Lab 02: ANOVA

Trang Nguyen, Dennis Liu

3/9/2017

## Objective

Today's lab is the second one in the lab series *Lab: Bad Practices of Experimental Design and Statistical Analysis*. As mentioned in the first lab (*Lab 01: How is T-test Unreliable?*), this lab aims at introducing *ANOVA* as a better alternative to t-test in answering the question **Does Gender influence the game completion time?**

## About ANOVA

**ANOVA** is an acronym for *Analysis of Variance*. This technique is applied to categorical explanatory variables (called *factor*) and continuous response variable. The advantage of ANOVA is that it can take multiple explanatory variables into consideration. Therefore, compared to *t-test*, *ANOVA* can minimize the influence of lurking variables on analysis results.

## Lab Activity

### 1. Getting Data

library(readr)  
library(dplyr)   
library(ggplot2)  
library(data.table)

To make sure the effectiveness of the analysis, we use the data attached the lab. This data set is the cleaned version of the orginal tangram data set. If you would like to know data cleaning details, please refer to the first lab.

gen\_tangram <- read\_csv("gentangram.csv")

### 2. Setting up data for ANOVA

There are a lot of categorical variables in the data set. Some variables, like *Puzzle Name*, have multiple levels. Since *ANOVA* is basically a between-group comparison, the intuition is that we need to have a good number of observations in each group. The fact that a variable has too many levels will break down the data set into many groups with a few observations only. Therefore, it's important to collapse those levels into some key ones. The following code collapses *Puzzle Name* into 5 main levels.

puzzlename <- c("A Nice Lighthouse", "Diamond", "House of Tangrams", "The Hook")  
gen\_tangram$PuzzleName2 <- ifelse(gen\_tangram$PuzzleName %in% puzzlename, gen\_tangram$PuzzleName, "Others")  
gen\_tangram$PuzzleName2 <- as.factor(gen\_tangram$PuzzleName2)

Recall that in the first lab, we observe that *Time Used* has a widely spread distribution. Therefore, we need to use log transformation to reduce its variance. The response variable we will use in the lab will be *TimeUsed\_log*.

gen\_tangram$TimeUsed\_log <- log(gen\_tangram$TimeUsed)  
gen\_tangram$level\_gender <- as.factor(gen\_tangram$level\_gender)

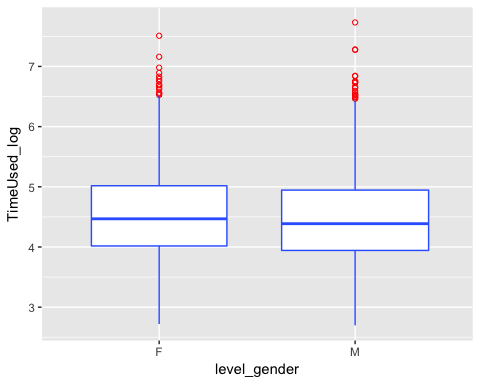
### 3. Using ANOVA

Question 1: Conduct ANOVA analysis with 1 factor only - Gender. Compare the result with the t-test result in the first lab. Is there any difference?

test <- aov(TimeUsed\_log ~ level\_gender, data=gen\_tangram)  
summary(test)

## Df Sum Sq Mean Sq F value Pr(>F)   
## level\_gender 1 8 7.996 15.97 6.5e-05 \*\*\*  
## Residuals 7083 3546 0.501   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

ggplot(gen\_tangram,aes(level\_gender,TimeUsed\_log)) +  
 geom\_boxplot(varwidth = TRUE,fill = "white", colour = "#3366FF",  
 outlier.colour = "red", outlier.shape = 1)



Question 2: Among all categorical variables given in the data set, besides Gender, what are other factors that might influence the game completion time?

As mentioned above, the advantage of *ANOVA* is that the model can take multiple factors into consideration. In the following example, we conduct a two-way ANOVA analysis by using 2 explanatory variables: *Gender* (2 levels) and *Puzzle Name* (5 levels).

Question 3: Use the following code to run a two-way ANOVA analysis. How does the p-value for Gender change compared to the one-way ANOVA test? What are the additional conclusions you can get from this model?

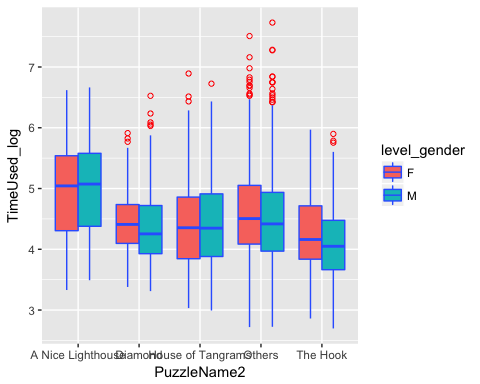
test2 <- aov(TimeUsed\_log ~ level\_gender + PuzzleName2,data=gen\_tangram)  
summary(test2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## level\_gender 1 8 8.00 16.85 4.09e-05 \*\*\*  
## PuzzleName2 4 187 46.77 98.57 < 2e-16 \*\*\*  
## Residuals 7079 3359 0.47   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

