STAT 425 - Statistical Design and Analysis of Experiments

Assignment 3

Name: Jana Osea
Student ID: 30016679

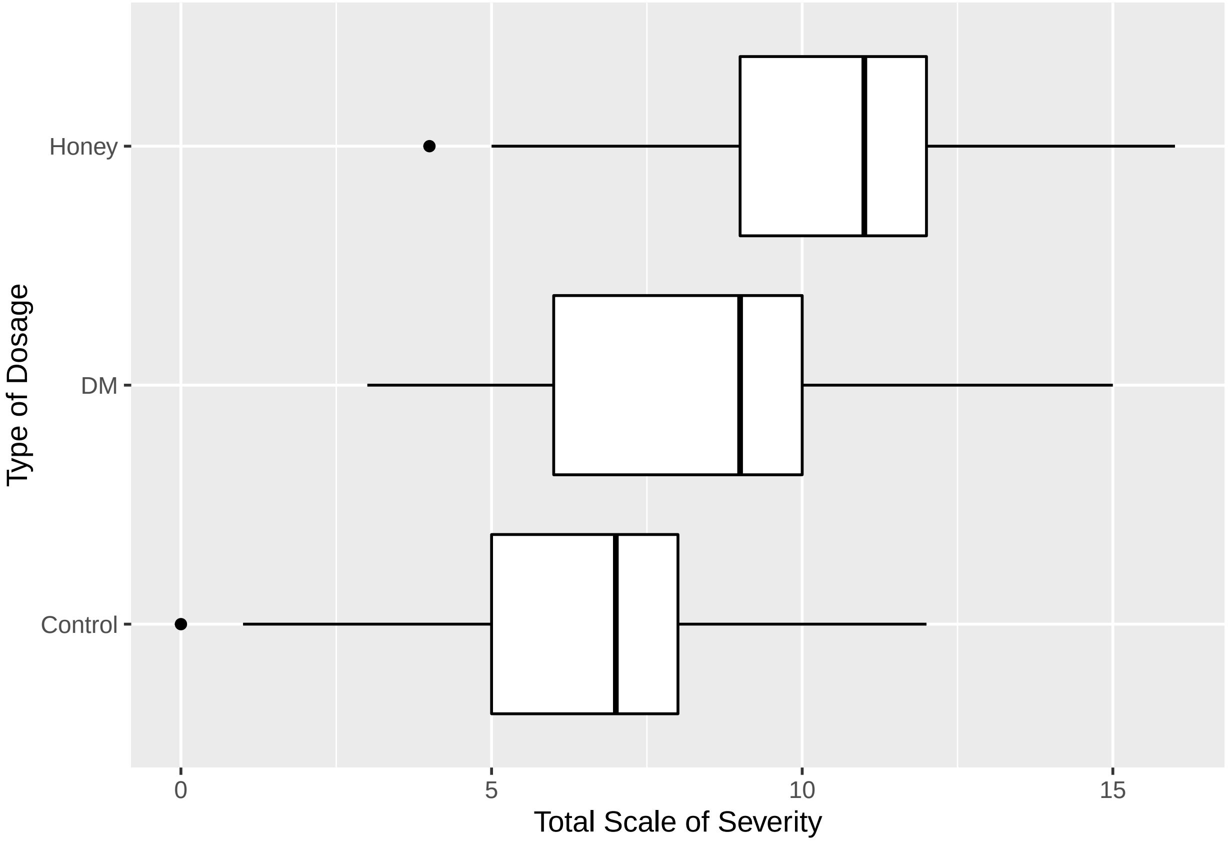
Problem 1 Pediatric researchers carrried out an experiment do investigate if a teaspoon on honey before bed calms a child's cough. A sample of n = children who were sick with an upper respiratory tract infection and their parents participated in this study. On the first night, parents rated their child's symptoms on a scale from 0 (no problems) to 6 (extremely severe) for five different areas. On the second night, the parents were instructed to give their sick child a dosage of liquid medicine prior to the child's bedtime. Unknown to the parents, some were given a dosage of dextromethorphan (DM) - an over-the-counter cough medicine, while others were given a similar dosage of honey. A third group of parents were giving their children nothing at all. Again, the parents rated their children's cough symptoms, and the improvements in total cough symptoms score was determined for each child. These data appear in the COUGH.csv file.

(a) Identify the treatment, the placebo, and the control group in this experiment.

Solution: In this experiment, the treatment group is the group of children given the honey. The placebo group is the group of children given the dextromethophan (DM). The control group is the group of children given nothing at all.

(b) Create boxplots of these data. What do your boxplots tell you about the "'cough score"?

Solution:



From the boxplots above, the the median total scale of severity is lowest for the control group, slightly higher for the placebo group, and highest for the treatment group. Also, the variance between the three groups appear to be the same because of the similar sized boxes.

(c) Assuming the conditions of (i) Normality of the e_{ij} terms and (ii) the variation of the cough score is the same, do these data indicate that there exists a treatment effect? Should your null hypotheses be rejected, rarry out one multiple comparison method to identify where the differences. Ensure you summarize and interpret your findings.

Solution: We are testing the following test of hypothesis:

$$H_0 : \mu_{Treatment} = \mu_{Placebo} = \mu_{Control}$$
$$H_a : \text{at least one } \mu \text{ is not the same for the } k = 3 \text{ populations}$$

Since, we can assume the residuals are normally distributed and the variation of the cough sore is the same, we can use R to perform the hypothesis test as follows:

```
summary(coughaov <- aov(severity ~ dosage, data=p1))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## dosage        2  310.9  155.47    17.2 3.29e-07 ***
## Residuals   108  976.4    9.04                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results above, we get that

$$F_{obs} = 17.2$$
$$P - value = P(F_{2,3} > F_{obs}) = 3.29e - 07$$

Based on this data, the $P - value = P(F_{2,108} > F_{obs}) = 3.29e - 07$ which is less than 0.05 so we reject the null hypothesis. One can conclude that on average the three populations are not equal with respect to the response variable (total scale of severity).

Because we rejected the global null hypothesis, we perform a Dunnett's comparison (since there is a control group) using the R code below:

```
DunnettTest(severity ~ dosage, data = p1, control="Control")
```

```
## 
##   Dunnett's test for comparing several treatments with a control :  
##     95% family-wise confidence level
## 
## $Control
##                   diff    lwr.ci   upr.ci    pval    
## DM-Control    1.702703 0.1357586 3.269647  0.0309 *  
## Honey-Control 4.081081 2.5141370 5.648025 1.1e-07 ***
## 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can summarize the results using the table below:

Lower Bound	$\mu_i - \mu_j$	Upper Bound	Finding
0.1357586	$\mu_{DM} - \mu_{Control}$	3.269647	$\mu_{DM} > \mu_{Control}$
2.5141370	$\mu_{Honey} - \mu_{Control}$	5.648025	$\mu_{Honey} > \mu_{Control}$

From the summary above, we see can conclude that:

$$\mu_{DM} > \mu_{Control}$$
$$\mu_{Honey} > \mu_{Control}$$

Based on the data and the table above derived from the Dunnett's multiple comparison method, we get that children who receive honey and DM result in higher mean total scale of severity scores compared to the controlled.

