Midterm Exam

Jason Lu

11/2/2020

Instruction

This is your midterm exam that you are expected to work on it alone. You may NOT discuss any of the content of your exam with anyone except your instructor. This includes text, chat, email and other online forums. We expect you to respect and follow the GRS Academic and Professional Conduct Code.

Although you may NOT ask anyone directly, you are allowed to use external resources such as R codes on the Internet. If you do use someone's code, please make sure you clearly cite the origin of the code.

When you finish, please compile and submit the PDF file and the link to the GitHub repository that contains the entire analysis.

Introduction

In this exam, you will act as both the client and the consultant for the data that you collected in the data collection exercise (20pts). Please note that you are not allowed to change the data. The goal of this exam is to demonstrate your ability to perform the statistical analysis that you learned in this class so far. It is important to note that significance of the analysis is not the main goal of this exam but the focus is on the appropriateness of your approaches.

Data Description (10pts)

Please explain what your data is about and what the comparison of interest is. In the process, please make sure to demonstrate that you can load your data properly into R.

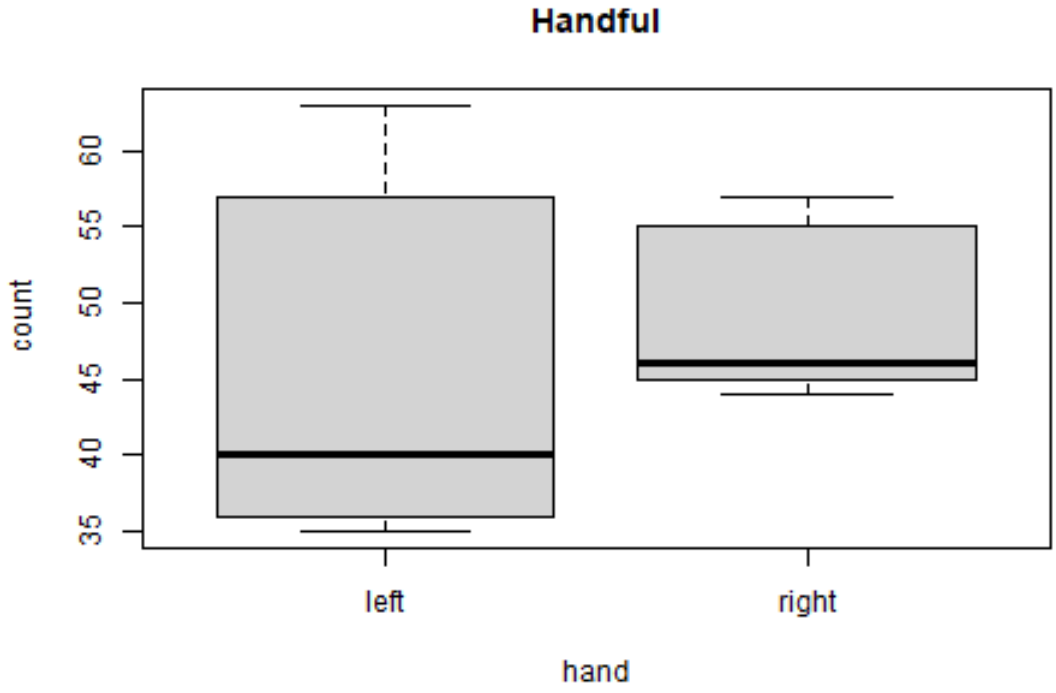
```
# My data is about gathering spinach leaves with both hands with different methods to see if
#   my left hand equally approximates quantities as my right hand.
# The comparison of interest is left hand vs right hand.

setwd("C:/Users/lujas/Desktop/MA 678")
spinach <- read.csv("Data Collection Assignment Data Spinach.csv")
```

EDA (10pts)

