STAT 222 Spring 2022 HW3

Matthew Zhao

Question 1

```
pr3.2 = read.table("http://users.stat.umn.edu/~gary/book/fcdae.data/pr3.2",h=T)
pr3.2$trt = factor(pr3.2$trt, labels=c("0", "1p", "1v","8p", "8v"))
```

Q1a — 5 points

Calculate Tukey's HSD (honest significant difference) required for two groups to be significantly different in mean at 5% level. Use Tukey's HSD to do pairwise comparisons and summarize the result with an underline diagram.

$$\frac{q_\alpha(g, N-g)}{\sqrt{2}} \times \sqrt{\text{MSE}(\frac{1}{n}+\frac{1}{n})}$$

```
anova(lm(days ~ as.factor(trt), data = pr3.2))
## Analysis of Variance Table
##
## Response: days
##                 Df Sum Sq Mean Sq F value       Pr(>F)
## as.factor(trt)   4  11939    2985    13.6 0.0000000035
## Residuals      120  26314     219
```

```
left <- sqrt(219.28*((1/25)+(1/25)))
right <- qtukey(0.95,5,125-5)/sqrt(2)
left * right
## [1] 11.6
```

```
## [1] "1-2 0.2   Not Significant"
## [1] "1-3 1.44   Not Significant"
## [1] "1-4 6.6   Not Significant"
## [1] "1-5 24.64   Significant"
## [1] "2-3 1.24   Not Significant"
## [1] "2-4 6.8   Not Significant"
## [1] "2-5 24.84   Significant"
## [1] "3-4 8.04   Not Significant"
## [1] "3-5 26.08   Significant"
## [1] "4-5 18.04   Significant"
```

Underline Diagram

```
  5 (8v)      4 (8p)       1 (none)      2 (1p)      3 (1v)
   38.72       56.76         63.36        63.56      64.80
          ------------------------------------------------
```

Q1b — 5 points

Bonferroni is the best for pre-planned contrasts.

```
mod1 = aov(days ~ as.factor(trt), data=pr3.2)
mod1emm = emmeans(mod1,specs = "trt",adjust="bonf")
contrast(mod1emm,
  list(o1 = c(1,-1,0,0,0), o2 = c(1,0,-1,0,0),o3 = c(1,0,0,-1,0),
                    o4 = c(1,0,0,0,-1),c1 = c(1,-0.5,-0.5,0,0),
                    c2 = c(1,0,0,-0.5,-0.5),c3 = c(0,-0.5,0.5,-0.5,0.5),
                    c4 = c(0,-1,1,1,-1)),level=0.95,infer=c(T,F))
##  contrast estimate   SE  df lower.CL upper.CL
##  o1           -0.20 4.19 120    -8.49     8.09
##  o2           -1.44 4.19 120    -9.73     6.85
##  o3            6.60 4.19 120    -1.69    14.89
##  o4           24.64 4.19 120    16.35    32.93
##  c1           -0.82 3.63 120    -8.00     6.36
##  c2           15.62 3.63 120     8.44    22.80
##  c3           -8.40 2.96 120   -14.26    -2.54
##  c4           19.28 5.92 120     7.55    31.01
##
## Confidence level used: 0.95
```

Critical value is:

```
qt(0.05/2/8,125-5,lower.tail = F)
## [1] 2.7835
```

Q1c — 4 points

The investigator can use Scheffe's Method for contrasts. $|t_o| = \dfrac{|\hat{C}|}{SE(\hat{C})}$

```
mod1emm = emmeans(mod1,specs = "trt")
contrast(mod1emm, list(C = c(0.25,0.25,0.25,0.25,-1)),infer=c(F,F))
##  contrast estimate   SE  df
##  C             23.4 3.31 120
```

df $= 120$ $\hat{C} = \sum_{i=1}^{g} c_i \bar{y}_{i\cdot} = 23.4$ $\text{SE}(\hat{C}) = \sqrt{MSE \times \sum_{i=1}^{g} \frac{c_i^2}{n_i}} = 3.31$ Test statistic: $|t_o| = \dfrac{|\hat{C}|}{\text{SE}(\hat{C})} = \dfrac{23.4}{3.31} = 7.0695$

Critical value: $\sqrt{(g-1)F_{\alpha,g-1,N-g}}$:

```
sqrt((5-1)*qf(0.05,df1=5-1,df2=125-5,lower.tail = F))
## [1] 3.1287
```

We can conclude that since the test statistic is larger than the critical value, we can reject the null hypothesis that the contrast equals zero, meaning that there is an effect.

Question 2

```
milk = read.table("http://www.stat.uchicago.edu/~yibi/s222/milkbacteria.txt", h=T)
```

Q2a — 3 points

Experimental unit: a bottle Measurement unit: a single 5 ml sample taken from a bottle

Q2b — 1 point

```
anova(lm(log(count) ~ as.factor(hours), data = milk))
## Analysis of Variance Table
##
## Response: log(count)
##                   Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(hours)  3  154.2    51.4     430 <2e-16
## Residuals        76    9.1     0.1
```

