STA303/1002H1S: Methods of Data Analysis II **Winter 2017**

Assignment 1 : ANOVA

Out: Jan. 23, 2017 *Due: 10pm, Feb 5, 2017*

Reminder : You MUST write your solution independently and turn in your own write-up.

This assignment is due 10 :00pm, Sunday, Feb. 5th, 2017. Submit your solution as instructed by Crowdmark, namely, one pdf file for each question.

Late assignments will be subject to a deduction of 5% of the total marks for the assignment for each day late. Any late assignment after the day I post the solution will get zero mark.

Presentation of solutions is very important. A Rmarkdown template for the solution is provided. Please following the instruction in that template to complete your solution. You should produce a PDF file, split the PDF into different files to get your solution PDF for each question. Also, make sure the source R code at the end is complete. Marks will be deducted if the instructions herein are not followed.

Data 1 : Working output data

The file work.csv contains daily part output for 10 workmen over 20 days each.

The variables in the dataset are :

- workman : from 1 to 10. You should turn it into factor variable using *as.factor()*
- y : the daily part output.

Data 2 : Tasting grade of beers data

The file beers.csv tasting grades of different beers as a function of their country of origin and beer style.

The variables in the dataset are :

- name : beer name
- type : factor variable with levels of IPA and Lager.
- country : factor variable with levels of Belgium, UK and USA.
- rating : response variable.

This data set comes with *R*. To access this data set,

```
# load in the data
work=read.table(file.choose(),sep=",",header=T)
str(work)
work$workman=as.factor(work$workman)
str(work) # double check the type of workman again
```

Listing 1 – Install package and read in data

Questions

Using R to do all the analysis on **Data 1** for the following questions.

Q1 (10=3+3+2+2 points)

(a) Calculate the means and standard deviations of output for each workman. Also make a boxplot comparing the part output for the 10 workmen, give a short comment for the boxplot produced. (Show R code and output as your answer)

(b) Applying one-way ANOVA to this data, testing the equality of the output means for the ten workmen. State the null and alternative hypothesis for the p-value in ANOVA output. How significant is the result ? (Show your R code and ANOVA output as well).

(c) ANOVA assumes that the data in each group are distributed normally. This assumption is equivalent tosaying that the residuals of the best-fitting model are distributed normally. Check the normality assumptionby doing a qqnorm plot in conjunction with qqline based on the residuals from the linear regression model fitting. What conclusion do you have from the plot ? (Show Q-Q plot too)

(d) Examine the output variability for the ten workmen using the Bartlett test. What is your conclusion ? (Show your R code and R output).

Q2 (10=3+2+3+2 points)

(a) To stabilizing the variance, we apply Box-cox power transformation, it suggests a simple variance stabilizing of the data. What is the simple transformation on Y suggested from boxcox() ? (Show R code and the optimal optimal λ value from box-cox R output)

(b) Examine the **transformed Y (from Q2-a)** variability for the ten workmen using the Bartlett test. What is your conclusion ? Does it agree or disagree with Q1-d ? (Show R code and output)

(c) Applying one-way ANOVA to this data, testing the equality of the output means for the ten workmen. How significant is the result ? Does it agree with result you have in (Q1-b). Also repeat Q1-c to check the normality assumption for the transformed data, compare to Q1-c, what comment do you have ?

(d) Why would we want to prefer the second ANOVA over the first one, even though both give roughly the same significance ?

Using R to do all the analysis on **Data 2** for the following questions.

Q3 (10=3+3+4 points)

(a) Find the rating mean for each country and type. Find also the cell mean for each treatment combination (county and type combination). You could use R code and R output as our answer.

(b) Create box-plot of rating with respect to two factors, type and country. What can you say about the difference of rating mean for each factor ?

(c) Create the interaction plot. What could you say about the main effect and interaction effect ? (Need also R code and plot)

Q4 (10=4+2+2+2 points)

(a) Perform a two-way ANOVA to test the main effect of country and type , and for the interactions upon the rating. What conclusion do you have from this two-way ANOVA analysis ? How does this result connect to Q3-b. (Need R code and ANOVA output)

(b) Refit the data with a two-way ANOVA without the interaction term, give the ANOVA output. Checkingthe normality assumption before and after refitting as in Q1-c and state your conclusion.

(c) nstead of examining the normal qq plot, now we consider to use the Shapiro-Wilk Normality Test (Rbuilt-in function : shapiro.test()) to evaluate the normality assumption for model without interaction term. Give a brief conclusion for this test.

(d) Find 95% TukeyHSD family-wise confidence interval for the difference of means of county. Try R command TukeyHSD(aov(rating∼type+country,data=beers), which=”country”). Does this result agree with the significance you have in Q4-b ?