MAT 3378 Final Exam (Spring 2020)

Gilles Lamothe

July 28, 2020 at 9:30 am

Name: _____

Student #: _____

Instructions

1) Please submit your solutions to this assignment in one PDF file in brightspace. Only one file will be accepted.

2) Only hand written (with a pen or pencil) solutions on paper are going to be accepted. For example, do ***not*** write your solutions on an ipad.

3) You can submit a PDF file more than once. However, only the last submission will be saved. If you want to modify your submission, that is fine as long as it is before the deadline.

4) Late submissions of the final exam are not going to be marked.

5) You will need to use R to compute a few quantiles and probabilities from t distributions, F distributions, and Tukey's studentized range distribution.

6) Print the midterm, and provide your answers in the provided space. To submit your answers, scan the document as a PDF, and submit it in Brightspace.

7) If you do not have a printer/scanner, please combine images of your hand-written solutions as one PDF. (See https://imagetopdf.com/ as a possible solution to combine images as one PDF).

8) The duration of the exam is 3 hours, and an extra 30 minutes is given to prepare and upload your PDF document. Please submit your solutions before 1 pm.

1. A university medical center urology group was interested in the association between prostate-specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer. We will study the association between the PSA level (the response), and two explanatory variables: the Gleason Score and the seminal vesicle invasion (presence or absence of seminal vesicle invasion: 1 if yes, 0 otherwise).

We imported the data and coerced each explanatory variable as a factor.

```
PSA<-read.csv("ProstateCancer.csv")
str(PSA)
```

```
## 'data.frame':    97 obs. of  3 variables:
##  $ PSA.level                : num  0.651 0.852 0.852 0.852 1.448 ...
##  $ Seminal.Vesicule.Invasion: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Gleason.Score            : int  6 7 7 6 6 6 6 6 7 6 ...
```

```
cols<-c("Seminal.Vesicule.Invasion","Gleason.Score")
PSA[cols] <- lapply(PSA[cols], factor)
```

(a) Here are cell statistics. Is the study balanced?

```
source("MyFunctions.r")
library(plyr)
stats<-ddply(PSA, .(Seminal.Vesicule.Invasion,Gleason.Score), summarize,
 Mean = my.mean(PSA.level),
 StdDev = my.sd(PSA.level),
  n = my.size(PSA.level))
stats
```

```
##   Seminal.Vesicule.Invasion Gleason.Score   Mean StdDev  n
## 1                         0             6  9.900 11.467 32
## 2                         0             7 11.659  8.096 34
## 3                         0             8 23.348 10.881 10
## 4                         1             6 28.219     NA  1
## 5                         1             7 28.467 16.131  9
## 6                         1             8 97.339 89.099 11
```

(b) An analyst fit an ANOVA model to describe the PSA level according to the two factors, and displayed the corresponding ANOVA table. The analyst changed the order of the variable, refit the model, and and displayed the corresponding ANOVA table. He noticed surprising results. According to the first table the Gleason main effects are significant ($p = 0.0005$), but they are much more significant ($p = 4.6 \times 10^{-8}$) according to the second table. Can you explain to the young statistician how to properly interpret these tables and give the correct p-value concerning the significance of the Gleason score main effects.

```
model<-lm(PSA.level~Seminal.Vesicule.Invasion*Gleason.Score,data=PSA)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: PSA.level
```

```
##                                          Df Sum Sq Mean Sq F value    Pr(>F)
## Seminal.Vesicule.Invasion                 1  44618   44618 45.7381 1.269e-09 ***
## Gleason.Score                             2  15950    7975  8.1754 0.0005429 ***
## Seminal.Vesicule.Invasion:Gleason.Score   2  10331    5165  5.2951 0.0066776 **
## Residuals                                91  88772     976
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

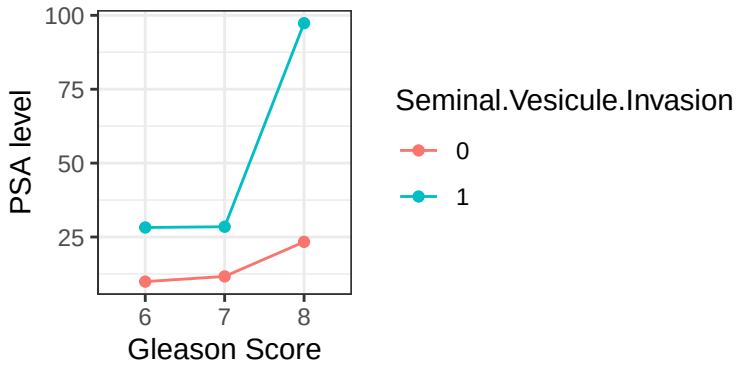
```
model<-lm(PSA.level~Gleason.Score*Seminal.Vesicule.Invasion,data=PSA)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: PSA.level
##                                          Df Sum Sq Mean Sq F value    Pr(>F)
## Gleason.Score                             2  39887 19943.3 20.4438 4.646e-08 ***
## Seminal.Vesicule.Invasion                 1  20682 20682.1 21.2012 1.338e-05 ***
## Gleason.Score:Seminal.Vesicule.Invasion   2  10331  5165.5  5.2951  0.006678 **
## Residuals                                91  88772   975.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) Compute the partial eta squared for the interaction effects, for the main Gleason score effects, and for the main Seminal Vesicule Invasion effects.