Problem 2 Refer to Assignment 2, Question 5: Can you conclude from the data collected that on average, people recall more commercials when they are watching a “’Neutral” program when compared to television programs with content that is either “Violent” or “Sexual”. Formulate the statistical hypothesis, then carry out the necessary statistical test. \
Solution: We are testing the null hypothesis that

$$\begin{aligned}
\mathrm{H}_0 &: L_1 = 0 \text{ or } \mu_{Neutral} - \frac{\mu_{Sex} + \mu_{Violent}}{2} = 0 \\
\mathrm{H}_A &: L_1 \neq 0 \text{ or } \mu_{Neutral} - \frac{\mu_{Sex} + \mu_{Violent}}{2} \neq 0
\end{aligned}$$

From the above, we can conclude that

$$c_{Neutral} = 1, c_{Sex} = -\frac{1}{2}, c_{Violent} = -\frac{1}{2}$$

We compute the value of $\widehat{L}_1^2$:

$$\begin{aligned}
\widehat{L}_1^2 &= \left(c_{Neutral}\overline{X}_{Neutral} + c_{Sex}\overline{X}_{Sex} + c_{Violent}\overline{X}_{Violent}\right)^2 \\
&= \left((1 * 3.166667) + \left(-\frac{1}{2} * 1.712963\right)\left(-\frac{1}{2} * 2.083333\right)\right) \\
&=1.609139
\end{aligned}$$

We also compute

$$\sum_{i=1}^{k=3} \frac{c_i^2}{n_i} = \frac{(1)^2}{108} + \frac{(-\frac{1}{2})^2}{108} + \frac{(-\frac{1}{2})^2}{108} = 0.01388889$$

Then, we find SS_{L_2} by

$$SS_{L_1} = \frac{\widehat{L}_1^2}{\sum_{i=1}^{k=3} \frac{c_i^2}{n_i}} = \frac{1.609139}{0.01388889} = 115.858$$

We use the above sum of squares to find the test statistic using the R code below:

```
p2 <- read.csv("p2.5.csv")
cneut <- 1
csex <- -1/2
cviol <- -1/2
summary(p2aov <- aov(SCORE~GROUP, data=p2))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## GROUP         2  123.3   61.63   20.45 4.36e-09 ***
## Residuals   321  967.4    3.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
xbarn <- favstats(SCORE~GROUP, data=p2)$mean
L1squared <- sum(xbarn*c(cneut, csex, cviol))^2
L1squared
```

```
## [1] 1.609139
```

```
denom <- sum(c(cneut, csex, cviol)^2/108)
SSL1 <- L1squared/denom
df1 <- 1
MSL1 <- SSL1/df1
MSE <- 3.01
teststat <- MSL1/MSE
teststat
```

```
## [1] 38.49104
```

```
pval <- 1-pf(teststat, 1,321)
pval
```

```
## [1] 1.69459e-09
```

From the results above, we get that

$$F_{obs} = 38.49104$$
$$P - value = P(F_{1,321} > F_{obs}) = 1.69459e - 09$$

Based on this data, the $P - value = P(F_{1,321} > F_{obs}) = 1.69459e - 097$ which is less than 0.05 so we reject the null hypothesis. One can conclude that on average people recall more commercials when they are watching a “’Neutral” program when compared to television programs with content that is either “Violent” or “Sexual”.

Problem 3 The data appearing in the data file GOLFRBD.csv and found in the link. The result of a random sample of four different brands of golf balls. For each brand (A, B, C, and D), a robotic golfer named “Iron Byron” was equipped with a 3-iron (a specific type of golf club) and hit a random sample of 10 balls of each of Brand A, Brand B, Brand C, and Brand D.

(a) Do these data suggest that there is variation in the distance of a golf ball (after it is hit) between the different brands of golf ball? State your statistical hypotheses, test statistic, P-value, and conclusion.

Solution: Since these are four golfball brands randomly chosen from a larger group of N-golfball brands, we treat the four brands A, B, C, and D as random levels of the golfball brand-factor.

$$\text{H}_0 : \sigma^2_\tau = 0 (\text{no variation between the N-golfball brands})$$
$$\text{H}_A : \sigma^2_\tau > 0 (\text{variatiion between the N-golfball brands exists})$$

We perform the test by using R below

```
p3 <- read.csv("p3.csv")
summary(p3aov <- aov(DISTANCE ~ BRAND, data=p3))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## BRAND        3   3299  1099.6   3.136 0.0372 *
## Residuals   36  12621   350.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above, we get the following results

$$F_{obs} = 3.136$$
$$P - value = P(F_{3,36} > F_{obs}) = 0.0372$$

Based on this data, the $P - value = P(F_{3,36} > F_{obs}) = 0.0372$ which is less than 0.05 so we reject the null hypothesis. One can conclude that on average there is a difference between the variation in the distance of a golf ball (after it is hit) between the different brands of golf ball.

(b) Compute the 95% confidence interval for σ_{Common}, the standard deviation of the distance of a golf ball hit by Iron Byron with a 3-iron.

Solution: Using the fact that

$$\frac{(\overbrace{n_1 + n_2 + \cdots + n_k}^{n} - k) * MSW}{\sigma^2_{Common}} \sim \chi^2_{df=n-k}$$

Then by rearranging, we can calculate the 95% confidence interval using R below

```
n <- 40
MSW <- 350.5695
top.value  = (n - 4)*MSW #computes numerator
lb.sigma = top.value/(qchisq(0.975, n - 4))
ub.sigma = top.value/(qchisq(0.025, n - 4))
sqrt(c(lb.sigma, ub.sigma))
```

```
## [1] 15.22615 24.32109
```

From the above, we get that

$$15.22615 \leq \sigma_{Common} \leq 24.32109.$$

Based on this data, the 95% confidence interval for σ is between 15.22615 and 24.32109. This means the if we perform this experiment many times, 95% of the common standard deviations calculated will be captured by this interval.

(c) From these data, compute the estimate for the $Var(X_{ij})$.

Solution: We estimate the total variation using the formula

$$S_p^2 + \widehat{\sigma}_\tau^2$$

We know that

$$S_p^2 = MSW = 350.5695$$

and also using R,

```
MSB <- 1099.5523
(MSB-MSW)/10
```

```
## [1] 74.89828
```

$$\widehat{\sigma}_\tau^2 = \frac{MSB - MSW}{10} = 74.89828$$

Putting it together, we get

$$S_p^2 + \widehat{\sigma}_\tau^2 = 350.5695 + 74.89828 = 425.4678$$

Based on the data, the estimate for the total variation $Var(X_{ij})$ is approximately 425.4678.

(d) Compute the 95% confidence interval for the intracluster correlation coefficient. Interpret the meaning of your finding(s) in the context of these data?

Solution: We know that the intraclass correlation coefficient is defined as

$$ICC = \frac{\sigma_\tau^2}{\sigma_{Common}^2 + \sigma_\tau^2}$$

The 95% confidence interval is computed as

$$\frac{L}{1+L} \leq ICC \leq \frac{U}{1+U}$$

where L and U are computed with

$$L = \frac{1}{n_i}\left[\frac{MSB}{MSW} * \frac{1}{F_{1-\frac{\alpha}{2},k-1,n-k}} - 1\right] \qquad U = \frac{1}{n_i}\left[\frac{MSB}{MSW} * \frac{1}{F_{\frac{\alpha}{2},k-1,n-k}} - 1\right]$$

First, we compute the value of L and U with R:

```
msbetween = MSB
mswithin = MSW
numerdf = 4 - 1
denomdf = 40 -4
ni = 10
#
L = (1/ni)*(((msbetween)/(mswithin)*(1/qf(0.975, numerdf, denomdf)) - 1)
U = (1/ni)*(((msbetween)/(mswithin)*(1/qf(0.025, numerdf, denomdf)) - 1)
icc.lb = L/(1 + L)
icc.ub = U/(1 + U)
c(icc.lb, icc.ub)
```

```
## [1] -0.01061778  0.81157427
```

The 95% confidence interval for ICC is

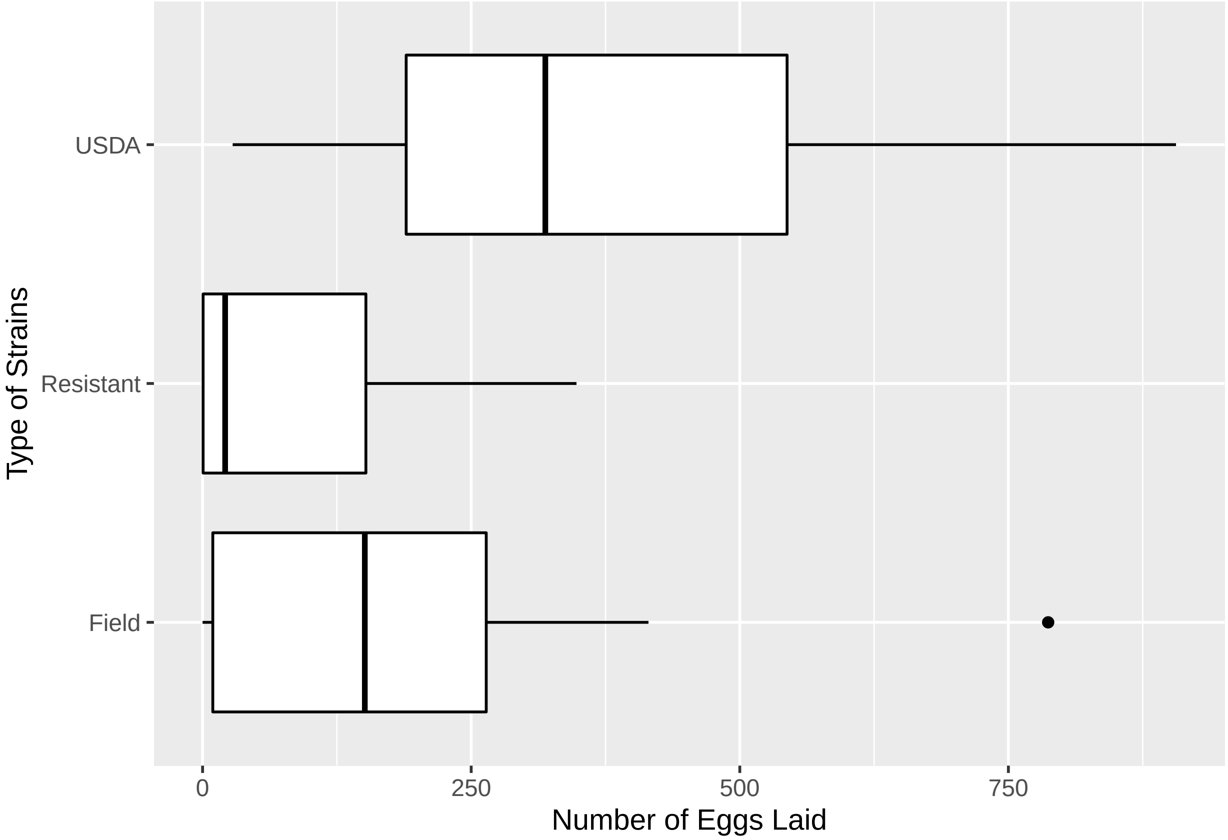
$$-0.01061778 \leq ICC \leq 0.81157427 \rightarrow 0 \leq ICC \leq 0.81157427$$

The variation between the different distance accounts for anywhere between 0% to 81.1574% of the variation in the golfball brands.

Problem 4 An entomologist counted the number of eggs laid by female moths on successive days in three strains of tobacco budworm from each of 15 matings. The data below are the number of eggs laid on the third day after the mating for each female in each of the strains.

(a) What do boxplots of these data tell you about the variation in the number of eggs laid between the three different strains of tobacco budworm? If it helps, carry out the relevant statistical test as well.

Solution:



From the boxplots above, the variation in the number of eggs laid between the three different strains of tobacco budworm appears to be the same. However, it is difficult to make any conclusion concerning the variance between the populations to be the same or not. Hence, we perform a Levene's test with the following hypothesis:

$$H_0 : \sigma_{USDA} = \sigma_{Field} = \sigma_{Resistant} = \sigma_{Common}$$
$$H_a : \text{at least one } \sigma \text{ is different for all } k = 3 \text{ strains}$$

Using R, we can perform a Levene Test to test for equal variances below:

```
leveneTest(eggs~strain, data=p4)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  2  2.9029 0.06594 .
##       42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above, we get that

$$F_{obs} = 2.9029$$
$$P - value = P(F_{2,42} > F_{obs}) = 0.06594$$

Based on this data, the $P - value = P(F_{2,42} > F_{obs}) = 0.06594$ which is greater than 0.05 and so we do not reject the null hypothesis. One can conclude that on average the standard deviation in the number of eggs laid by the different strains of tobacco budworms is the same. However, this p-value is quite low and so it is good practice to transform the data to stabilize the variance.

(b) Consider the nature of the variable. What transformation method would you suggest to stabilize the variation in the number of eggs laid?

Solution: We wish to discover a transformation which will produce a more normal response variable using the transformation:

$$x_{ij}^{'} = x_{ij}^{\lambda}$$

where

$$\sigma_{x'_{ij}} \text{ is proportional to } \mu^{\lambda+\text{some power}-1}$$

