MATH 531T-A : Exam 1

Louis Bensard

June 4, 2018

Problem 1:

(a)

To quantify the gains from using randomization and from using balanced randomization, we need to incorporate a learning effect caused by using one keyboard before another for a given manuscript. Let's formally define that model:

$$y_{ij} = \eta + \alpha_i + \tau_j + \delta ij \cdot l_i + \epsilon_{ij}$$

With:

$i = 1, ..., 6$ (manuscripts or blocks)
$j = 1, 2$ (keyboards or treatments A, B)
α_i is the blocking effect
$\eta = \bar{y_{..}}$ (grand mean)
$\tau_j = \bar{y_{j.}} - \bar{y_{..}}$ (treatment effect)
$\delta_{ij} = \begin{cases} 1, & \text{if keyboard j is used second for manuscript i} \\ -1 & \text{otherwise} \end{cases}$
l_i is learning effect for the ith manuscript
$L = \sum_i \sum_j \delta_{ij} \cdot l_i,$ overall learning effect

To quantify the gains of the different type experiment, we assume that $l = l_1 = l_2 = l_3 = l_4 = l_5 = l_6$. First, let's determine the learning effect for the first method that does not use any randomization, and then compare it to the randomization and balanced randomization to quantify the decrease in learning effect resulting from those 2 methods.

Without randomization, here is a kind of sequence we would have: AB, AB, AB, AB, AB, AB. Here, it is clear that the learning effect would be as bad as it can get considering our assumptions and thus $L = 6l$.

Now if we use simple randomization, here is a kind of sequence we would have: AB, BA, BA, AB, AB, AB. The learning effect is now equal to $L = 4l - 2l = 2l$. Therefore, by using this method, we decreased the learning effect by $g_1 = \frac{6l-2l}{6l} = 67\%$ (for that specific randomly generated sequence, other sequences obtained would give us a difference decrease in learning effect).

Finally, if we use balanced randomization, here is a kind of sequence we would have: AB, AB, AB, BA, BA, BA. The learning effect is now equal to $L = 3l - 3l = 0$. Therefore, by using this method, we decreased the learning effect by $g_2 = \frac{6l-0}{6l} = 100\%$.

(b)

I would not use this sequence for the study. In my opinion, there are two pieces of information that are missing here and that make this sequence not good to use.

First, we don't know whether each manuscript is typed by the same person or if they are all typed by someone different each time. If it is the same person for all manuscript, we can easily understand that the learning effect of typing with one keyboard before another will be high at first, but as the person becomes more familiar with both keyboards, the learning effect will get smaller and smaller such that $l_1 > l_2 > l_3 > l_4 > l_5 > l_6$.

Second, we have no information about the manuscripts, one manuscript might be rich in words and make the typer learn the keyboard very fast while another might be very poor and redondant which would make the typer learn the keyboards slower.

So with that much uncertainty we must randomize the order of the elements of the sequence AB, AB, AB, BA, BA, BA to minimize the overall learning effect or "manuscript effect".

Problem 2:

```
## No id variables; using all as measure variables
```

(Code in Appendix 1)

Here are the 95% simultaneous confidence interval for the six pairs of treatment differences using the Bonferroni method:

$$
\begin{aligned}
B - A &: [-0.8002, 0.4402] \\
C - A &: [-0.2402, 1.0002] \\
D - A &: [-0.1802, 1.0602] \\
C - B &: [-0.0602, 1.1802] \\
D - B &: [-1.8391 \times 10^{-4}, 1.2402] \\
D - C &: [-0.5602, 0.6802]
\end{aligned}
$$

We can see that only D-B is very close to have significance. Indeed, the CI of D-B is the only one very close to not containing 0.

The length of the Bonferroni intervals is 1.2404.

Here are the 95% simultaneous confidence interval for the six pairs of treatment differences using the Tukey method:

$$
\begin{aligned}
B - A &: [-0.7698, 0.4098] \\
C - A &: [-0.2098, 0.9698] \\
D - A &: [-0.1498, 1.0298] \\
C - B &: [-0.0298, 1.1498] \\
D - B &: [0.0302, 1.2098] \\
D - C &: [-0.5298, 0.6498]
\end{aligned}
$$

Now we can see that only D-B has significance. Indeed, the CI of D-B is the only one not containing 0.

The length of the Bonferroni intervals is 1.1796.

The Tukey method gives shorter intervals.

