Stat 21 Homework 10

Rubric and Solutions

Part I: Concept problems

For Problems 1-3 consider the following setting.

A researcher is interested in studying the size of hatchling Ornate Box Turtles based on the state in which they are found. He gathers a random sample of turtles in each of three different states: Nebraska, Oklahoma, and Texas, wondering if the size changes from North to South.

Problem 1

(a) State the appropriate null hypothesis, both in symbols and in words.

(b) What additional information do you need about these three samples in order to conduct ANOVA to determine if there is a statistically significant difference among these three means?

(c) What additional information do you need in order to assess whether the conditions for ANOVA are satisfied?

Solution to Problem 1

(a) The null hypothesis is that the average size of turtles is the same in all three states. In symbols, $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ (or$H_0 : \beta_2 = \beta_3 = 0$ or $H_0 : \mu_1 = \mu_2 = \mu_3 = 0$) where group 1 represents those in Nebraska, group 2 represents those in Oklahoma, and group 3 represents those in Texas.

(b) We need to have the actual values of the size of the turtles for all turtles in each group in order to conduct the ANOVA.

(c) We need to compute the residuals and examine the widths of boxplots for each group in order to assess whether the conditions are met for ANOVA.

Rubric for Problem 1

- 1 point - made an attempt at all parts
- 0 points - left blank or wrote something nonsensical

Problem 2

Suppose that the sample sizes are 15 for each of the three groups, and also suppose that the standard deviations of the turtle sizes are very similar in the three groups. Assume that all three populations do, in fact, have the same standard deviation. Suppose that the three sample means turn out to be 18.3, 20.1, and $22.4mm$.

Without doing any ANOVA F-test calculations, state a value for the standard deviation that would lead you to reject. Explain your answer, as if to a peer who has not taken a statistics course, without resorting to formulas or calculations.

Solution to Problem 2

Answers will vary. One possible answer is a common standard deviation of 0.2. With a standard deviation of this size, there will be very little overlap between observations in the three groups. This would suggest that the means are, in fact, different.

Rubric for Problem 2

- 1 point - made an attempt at all parts
- 0 points - left blank or wrote something nonsensical

Problem 3

Under the same assumptions as Problem 1, now repeat Problem 1, but state a value for the standard deviation that would lead you to fail to reject.

Solution to Problem 3

Answers will vary. One possible answer is a common standard deviation of 8.0. Since the difference between the largest mean and the smallest mean is only 4.1, with a standard deviation of 8.0 in each group, there will be lots of overlap between observations in the three groups, which makes it harder to conclude that there is a true difference in the population means.

Rubric for Problem 3

- 2 points - made an attempt at all parts
- 0 points - left blank or wrote something nonsensical

Part II: R Problems

Amyloid (Abeta) is a protein fragment that has been linked to Alzheimer's disease. Autopsies from a sample of Catholic priests included measurements of Abeta (pmol/g tissue from the posterior cingulate cortex) from three groups: subjects who had exhibited no cognitive impairment before death, subjects who had exhibited mild cognitive impairment, and subjects who had mild to moderate Alzheimer's disease. The data are in the file Amyloid.

```
library(Stat2Data)
data(Amyloid)
Amyloid %>% head
```

```
##   Group Abeta
## 1   NCI   114
## 2   NCI    41
## 3   NCI   276
## 4   NCI     0
## 5   NCI    16
## 6   NCI   228
```

Problem 4

(a) Report the sample sizes, sample means, and sample standard deviations for each group.

(b) Make parallel boxplots of the data by group. What do these plots indicate about whether an analysis of variance model would be appropriate?

Solution to Problem 4

```
Amyloid$Group %>% summary
```

```
## mAD MCI NCI
##  17  21  19
```

```
mAD_group = Amyloid %>% filter(Group=="mAD")
  mAD_group$Abeta %>% mean
```

```
## [1] 761.2941
```

```
  mAD_group$Abeta %>% sd
```

```
## [1] 426.6942
```

```
MCI_group = Amyloid %>% filter(Group=="MCI")
  MCI_group$Abeta %>% mean
```

```
## [1] 341.0476
```

```
  MCI_group$Abeta %>% sd
```

```
## [1] 406.4092
```