Furthermore, we use the relation

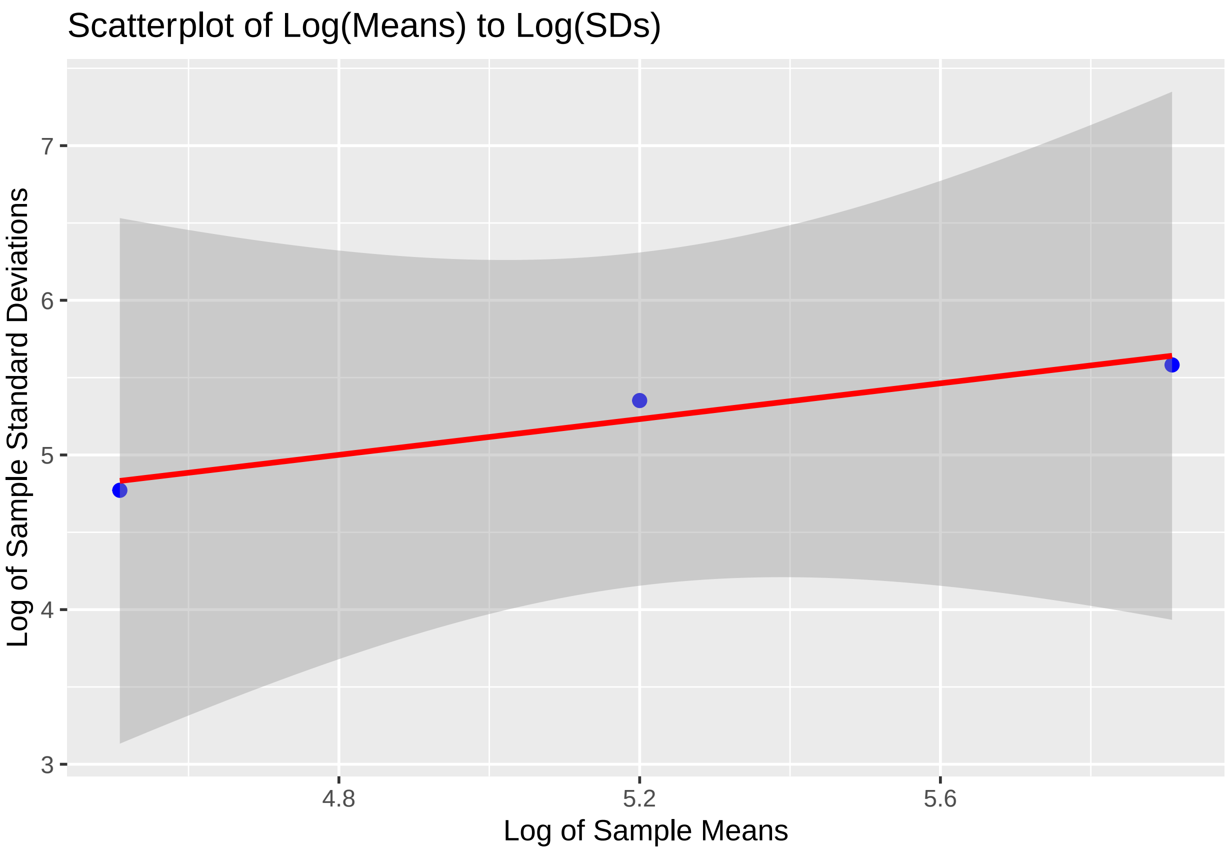
$$log(\sigma_{X_{ij}}) = log(\text{some OTHER constant}) + (\text{some constant}) * log(\mu_i)$$

which means that a plot of the log of the $\sigma_{X'_{ij}}$ to the $log(\mu_i)$ should be a straight line. Using the R code below, we produce this plot:

```r
a <- favstats(eggs~strain, data=p4)
logsds = log(a$sd)
logmeans = log(a$mean)
logstats.df = data.frame(a$strain, logsds, logmeans)
logstats.df
```

```
##    a.strain   logsds logmeans
## 1     Field 5.351816 5.199969
## 2 Resistant 4.771588 4.508659
## 3      USDA 5.582072 5.908083
```

```r
ggplot(data=logstats.df, aes(x = logmeans, y = logsds)) + geom_point(col="blue", size=2) + xlab("Log of
```



Using the moethod of least squares, we compute the estimate for **some OTHER constant** and **some constant**:

```
summary(lm(logsds ~ logmeans, data=logstats.df))
```

```
##
## Call:
## lm(formula = logsds ~ logmeans, data = logstats.df)
##
## Residuals:
##        1        2        3
##  0.11990 -0.06067 -0.05923
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2257     0.7771   2.864    0.214
## logmeans      0.5781     0.1484   3.896    0.160
##
## Residual standard error: 0.1468 on 1 degrees of freedom
## Multiple R-squared:  0.9382, Adjusted R-squared:  0.8764
## F-statistic: 15.18 on 1 and 1 DF,  p-value: 0.16
```

From the above, **some OTHER constant** is estimated by 2.2257 and **some constant** is estimated by 0.5781. Hence, we can find λ below:

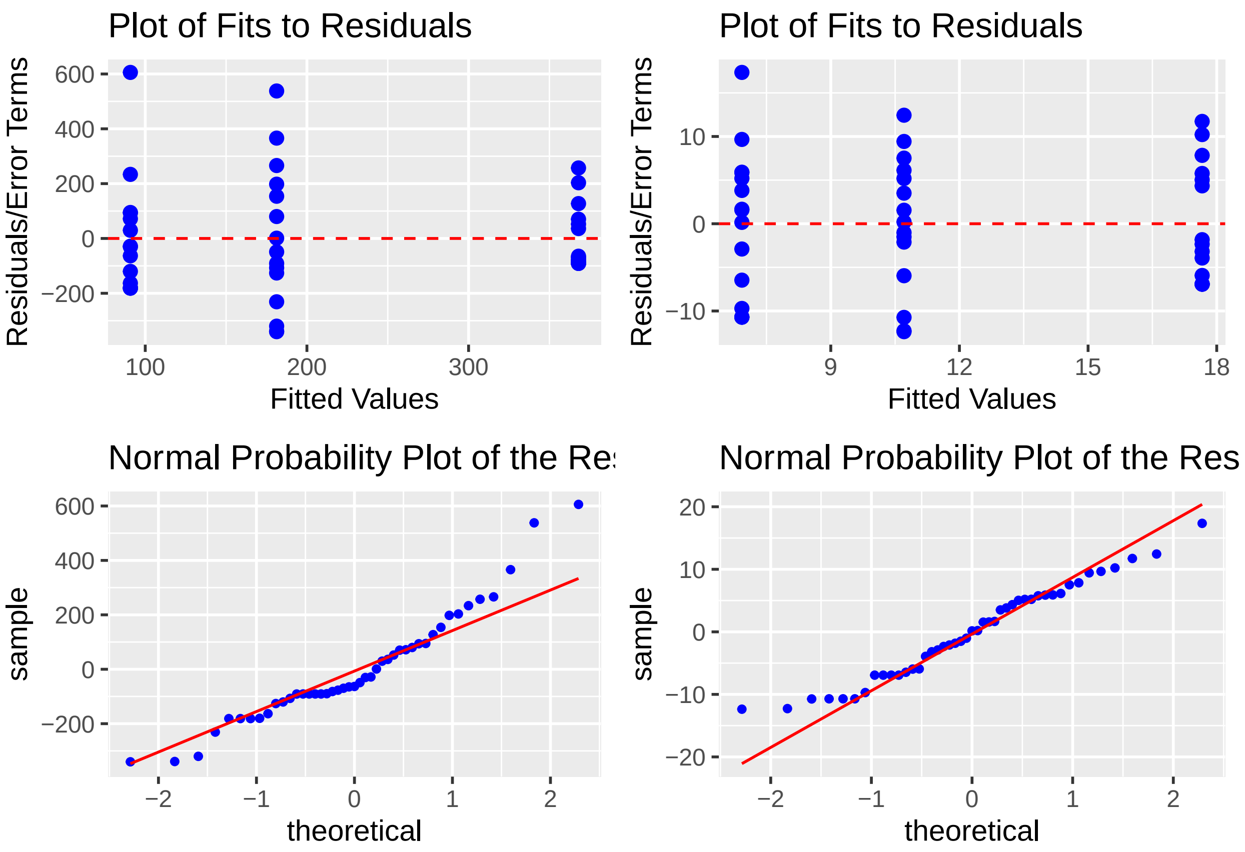
$$\begin{aligned}\lambda &= 1 - \textbf{some constant}\\ &= 1 - 0.5781\\ &= 0.4219\\ &\approx 0.5\end{aligned}$$

Hence, the transformation of the original data X_{ij} is

$$X'_{ij} = X_{ij}^{0.5} = \sqrt{X_{ij}}$$

(c) Apply the transformation you suggest in part (b). What are your finding(s)?