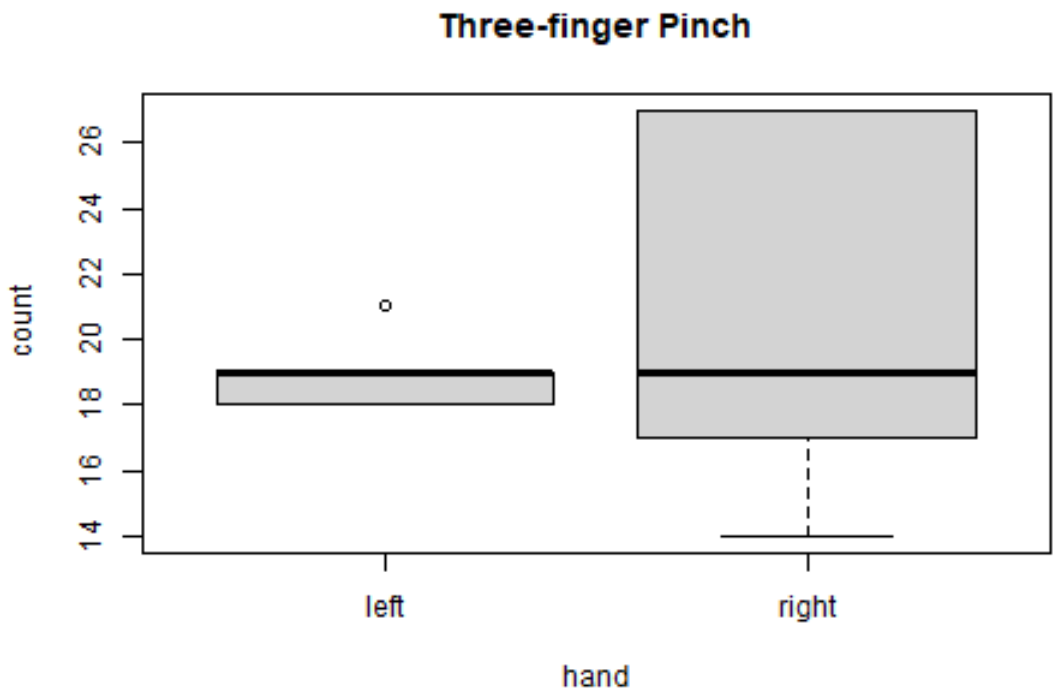
Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

```
# need to reformat my data
left <- subset(spinach, select=-c(Right))
left$hand <- "left"
colnames(left) <- c("Method","count","hand")

right <- subset(spinach, select=-c(Left))
right$hand <- "right"
```

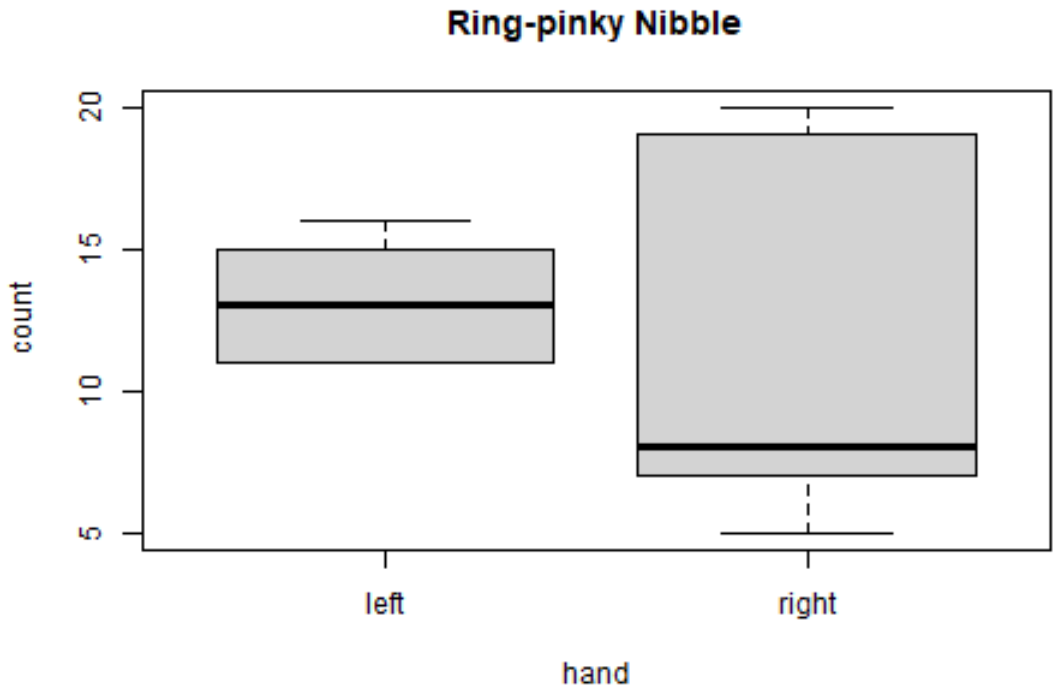
```r
colnames(right) <- c("Method","count","hand")

data <- rbind(left, right)
handful <- subset(data, Method=="Handful")
pinch <- subset(data, Method=="Three-finger Pinch")
nibble <- subset(data, Method=="Ring-pinky Nibble")

# Figures
boxplot(count~hand, data=handful, main="Handful")
```




```r
boxplot(count~hand, data=pinch, main="Three-finger Pinch")
```




```
boxplot(count~hand, data=nibble, main="Ring-pinky Nibble")
```