Q2c — 1 point

```
milk2 = with(milk, aggregate(count, by=list(hours, bottle), mean))
names(milk2) = c("hours", "bottle", "count")
milk2
##    hours bottle   count
## 1      1      1   645.0
## 2      1      2   902.5
## 3      1      3   792.5
## 4      1      4   665.0
## 5      1      5   925.0
## 6      6      6  1120.0
## 7      6      7  2250.0
## 8      6      8  2150.0
## 9      6      9  1300.0
## 10     6     10  3925.0
## 11    12     11  7375.0
```

```
## 12    12     12  9050.0
## 13    12     13  7525.0
## 14    12     14  5550.0
## 15    12     15 13250.0
## 16    18     16 34000.0
## 17    18     17 30250.0
## 18    18     18 33750.0
## 19    18     19 15750.0
## 20    18     20 39750.0
```

```
anova(lm(log(count) ~ as.factor(hours), data = milk2))
## Analysis of Variance Table
##
## Response: log(count)
##                  Df Sum Sq Mean Sq F value        Pr(>F)
## as.factor(hours)  3   38.3   12.77     100 0.00000000014
## Residuals        16    2.0    0.13
```

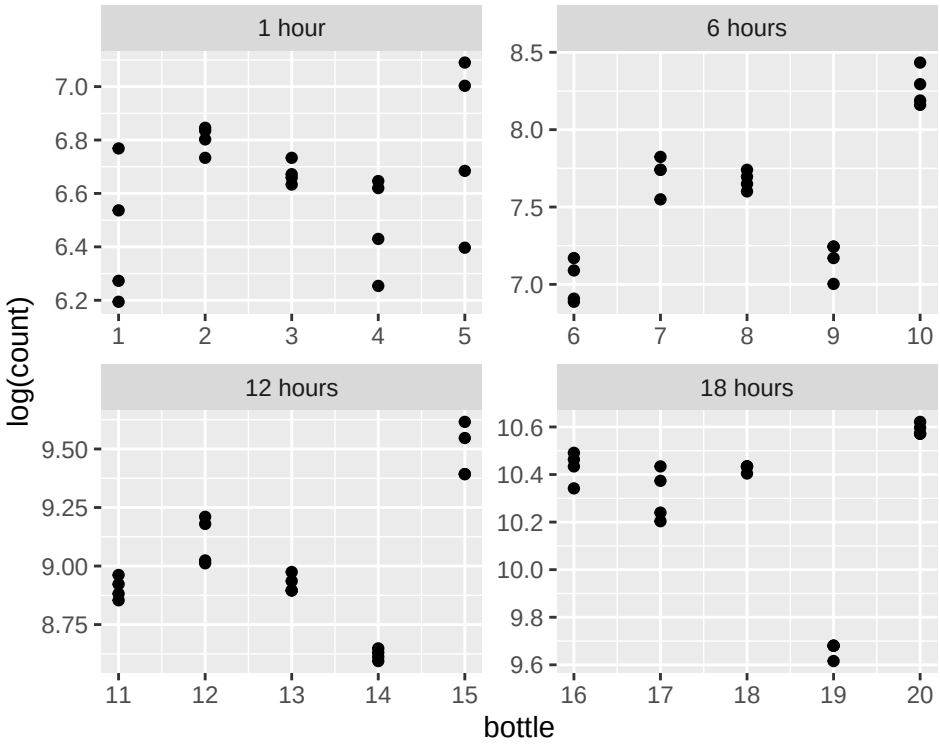
The result using the 5 mL sample is more significant.

Q2d — 5 Points

```
anova(lm(log(count) ~ as.factor(bottle), data=subset(milk, hours == 1)))
## Analysis of Variance Table
##
## Response: log(count)
##                   Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(bottle)  4  0.455  0.1138    2.75  0.067
## Residuals         15  0.620  0.0414
anova(lm(log(count) ~ as.factor(bottle), data=subset(milk, hours == 6)))
## Analysis of Variance Table
##
## Response: log(count)
##                   Df Sum Sq Mean Sq F value        Pr(>F)
## as.factor(bottle)  4   3.97   0.993    76.9 0.00000000083
## Residuals         15   0.19   0.013
anova(lm(log(count) ~ as.factor(bottle), data=subset(milk, hours == 12)))
## Analysis of Variance Table
##
## Response: log(count)
##                   Df Sum Sq Mean Sq F value       Pr(>F)
## as.factor(bottle)  4  1.624   0.406    73.8 0.0000000011
## Residuals         15  0.083   0.006
anova(lm(log(count) ~ as.factor(bottle), data=subset(milk, hours == 18)))
## Analysis of Variance Table
##
## Response: log(count)
```

```
##                    Df Sum Sq Mean Sq F value          Pr(>F)
## as.factor(bottle)  4  2.084   0.521     145 0.0000000000084
## Residuals         15  0.054   0.004
```

```
library(ggplot2)
milk$hoursfac = factor(milk$hours, labels=c("1 hour", "6 hours","12 hours","18 hours"))
ggplot(milk, aes(x=bottle, y=log(count)))+
  geom_point() +
  facet_wrap(~hoursfac,scale="free")
```



i) The results from the tests should be insignificant if the samples from the bottle are independent.

ii) The four samples are positively correlated. In words, if a sample is taken from a specific bottle with a certain treatment, log(count) of that bottle will most likely be similar to log(count) of samples from the same bottle and treatment.

iii) If we use samples as the experimental unit, the significant result that we get from looking at the relationship between log(count) and hours is also capturing the statistically significant effect of samples being from the same bottle. This confounding variable of bottle causes our result to appear more significant than it actually is and also means that the samples are not iid, specifically they are not independent.