Solution: Using R, we apply the suggested transformation in (b). Observe the difference between the plots below:



The left side of the plot above is the residual vs fitted and QQplot of the original response variable (eggs) and the right side is for the transformed response variable. Notice that the right residual vs fitted plot has a less apparent cone shape indicating that the transformed response variable provides stronger homoscedasticity. As well, notice that the right QQplot follows the straight line more than the left QQplot–this means that the residuals of the transformed response variable follow a normal distribution better.

As well, we perform a Levene Test to compare with (a).

```r
leveneTest(eggsT~strain, data=p4)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.2162 0.8065
##       42
```

From the above, we get that

$$F_{obs} = 0.2162$$
$$P - value = P(F_{2,42} > F_{obs}) = 0.8065$$

Based on this data, the $P - value = P(F_{2,42} > F_{obs}) = 0.8065$ which is greater than 0.05 and so we do not reject the null hypothesis. Notice how the transformed data produces a higher p-value compared to (a) indicating that the variance is more stable with the transformed data.

(d) Refer to part (c): Carry out a test of equal means to the transformed data from part (c). If you find a statistically significant difference, apply Tukey's HSD method to identify where the mean difference in your transformed data are.

Solution: We assume that the conditions (homoscedasticity and normality of residuals) are met. We test the following hypothesis:

$$H_0 : \mu_{USDA} = \mu_{Field} = \mu_{Resistant}$$
$$H_a : \text{at least one of the k=3 strains is not the same}$$

Using R to perform the test:

```r
summary(aov1 <- aov(eggsT~strain, data=p4))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## strain       2  889.2   444.6   7.522 0.00161 **
## Residuals   42 2482.7    59.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From which, we get

$$F_{obs} = 7.522$$
$$P - value = P(F_{2,42} > F_{obs}) = 0.00161$$

Based on this data, the $P-value = P(F_{2,42} > F_{obs}) = 0.00161 $ which is less than 0.05 and so we reject the null hypothesis. One can conclude that at least one k strains of tobacco budworm lays on average a different number of eggs. Since we rejected the null hypothesis, we will perform Tukey's HSD method to identify where the mean differences in the transformed data using the R code below:

```r
PostHocTest(aov(eggsT ~ strain, data=p4), method="hsd", conf.level=0.95, ordered = TRUE)
```

```
##
##   Posthoc multiple comparisons of means : Tukey HSD
##     95% family-wise confidence level
##     factor levels have been ordered
##
## $strain
##                       diff     lwr.ci   upr.ci   pval
## Field-Resistant  3.781395 -3.0391855 10.60197 0.3777
## USDA-Resistant  10.733657  3.9130769 17.55424 0.0012 **
## USDA-Field       6.952262  0.1316823 13.77284 0.0449 *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can summarize the results using the table below:

Lower Bound	$\mu_i - \mu_j$	Upper Bound	Finding
-3.0392	$\mu_{Field} - \mu_{Resistant}$	10.60	$\mu_{Field} = \mu_{Resistant}$
3.9131	$\mu_{USDA} - \mu_{Resistant}$	0.0012	$\mu_{USDA} > \mu_{Resistant}$
0.1317	$\mu_{USDA} - \mu_{Field}$	13.77	$\mu_{USDA} > \mu_{Field}$

From the summary above, we see can conclude that:

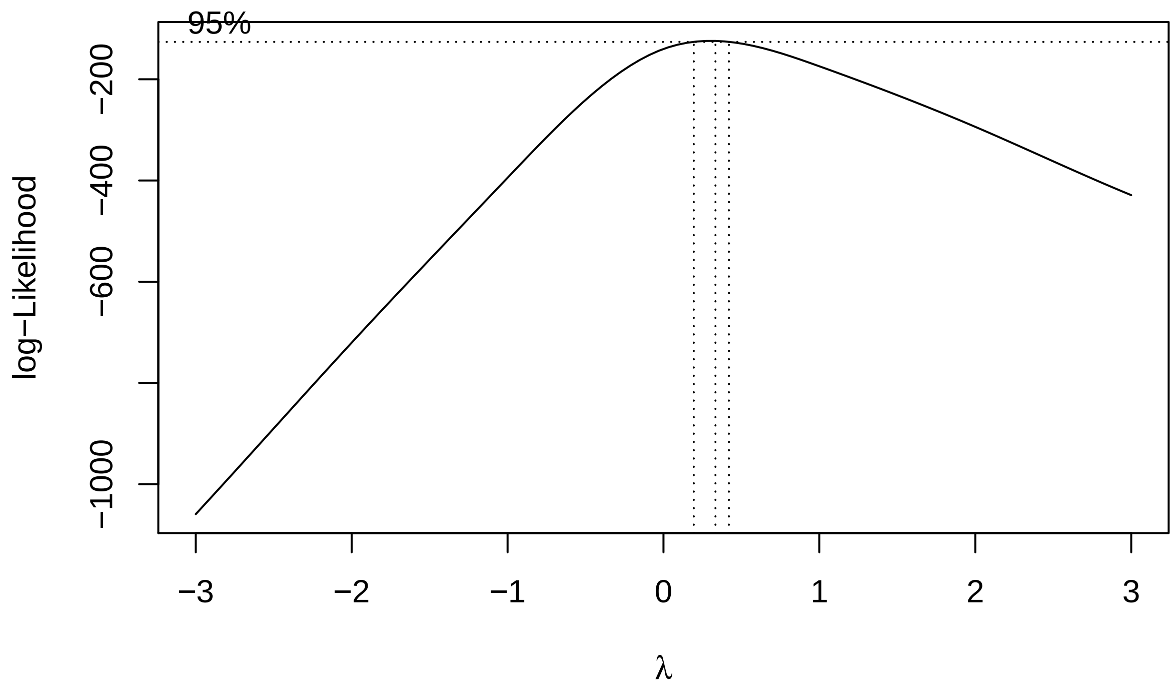
$$\mu_{USDA} > \mu_{Resistant} = \mu_{Field}$$

Based on the data and the table above derived from the Tukey's HSD method, we get that the USDA strain of the tobacco budworm lays the high number of eggs and that the resistant and field strains lays approximately the same number of eggs.

(e) Consider the Box-Cox transformation method. Apply this method to these data. What transformation method does the Box-Cox method suggest? It is the same as in part (c) or different?

Solution: Using the Box-Cox transformation to compute the value of λ for the variance stabilization transformation. It is good to note that the response varaible contains some 0 values. Hence, prior to running the Box-Cox transformation, we add 0.01 to each response value since Box-Cox uses strictly positive values.

```
p4$eggsPlus <-p4$eggs + 0.01
originalmodel = lm(eggsPlus ~ strain, data=p4) #assign the orginal model
bcoutput = boxcox(originalmodel, lambda=seq(-3, 3))
```



```
best.lambda=bcoutput$x[which(bcoutput$y == max(bcoutput$y))]
best.lambda
```

```
## [1] 0.3333333
```

From the output above, the suggest transformation is as follows:

$$X_{ij}^{\lambda} = X_{ij}^{0.33} \approx X_{ij}^{0.5} = \sqrt{X_{ij}}$$

Compared to part (c) where $\lambda = 0.5$, the Box-Cox transformation suggests to perform the same transformation to the response variable.