Power Analysis (10pts)

Please perform power analysis on the project. Use 80% power, the sample size you used and infer the level of effect size you will be able to detect. Discuss whether your sample size was enough for the problem at hand. Please note that method of power analysis should match the analysis. Also, please clearly state why you should NOT use the effect size from the fitted model.

```
pwr.t.test(n=5,d=NULL,sig.level=0.05,power=0.8,type = "paired")
```

```
##
##      Paired t test power calculation
##
##               n = 5
##               d = 1.682001
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

```
# d = 1.682
# I do not think my sample size was large enough since 5 is a small size alongside a lot of
# variance already in grabbing spinach leaves with a high variance corresponding to
# sample size of 5 in each category.

# I should not use the effect size from the fitted model since it is overestimated.
```

Modeling (10pts)

Please pick a regression model that best fits your data and fit your model. Please make sure you describe why you decide to choose the model. Also, if you are using GLM, make sure you explain your choice of link function as well.

```
fit <- glm(count ~ hand + Method, data=data)
summary(fit)
```

```
##
## Call:
## glm(formula = count ~ hand + Method, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -12.200   -4.175   -0.900    5.450   15.800
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 47.200      2.495  18.915  < 2e-16 ***
## handright                    1.200      2.495   0.481    0.635
## MethodRing-pinky Nibble    -35.300      3.056 -11.551 9.72e-12 ***
## MethodThree-finger Pinch   -27.900      3.056  -9.129 1.36e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 46.7)
##
##     Null deviance: 8155.9  on 29  degrees of freedom
```

```
## Residual deviance: 1214.2  on 26  degrees of freedom
## AIC: 206.16
##
## Number of Fisher Scoring iterations: 2
```

