Name - vvgupta2

Homework 7

Homework Instructions

For questions that require code, please create a code chunk directly below the question and type your code there. Your knitted pdf will show both your code and your output. You are encouraged to knit your file as you work to check that your coding and formatting is done so appropriately.

For written responses or multiple choice questions, please bold your (selected) answer.

Grading Details

All questions will be graded full credit (1 point), half credit (0.5 point) or no credit (0 points).

Full credit responses should have the correct response and appropriate code (if applicable). Half credit responses will have a reasonable attempt (typically no more than one small error or oversight), and no credit responses will be either non-attempts or attempts with significant errors.

Exercise 1

For most of this homework, we're going to be looking at the Flint data. A few years ago, reports showed that the water in many households in Flint, Michigan had dangerously high lead levels. A group of researchers conducted an independent investigation where they contacted 300 homes around Flint and asked them to provide and mail back three water samples from a kitchen faucet.

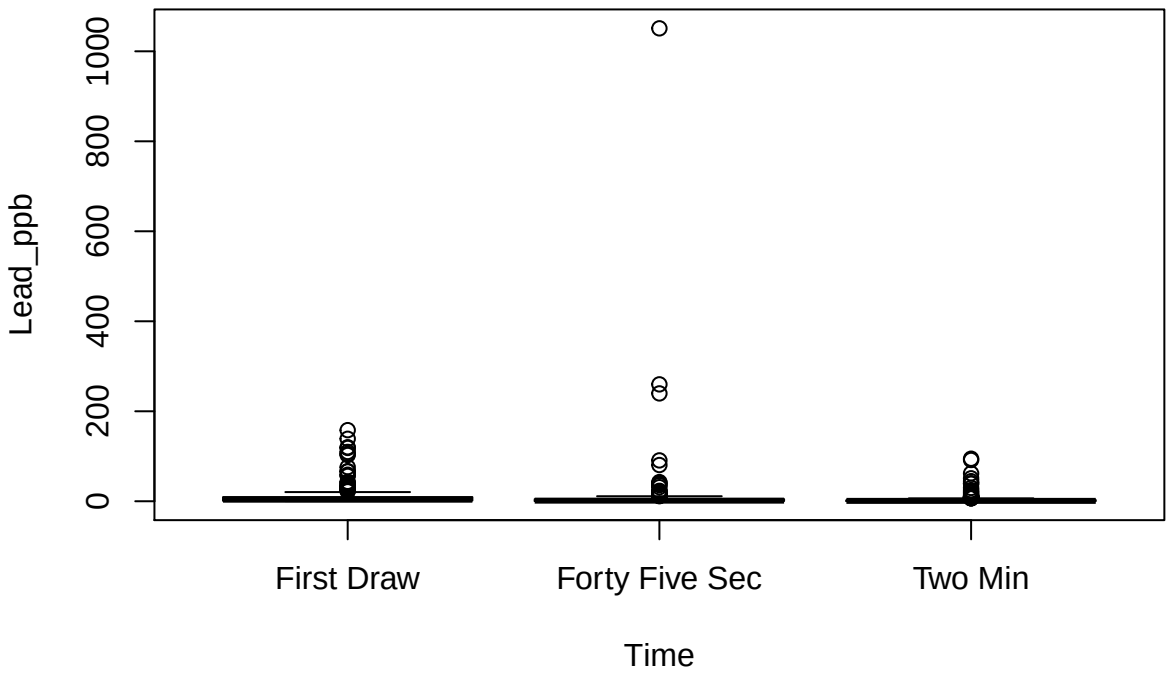
- The first sample labeled "First_Draw" (when the faucet was first turned on).
- The second sample was labeled "Forty_Five" (when the faucet had been running 45 seconds)
- The third sample was labeled "Two_Min" (when the faucet had been running 2 minutes).

269 of these homes returned their water samples, and the Zip code and Ward (district) of the household was also recorded. The lead content (measured in parts per billion) was recorded.

