Math 328 Chapter 8 HW - Part A

David Oniani

April 11, 2021

## Setup

library(Stat2Data)  
library(agricolae)  
library(emmeans)  
library(mosaic)

## Registered S3 method overwritten by 'mosaic':  
## method from   
## fortify.SpatialPolygonsDataFrame ggplot2

##   
## The 'mosaic' package masks several functions from core packages in order to add   
## additional features. The original behavior of these functions should not be affected by this.

##   
## Attaching package: 'mosaic'

## The following objects are masked from 'package:dplyr':  
##   
## count, do, tally

## The following object is masked from 'package:Matrix':  
##   
## mean

## The following object is masked from 'package:ggplot2':  
##   
## stat

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,  
## quantile, sd, t.test, var

## The following objects are masked from 'package:base':  
##   
## max, mean, min, prod, range, sample, sum

library(multcomp)

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

##   
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':  
##   
## geyser

## Exercise 8.2

In case of two-way ANOVA, we must apply Levene’s test to groups of data formed by the cells (i.e., the combinations of levels of factors), rather than to groups formed by levels of a single factor. In this case, Levene’s test was applied to groups by levels of Factor A (single factor) and Factor B (single factor) separately and thus, such application of Levene’s test would not be appropriate.

## Exercise 8.14

# Load the data  
data("SeaSlugs")  
  
# a) Use Fisher’s LSD intervals to find any differences that exist between the  
# percent of larvae that metamorphosed in the different water conditions.  
fit1 <- aov(Percent ~ factor(Time), SeaSlugs)  
anova(fit1)

## Analysis of Variance Table  
##   
## Response: Percent  
## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Time) 5 0.63091 0.126182 5.9648 0.0006067 \*\*\*  
## Residuals 30 0.63464 0.021155   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

print(agricolae::LSD.test(fit1, "factor(Time)"))

## $statistics  
## MSerror Df Mean CV t.value LSD  
## 0.02115452 30 0.2716667 53.53838 2.042272 0.1714963  
##   
## $parameters  
## test p.ajusted name.t ntr alpha  
## Fisher-LSD none factor(Time) 6 0.05  
##   
## $means  
## Percent std r LCL UCL Min Max Q25 Q50  
## 0 0.5356667 0.1687859 6 0.41440050 0.6569328 0.357 0.857 0.47525 0.5000  
## 10 0.1776667 0.1238881 6 0.05640050 0.2989328 0.067 0.333 0.08350 0.1330  
## 15 0.1833333 0.1470397 6 0.06206716 0.3045995 0.000 0.333 0.05350 0.2405  
## 20 0.2191667 0.1383914 6 0.09790050 0.3404328 0.067 0.437 0.10775 0.2335  
## 25 0.1686667 0.1484650 6 0.04740050 0.2899328 0.000 0.412 0.08350 0.1330  
## 5 0.3455000 0.1423921 6 0.22423383 0.4667662 0.125 0.467 0.26050 0.4000  
## Q75  
## 0 0.52475  
## 10 0.28300  
## 15 0.28125  
## 20 0.26700  
## 25 0.23350  
## 5 0.45025  
##   
## $comparison  
## NULL  
##   
## $groups  
## Percent groups  
## 0 0.5356667 a  
## 5 0.3455000 b  
## 20 0.2191667 bc  
## 15 0.1833333 bc  
## 10 0.1776667 bc  
## 25 0.1686667 c  
##   
## attr(,"class")  
## [1] "group"