```
# I chose this model since I want to see if the hand I used significantly altered how many
#   leaves I picked, so a linear regression seems appropriate.
# Used an identity link (default) because I want a linear regression
```

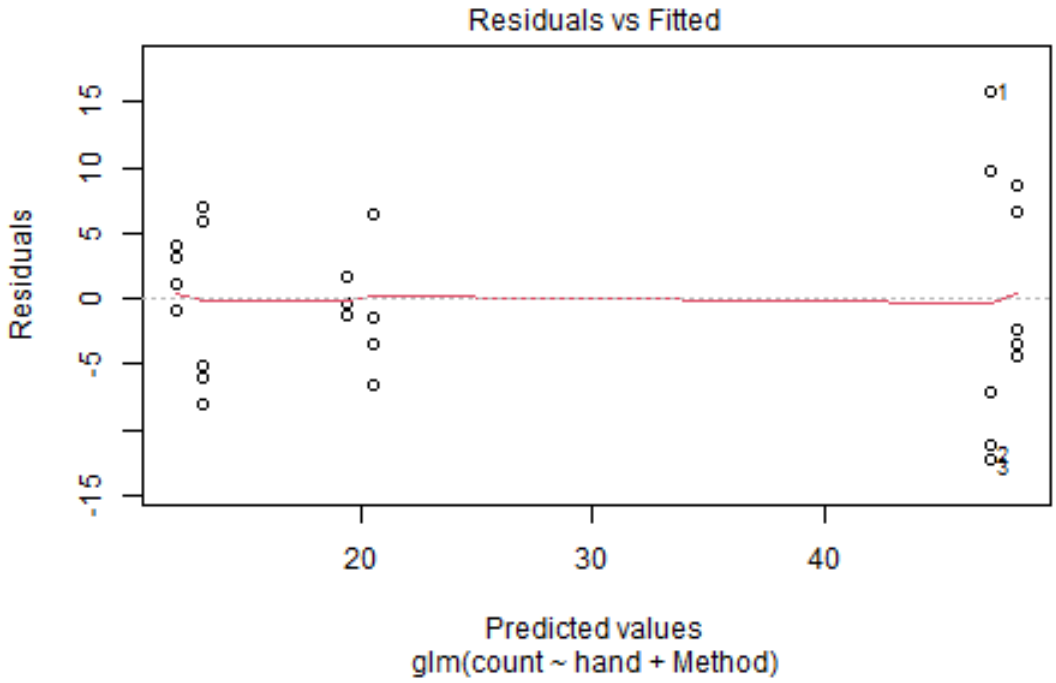
Validation (10pts)

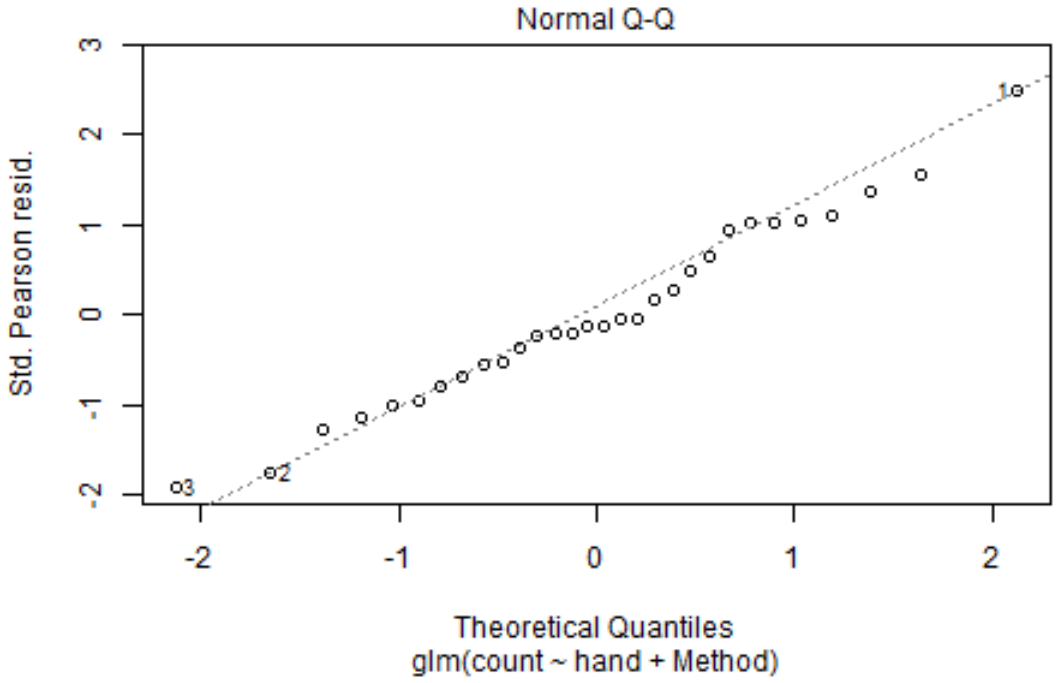
Please perform a necessary validation and argue why your choice of the model is appropriate.

