportion of males and females in each of honey and control groups, data shows that there are approximately but not exactly an equal proportion of males and females in each group. But definitely both groups are not highly imbalanced in terms of the `gender` variable.]

c. (9pts) Now we'll produce some bivariate non-graphical EDA for comparing the `coughChange` across the two groups. For this part, answer the following questions.

- (5pts) What is the mean `coughChange` for the honey group and for the control group? Also, what is the *variance* of `coughChange` within the honey group and within the control group? Be sure to include the code you used to compute this, but also be sure to write your final answer in text outside of your code! (**Hint**: To answer both of these questions, you can use the `aggregate()` function. To remember how to do this, go back to Lab2.)

Hide

```
aggregate(coughChange~group, data = honeyDataSubset, FUN = mean)
```

```
##   group coughChange
## 1 honey    1.828571
## 2  none    1.054054
```

Hide

```
aggregate(coughChange~group, data = honeyDataSubset, FUN = var)
```

```
##   group coughChange
## 1 honey   1.0873950
## 2  none   0.8858859
```

[The honey group has the mean of 1.828571 and variance of 1.0873950. the control group has the mean of 1.054054 and variance of 0.8858859 .]

- (4pts) If you did some extra calculations, you could find that the 95% confidence interval for the mean `coughChange` in the honey group is (1.47, 2.19) and the 95% confidence interval for the mean `coughChange` in the control group is (0.74, 1.37). Given *only* these confidence intervals, can you conclude whether or not there is a significant mean difference in `coughChange` between the two groups? State Yes or No, and explain your reasoning in 1-2 sentences.

[I think Yes, there is a significant mean difference in `coughChange` between the two groups because their confidence intervals in this dataset do not overlap.]

d. (5pts, ONLY REQUIRED FOR 36-749 STUDENTS; BONUS QUESTION FOR 36-309 STUDENTS) In Part C, we stated that the 95% confidence interval for the mean `coughChange` in the honey group is (1.47, 2.19) and the 95% confidence interval for the mean `coughChange` in the control group is (0.74, 1.37). Professor Seba asks his graduate student P. Stickles (https://en.wikipedia.org/wiki/Titus_Andronicus_(band)) to actually compute these intervals; Stickles is hoping that you, as a fellow student, can just compute them for him. Using the numbers you computed in Parts B and C, write code that computes these confidence intervals. Your final code should output these confidence intervals, which is all you have to do for this part. (**Hint**: In this case, 95% confidence intervals will rely on 0.975 quantiles of t-distributions; to compute this, you can use `qt(p = 0.975, df = ?)` , where you must specify the correct degrees of freedom for `df` . To help you get started, Stickles has provided template code below. Please include comments in your code, so that it's clear what you are trying to compute.)

Hide

```
# the control list
control = subset(honeyDataSubset, group == "none")
#the 0.975 quantiles of t-distributions with the degree freedom of sample size- 1
none.quant = qt(p = 0.975, df = 36)
#The maximum value of confidence interval: mean + 0.975 quantiles*(square root (variance/s
ample size))
(mean(control$coughChange)) + (none.quant*sqrt (var(control$coughChange)    / 37))
```

```
## [1] 1.367871
```

Hide

```
#The minimum value of confidence interval: mean - 0.975 quantiles*(square root (variance/s
ample size))
(mean(control$coughChange)) - (none.quant*sqrt (var(control$coughChange)    / 37))
```

```
## [1] 0.7402373
```

Hide

```
# the honey list
honey = subset(honeyDataSubset, group == "honey")
#the 0.975 quantiles of t-distributions with the degree freedom of sample size- 1
honey.quant = qt(p = 0.975, df = 34)
#The maximum value of confidence interval: mean + 0.975 quantiles*(square root (variance/s
ample size))
(mean(honey$coughChange)) + (honey.quant* sqrt(var(honey$coughChange)   / 35))
```

```
## [1] 2.18678
```

Hide

```
#The minimum value of confidence interval: mean - 0.975 quantiles*(square root (variance/s
ample size))
(mean(honey$coughChange)) - (honey.quant* sqrt(var(honey$coughChange)   / 35))
```

```
## [1] 1.470363
```

e. (10pts) We will use the independent samples t-test to formally analyze this dataset. But first, for this part, we will assess assumptions for the t-test. As we discussed in class, the two-samples t-test makes three modeling assumptions: (1) The outcome is Normally distributed in both treatment groups, (2) the variance of the outcome is the same in both treatment groups, and (3) subjects' outcome measurements are independent of each other. Use graphical EDA to assess the first two assumptions, and use your own practical judgment to assess the third assumption. You may only use 1-3 graphs in your answer.