Problem 5 The potential of solar panels installed on the roofs build above national highways as a source of energy was investigated. Compute simulation was used to estimate the monthly solar energy, in kilowatt-hours, generated from solar panels installed across 200-km stretch of highway in India. Each month, the simulation was run under each of four conditions: single-layer solar panels, double-layer solar panels that are 1 metre apart; double-layer solar panels that are 2metres apart, and double-layer solar panels that are 3 metres apart. The data collected for 12 months are found in the data file

(a) State the statistical hypothesis that is required to investigate the mean amount of energy generated by the four solar panel configurations.

Solution: We consider the model

$$\begin{aligned} X_{ij} =& \mu + (\mu_{i.} - \mu) + (\mu_{.j} - \mu) + e_{ij} \\ X_{ij} =& \mu + \tau_i + \beta_j + e_{ij} \qquad \text{for } i = S, D1, D2, D3; j = 1, 2, 3, \cdots, 12 \end{aligned}$$

where

$$\begin{aligned} &\tau_i = \mu_{i.} - \mu && \text{(effect of population } i = S, D1, D2, D3\text{)} \\ &\beta_j = \mu_{.j} - \mu && \text{(effect of month block)} \\ &e_{ij} = \text{random error term} \end{aligned}$$

Now, we test the following hypothesis where μ_i = mean amount of energy generate for $i = S, D1, D2, D3$ where S = Single, $D1$=double layer 1 metre apart, $D2$= double layer 2 metres apart, and $D3$= double layer 3 metres apart.

$$\begin{aligned} &\text{H}_0 : \tau_S = \tau_{D1} = \tau_{D2} = \tau_{D2} = 0 && \text{(the mean response is the same for all } 4-\text{condition levels)} \\ &\text{H}_A : \text{H}_0 \text{ is false} \end{aligned}$$

(b) Carry out the appropriate statistical test to address the inquiry in part (a). Ensure you provide the value of the test statistic, P-value, decision and conclusion.

Solution:

Assuming that the conditions for the ANOVA test are satisfied, we use the following R code to perform an ANOVA test for the hypothesis stated in (a).

```r
summary(aov(Energy~Condition+Month, data=p5))
```

```
##             Df   Sum Sq  Mean Sq F value Pr(>F)    
## Condition    3 49730750 16576917  115.54 <2e-16 ***
## Month       11 90618107  8238010   57.42 <2e-16 ***
## Residuals   33  4734730   143477                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From which, we get

$$F_{obs} = 115.54$$
$$P - value = P(F_{3,33} > F_{obs}) = 2e - 16$$

Based on this data, the $P - value = P(F_{3,33} > F_{obs}) \approx 0$ which is less than 0.05 and so we reject the null hypothesis. One can conclude that on average, the amount of energy generated is different for at least one of the conditions.

(c) Is there a “Month” effect here? State the appropriate statistical hypotheses, then carry out the test. What can you infer?

Solution: Given the model in (a) above, We wish to test if there is a “month” effect by using the following test of hypothesis

$$\begin{aligned} &\text{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_{12} = 0 && \text{(the mean energy generated is the same for all } 12-\text{month levels)} \\ &\text{H}_A : \text{H}_0 \text{ is false} \end{aligned}$$

We use R to perform the appropriate ANOVA test:

```r
summary(block <- aov(Energy ~ Condition + Month, data=p5))
```

```
##             Df   Sum Sq  Mean Sq F value Pr(>F)    
## Condition    3 49730750 16576917  115.54 <2e-16 ***
## Month       11 90618107  8238010   57.42 <2e-16 ***
## Residuals   33  4734730   143477                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From which, we get

$$F_{obs} = 57.42$$
$$P - value = P(F_{11,33} > F_{obs}) = 2e - 16$$

Based on this data, the $P - value = P(F_{11,33} > F_{obs}) \approx 0$ which is less than 0.05 and so we reject the null hypothesis. One can conclude that on average, the mean amount of energy produced for at least one month is different. Thus, there exists a "month" effect.

(d) Consider your finding in part (b). Carry out the appropriate multiple comparison method to identify the "best" solar panel configuration.

Solution: We will perforam a Tukey's Honestly Significant difference multiple comparison test:

~~~~
```
Tukey multiple comparisons of means
95% family-wise confidence level
factor levels have been ordered

Fit: aov(formula = Energy ~ Condition + Month, data = p5)

$Condition
           diff        lwr       upr     p adj
D1-S  1077.9167  659.63016 1496.2032 0.0000003
D2-S  2153.9167 1735.63016 2572.2032 0.0000000
D3-S  2637.0000 2218.71349 3055.2865 0.0000000
D2-D1 1076.0000  657.71349 1494.2865 0.0000003
D3-D1 1559.0833 1140.79682 1977.3698 0.0000000
D3-D2  483.0833   64.79682  901.3698 0.0184786
```
~~~~

We can summarize the results using the table below:

Lower Bound	$\mu_i - \mu_j$	Upper Bound	Finding
659.63016	$\mu_{D1} - \mu_S$	1496.2032	$\mu_{D1} > \mu_S$
1735.63016	$\mu_{D2} - \mu_S$	2572.2032	$\mu_{D2} > \mu_S$
2218.71349	$\mu_{D3} - \mu_S$	3055.2865	$\mu_{D3} > \mu_S$
657.71349	$\mu_{D2} - \mu_{D1}$	1494.2865	$\mu_{D2} > \mu_{D1}$
1140.79682	$\mu_{D3} - \mu_{D1}$	1977.3698	$\mu_{D3} > \mu_{D1}$
64.79682	$\mu_{D3} - \mu_{D2}$	901.3698	$\mu_{D3} > \mu_{D2}$

From the summary above, we see can conclude that:

$$\mu_{D3} > \mu_{D2} > \mu_{D1} > \mu_S$$

Based on the data and the table above derived from the Tukey's HSD method, we get $i = D3$=double layer with 3 metres apart is the best configuration for generating energy.

Problem 6 A soft-drink manufacturer uses five agents (1, 2, 3, 4, 5) to handle 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. The results are given below.

(a) Agents 1 and 2 distribute merchandise only, agents 3 and 4 distribute cash-value coupons only, and agent 5 distributes both merchandise and coupons. Does the data indicate that there is no difference in the time lapse of the premium distributions between a "merchandise only" approach and a "cash-value coupon" approach? Use $\alpha = 0.05$.