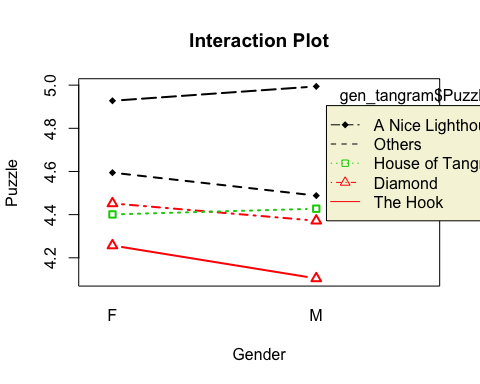
As you can see from the model, besides 2 main factors, *Gender* and *Puzzle Name*, we have an additional interaction term between *Gender* and *Puzzle Name*. This term demonstrates how each factor influences *Time Used* in each level of the other factor. This is called **interaction effect**. The best way to understand **interaction effect** is by visualizing it on an interaction plot.

Question 4: Use the following code to produce relevant plots that visualize the ANOVA model. What can you tell from these plots?

pd = position\_dodge(.2)  
ggplot(gen\_tangram,aes(PuzzleName2,TimeUsed\_log)) +  
 geom\_boxplot(aes(fill=level\_gender),colour = "#3366FF",  
 outlier.colour = "red", outlier.shape = 1)



interaction.plot(gen\_tangram$level\_gender, gen\_tangram$PuzzleName2, gen\_tangram$TimeUsed\_log, type="b", col=c(1:3),   
 leg.bty="o", leg.bg="beige", lwd=2, pch=c(18,24,22),   
 xlab="Gender",   
 ylab="Puzzle",   
 main="Interaction Plot")



Question 5: The model shows that Puzzle Name can influence the game completion time. Now assume that you are a researcher who has to design a study to investigate the influence of Gender on game completion time. Will you have all subjects do the same puzzle or a variety of puzzles? What is the concern for each approach?

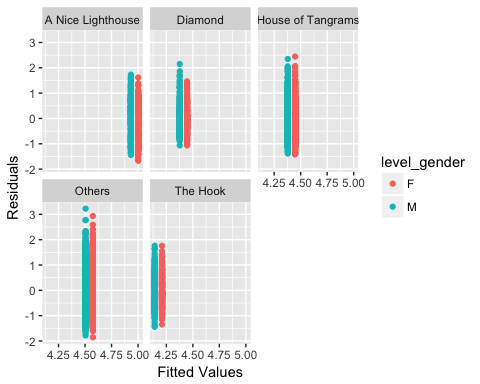
Question 6 (More Practice with ANOVA): Choose a third explanatory variable and write code to conduct a three-way ANOVA test. Graph some relevant plots, and write up your results and conclusion.

### 4. Checking Model Assumption

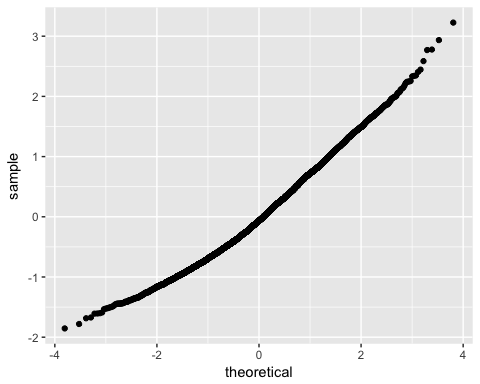
Two most important model assumptions of ANOVA is the normality and homogeneity of variances of residuals. Therefore, it's necessary to look at the distribution of the model's residuals to assess the reliability of our analysis.

Question 7: Use the following code to investigation to distribution of residuals? Is there any significant evidence of model assumption violation?

test\_data <- gen\_tangram  
test\_data$fit <- fitted(test2)  
test\_data$res <- resid(test2)  
ggplot(test\_data, aes(fit, res, colour = level\_gender)) + geom\_point() +  
 xlab("Fitted Values") + ylab("Residuals") + facet\_wrap(~ PuzzleName2)



ggplot(test\_data, aes(sample = res)) + stat\_qq()



## Some limitations of ANOVA

Even though you can have as many factors in an ANOVA model as you want, keep in mind that the model is only reliable when you have a good number of observations in each group. This implies that students must choose a good number of relevant factors to put in the model. This process is often technically called **feature selection**, the selection of the best predictors.  
Another disadvantage of using ANOVA is that explanatory variables are limited to categorical ones only. You can see that there are some continuous variables in the data set that might influence the game completion time but we cannot put them into the ANOVA model.  
Fortunately, there are a couple of alternatives to make up for these two limitations of ANOVA. The next lab, *Lab 03: CART analysis*, will introduce a new analysis technique, called CART analysis or classification tree. This method is not only good for **feature selection** but can also take both continuous and categorical variables into account.