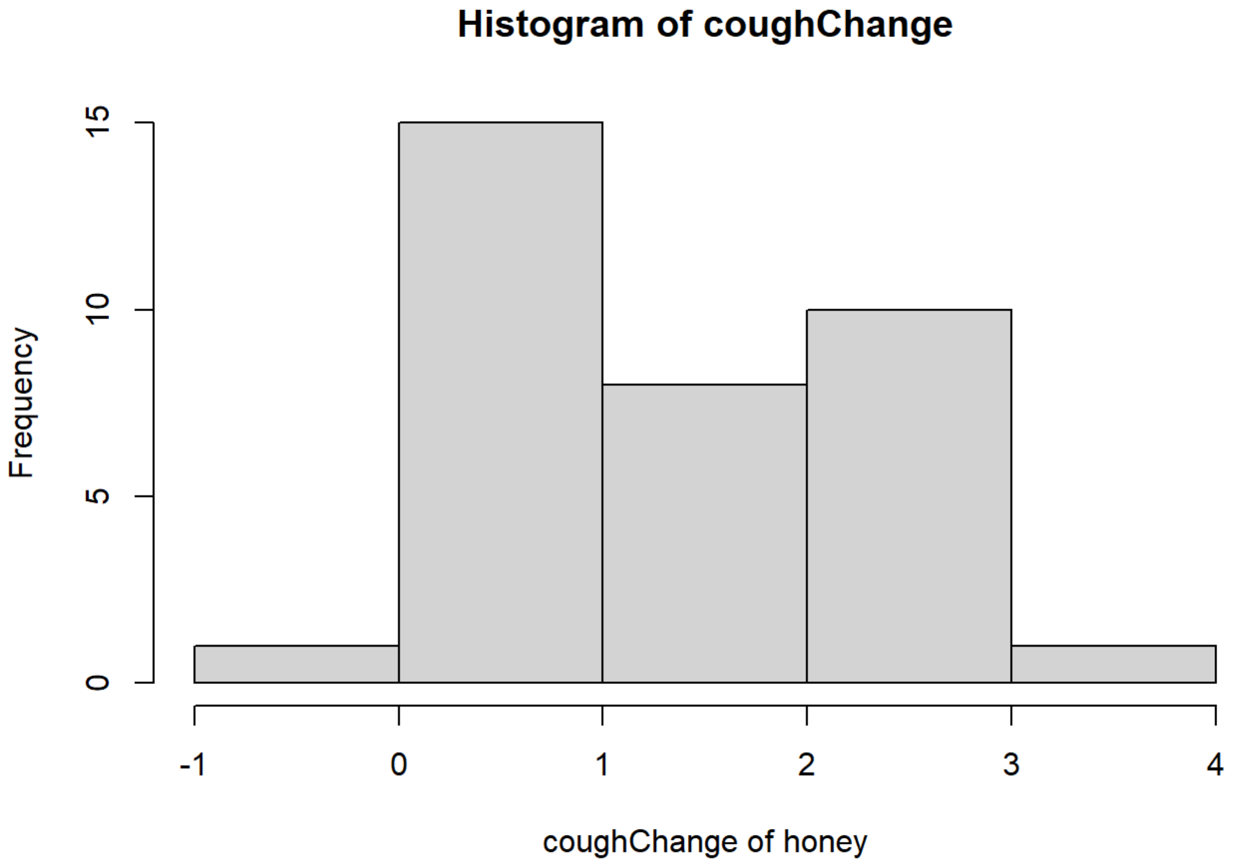
Hint: It may be helpful to have a dataset of the *honey group specifically* as well as a dataset of the *control group specifically*. To help you, here is code defining these subsets:

Hide

```
#data on just the honey group
honey = subset(honeyDataSubset, group == "honey")
#data on just the control group
control = subset(honeyDataSubset, group == "none")
```