Solution: We wish to test the following hypothesis:

$$\text{H}_0 : L_1 = 0 \text{ or } \frac{\mu_{A1} + \mu_{A2}}{2} - \frac{\mu_{A3} + \mu_{A4}}{2} = 0$$
$$\text{H}_A : L_1 \neq 0 \text{ or } \frac{\mu_{A1} + \mu_{A2}}{2} - \frac{\mu_{A3} + \mu_{A4}}{2} \neq 0$$

From the above, we can conclude that

$$c_{A1} = \frac{1}{2}, c_{A2} = \frac{1}{2}, c_{A3} = -\frac{1}{2}, c_{A4} = -\frac{1}{2}$$

Using these contants, we perform the following calculations on R to produce the test statistics and p-value:

```
p6 <- read.csv("p6.csv")

timelapse <- c(p6[,2], p6[,3], p6[,4], p6[,5], p6[,6])
agent <- as.factor(c(rep(1, 20), rep(2, 20), rep(3, 20), rep(4, 20), rep(5, 20)))
p6 <- data.frame(agent, timelapse)

c1 <- 1/2
c2 <- 1/2
c3 <- -1/2
c4 <- -1/2
summary(p6aov <- aov(timelapse~agent, data=p6))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## agent        4   4430  1107.5   147.2 <2e-16 ***
## Residuals   95    715     7.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
xbarn <- favstats(timelapse~agent, data=p6)$mean[1]
L1squared <- sum(xbarn*c(c1, c2, c3, c4))^2
L1squared
```

```
## [1] 0
```

```
denom <- sum(c(c1, c2, c3, c4)^2/20)
denom
```

```
## [1] 0.05
```

```
SSL1 <- L1squared/denom
SSL1
```

```
## [1] 0
```

```
  df1 <- 1
  MSL1 <- SSL1/df1
  MSL1
```

```
## [1] 0
```

```
  MSE <- 7.5226
  fObs <- MSL1/MSE
  fObs
```

```
## [1] 0
```

```
  1-pf(fObs, 1,98)
```

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

From the results above, we get that

$$F_{obs} = 280.6892$$
$$P - value = P(F_{1,93} > F_{obs}) \approx 0$$

Based on this data, the $P - value = P(F_{1,93} > F_{obs}) \approx 0$ which is less than 0.05 so we reject the null hypothesis. One can conclude that on average there is a difference in the time lapse of the premium distributions between a "merchandise only" approach and a "cash-value coupon" approach.

(b) Using Scheffe's procedure, construct a 99% family of confidence interval estimates for the above constrasts. Interpret your results.

Solution: We will use the following formula to compute the 99% family of confidence interval estimates:

$$\widehat{L} \pm S * \sqrt{MSW\left(\sum_{i=1}^{k} \frac{c_i^2}{n_i}\right)}$$

where

$$\widehat{L}_1 = c_1\overline{X}_1 + c_2\overline{X}_2 + \cdots + c_k\overline{X}_k$$

As well, since we are computing a family wise confidence, we use the following alpha to calculate each margin or error:

```
alpha <- 1 - 0.99^(1/5)
alpha
```

```
## [1] 0.002008048
```

We will use the R function below to produce the corresponding margin of errors:

```
contrastError = function(alpha, MSW, c1, c2, c3, c4, k, ni) #create a function that computes the MOE of
{
  svalue = sqrt((k - 1)*qf(1 - alpha, k - 1, (k*ni - k)))
  contrast.sd = sqrt(MSW*((c1^2/ni) + (c2^{2}/ni) + (c3^2/ni) + (c4^{2}/ni)))
  contrast.moe = svalue*contrast.sd
  contrast.moe
}
```

```
moecontrast1 = contrastError(alpha, MSE, 1, -1, 0, 0, 5, 20)
moecontrast2 = contrastError(alpha, MSE, 1, -1, 0, 0, 5, 20)
moecontrast3 = contrastError(alpha, MSE, 1/2, 1/2, -1, 0, 5, 20)
moecontrast4 = contrastError(alpha, MSE, 1/2, 1/2, -1, 0, 5, 20)
moecontrast5 = contrastError(alpha, MSE, 1/2, 1/2, -1/2, -1/2, 5, 20)
```

We will use the R code below to produce the corresponding L estimates:

```
lEstimate <- function(c1, c2, c3, c4, xbar1, xbar2, xbar3, xbar4)
{
  a <- c(c1, c2, c3, c4)
  b <- c(xbar1, xbar2, xbar3, xbar4)
  sum(a*b)
}



x <- favstats(timelapse~agent, data=p6)
a1 <- x$mean[1]
a2 <- x$mean[2]
a3 <- x$mean[3]
a4 <- x$mean[4]
a5 <- x$mean[5]

L1 <- lEstimate(1, -1, 0, 0, a1, a2, 0, 0)
L2 <- lEstimate(1, -1, 0, 0, a3, a4, 0, 0)
L3 <- lEstimate(1/2, 1/2, -1, 0, a1, a2, a5, 0)
L4 <- lEstimate(1/2, 1/2, -1, 0, a3, a4, a5, 0)
L5 <- lEstimate(1/2, 1/2, -1/2, -1/2, a1, a2, a3, a4)
```

Using the formula above, we can calculate the family wise 99% confidence intervals below:

```
a <- L1 + c(-1,1)*moecontrast1 #CI for L1
b <- L2 + c(-1,1)*moecontrast2 #CI for L2
c <- L3 + c(-1,1)*moecontrast3 #CI for L3
d <- L4 + c(-1,1)*moecontrast4 #CI for L4
e <- L5 + c(-1,1)*moecontrast5 #CI for L5



lower <- as.numeric(rbind(a, b, c, d, e)[,1])
upper <- as.numeric(rbind(a, b, c, d, e)[,2])
combo <- c("L1", "L2", "L3", "L4", "L5")
data.frame(combo, lower, upper)
```

```
##   combo      lower       upper
## 1    L1  -1.711106   5.7111062
## 2    L2  -6.761106   0.6611062
## 3    L3  -9.763912  -3.3360878
## 4    L4 -20.038912 -13.6110878
## 5    L5   7.650852  12.8991484
```

The output above can be summarized below:

Lower Bound	$\mu_i - \mu_j$	Upper Bound	Finding
-1.7111	$\mu_1 - \mu_2$	5.71111	$\mu_1 = \mu_2$
-6.7611	$\mu_3 - \mu_4$	0.66111	$\mu_3 = \mu_4$
-9.7639	$\frac{\mu_1+\mu_2}{2} - \mu_5$	-3.33609	$\frac{\mu_1+\mu_2}{2} < \mu_5$
-20.0389	$\frac{\mu_3+\mu_4}{2} - \mu_5$	-13.61109	$\frac{\mu_3+\mu_4}{2} < \mu_5$
7.6509	$\frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2}$	12.89915	$\frac{\mu_1+\mu_2}{2} > \frac{\mu_3+\mu_4}{2}$

- Based on the data with a family confidence of 99%, the confidence interval for L_1 is between -0.8774332 and 4.8774332. One can infer that on average, there is no difference between mean transaction days between agent 1 and 2.
- Based on the data with a family confidence of 99%, the confidence interval for L_2 is between -5.9274332 and -0.1725668. One can infer that on average, the mean transaction days for agent 4 and for agent 3 are the same.
- Based on the data with a family confidence of 99%, the confidence interval for L_3 is between -9.3482433 and -3.7517567. One can infer that on average, the merchandise only method produces lower mean transaction days than a combination of merchandise and coupons.
- Based on the data with a family confidence of 99%, the confidence interval for L_4 is between -19.6232433 and -14.0267567. One can infer that on average, the coupon only method produces lower mean transaction days than a combination of merchandise and coupons.
- Based on the data with a family confidence of 99%, the confidence interval for L_5 is between 7.8058761 and 12.7441239. One can infer that on average, the coupon only method produces lower mean transaction days than merchandise only method.

