Stat 21 Homework 9

Solutions and Rubric

Part I: Concept problems

Problem 1

Aphids (a type of small insect) produce a form of liquid waste, called honeydew, when they eat plant sap. An experiment was conducted to see whether the amounts of honeydew produced by aphids differ for different combinations of type of aphid and type of host plant. The following ANOVA table was produced with the data from this experiment

Analysis of variance table

Source	DF	SS	MS	F-value	P-value
Aphid/host combination	5	24.9	4.9807	5.75	0.000
Error	46	39.87	0.8667		
Total	51	64.77			

(a) Fill in the three missing values in this ANOVA table. Also show how you calculate them.

(b) How many different aphid/plant combinations were considered in this analysis? Explain how you know.

(c) Summarize the conclusion from this ANOVA (in context).

Solution to Problem 1

(a) The degrees of freedom for aphid/plant combination can be found by subtracting the degrees of freedom for Error from the degrees of freedom for Total DF = 51-46 = 5. The sum of squares for aphid/plant combination can be found by subtracting the sum of squares for Error from the sum of squares for Total. SSGroups = 64.77-39.87 = 24.90. The F-value can be found by dividing the mean square for aphid/plant combination by the mean square for Error: F 4.9807/0.8667 = 5.75.

(b)Since the degrees of freedom for aphid/plant combination is 5 and the degrees of freedom for treatments is the number of groups minus 1, the number of different occupations considered is 6.

(c)Since the P-value is so small, we reject the null hypothesis. We have strong evidence that mean amount of honeydew produced by aphids is different for at least one aphid/plant combination.

Rubric for Problem 1

(a) 1 point - correct answers in table and show/explain reasoning 0.5 point - correct answers in table but no work shown

(b) 1 point - correct answer with some mathematical or written explanation 0.5 point - correct answer but no justification

(c) 1 point - correct conclusion with some statement interpreting in context 0.5 points - correct conclusion but no interpretation within context

Problem 2

Nationally, the abuse of methamphetamine has become a concern, not only because of the effects of drug abuse, but also because of the dangers associated with the labs that produce them. A stratified random sample of a total of 12 counties in Iowa (stratified by size of county—small, medium, or large) produced the following ANOVA table relating the number of methamphetamine labs to the size of the county. Use this table to answer the following questions.

Analysis of variance

Source	DF	SS	MS	F-value
Type	2	37.51	18.755	5.101
Error	9	33.09	3.677	
Total	11	70.60		

(a) Fill in the values missing from the table and show your work.

(b) What does the MS for county type tell you?

(c) Find the P-value for the F-test in the table.

(d) Describe the hypotheses tested by the F-test in the table, and using the P-value from part (c), give an appropriate conclusion.

Solution to Problem 2

(a) See table above.

(b) The MS for county type measures the amount of variability between the means of the three county types.

(c) Using an F-distribution with 2 and 9 degrees of freedom, the P-value is 0.033.

(d) The null hypothesis is that the three sizes of counties have the same average number of meth labs. In symbols, this is $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$. The alternative hypothesis is that at least one type of county has a different average number of meth labs than the other two sizes of county. With a P-value of 0.033, we reject the null hypothesis and conclude that at least one type of county has a different average number of meth labs.

Rubric for Problem 2

(a) 0.5 points - correct answer and shows work and/or provides explanation 0 points - correct answer but does not show work

(b) 0.5 points - provides some description mentioning variability among the group levels

(c) 0.5 points - correct answer

(d) 0.5 points - correct null hypothesis 0 points - incorrect null hypothesis

Part II: R Problems

Context for Problems 3-5:

The city of New Haven, Connecticut, administered exams (both written and oral) in November and December of 2003 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
```

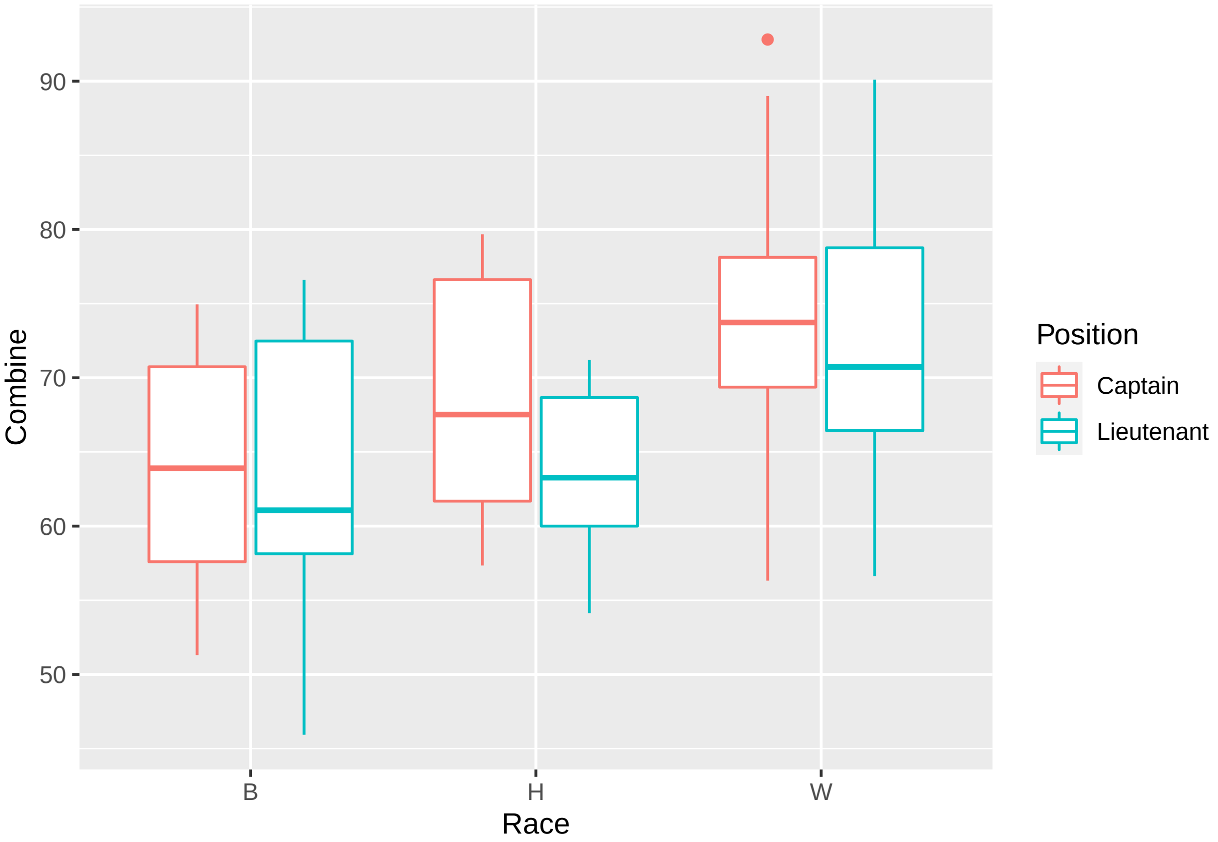
For each person who took the exams, there are measurements on their race or ethnicity (Black, white, or Hispanic), which position they were trying for (Lieutenant, Captain), scores on the oral and written exams, and the combined score.

The concern over the exams administered by the city was that they were discriminatory based on race. Here we concentrate on the overall, combined score on the two tests for these people seeking promotion and we analyze the average score for the three different races.

Problem 3

Use a graphical approach to answer the question of whether the average combined score is different for the three races. What do the graphs suggest about any further analysis that could be done? Explain.