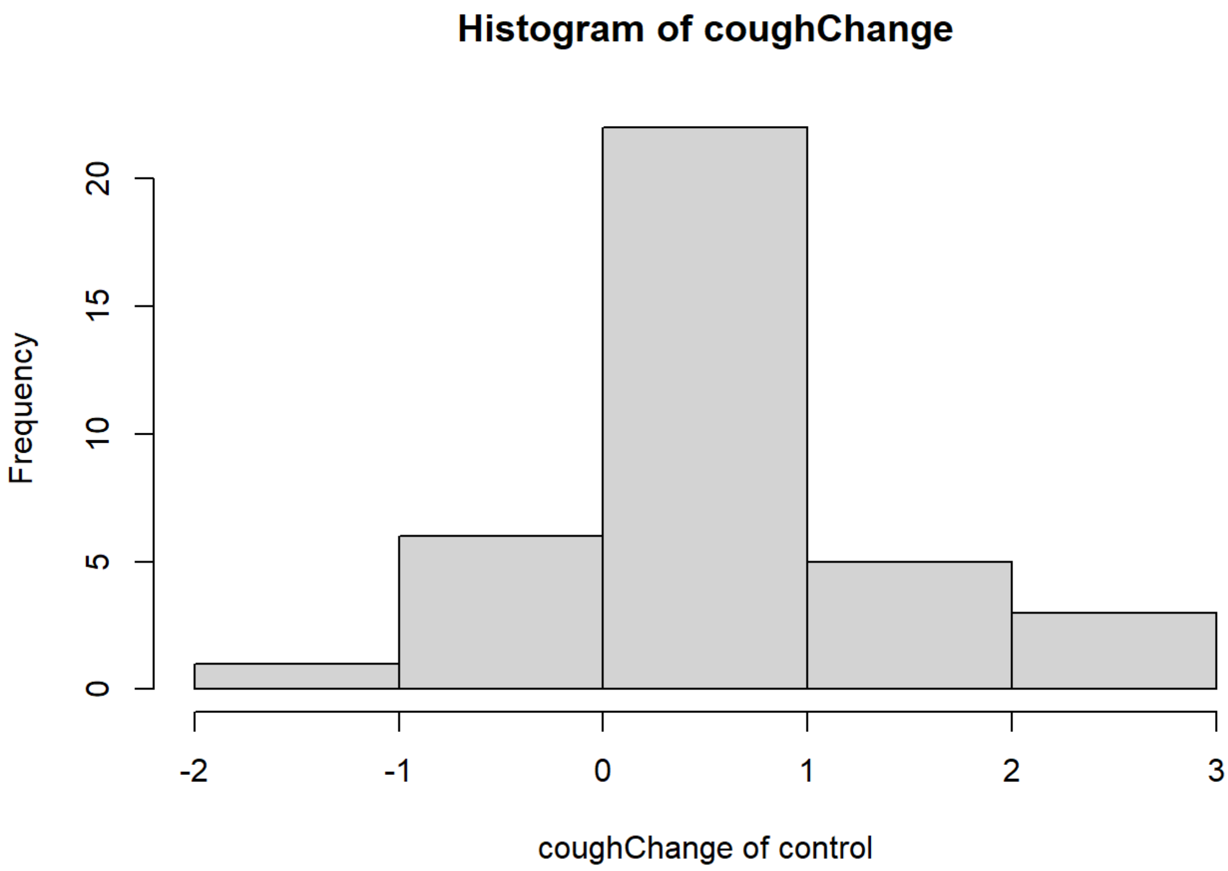
Hide

```
hist(honey$coughChange,
xlab = "coughChange of honey",
main = "Histogram of coughChange",
breaks = 7)
```