Problem 3:

(a)

A one-way layout is a proper design here. That's too easy, let's add some possibly irrelvant answer to the question by explaining how I would perform such an experiment:

Let's assume that we have closely identical cells containing the virus causing the illness. Those cells would be our experimental unit and we would have 4 factors (drugs A, B, C, D) with 1 level so 4 possible treatments.

We are not really sure where the cells exactly come from so we would randomly assign one cell to each treatment every trial. The trial would consist of using each drug on a different cell and measure the effect of the drug on the virus of that cell. Then, I would replicate this trial n times on a new set of cells each time.

(b)

$$F = \frac{MSTr}{MSE} = \frac{21.47}{2.39} = 8.98 \sim F_{k-1,N-k} = F_{3,26}$$

Thus the corresponding p-value is $p = 0.0003 < 0.01$ and thus there is evidence in the data that at least one pair of treatments have significantly different bioactivity.

(c)

First, let's compute $q^* = \frac{1}{\sqrt{2}} \cdot q_{k,N-k,\alpha}$ where $q_{k,N-k,\alpha}$ is the upper α point of the studentized range distribution.

$$q^* = \frac{1}{\sqrt{2}} \cdot q_{3,26,0.01} = 3.1893$$

Now let's obtain the t statistics corresponding to the test of significant difference for each pair of drugs:

(Code in Appendix 2)

$$\begin{aligned} t_{B-A} &= -0.4374 \\ t_{C-A} &= -4.4539 \\ t_{D-A} &= -2.616 \\ t_{C-B} &= -4.1533 \\ t_{D-B} &= -2.2757 \\ t_{D-C} &= 1.4973 \end{aligned}$$

The Tukey method tells us that the pair of treatment AC (|-4.4539|>3.1893) and the pair BC (|-4.1533|>3.1893) have significant differences in their bioactivity.

(d)

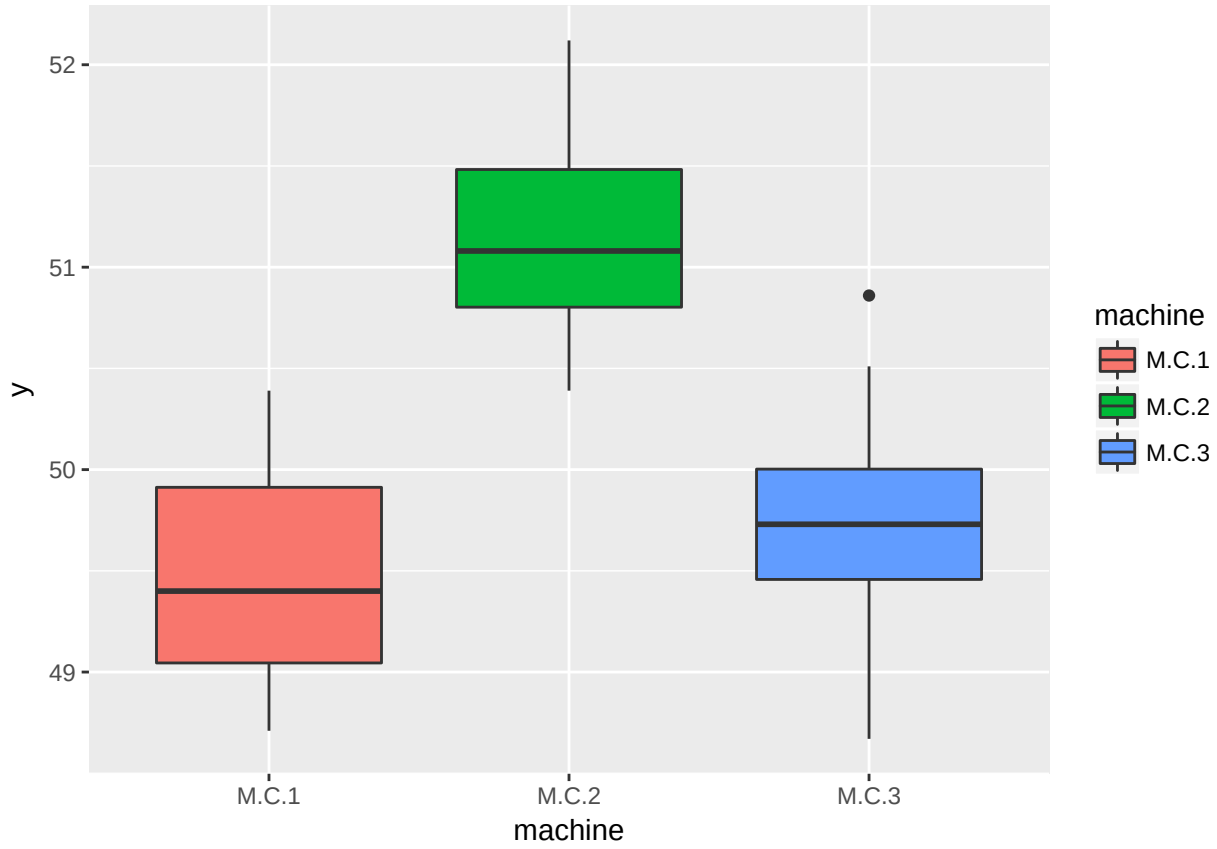
Here is the t statistic if we want to compare the bioactivity of brand-name versus generic drugs:

$$t_{AB-CD} = \frac{(\bar{y_A} + \bar{y_B}) - (\bar{y_C} + \bar{y_D})}{\sqrt{MSE \cdot (\frac{1}{n_A} + \frac{1}{n_B} + \frac{1}{n_C} + \frac{1}{n_D})}} = 4.7025 \sim t_{N-k} = t_{26}$$

The p-value $p = 7.3721 \times 10^{-5} < 0.01$ thus there is evidence that there is a significant difference between the bioactivity of brand-name drugs and generic drugs.

Problem 4:

First, let's boxplot the data to get an overview of what differences between machines we could expect:



Machine 1 and Machine 3 seem don't seem to pack bags very differently, but we should expect our further analysis to outline a difference between M/C 1 and 3 and M/C 2 and 3. Now let's perfom a more rigorous analysis.

We are going to fit the following linear model:

$$y_{ij} = \eta + \alpha_i + \tau_j + \epsilon_{ij}$$

With:

$i = 1, ..., 20$ (Bags or blocks)
$j = 1, 2, 3$ (Machines or treatments)
α_i is the blocking effect
$\eta = \bar{y_{..}}$ (grand mean)
$\tau_j = \bar{y_{j.}} - \bar{y_{..}}$ (treatment effect)

Now we perfrom a one-way layout ANOVA on this model to test the null hypothesis H_0: There is no difference in packing between machines.

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## machine    2   31.7   15.83    59.5 1.1e-14 ***
## Residuals 57   15.2    0.27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is very small, $p = 1.1 \cdot 10^{-14} < 0.01$, so there is significant evidence that at least one pair of machines pack bags differently. Now let's see whose pair(s) it is using a Bonferroni method and Tukey method.

Bonferroni method:

(Code in Appendix 3)

The critical value t^* for Boneferroni is 3.0639.

Now let's obtain the t statistics corresponding to the test of significance of each pair of Machine:

$$t_{2-1} = 10.1975$$
$$t_{3-1} = 1.7415$$
$$t_{3-2} = -8.456$$

The Bonferroni method tells us that the pair of Machine 1 & 2 (10.1975>3.0639) and the pair of Machines 2 & 3 (|-8.4560|>3.0639) have significant differences in their packing. (as expected from the boxplots).

Tukey method:

Here is the output of the function "TukeyHSD" that takes an argument of class aov() and after changing the argument "conf.level" to 0.99. It gives the 99% confidence intervals of the difference of the treament effect between the machines.

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = y ~ machine, data = data.m)
##
## $machine
```

```
##                diff     lwr     upr  p adj
## M.C.2-M.C.1   1.663  1.1682  2.1578 0.0000
## M.C.3-M.C.1   0.284 -0.2108  0.7788 0.1988
## M.C.3-M.C.2 -1.379 -1.8738 -0.8842 0.0000
```

The only confidence intervals not containing 0 are the ones from the pair of Machines 1 & 2 and the pair of Machines 2 & 3.

Therefore, using the Tukey method and Bonferroni method, we are 99% confident that the Machines 1 and 2 and the Machines 2 and 3 significantly pack bags differently.

Appendix 1:

```
data<-read.table("http://www2.isye.gatech.edu/%7Ejeffwu/book/data/pulp.dat", h=T)

data.m = melt(data)
names(data.m) =  c("operator", "y")

## ANOVA Decomposition ##
Y = as.matrix(data)
n = nrow(Y); k = ncol(Y)
n_vect = rep(n,4)
N = sum(n_vect)
alpha = 0.05

Yidot_bar <- apply(Y, 2, mean) # sample mean for each treatment
avgY <- mean(Yidot_bar) # grand mean

g<-lm(y ~ operator, data = data.m)

mse <- anova(g)[["Mean Sq"]][2]