```
pacman::p_load("boot")
library(boot)

cv.glm(data,fit)$delta[1]
```

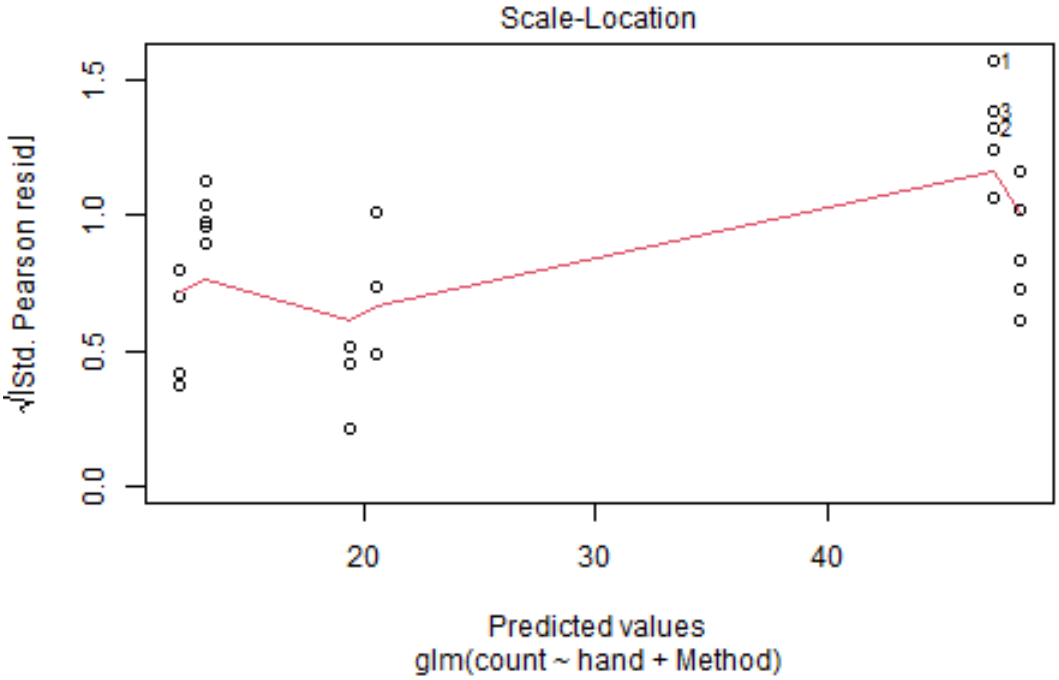
```
## [1] 53.88462
```

```
# Although my prediction error is seemingly large, I believe my choice of model is
#   appropriate because I am testing for whether hand choice alters the count.
# I am not trying to find probability of hand (in which a logistic regression is needed).

plot(fit)
```






```
## hat values (leverages) are all = 0.1333333
##  and there are no factor predictors; no plot no. 5
```



```
# residual plot (no fan shape) and normality (follows the line) not violated
```

Inference (10pts)

Based on the result so far please perform statistical inference to compare the comparison of interest.

```
confint(fit)
```

```
## Waiting for profiling to be done...
```

```
##                                 2.5 %     97.5 %
## (Intercept)                42.309245  52.090755
## handright                  -3.690755   6.090755
## MethodRing-pinky Nibble  -41.289927 -29.310073
## MethodThree-finger Pinch -33.889927 -21.910073
```

```
# Because the 95% CI contains 0, I fail to reject the null hypothesis.
# H0: left and right hands equally approximate quantities
```

Discussion (10pts)

Please clearly state your conclusion and the implication of the result.

There is no difference in me approximating ingredients with my right hand than my left. This means that I can use them interchangeably when cooking without having to worry about self-measurements being different when using either hand.

Limitations and future opportunity. (10pts)

Please list concerns about your analysis. Also, please state how you might go about fixing the problem in your future study.

My sample size was small which led to a larger necessary effect size for significance. It also led to higher standard deviations, limiting the validity of the experiment. I can fix this by doing more samples. My validation error was pretty large, also can be fixed by more samples in each method. I can only limit this to spinach leaves, however further experiments done should involve seasonings such as salt or pepper, as I did not have a precise scale to measure them.

Comments or questions

If you have any comments or questions, please write them here.