First, load `Flint.csv` here. You may wish to view the data to observe the variable names.

Create a boxplot to compare the water samples from first draw, forty five seconds, and 2 minutes. No formatting/subsetting required for now... we'll make a few adjustments in Exercise 3!

```
#solution
boxplot(Lead_ppb~Time, data=Flint)
```



Exercise 2

Use the `dplyr` package to create a summary statistic table to compare the the levels under `Time`. Your summary table should report the means, medians, and variance of `Lead_ppb` for each of the three Time points.

```
#solution
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Flint%>%
  group_by(Time)%>%
  summarise (
    mean_L = mean (Lead_ppb, na.rm = TRUE),
```

```
    median_L = median (Lead_ppb, na.rm = TRUE),
    var_L = var (Lead_ppb, na.rm = TRUE)
  )
```

```
## # A tibble: 3 x 4
##   Time           mean_L median_L var_L
## * <chr>           <dbl>    <dbl> <dbl>
## 1 First Draw      10.7     3.48   468.
## 2 Forty Five Sec  10.3     1.4   4593.
## 3 Two Min          3.68    0.831  112.
```

Anecdote: One thing you might notice from the table is that the means for the first draw and 45 second group are quite close, but the medians are noticeably different, and the variance for the 45 second group is extremely high. This is primarily because the 45 second group has a mega outlier at 1,051 ppb!

Exercise 3

Two assumption issues we should consider before proceeding with an ANOVA.

1) Our variances are wildly different (in particular in the 45 second group)
2) Our data is not normally distributed for each time point, but instead quite right skewed.

We can address the first issue by removing the most extreme outlier–the value above 1000. Create a subset of data that only includes values of Lead ppb that are below 500ppb. Our variances will still be quite different, but the more relaxed assumption (no sample variance is 5x larger than another) is approximately met.

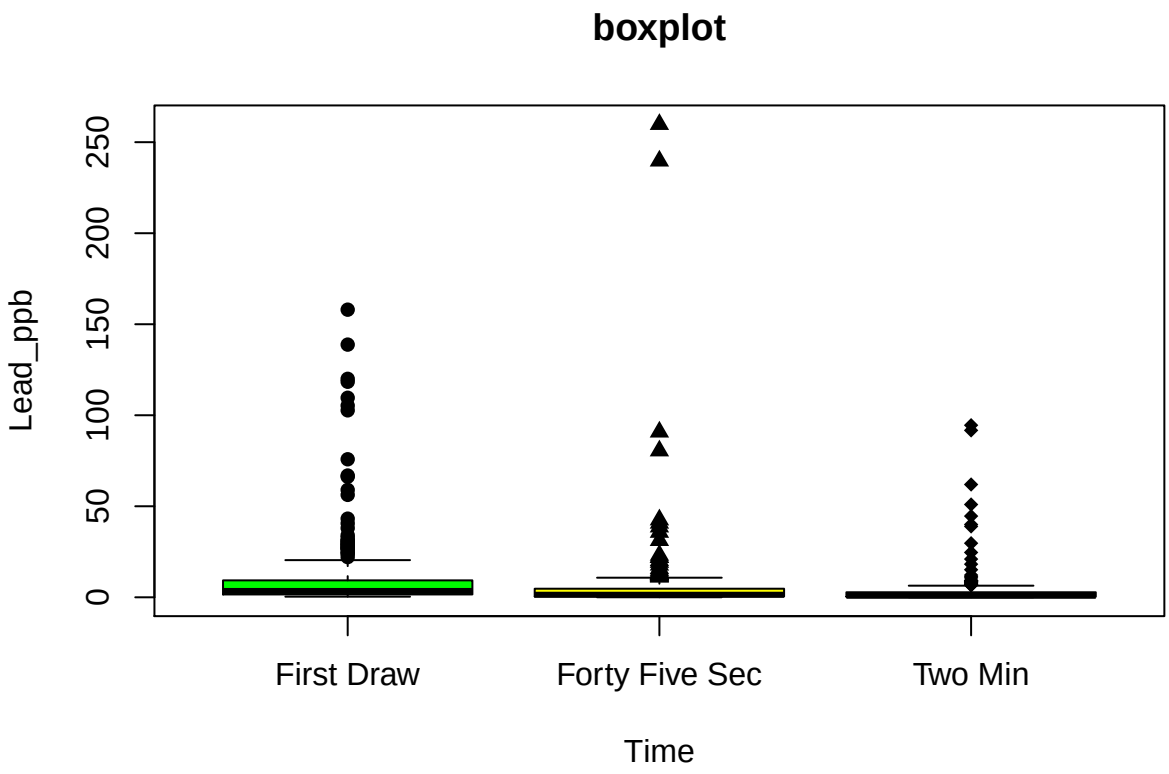
A note: we might also consider just removing all three data points of that household from this data as a more consistent choice, but we won't worry about that for now.