(d) Here is an interaction plot for these data. Does the plot suggest that there are interactions between the Gleason score factor and the Seminal Vesicule Invasion factor? Discuss. Is it consistent with the test for the significance of the interaction effects from part (b)?

```
library(ggplot2)
ggplot(PSA) +
  aes(x = Gleason.Score, color = Seminal.Vesicule.Invasion, group = Seminal.Vesicule.Invasion,
      y = PSA.level) +
  stat_summary(fun  = mean, geom = "line") +
  stat_summary(fun  = mean, geom = "point") +
  labs(x = "Gleason Score")+
  labs(y = "PSA level")+
  theme_bw()
```



(Question 1 cont.)

(Question 1 cont.)

2. Consider a two factor nested design, where A has fixed effects and B has random effects. Let Y_{ijk} be the response for the kth unit at level j of factor B within level i of factor A.

The covariance structure is

$$\mathrm{Cov}(Y_{ijk}; Y_{i'j'k'}) = \begin{cases} \sigma^2 + \sigma^2_\beta, & i = i', j = j', k = k' \\ \sigma^2_\beta, & i = i', j = j', k \neq k' \\ 0, & \text{else} \end{cases}$$

(a) Show that

$$V(\bar{Y}) = \frac{\sigma^2 + n\,\sigma^2_\beta}{n\,a\,b}.$$

where $\bar{Y} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} Y_{ijk}/(n\,a\,b)$.

(b) Suppose that we fit an nested ANOVA model with fixed effects, and produced the corresponding ANOVA table. How, would you compute the estimated standard error for the global mean $\bar{Y}$ under the assumption that A is fixed and B has random effects.

(Question 2 cont.)

3. Solutions of alcohol are used for calibrating Breathalyzers. Six bottles of alcohol solution were randomly chosen to study the variation of alcohol concentration among a large batch of bottled. Four determinations of alcohol concentrations were determined by gas chromatography from each of the 6 bottles.

We imported the data, and displayed the structure of the dataframe.

```
alcohol<-read.csv("AlcoholSolution.csv")
str(alcohol)
```

```
## 'data.frame':    24 obs. of  2 variables:
##  $ Bottle       : int  1 1 1 1 2 2 2 2 3 3 ...
##  $ Concentration: num  1.44 1.43 1.43 1.43 1.42 ...
```

We coerced the explanatory variable as a factor.

```
alcohol$Bottle<-factor(alcohol$Bottle)
```

We fit a one-factor ANOVA model with fixed effects to describe the concentration according to the bottle factor. We also displayed the corresponding ANOVA table.

```
model<-lm(Concentration~Bottle,data=alcohol)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Concentration
##           Df     Sum Sq   Mean Sq F value    Pr(>F)
## Bottle     5 0.00095052 1.901e-04  61.359 1.127e-10 ***
## Residuals 18 0.00005577 3.098e-06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here are bottle statistics.

```
# group descriptive statistics
library(plyr)
source("MyFunctions.r")
stats<-ddply(alcohol,.(Bottle),summarise,
Mean=my.mean(Concentration),
StdDev=my.sd(Concentration),
n=my.size(Concentration))
stats
```