#Bonferroni

kprime <- choose(k,2)
t_std = qt(1-(alpha/(2*kprime)), N-k)*(sqrt(mse*(1/n+1/n)))

Lb12 = Yidot_bar[2] - Yidot_bar[1] - t_std
Lb13 = Yidot_bar[3] - Yidot_bar[1] - t_std
Lb14 = Yidot_bar[4] - Yidot_bar[1] - t_std
Lb23 = Yidot_bar[3] - Yidot_bar[2] - t_std
Lb24 = Yidot_bar[4] - Yidot_bar[2] - t_std
Lb34 = Yidot_bar[4] - Yidot_bar[3] - t_std

#vector of lower bounds of the CI's
L_bon = c(Lb12, Lb13, Lb14, Lb23, Lb24, Lb34)

Ub12 = Yidot_bar[2] - Yidot_bar[1] + t_std
Ub13 = Yidot_bar[3] - Yidot_bar[1] + t_std
Ub14 = Yidot_bar[4] - Yidot_bar[1] + t_std
Ub23 = Yidot_bar[3] - Yidot_bar[2] + t_std
Ub24 = Yidot_bar[4] - Yidot_bar[2] + t_std
Ub34 = Yidot_bar[4] - Yidot_bar[3] + t_std
```

```
#vector of upper bounds of the CI's
U_bon = c(Ub12, Ub13, Ub14, Ub23, Ub24, Ub34)

length_bon = t_std*2

#Tukey
q_std = qtukey(1-alpha,k,N-k)/sqrt(2)*(sqrt(mse*(1/n+1/n)))

Lt12 = Yidot_bar[2] - Yidot_bar[1] - q_std
Lt13 = Yidot_bar[3] - Yidot_bar[1] - q_std
Lt14 = Yidot_bar[4] - Yidot_bar[1] - q_std
Lt23 = Yidot_bar[3] - Yidot_bar[2] - q_std
Lt24 = Yidot_bar[4] - Yidot_bar[2] - q_std
Lt34 = Yidot_bar[4] - Yidot_bar[3] - q_std

#vector of lower bounds of the CI's
L_tuk = c(Lt12, Lt13, Lt14, Lt23, Lt24, Lt34)

Ut12 = Yidot_bar[2] - Yidot_bar[1] + q_std
Ut13 = Yidot_bar[3] - Yidot_bar[1] + q_std
Ut14 = Yidot_bar[4] - Yidot_bar[1] + q_std
Ut23 = Yidot_bar[3] - Yidot_bar[2] + q_std
Ut24 = Yidot_bar[4] - Yidot_bar[2] + q_std
Ut34 = Yidot_bar[4] - Yidot_bar[3] + q_std

#vector of upper bounds of the CI's
U_tuk = c(Ut12, Ut13, Ut14, Ut23, Ut24, Ut34)

length_tuk = q_std*2
```

Appendix 2:

```
alpha = 0.01 ; k=3; N_k = 26; mse = 2.39
nA = 7; nB = 8; nC = 9; nD = 6
y_barA = 66.10; y_barB = 65.75; y_barC = 62.63; y_barD = 63.85

q_star = qtukey(1-alpha,k,N_k)/sqrt(2)

tAB = (y_barB - y_barA)/sqrt(mse*(1/nB + 1/nA))
tAC = (y_barC - y_barA)/sqrt(mse*(1/nC + 1/nA))
tAD = (y_barD - y_barA)/sqrt(mse*(1/nD + 1/nA))
tBC = (y_barC - y_barB)/sqrt(mse*(1/nC + 1/nB))
tBD = (y_barD - y_barB)/sqrt(mse*(1/nD + 1/nB))
tCD = (y_barD - y_barC)/sqrt(mse*(1/nD + 1/nC))

abst_vect = abs(c(tAB,tAC,tAD,tBC,tBD,tCD))
```

Appendix 3:

```
n = 20; N = 60; k = 3
alpha = 0.01

Y = as.matrix(data)
y_bar <- apply(Y, 2, mean)

mse <- anova(model)[["Mean Sq"]][2]

kprime <- choose(k,2)

#bonferroni
t_star = qt(1-(alpha/(2*kprime)), N-k)

t12 = (y_bar[2] - y_bar[1])/sqrt(mse*(1/n + 1/n))
t13 = (y_bar[3] - y_bar[1])/sqrt(mse*(1/n + 1/n))
t23 = (y_bar[3] - y_bar[2])/sqrt(mse*(1/n + 1/n))

#tukey
fit = aov(y ~ machine, data = data.m)
TukeyHSD(fit, conf.level = 0.99)
```