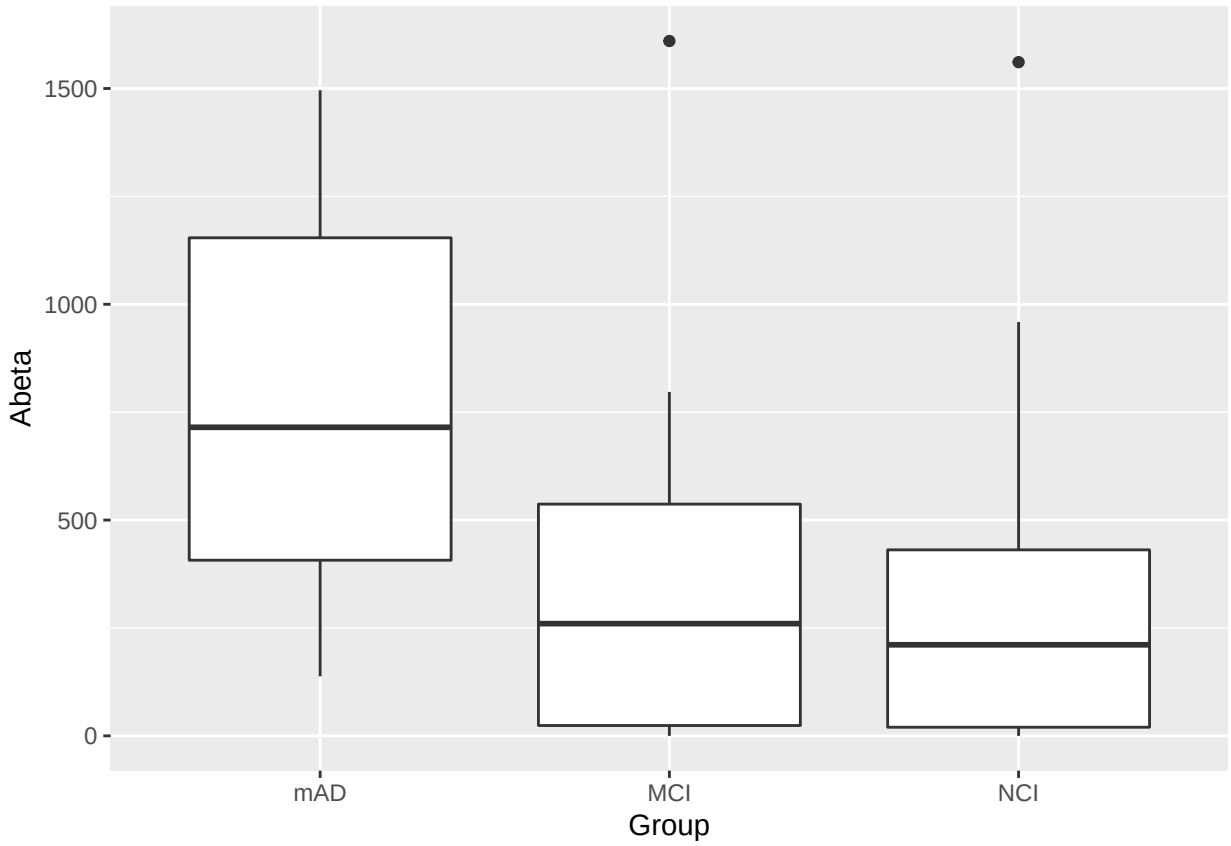
```
NCI_group = Amyloid %>% filter(Group=="NCI")
  NCI_group$Abeta %>% mean
```

```
## [1] 336.2632
```

```
  NCI_group$Abeta %>% sd
```

```
## [1] 435.6096
```

```
ggplot() + geom_boxplot(aes(x=Group, y=Abeta), Amyloid)
```



Variable	Group	N	Mean	StDev
ABeta-42	mAD	17	761.3	426.7
	MCI	21	341.0	406.4
	NCI	19	336.3	435.6

(a)

(b) Parallel boxplots show that the MCI and NCI groups have skewed distributions, so the normality condition for ANOVA is not met.

Rubric for Problem 4

- 2 points - provides all correct information in part (a) and has correct boxplots and code
- 1 point - at least either all aswers in (a) are correct or the code and boxplot in (b) are correct but not both

Problem 5

(a) Transform the data by taking the square root of each observation. Report the sample means and sample standard deviations for each group in the transformed scale.