The second issue we can't address directly, but we can depend on the more relaxed assumption that the CLT will apply here. The distributions of the possible sample means generated from each time point should be normally distributed with samples of size 269.

Now, let's subset the data to only include readings where lead content is less than 500ppb and make a new boxplot with this data. Your boxplot should include

- An appropriate title
- Appropriate axes labels
- A different fill color for each boxplot (hint: you can do this by assigning col to be equal to a vector of three color names. Look online for assistance if you are unsure!)
- Change the dot style (`pch`) to a different value of your choice.

```
#solution
new_table = Flint%>%
  filter(Lead_ppb < 500)
boxplot(Lead_ppb~Time, data=new_table, ylab='Lead_ppb', xlab = 'Time', main = 'boxplot', col = c ('gree
```



Exercise 4

We'd like to test to see if running the faucet might reduce the lead content in the water.

Which statement represents the **null** hypothesis if we were to test this with a one-way ANOVA? In the below statements, assume μ_j and $\bar{x}_j$ represent the average lead content in the jth group, with μ being a parameter and $\bar{x}$ being a statistic.

Please delete the incorrect answer choices and leave only your answer.

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

Exercise 5

Run a one-way ANOVA using our **subsetted** data (with values less than 500 lead ppb).

Then, make a conclusion **in context** about what this result communicates to us, using $\alpha = 0.01$ as a benchmark for consideration.

```
#solution
summary(aov(Lead_ppb~Time,data=new_table))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Time          2   6704    3352   8.997 0.000137 ***
## Residuals   803 299176     373
```

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

Since the alpha value is larger than the p value we will have to reject the null hypothesis

Exercise 6

Complete a Tukey HSD Test to report on the pairwise differences in lead content between the different time points.

Then choose the most appropriate conclusion from the results. (Use $\alpha = 0.05$ as a benchmark).

```
#solution
TukeyHSD(aov(Lead_ppb~Time,data=new_table))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Lead_ppb ~ Time, data = new_table)
##
## $Time
##                                diff        lwr        upr     p adj
## Forty Five Sec-First Draw -4.296112  -8.207748 -0.3844747 0.0272270
## Two Min-First Draw        -7.000208 -10.908201 -3.0922151 0.0000857
## Two Min-Forty Five Sec    -2.704097  -6.615734  1.2075403 0.2363991
```

- There is evidence that the true mean lead content for first draw measures is different than that of 45 second and 2 minute measures.

- There is evidence that the true mean lead content for all three timings are different from one another.
- There is no evidence for a difference in means based on the timing of the measure.

Exercise 7

Now, let's turn to a different dataset named `Plants.csv`. This data was collected as a result of an experiment to grow hydrangea bushes. This experiment includes data from 100 plants.