# b) Use Tukey’s HSD intervals to find any differences that exist between the  
# percent of larvae that metamorphosed in the different water conditions.  
fit2 <- aov(Percent ~ factor(Time), SeaSlugs)  
TukeyHSD(fit2)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Percent ~ factor(Time), data = SeaSlugs)  
##   
## $`factor(Time)`  
## diff lwr upr p adj  
## 5-0 -0.190166667 -0.4455792 0.06524590 0.2397208  
## 10-0 -0.358000000 -0.6134126 -0.10258743 0.0023231  
## 15-0 -0.352333333 -0.6077459 -0.09692077 0.0027831  
## 20-0 -0.316500000 -0.5719126 -0.06108743 0.0085222  
## 25-0 -0.367000000 -0.6224126 -0.11158743 0.0017407  
## 10-5 -0.167833333 -0.4232459 0.08757923 0.3666256  
## 15-5 -0.162166667 -0.4175792 0.09324590 0.4038772  
## 20-5 -0.126333333 -0.3817459 0.12907923 0.6641386  
## 25-5 -0.176833333 -0.4322459 0.07857923 0.3114499  
## 15-10 0.005666667 -0.2497459 0.26107923 0.9999998  
## 20-10 0.041500000 -0.2139126 0.29691257 0.9960188  
## 25-10 -0.009000000 -0.2644126 0.24641257 0.9999978  
## 20-15 0.035833333 -0.2195792 0.29124590 0.9980127  
## 25-15 -0.014666667 -0.2700792 0.24074590 0.9999748  
## 25-20 -0.050500000 -0.3059126 0.20491257 0.9901287

# c) Were your conclusions to (a) and (b) different? Explain. If so, which  
# would you prefer to use in this case and why?  
#   
# No code needed, see the response below

1. Means with the same letter seem to be different at intervals 0, 5, and 25 as compared with those at intervals 20, 15, and 10. Other than this, there does not seem to be any significant differences between the time intervals 20, 15, and 10 and other time intervals.
2. It seems like intervals 0 - 10, 0 - 15, 0 - 20, and 0 - 25 are the most significant. Hence, there are differences between the percent of larvae that metamorphosed in the different water conditions.
3. Conclusions in (a) and (b) are indeed different. Fisher’s LSD seems to be narrower than Tukey’s HSD. Since in Fisher’s LSD we are likely to have a higher risk of making Type I error, I would prefer Tukey’s HSD in this case.

## Exercise 8.18

# Load the data  
data("Blood1")  
  
# a) What are the hypotheses that describe the test we would like to  
# perform?  
fit1 <- aov(SystolicBP ~ Overwt, Blood1)  
summary(fit1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Overwt 1 27788 27788 38.09 1.4e-09 \*\*\*  
## Residuals 498 363286 729   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# b) Write the contrast of interest in symbols and compute its estimated  
# value.  
ms = mean(SystolicBP ~ Overwt, data=Blood1)  
cc = c(1, -0.5, -0.5)  
contrast = sum(ms \* cc)  
contrast

## [1] -12.45866

# c) What is the standard error of the contrast?  
mse = summary(fit1)[[1]]$"Mean Sq"[2]  
ns = tally(~Overwt, data=Blood1)  
se = sqrt(mse \* sum(cc^2 / ns))  
se

## [1] 2.543255

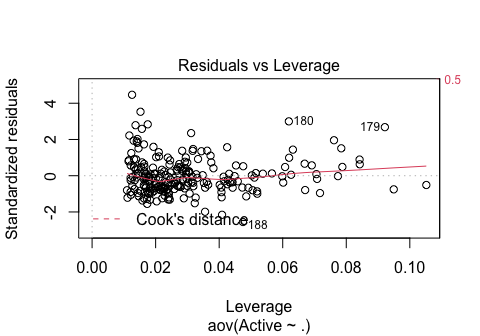
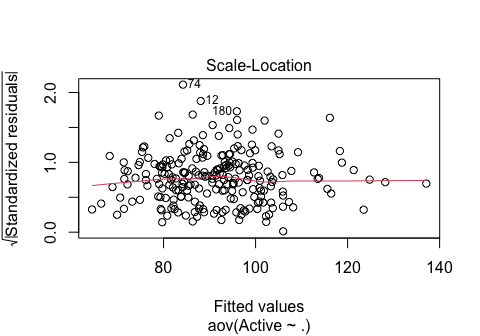
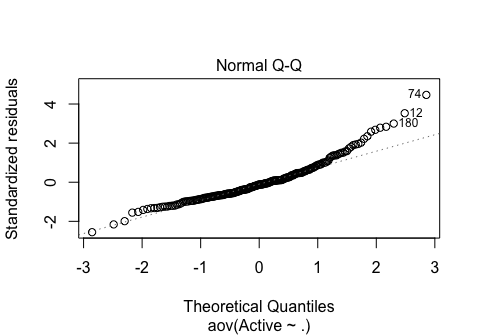
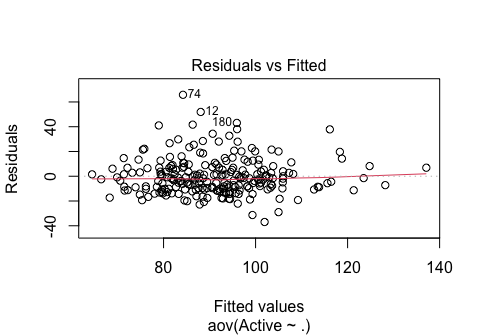
# d) Perform the hypothesis test to test the alternative hypothesis that the  
# mean systolic blood pressure is different for people of normal weight, as  
# compared to people who are overweight or obese. Be sure to give your  
# conclusions.  
t1 = contrast / se  
errordf = fit1$df.residual  
pvalue1 = 2 \* pt(t1, errordf)  
pvalue1