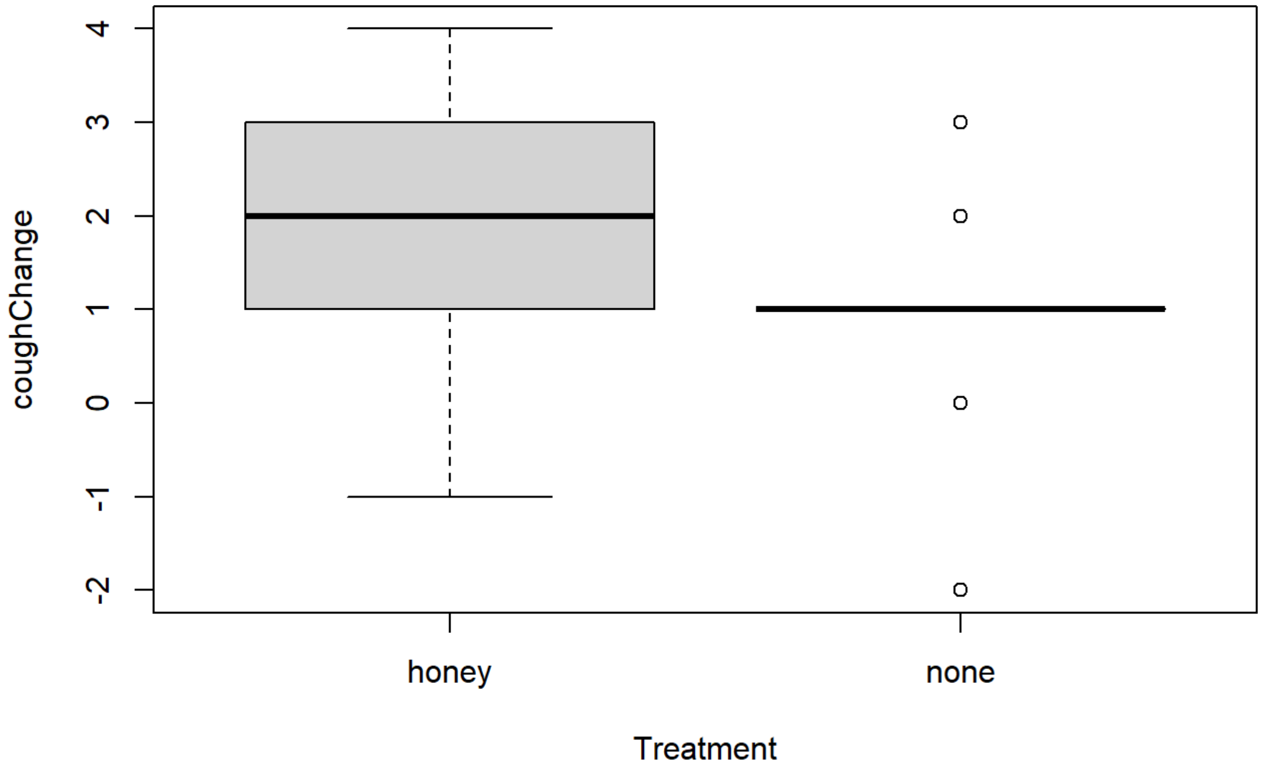
Hide

```
hist(control$coughChange,
xlab = "coughChange of control",
main = "Histogram of coughChange",
breaks = 7)
```



Hide

```
boxplot(honeyDataSubset$coughChange~honeyDataSubset$group,
xlab = "Treatment", ylab = "coughChange")
```



[To check the normality I used histograms. The histogram of control group is very normally distributed. The histogram of Honey group is roughly normally distributed, however, the boxplot shows that we can assume that honey group is normally distributed as the mean is actually in the middle of IQR.To check the variance, I used the boxplot. The results showed that these two samples do not have similar variances, As the IQR of honey group is way bigger than IQR of the control group. In terms of independency, I assume that the sample data is independent of each other because kids are not related to each other. They all separately were participated in the study.]

f. (10pts) Now we'll conduct the t-test. Answer the following questions:

- (5pts) Run the t-test using the `t.test()` function; use your answer from Part E to determine whether you should set `var.equal = TRUE` within the `t.test()` function. After running the test, state the 95% confidence interval, with a careful statement about what it is an interval for.

Hide

```
t.test(coughChange~group, data = honeyDataSubset, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  coughChange by group
## t = 3.3022, df = 68.292, p-value = 0.001528
## alternative hypothesis: true difference in means between group honey and group none is
not equal to 0
## 95 percent confidence interval:
##  0.3065266 1.2425082
## sample estimates:
## mean in group honey  mean in group none 
##            1.828571            1.054054 
```

[The confidence interval is (0.3065266, 1.2425082) it shows that we are 95% sure that the difference between cough changes of honey group and cough changes of control group falls within this range (0.3065266, 1.2425082). it also shows that honey group’s mean is larger than control groups mean in the (0.3065266, 1.2425082) range with 95% confidence.]

- (5pts) State your scientific conclusion according to the t-test. To follow common practice, in your answer, be sure to include a *p-value* and a *point estimate* for the difference in effects between the honey group and control group. When writing your conclusion, keep in mind the interests of Professor Seba, who conducted this study to assess how different treatments affected children’s nighttime cough.