```
##   Bottle  Mean StdDev n
## 1      1 1.434  0.002 4
## 2      2 1.424  0.002 4
## 3      3 1.416  0.002 4
## 4      4 1.432  0.002 4
## 5      5 1.427  0.002 4
## 6      6 1.420  0.002 4
```

(a) Formulate a one-factor ANOVA model with random effects for this study.

(b) Estimate the intra-class correlation, and interpret its value within the context of the study.

(Question 3 cont.)

(Question 3 cont.)

4. Is diet or exercise effective in combating insomnia ? Some believe that cutting our desserts can help alleviate the problem, while others recommend exercise. Forty volunteers suffering from insomnia agreed to participate in a month-long test. Half were randomly assigned to a special no-desserts diet ; the others continued desserts as usual. Half of the people in each of these groups were randomly assigned to an exercise program, while the others did not exercise. Those who ate no desserts and engaged in exercise showed the most improvement. Identify (if possible) :

(a) the factor(s) in the experiment and the number of levels for each.

(b) the number of treatments.

(c) the experimental units.

(d) the blocks.

(e) the design of the experiment.

(Question 4 cont.)

5. Four plants of the same variety were randomly selected in an experiment to investigate the concentration of a particular acid. Three leaves per plant were randomly selected and tree separate determinations of the acid concentration were obtained per leaf.

We imported that data and coerced each explanatory variable as a factor.

```
plants<-read.csv("plants.csv")
str(plants)
```

```
## 'data.frame':    36 obs. of  3 variables:
##  $ concentration: num  11.2 11.6 12 16.5 16.8 16.1 18.3 18.7 19 14.1 ...
##  $ plant        : int  1 1 1 1 1 1 1 1 1 2 ...
##  $ leaf         : int  1 1 1 2 2 2 3 3 3 1 ...
```

```
cols<-c("plant","leaf")
plants[cols] <- lapply(plants[cols], factor)
```

(a) Formulate an appropriate ANOVA model for this study. Assume that the plan factor is random and the leaf factor is also random.

(b) Fit an ANOVA model to describe the acid concentration according to the plant and the leaf factor.

```
model<-aov(concentration~plant+leaf%in%plant,data=plants)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: concentration
##            Df Sum Sq Mean Sq F value    Pr(>F)
## plant       3 343.18 114.393  905.09 < 2.2e-16 ***
## plant:leaf  8 187.45  23.432  185.39 < 2.2e-16 ***
## Residuals  24   3.03   0.126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

+(i) Test for the significance of the random effect of the plant. Formulate the hypotheses, give the observed value of the test statistic, compute the p-value, and give the conclusion.

+(ii) Test for the significance of the random effect of the leaf. Formulate the hypotheses, give the observed value of the test statistic, compute the p-value, and give the conclusion.

+(iii) Obtain estimates for the variance σ^2_α of the random effect of the plant, for the variance σ^2_β of the random effect of the leaf, and for the variance σ^2 of the random error. Which component of variance appears to be most important in the total variance σ^2_Y?

(Question 5 cont.)

(Question 5 cont.)

6. The following is a partial analysis of variance table for four treatments and a total of 20 observations of the response variable. Complete the table and test the null hypothesis of equality of the cell means. Give your conclusion at a level of significance of 5%. The study is balanced.

Source	df	SS	MS	F	p-value
Treatments		120			
Error			20		
Total		440			

(Question 6 cont.)

7. Suppose you wish to determine whether the amount of carbon used in the manufacture of steel has an effect on the tensile strength of the steel. Four different percentages of carbon are investigated: 0.3%, 0.4%, 0.5%, and 0.6%. For each percentage of carbon, five steel specimens are randomly selected from the same batch and their strengths are measured.

We imported the data and displayed the structure of the dataframe.

```
steel<-read.csv("StrengthSteel.csv")
str(steel)
```

```
## 'data.frame':    20 obs. of  2 variables:
##  $ Carbon.Content: chr  "0.30%" "0.30%" "0.30%" "0.30%" ...
##  $ Strength      : int  1420 1510 1410 1530 1470 1480 1470 1520 1540 1510 ...
```

We coerced the explanatory variable as a factor.

```
steel$Carbon.Content<-factor(steel$Carbon.Content)
```

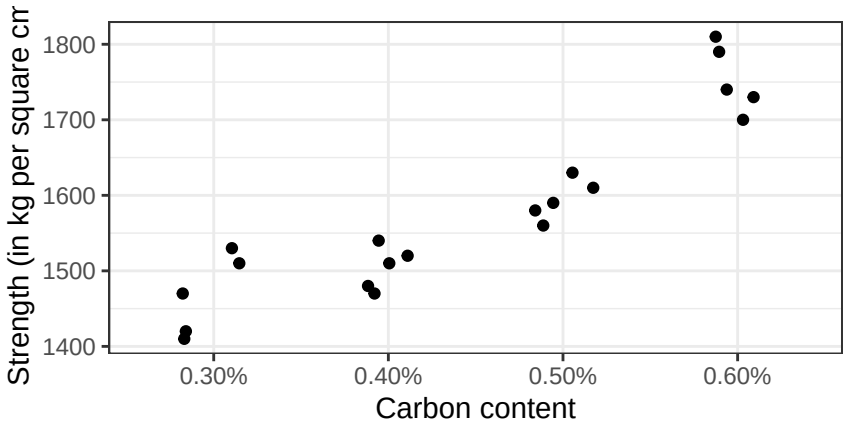
Here are group statistics.

```
# group descriptive statistics
library(plyr)
source("MyFunctions.r")
stats<-ddply(steel,.(Carbon.Content),summarise,
Mean=my.mean(Strength),
StdDev=my.sd(Strength),
n=my.size(Strength))
stats
```

```
##   Carbon.Content Mean StdDev n
## 1          0.30% 1468 53.104 5
## 2          0.40% 1504 28.810 5
## 3          0.50% 1594 27.019 5
## 4          0.60% 1754 45.056 5
```