Load the data here

Three variables of interest to us are listed below

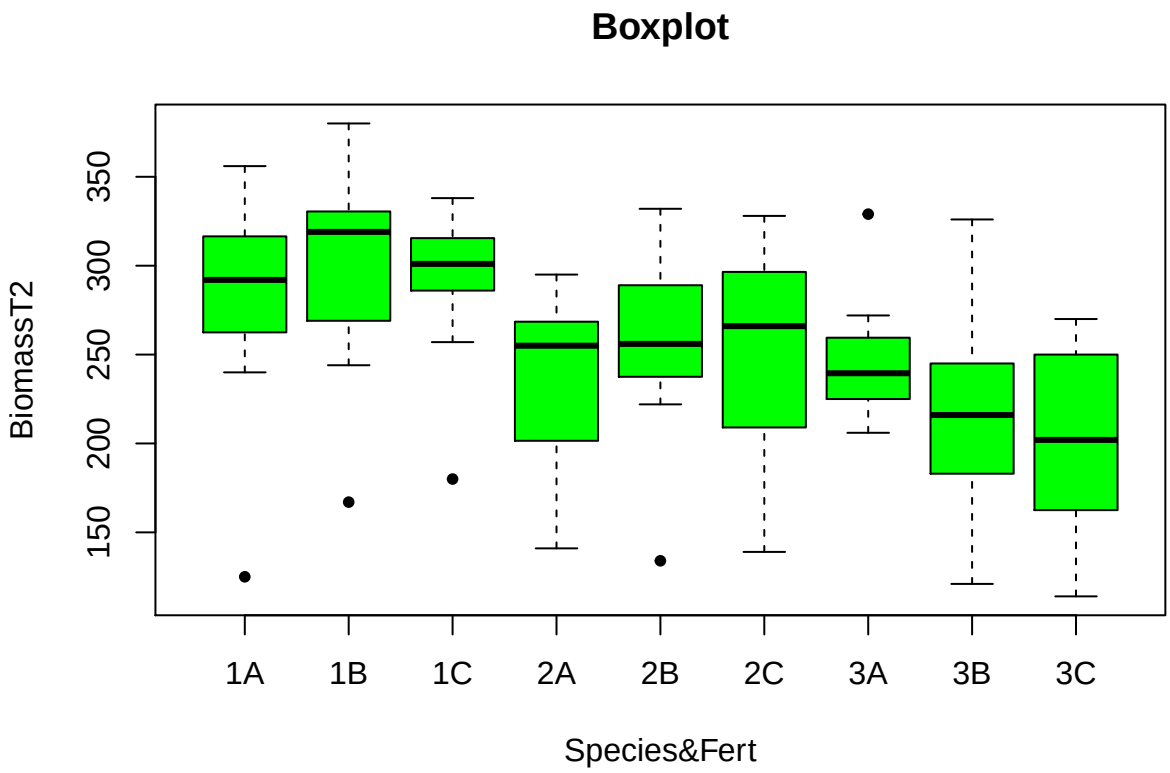
- BiomassT2: Biomass of the plant by the end of the experiment
- Fert: Fertilizer used on that plant during course of the experiment (A, B, or C)
- Species: Species of the plant (1, 2, or 3)

Create a set of boxplots that compares the BiomassT2 readings of each combination of Species and Fert. In other words, you should have 9 side-by-side boxplots.

Hint: When creating a boxplot, you can simply add + in your listing of predictor variables to add both predictors at the same time!