[We see that the p-value is less than 0.05, and thus we reject the null hypothesis that the population mean parameters of μ(honey) and μ(control) are equal. In particular, by looking at the sample means for the two groups, we conclude that μ(honey) > μ(control), therefore the honey group’s treatment was more effective than control group as it reduced the cough changes significantly compared to the control group. We make this conclusion because we see that the sample mean for the treatment group is greater than that of the control group. Also the point estimate for difference in effects between the honey group and control group is 1.828571-1.054054 = 0.774517, which falls withing the confidence interval range.]

Problem 2: Comparing All Three Groups (51 points)

Impressed by your work so far, Professor Seba now provides you with this dataset:

Hide

```
honeyData = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/master/h
oneyData.csv")
```

This dataset contains children assigned to the honey group, the dextromethorphan (“DM”) group, and the control group. In other words, the above dataset is the same as the dataset in Problem 1, but with some additional subjects assigned to the DM group.

a. (10pts) For this part, we’ll do some more non-graphical EDA. Answer the following questions:

- (2pts) First, similar to Problem 1A, write code that creates the `coughChange` variable for the `honeyData` dataset. (This should just involve copy-and-pasting the Problem 1A code and editing it slightly.)

Hide

```
honeyData$coughChange = honeyData$cough1 - honeyData$cough2
```

- (5pts) Now Professor Seba asks the following questions: “What is the number of subjects assigned to each of the three groups? Also: For the t-test and one-way ANOVA, is there any assumption about equal group sizes?” Be sure to include any code you used to answer the first question. As for the second question, simply state Yes or No.

Hide

```
table(honeyData$group)
```

```
## 
##    DM honey  none 
##    33    35    37
```

[The subjects assigned to DM, honey, and control groups are 33, 35, 37 respectively. And NO, there is no assumption about equal group size.]

- (3pts) At this point, we’ve learned the t-test and one-way ANOVA, and in this problem we’ll use one-way ANOVA to analyze this dataset. However, Professor Seba asks, “Wait, why can’t we just use a t-test to analyze this dataset too?” Explain in 1-2 sentences why a t-test not a valid choice for analyzing this dataset.

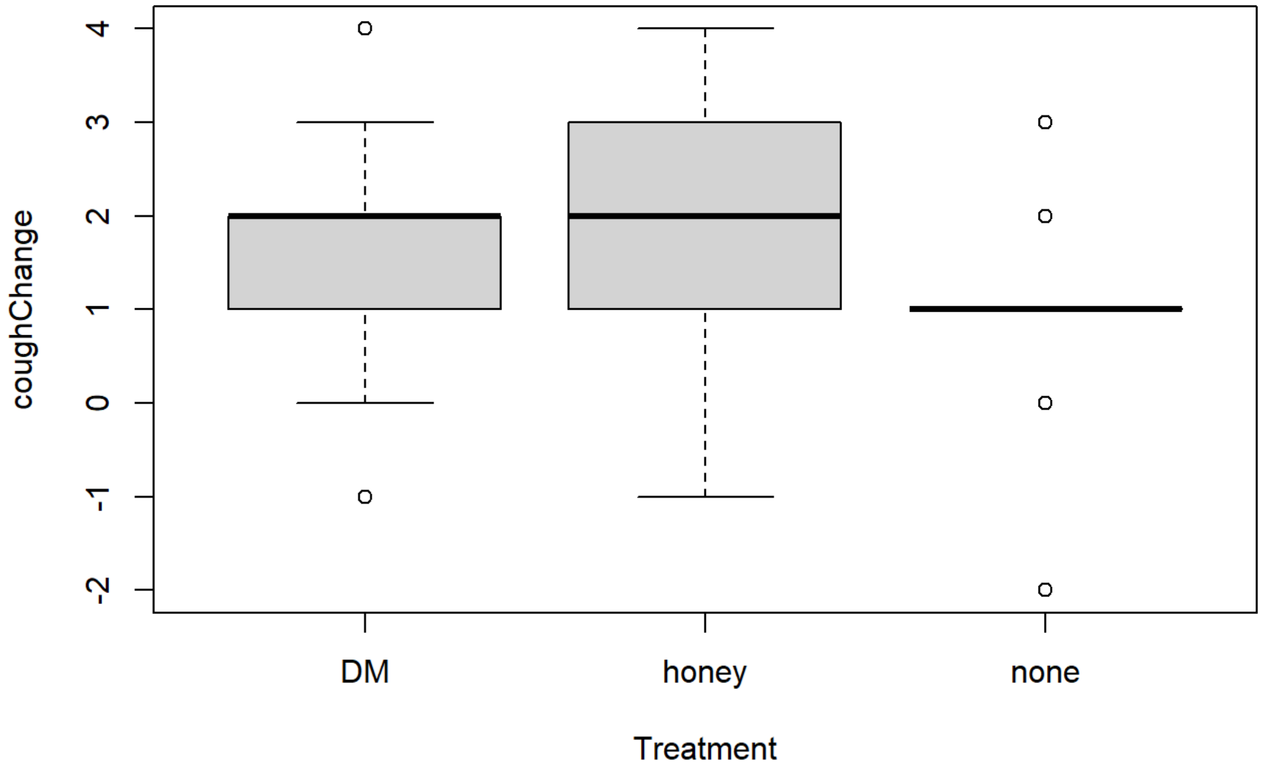
[We use t-tests for experiments where there is a two-level explanatory variable (treatment) and a quantitative outcome. However here we have three-level explanatory variable, so one-way ANOVA needs to be used.]