(b) Make parallel boxplots of the transformed data by group. What do these plots indicate about the amount of Abeta levels in the three groups?

(c) What do the parallel boxplots indicate about whether an analysis of variance model would be appropriate?

(d) Conduct an ANOVA using the transformed data. Interpret the results. What do you conclude about Abeta and cognitive impairment?

```
new_data <- Amyloid %>% mutate(trans_Abeta = sqrt(Abeta))
new_data$Group %>% summary
```

```
## mAD MCI NCI
##  17  21  19
```

```
new_mAD_group = new_data %>% filter(Group=="mAD")
  new_mAD_group$Abeta %>% mean
```

```
## [1] 761.2941
```

```
  new_mAD_group$Abeta %>% sd
```

```
## [1] 426.6942
```

```
new_MCI_group = new_data %>% filter(Group=="MCI")
  new_MCI_group$Abeta %>% mean
```

```
## [1] 341.0476
```

```
  new_MCI_group$Abeta %>% sd
```

```
## [1] 406.4092
```

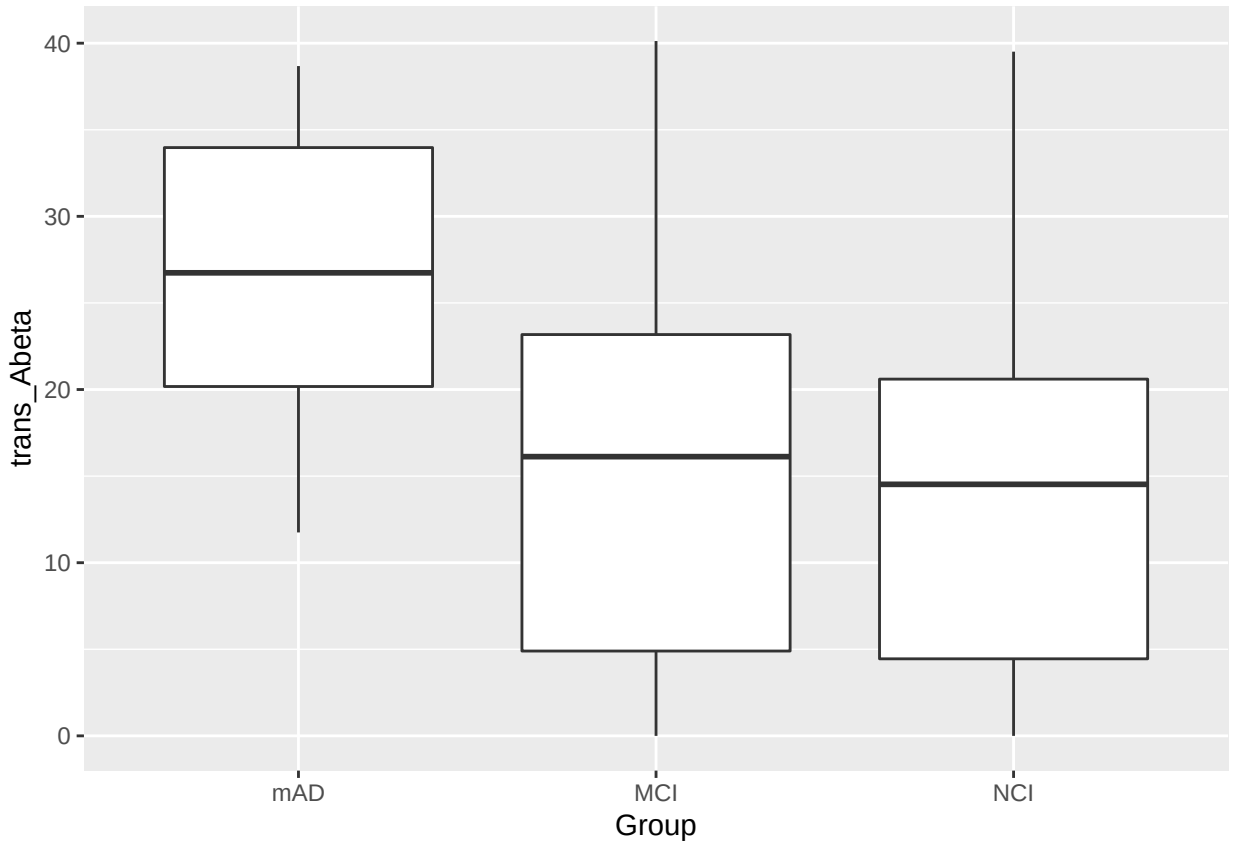
```
new_NCI_group = new_data %>% filter(Group=="NCI")
  new_NCI_group$Abeta %>% mean
```

```
## [1] 336.2632
```

```
  new_NCI_group$Abeta %>% sd
```

```
## [1] 435.6096
```

```
ggplot() + geom_boxplot(aes(x=Group, y=trans_Abeta), new_data)
```




```
trans_anova <- lm(trans_Abeta ~ Group, new_data)
trans_anova %>% summary
```

```
##
## Call:
## lm(formula = trans_Abeta ~ Group, data = new_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6890  -8.9444   0.3219   7.8023  25.4358
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   26.418      2.607  10.133 4.28e-14 ***
## GroupMCI     -11.729      3.507  -3.344  0.00151 **
## GroupNCI     -12.168      3.589  -3.391  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.75 on 54 degrees of freedom
## Multiple R-squared:  0.2143, Adjusted R-squared:  0.1852
## F-statistic: 7.364 on 2 and 54 DF,  p-value: 0.001486
```