Add an appropriate title and axes labels. All other formatting optional.

```r
#solution
boxplot(BiomassT2~Species+Fert, data=Plant, names = c('1A','1B','1C','2A','2B','2C','3A','3B','3C'), co
```



Exercise 8

Create a two-way anova model (with interaction term) to predict BiomassT2 using Species and Fert as predictors. Report the summary of that model.

```r
#solution
summary(aov(BiomassT2~Species*Fert, data=Plant))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Species       1    895     895   0.324   0.5707
## Fert          2  85387   42693  15.447 1.58e-06 ***
## Species:Fert  2  13206    6603   2.389   0.0973 .
## Residuals    94 259809    2764
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

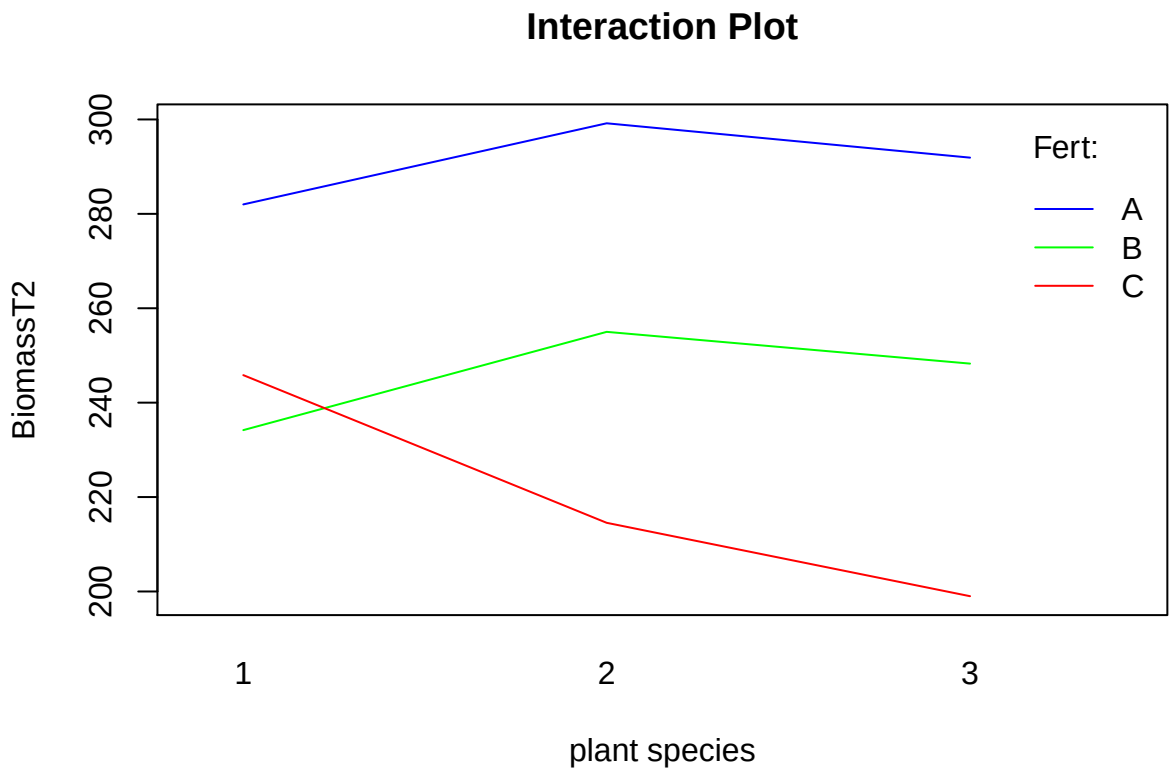
Exercise 9

Create an interaction plot for your two categorical predictors with respect to BiomassT2 as the response.

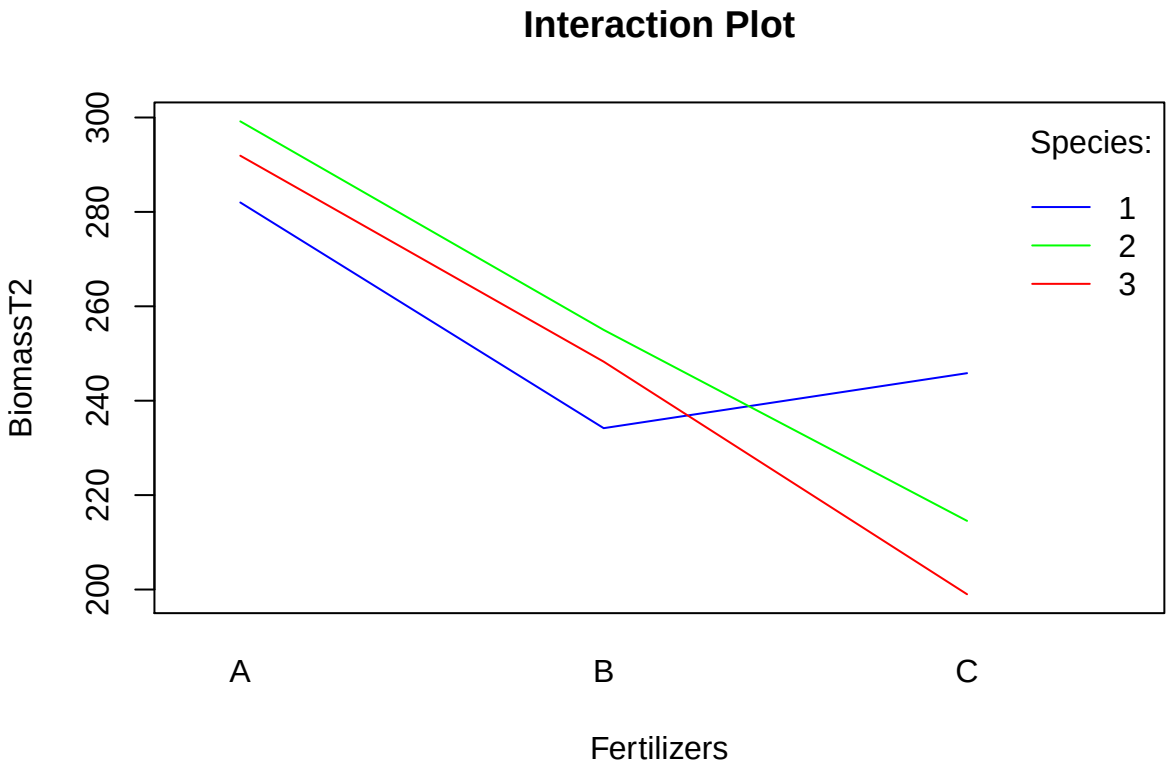
Please include a title, axes labels and a label for your legend (`trace.label`). Also give each line a different color. All other customization optional.

Hint: Look at the documnation for ***`interaction.plot`*** *to see specific names for each feature.*