Here are dotplots boxplots

```
# comparative boxplots
library(ggplot2)
ggplot(steel,
       aes(x =Carbon.Content , y = Strength)) +
  theme_bw()  +
  geom_jitter(height=0,width=0.2) +
  labs(y = "Strength (in kg per square cm)",x="Carbon content")
```



(a) Based on the above plot, are any differences in average tensile strength apparent to you?

(b) We used the max |T| test to adjust the p-values to control the FWER in testing for the significance of the Tukey contrasts to compare the carbon content effects pairwise. Use the Insert and Absorb Algorithm to obtain the letters for the Tukey groups of non-significantly different treatments.

```
model<-lm(Strength~Carbon.Content,data=steel)
library(multcomp)
test<-glht(model,mcp(Carbon.Content="Tukey"))
summary(test)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Strength ~ Carbon.Content, data = steel)
##
## Linear Hypotheses:
##                     Estimate Std. Error t value Pr(>|t|)
## 0.40% - 0.30% == 0    36.00      25.32   1.422   0.5045
## 0.50% - 0.30% == 0   126.00      25.32   4.977   <0.001 ***
## 0.60% - 0.30% == 0   286.00      25.32  11.296   <0.001 ***
## 0.50% - 0.40% == 0    90.00      25.32   3.555   0.0126 *
## 0.60% - 0.40% == 0   250.00      25.32   9.874   <0.001 ***
## 0.60% - 0.50% == 0   160.00      25.32   6.320   <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

(Question 7 cont.)

(Question 7 cont.)

8. A soft-drink manufacturer uses five agents (1, 2, 3, 4, 5) to hand premium distributions for its various products. The marketing director desired to study the timeliness with which the premiums are distributed. Twenty transactions for each agent were selected at random, and the time lapse (in days) for handling each transaction was determined. Assume that it is appropriate to analyze this study with a one-factor ANOVA model.

We imported the data and displayed the structure of the dataframe.

```
premium<-read.csv("Premium.csv")
str(premium)
```

```
## 'data.frame':    100 obs. of  2 variables:
##  $ Time..in.days.: int  24 24 29 20 21 25 28 27 23 21 ...
##  $ Agent         : int  1 1 1 1 1 1 1 1 1 1 ...
```

We coerced the explanatory variable as a factor, and displayed its coding matrix.

```
premium$Agent<-factor(premium$Agent)
contrasts(premium$Agent)<-contr.sum(5)
contrasts(premium$Agent)
```

```
##   [,1] [,2] [,3] [,4]
## 1    1    0    0    0
## 2    0    1    0    0
## 3    0    0    1    0
## 4    0    0    0    1
## 5   -1   -1   -1   -1
```

(a) Is it deviation coding or dummy coding that is being used to code the Agent factor?

(b) Define $\tau_i = \mu_i - \mu$ as the effect for agent i, where $\mu_i = \sum_{i=1}^{5} \mu_i/5$. Show that

$$\sum_{i=1}^{5} \tau_i = 0 \quad \Leftrightarrow \quad \mu = \sum_{i=1}^{5} \mu_i/5.$$

(c) We fit a one-factor ANOVA model to describe the time lapse according to the agent, and we displayed the estimated coefficients, and the estimate of the standard deviation of the random error.

```
model<-lm(Time..in.days.~Agent,data=premium)
coefficients(model)
```

```
## (Intercept)      Agent1      Agent2      Agent3      Agent4
##       20.75        3.80        1.80       -9.00       -5.95
```

```
summary(model)$sigma
```

```
## [1] 2.742742
```

Give a point estimate for each of the cell means $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$.

(d) Agents 1 and 2 distribute merchandise only, agents 3 and 4 distributed cash-value coupons only, and agent 5 distributes both merchandise and coupons. Estimate the contrast

$$\frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$

with a 95% confidence interval. Interpret your interval estimate.

(Question 8 cont.)