STAT 222 Spring 2022 HW4

Matthew Zhao

Question 1

Q1a — 4 points

Grand mean $\mu = \frac{\sum_{i=1}^{g} n_i \mu_i}{N} = \frac{45 * 3 + 33 * 3 + 60 * 4}{3 + 3 + 4} = 47.4$

```
(45 * 3 + 33* 3 + 60* 4)/(3 + 3 + 4)
## [1] 47.4
```

$\delta^2 = \frac{\sum_{i=1}^{g} n_i (\mu_i - \mu)^2}{\sigma^2} = 35.4$

```
((45-47.4)^2 * 3 + (33-47.4)^2* 3 + (60-47.4)^2* 4)/(36)
## [1] 35.4
```

The power can be calculated using the R code below:

```
pf(qf(1-0.01,3-1,10-3),3-1,10-3,ncp=35.4,lower.tail = F)
## [1] 0.882892
```

Q1b — 4 points

Grand mean $\mu = \frac{\sum_{i=1}^{g} n_i \mu_i}{N} = \frac{45 * n + 33 * n + 60 * n}{n + n + n} = \frac{138n}{3n} = 46$

$\delta^2 = \frac{\sum_{i=1}^{g} n_i (\mu_i - \mu)^2}{\sigma^2} = \frac{n(45-46)^2 + n(33-46)^2 + n(60-46)^2}{36} = \frac{n + 169n + 196n}{36} = \frac{366n}{36} = 10.167n$

```
# qf(alpha, g-1, N-g, lower.tail=F)
n=3
f_crit = qf(0.01, 3-1, 3*n-3, lower.tail=F)
pf(f_crit,3-1,3*n-3,ncp = 366*n/36,lower.tail=F)
## [1] 0.757129
```

```
n=4
f_crit = qf(0.01, 3-1, 3*n-3, lower.tail=F)
pf(f_crit,3-1,3*n-3,ncp = 366*n/36,lower.tail=F)
## [1] 0.967797
```

We only reach a power of at least 0.95 with 4 subjects per group, meaning that we need to use at least 4 subjects per group to get to a power of 0.95.

Question 2

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

Q2a — 4 points

```
library(mosaic)
mean(failtime ~ as.factor(material), data=insulate)
##       1       2       3       4       5
##  159.75    6.25 2941.75 5723.00   10.75
anova(lm(failtime ~ as.factor(material), data=insulate))
## Analysis of Variance Table
##
## Response: failtime
##                      Df    Sum Sq  Mean Sq F value  Pr(>F)
## as.factor(material)   4 103191489 25797872   6.191 0.00379
## Residuals            15  62505657  4167044
```

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

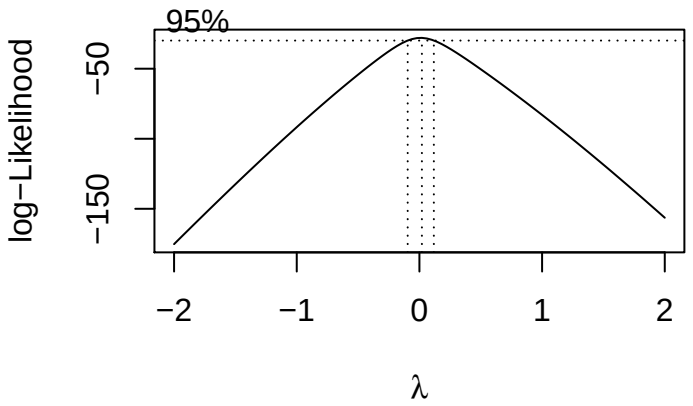
```
(qtukey(1-0.1,5,20-5)/sqrt(2))*(sqrt(4167044*((1/4)+(1/4))))
## [1] 3906.62
insl_means <- insulate %>% group_by(material) %>% summarise(group_mean = mean(failtime))
for (x in 1:(nrow(insl_means)-1)) {
  for (y in (x+1):nrow(insl_means)) {
    print(paste0(as.character(insl_means[x,1]),"-",
                 as.character(insl_means[y,1]), " ",
                 round(abs(insl_means$group_mean[x]-
                             insl_means$group_mean[y]),3),
                 "   ",
                 ifelse(abs(insl_means$group_mean[x]-
                              insl_means$group_mean[y]) < 3906.62,
                        "Not Significant", "Significant")))
  }
}
## [1] "1-2 153.5   Not Significant"
## [1] "1-3 2782   Not Significant"
## [1] "1-4 5563.25   Significant"
## [1] "1-5 149   Not Significant"
## [1] "2-3 2935.5   Not Significant"
## [1] "2-4 5716.75   Significant"
## [1] "2-5 4.5   Not Significant"
## [1] "3-4 2781.25   Not Significant"
## [1] "3-5 2931   Not Significant"
## [1] "4-5 5712.25   Significant"
```

Underline Diagram

```
     2       5        1          3          4
  6.25   10.75  159.75    2941.75    5723.00
 ---------------------------------
                          -------------------
```

Q2b — 2 points

```
library(MASS)
boxcox(failtime ~ as.factor(material), data=insulate)
```



Since our interval is centered around 0, we should use the log transformation for the response.