```
ggplot() + geom_boxplot(aes(y=Combine, x=Race, col=Position), Ricci)
```



Solution to Problem 3

The above are the boxplots of the combined scores, broken down by the race of the individual taking the exam. It appears that there may be a difference between the scores of the various races. Specifically, it appears that whites generally did a little bit better, that Hispanics had scores in the middle, and that blacks had the lowest scores, though the difference between Hispanics and blacks does not appear to be large and might turn out to be statistically insignificant. The spreads of the three samples appear to be similar so continuing by conducting an ANOVA analysis would likely be the next step.

Rubric for Problem 3

2 points - answer notes the possibility of difference in scores among racial/ethnic groups and plot indicates both score and race/ethnicity information (does not have to be the exact same plot as above)

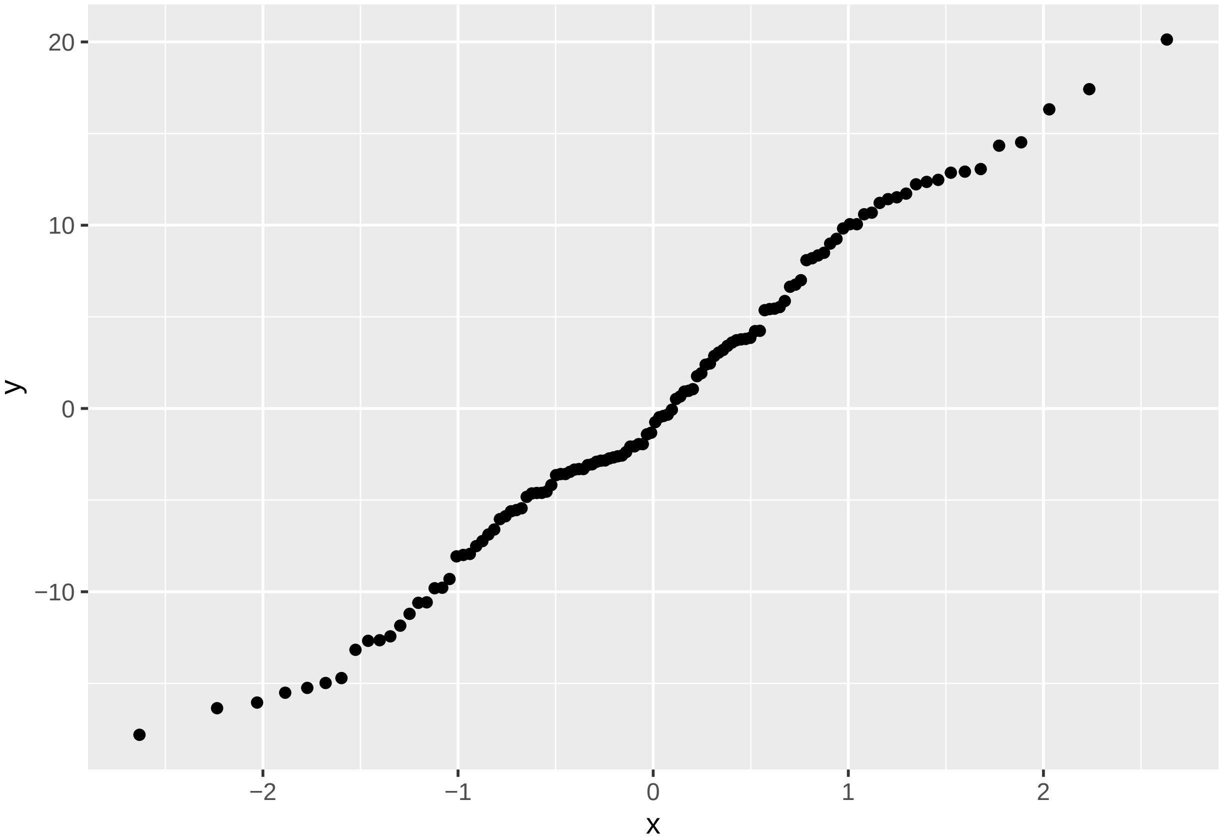
1 point - answer notes the possibility of difference in scores among racial/ethnic groups but plot does not display both score and race/ethnicity information

Problem 4

(a) Check the conditions necessary for conducting an ANOVA to determine if the combined score is significantly different for at least one race.

(b) Write out in words and in symbols the hypotheses that would be tested by an ANOVA model in this setting.