## [1] 1.30648e-06

1. Since we want to compare levels of systolic blood pressure based on weight status, one-way ANOVA would be appropriate. Hence, we would perform a one-way ANOVA for systolic blood pressure by weight status. The model and its summary statistics are shown above.
2. Since we are comparing normal to overweight and obese, we can set normal to 1 and both overweight and obese to -0.5. Hence, the equation will be 1 \* y\_normal - 0.5 \* y\_overweight - 0.5 \* y\_obese with 1 + (-0.5) + (-0.5) = 0 which satisfies the condition that all coefficients add up to 0. After calculating the contrast, we got the value of approximately -12.459.
3. After calculating the standard error of the contrast, we got that its value is approximately 2.543.
4. After performing the hypothesis test, we got the p-value of 1.30648e-06 which is a lot smaller than the 0.05 cutoff. This means that difference is significant and hence, we reject the null and accept alternative, effectively concluding that there is a significant difference of systolic blood pressure based on weight status (normal vs overweight or obese).

## Exercise 8.26

# Load the data  
data("Pulse")  
  
# a) Create one or more graphs to assess the conditions required by ANOVA.  
# Comment on why the Kruskal-Wallis test might be more appropriate  
# than an ANOVA.  
fit1 <- aov(Active ~ ., Pulse)  
plot(fit1)



# b) Run the Kruskal-Wallis test and report the results and conclusions  
kruskal.test(Active ~ Exercise, Pulse)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Active by Exercise  
## Kruskal-Wallis chi-squared = 29.826, df = 2, p-value = 3.337e-07

1. Residuals vs Fitted shows a constant variance and the red line follows the dotted line closely. Hence, based on Residuals vs Fitted plot, the equal variance condition does not seem to be violated (i.e., the equal variance condition is met). The normality condition does not seem to bet met. While most of the points in the middle of the Normal Q-Q do seem to follow the dotted line, there is a significant deviation at both ends. Thus, based on the Normal Q-Q plot, we can conclude that the normality assumption is not met and therefore, is violated. The Scale-Location plot does not show a decreasing trend. Residuals vs Leverage plot does not show any obvious outliers. Although the variance condition is met, the normality assumption is not. If the normality assumption was not violated, using ANOVA would be appropriate, but since it is, Kruskal-Wallis test is more appropriate. Hence, Kruskal-Wallis test is more appropriate since the normality assumption is not met.

* NOTE: We fitted the full model response vs all variables. The book was a bit ambiguous about the exact kind of model they wanted us to fit. That being said, I have also rechecked a model fit1 <- aov(Active ~ Exercise) and the normality assumption is still being violated (i.e., is not met).

1. After performing the Kruskal-Wallis test, we get that the Kruskal-Wallis chi-squared value is 29.826 with the p-value of 3.337e-07 and 2 degrees of freedom. Now, since the p-value is significantly lower than the 0.05 cutoff, we reject the null hypothesis of no difference in the pulse rates (Activity variable) and conclude that there are different pulse rates, depending on exercise given to students.