Q2c — 4 points

```
mean(log10(failtime) ~ as.factor(material), data=insulate)
##        1        2        3        4        5
## 2.193879 0.539591 3.351191 3.567298 0.826874
anova(lm(log10(failtime) ~ as.factor(material), data=insulate))
## Analysis of Variance Table
##
## Response: log10(failtime)
##                      Df Sum Sq Mean Sq F value      Pr(>F)
## as.factor(material)  4  31.13   7.783   37.66 0.000000118
## Residuals           15   3.10   0.207
```

```
(qtukey(1-0.1,5,20-5)/sqrt(2))*(sqrt(0.207*((1/4)+(1/4))))
## [1] 0.870709
insl_means <- insulate %>% mutate(failtime = log10(failtime)) %>% group_by(material) %>% summa
```

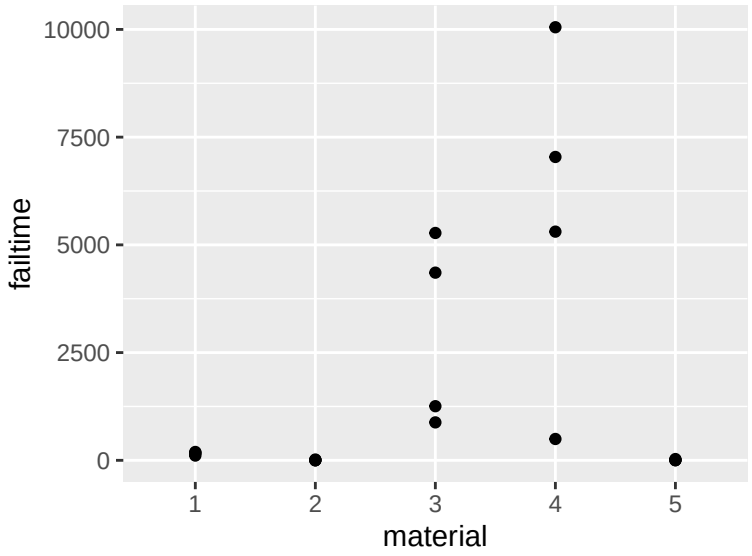
```
for (x in 1:(nrow(insl_means)-1)) {
  for (y in (x+1):nrow(insl_means)) {
    print(paste0(as.character(insl_means[x,1]),"-",
                 as.character(insl_means[y,1]), " ",
                 round(abs(insl_means$group_mean[x]-
                             insl_means$group_mean[y]),3),
                 "   ",
                 ifelse(abs(insl_means$group_mean[x]-
                              insl_means$group_mean[y]) < 0.870709,
                        "Not Significant", "Significant")))
  }
}
## [1] "1-2 1.654   Significant"
## [1] "1-3 1.157   Significant"
## [1] "1-4 1.373   Significant"
## [1] "1-5 1.367   Significant"
## [1] "2-3 2.812   Significant"
## [1] "2-4 3.028   Significant"
## [1] "2-5 0.287   Not Significant"
## [1] "3-4 0.216   Not Significant"
## [1] "3-5 2.524   Significant"
## [1] "4-5 2.74   Significant"
```

Underline Diagram

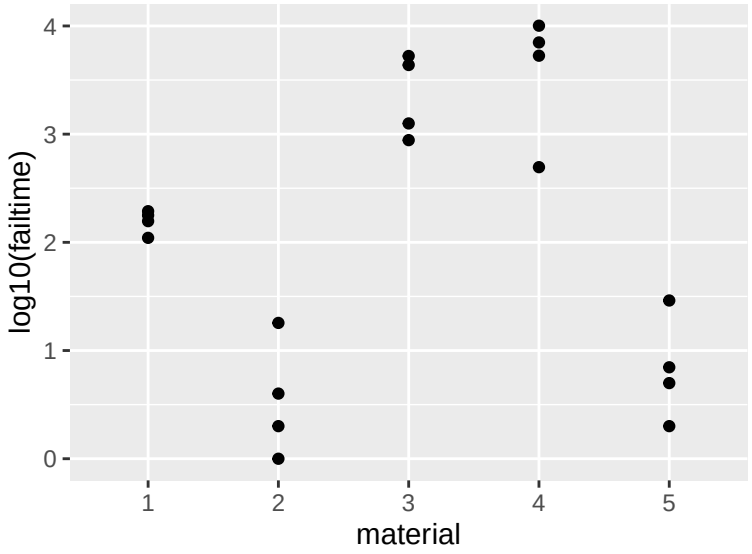
```
     2         5          1         3         4
  1.24245   1.90395    5.05159   7.71640   8.21401
  ------------------            -------------------
```

Q2d — 3 points

```
ggplot(data = insulate, aes(x=as.factor(material),
                            y=failtime)) +
  geom_point() +
  labs(x="material", y="failtime")
```




```
ggplot(data = insulate, aes(x=as.factor(material),
                            y=log10(failtime))) +
  geom_point() +
  labs(x="material", y="log10(failtime)")
```



After transforming the response, the standard deviation across groups appears to be more similar than before, with the standard deviations of the groups fairly close at a glance (besides group 1).

Q2e — 4 points

After the transformation, 1-2, 1-3, 1-5, 2-3, and 3-5 all become significant. All of the previously significant pairs are still significant.