```
#solution
interaction.plot(x.factor = Plant$Species, trace.factor = Plant$Fert, response = Plant$BiomassT2, col =
```




```
interaction.plot(x.factor = Plant$Fert, trace.factor = Plant$Species, response = Plant$BiomassT2, col =
```



Exercise 10

Which statement best interprets the interaction plot, as well as the results we saw in the full model from Exercise 7?

No alpha level is provided for you. . . choose the language that you think best describes the results you see!

- there is a strong interaction between Species and Fertilizer when predicting BiomassT2
- **There is mild interaction between Species and Fertilizer in predicting BiomassT2**
- there is a strong interaction between Species and BiomssT2 when predicting Species
- There is mild interaction between Species and BiomassT2 in predicting Fertilizer

Exercise 11

We running a one-way ANOVA on three groups that really do come from different distributions with different means.

The higher ther variability within each group on average, then how will that affect our power to detect a difference?

- This will make it easier to reject the null hypothesis (more power)

- This will make it harder to reject the null hypothesis (less power)

- This will not affect our ability to reject the null hypothesis

Exercise 12

Why might we choose to do an ANOVA rather than a series of two-sample t-tests to compare the means of each pair of groups I have?

- Because two-sample t-tests require assumptions that one-way ANOVA does not require
- Because we may greatly increase the chance of a Type I error when completing multiple separate tests.

- Because ANOVA tests whether the variance of each distribution is different, not simply the means.