b. (11pts) For this part, we’ll consider the modeling assumptions of one-way ANOVA. Answer the following questions:

- (6pts) First, create a side-by-side boxplot comparing the `coughChange` across the three groups. Then, in 2-4 sentences, describe the key points seen in the side-by-side boxplots.

Hide

```
boxplot(honeyData$coughChange~honeyData$group,
xlab = "Treatment", ylab = "coughChange")
```



[The mean of DM and Honey groups are similar; however the control group's mean is less than the others. The DM groups has two outliers (one on each side), and it is not normally distributed. The honey group is normally distributed and no outliers it has. The control group has several outliers (two on each side). Overall these box plots showed that DM and Honey both worked well and were more effective compared to control group.]

- (5pts) One of the modeling assumptions of one-way ANOVA is the equal variance assumption. Using your boxplot, do you think the equal variance assumption is violated for this dataset? State Yes or No, and explain your reasoning (being sure to refer to your boxplot in your explanation).

[Yes, as discussed previously, the IQR of control group is way less than the other two groups. Even the IQR of the DM is less than the honey group about twice its range.]

c. (5pts) State the null and alternative hypotheses for one-way ANOVA in the context of this study. For the null hypothesis, please use mathematical notation; for the alternative hypothesis, you may use mathematical notation or words. However, when using mathematical notation, be sure to explain what your mathematical notation means. (**Hint**: In Lab2 we discussed ways to incorporate mathematical notation within .Rmd files, and so it may be helpful to revisit Lab2.)

[H0 : μ(DM) = μ(honey) = μ(control) . the null hypothesis states that the means of all groups are equal; HA: The means of all groups are not all equal.]

d. (16pts) Now, finally, we'll run one-way ANOVA. For this part, answer the following questions.

- (4pts) First, run the one-way ANOVA for this dataset using the `aov()` function. In your code, apply the `summary()` function to the `aov()` function, such that you can see the output from the one-way

ANOVA analysis. (To remember how to do this, see Lab2 or the Lecture 2 R Demo.) For this part, you just need to successfully display the summary output.