Solution to Problem 5

(a) In square root scale, the sample means are 26.4, 14.7, and 14.2. The sample SDs are 8.2, 11.5, and 11.9.

(b) The parallel boxplots show that Abeta levels tend to be higher in the mAD group than in the other

two groups.

(c) Parallel boxplots of the transformed data show reasonable symmetry for each of the groups. The normality condition for ANOVA is now met.

(d) The ANOVA F-statistic is 7.36 and the P-value is 0.001. There is strong evidence for the alternative hypothesis that Abeta is related to group membership.

Rubric for Problem 5

- 2 points - answers are correct and supported with written statements
- 1 point - missing written statements for any of part (a) - (d) (e.g. just showing code and output) but otherwise answers are correct

Problem 6

Recall the data from HW 9 on exams administered to firefighters hoping to qualify for promotion to either Lieutenant or Captain in the city fire department. A final score consisting of a 60% weight for the written exam and a 40% weight for the oral exam was computed for each person who took the exam. Those people receiving a total score of at least 70% were deemed to be eligible for promotion.

In a situation where t openings were available, the people with the top $t + 2$ scores would be considered for those openings. A concern was raised, however, that the exams were discriminatory with respect to race and a lawsuit was filed. The data are given in the data file `Ricci`.

```
library(Stat2Data)
data(Ricci)
Ricci %>% head
```

```
##   Race Position  Oral Written Combine
## 1    W  Captain 89.52      95  92.808
## 2    W  Captain 80.00      95  89.000
## 3    W  Captain 82.38      87  85.152
## 4    W  Captain 88.57      76  81.028
## 5    W  Captain 76.19      84  80.876
## 6    H  Captain 76.19      82  79.676
```

In HW 9, you discovered that there was at least one significant difference in mean combined score for the three different groups of firefighters who took the promotion exam. Now, use Fisher's LSD to find the differences that are significant.

Solution to Problem 6

```
library(DescTools)
regmod <- lm(Combine ~ Race, Ricci)

ANOVA_mod <- regmod %>% aov
PostHocTest(ANOVA_mod, method="lsd")
```

```
##
##   Posthoc multiple comparisons of means : Fisher LSD
##     95% family-wise confidence level
##
## $Race
##         diff    lwr.ci    upr.ci    pval
## H-B 1.600077 -3.186150  6.386304 0.50917
## W-B 8.941203  5.104311 12.778095   1e-05 ***
```

```
## W-H 7.341125  3.272418 11.409832 0.00052 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The following output shows that, in fact, white people have a significantly higher mean score than people of either other group, but there is no significant difference between Black people and Hispanic people.

Rubric for Problem 6

- 2 points - correct code supported with a written interpretation of results
- 1 point - correct code output but missing written statement interpreting code output