```
ricci_mod <- lm(Combine ~ Race, Ricci)
dat_full <- Ricci %>% mutate(resids = ricci_mod$residuals,
                             fits = ricci_mod$fitted.values)
ggplot() + stat_qq(data=dat_full, aes(sample=resids)) + stat_qq_line()
```



Solution to Problem 4

(a) To check the conditions necessary for conducting an ANOVA we begin by checking the normality of the residuals. Following is a normal plot of the residuals in this case. While there is a small amount of curvature to this plot, there is not enough to be worried about.

Next we check whether the data are consistent with the idea that the variances for the three groups are the same. The residuals versus fitted values plot suggests that the data are comparable with the equal variances requirement. Also the ratio of the largest standard deviation to the smallest is 8.83-7.14 = 1.24 which is much less than the suggested cutoff of 2.

Then we need to evaluate the independence of the errors and the idea that they come from a population with mean 0. In this case, there is no reason to believe that any one individual's exam score is related to any other individual's score, so the independence of the errors seems reasonable.

Finally, we need to evaluate the constant, additive effects assumption. We can do this by asking whether or not it is reasonable to assume that the changes in scores for each racial or ethnic group are not functions of (or dependent on) some lurking variable that is relevant but not measured (e.g. depression, economic stability, age, etc.). This is something that can't really be assessed without domain knowledge. The additive effect part can be assessed in part by looking at boxplots of the grouped scores but since there are only three categories, it's not really possible to determine if the differences may be multiplicative or exponential.

(b) The null hypothesis is that the mean combined exam score is the same for the populations of white, Black, and Hispanic firefighters in New Haven, Connecticut. The alternative is that at least one group of firefighters has a different mean combined exam score. In symbols we should write $H_0 : \alpha_B = \alpha_H = \alpha_W = 0$ versus $H_A :$ At least one α_k is not 0

Rubric for Problem 4

(a) 1 point - mentioning constant effects and additive effects (or more general "linearity") assumption in addition to the constant variance, independence, and normality conditions 0.5 points - neglecting to mention the linearity (or equivalently, constant and additive effects) assumption or the constant variance assumption

(b) 1 point - valid null can be written with α_j or $\alpha_{B,H,W}$ notation but must also specify the alternative

Problem 5

What conclusions do you reach from the ANOVA analysis? Write a paragraph discussing your conclusions. Include the relevant statistics (F-value, P-value, etc.) as part of your discussion.

```
ricci_mod %>% summary
```

```
##
## Call:
## lm(formula = Combine ~ Race, data = Ricci)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.804  -5.290  -1.036   5.780  20.130
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   63.736      1.639  38.892  < 2e-16 ***
## RaceH          1.600      2.416   0.662    0.509
## RaceW          8.941      1.937   4.616 1.03e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.516 on 115 degrees of freedom
## Multiple R-squared:  0.1912, Adjusted R-squared:  0.1772
## F-statistic:  13.6 on 2 and 115 DF,  p-value: 5.014e-06
```

Solution to Problem 5

The ANOVA suggests that there is a difference for the mean combined exam score for at least one of the three groups of firefighters. The F-statistic is 13.60 (with 2 and 115 df), which results in a P-value near 0. Looking at the boxplots from earlier, it would appear that white firefighters have a higher mean score than the other two groups. Without further testing, we cannot tell whether there is a statistically significant difference between Black and Hispanic firefighters.

Rubric for Problem 5

1 point - Correct conclusion that references a relevant statistic (doesn't have to be just the F-statistic and p-value)

0.5 points - Minimal statement of interpretation without explicitly referencing the context of the problem