Hide

```
summary(aov(coughChange~group, data = honeyData))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)  
## group         2  10.82   5.410    4.57 0.0126 *
## Residuals   102 120.74   1.184                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- (4pts) In your output, you should see that the within mean sum of squares is 1.184, and the F statistic is 4.57. What calculations were used to compute these numbers? (**Hint**: These quantities are computed using other numbers you should see in your ANOVA table. Thus, for this part, you just need to state how the within mean sum of squares and F statistic are computed using other numbers in that ANOVA table.)

[within mean sum of squares equals to Residuals Sum of Squares divided by Residuals degree of freedom: hence, 120.74/ 102 = 1.184. Moreover, F statistic is between mean sum of squares divided by within mean sum of squares. Hence, 5.410/ 1.184 = 4.57.]

- (4pts) If we were to run this experiment many times, and if the model assumptions are correct, and if the null hypothesis were true, how often would we expect the F statistic to be bigger than 4.57? Please explain your answer in 1-2 sentences. (**Hint**: Your answer should be an actual number, but you don't have to do any additional computations to answer this question.)

[since the P-value is 0.0126 it means the probability of having the null hypothesis to be true (or in other words the F statistic to be bigger than 4.57) is 0.0126.]

- (4pts) State your scientific conclusions from the one-way ANOVA analysis in the context of this study. (Again, remember to keep in mind the interests of Professor Seba.)

[We see that the p-value for this test is less than 0.05, and thus we reject the null hypothesis that the population means of cough changes of three groups, μ(DM), μ(honey), and μ(control), are all equal, thereby concluding that they are not all equal. However, from this test alone, we do not know which of these population means are not equal to each other; we only know that they are not all equal to each other.]

e. (9pts) For this part, please answer the following questions.

- (1pt) In Problem 1, you ran a t-test between the honey group and the control group. Look back at the conclusion you made from that test, and state that conclusion again here. (This simply involves copy-and-pasting part of your answer for Problem 1F).

[We see that the p-value is less than 0.05, and thus we reject the null hypothesis that the population mean parameters of μ(honey) and μ(control) are equal. In particular, by looking at the sample means for the two groups, we conclude that μ(honey) > μ(control), therefore the honey group's treatment was more effective than control group as it reduced the cough changes significantly compared to the control group. We make this conclusion because we see that the sample mean for the treatment

group is greater than that of the control group. Also the point estimate for difference in effects between the honey group and control group is 1.828571-1.054054 = 0.774517, which falls withing the confidence interval range.]

- (4pts) Professor Seba would like to understand how the three different groups compare to each other, and so he asked his student P. Stickles to run a t-test between (1) the honey group and the DM group, and (2) the DM group and the control group. To do this, Stickles first made these datasets:

Hide

```
#dataset without control group
data.noControl = subset(honeyData, group != "none")
#dataset without honey group
data.noHoney = subset(honeyData, group != "honey")
```

Then, the code Stickles wrote for the two t-tests is provided below. Uncomment the below code; it should run if you have successfully defined `coughChange` for `honeyData` (and defined `data.noControl` and `data.noHoney` by running the above code).

Hide

```
t.test(coughChange ~ group, data = data.noControl)
```

```
##
##  Welch Two Sample t-test
##
## data:  coughChange by group
## t = -1.5348, df = 61.952, p-value = 0.1299
## alternative hypothesis: true difference in means between group DM and group honey is no
t equal to 0
## 95 percent confidence interval:
##  -1.0007277  0.1314636
## sample estimates:
##    mean in group DM mean in group honey
##            1.393939            1.828571
```

Hide

```
t.test(coughChange ~ group, data = data.noHoney)
```

```
##
##  Welch Two Sample t-test
##
## data:  coughChange by group
## t = 1.2574, df = 58.451, p-value = 0.2136
## alternative hypothesis: true difference in means between group DM and group none is not
equal to 0
## 95 percent confidence interval:
##  -0.2011212  0.8808918
## sample estimates:
##   mean in group DM mean in group none
##           1.393939           1.054054
```

Based on the above output, state the scientific conclusion that can be made from each of these t-tests.

[For the first t.test, We see that the p-value is greater than 0.05, and thus we cannot reject the null hypothesis that the population mean parameters of μ(honey) and μ(DM) are equal. Also the confidence interval for differences of mean cough changes of honey and DM groups includes zero, which shows that the population mean parameters of μ(honey) and μ(DM) might be equal. For the second t.test as well, We see that the p-value is greater than 0.05, and thus we cannot reject the null hypothesis that the population mean parameters of μ(control) and μ(DM) are equal. Also the confidence interval for differences of mean cough changes of control and DM groups includes zero, which shows that the population mean parameters of μ(control) and μ(DM) might be equal.]

- (4pts) Now Professor Seba looks at the conclusion you wrote in the first bullet point and the conclusions you wrote for the second bullet point. He says: “These test conclusions seem to contradict each other.” Do you agree or disagree? State Agree or Disagree, and then explain in 1-4 sentences.

[I disagree, because there is significant difference between the means of cough changes of control group and honey group. This leads to have a p-value less than 0.05 for the F-test. Because F.test results indicated that population means of groups are not equal to each other, however it did not say that which groups are actually different. After the last t.tests we found out that the different groups are control and honey, and there is not enough evidence to say that honey and DM as well as DM and control groups are different.]