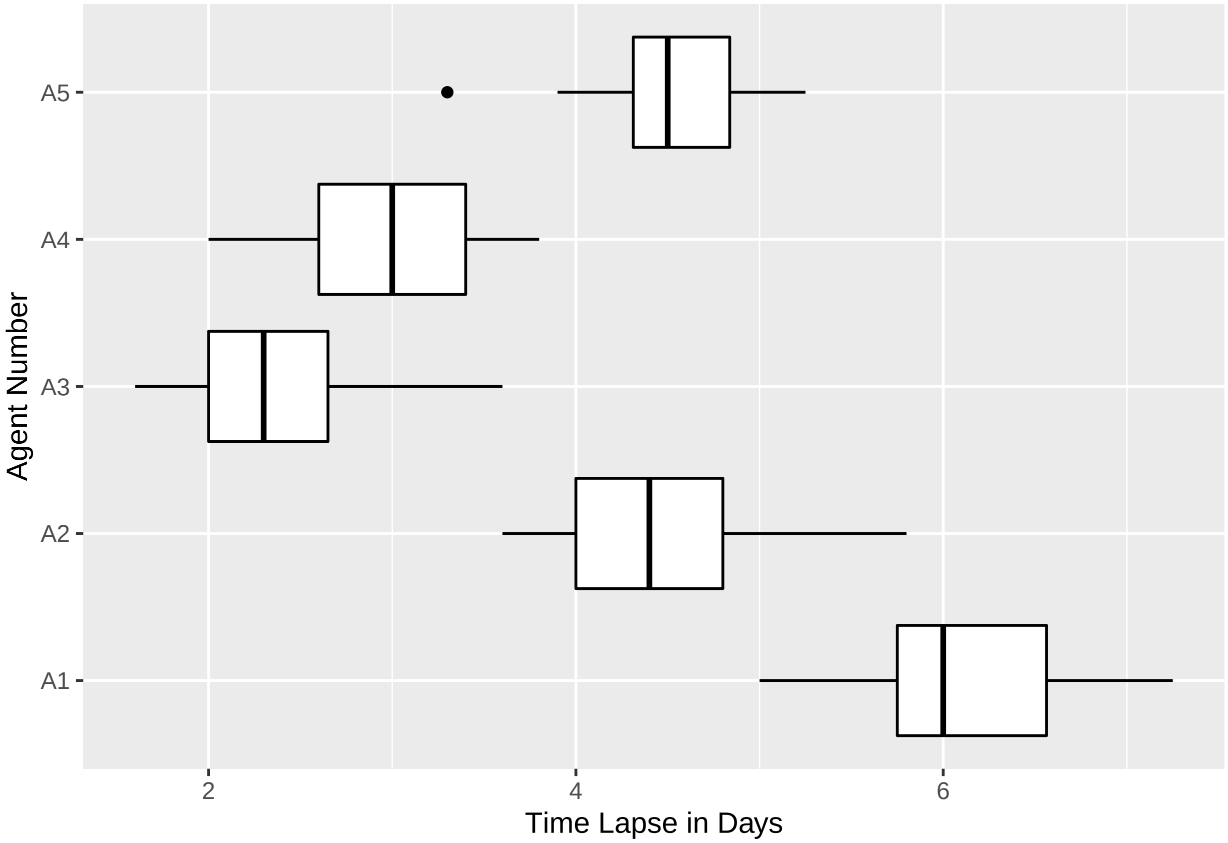
(c) Of all the premium distributions, 25% are handled by Agent 1, 20% by Agent 2, 20% by Agent 3, 20% by Agent 4, and 15% by Agent 5. Using this new information, is there a better (faster) method for distributing premiums ? Test at $\alpha = 0.05$.

Solution: Given that 25% are handled by Agent 1, 20% by Agent 2, 20% by Agent 3, 20% by Agent 4, and 15% by Agent 5. We will first multiply all the response variables by their corresponding percentages according as shown below:

```
p6 <- read.csv("p6.csv")
A1 <- p6$A1*0.25
A2 <- p6$A2*0.2
A3 <- p6$A3*0.2
A4 <- p6$A4*0.2
A5 <- p6$A5*0.15
p6T <- stack(data.frame(A1, A2, A3, A4, A5))
```

Next, we will determine if the variances are equal by looking at the boxplots below:

```
p6T <- stack(data.frame(A1, A2, A3, A4, A5))
ggplot(data=p6T, aes(x = ind, y = values)) + geom_boxplot(col="black", fill="white") + xlab("Agent Numbe
```



From the boxplots above, we can conclude that the variances for all agents are the same because the width of all the boxes appear to be similar. Next, we will use the following formula to compute the pairwise 95% confidence intervals from the multiplied data. Note that $\overline{X}_i$ is the average for the **multiplied** data.

$$(\overline{X}_i - \overline{X}_j) \pm t_{1-\frac{\alpha}{2}, df=n_i+n_j-2} * \sqrt{S_p^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

where $i \neq j$ and $i, j \in 1, 2, 3, 4, 5$. Using the t.test command in R, we produce

```
a <- t.test(A2,A1, alternative = "two.sided", var.equal=T)$conf.int[1:2]
b <- t.test(A3,A1, alternative = "two.sided", var.equal=T)$conf.int[1:2]
c <- t.test(A4,A1, alternative = "two.sided", var.equal=T)$conf.int[1:2]
d <- t.test(A5,A1, alternative = "two.sided", var.equal=T)$conf.int[1:2]
e <- t.test(A3,A2, alternative = "two.sided", var.equal=T)$conf.int[1:2]
f <- t.test(A4,A2, alternative = "two.sided", var.equal=T)$conf.int[1:2]
g <- t.test(A5,A2, alternative = "two.sided", var.equal=T)$conf.int[1:2]
h <- t.test(A4,A3, alternative = "two.sided", var.equal=T)$conf.int[1:2]
i <- t.test(A5,A3, alternative = "two.sided", var.equal=T)$conf.int[1:2]
j <- t.test(A5,A4, alternative = "two.sided", var.equal=T)$conf.int[1:2]

lower <- as.numeric(rbind(a,b,c,d,e,f,g,h,i,j)[,1])
upper <- as.numeric(rbind(a,b,c,d,e,f,g,h,i,j)[,2])

combo <- c("A2-A1","A3-A1", "A4-A1", "A5-A1", "A3-A2",
  "A4-A2", "A5-A2", "A4-A3", "A5-A3", "A5-A4")
data.frame(combo, lower, upper)
```

```
##    combo      lower      upper
## 1  A2-A1 -2.0170021 -1.2379979
## 2  A3-A1 -4.1522958 -3.4227042
## 3  A4-A1 -3.5395955 -2.8154045
## 4  A5-A1 -1.9731334 -1.2718666
## 5  A3-A2 -2.5165524 -1.8034476
## 6  A4-A2 -1.9037892 -1.1962108
## 7  A5-A2 -0.3370488  0.3470488
## 8  A4-A3  0.2836093  0.9363907
## 9  A5-A3  1.8513736  2.4786264
## 10 A5-A4  1.2445186  1.8654814
```

The output above can be summarized below:

Lower Bound	$\mu_i - \mu_j$	Upper Bound	Finding
-2.0170021	$\mu_2 - \mu_1$	-1.2379979	$\mu_2 < \mu_1$
-4.1522958	$\mu_3 - \mu_1$	-3.4227042	$\mu_3 < \mu_1$
-3.5395955	$\mu_4 - \mu_1$	-2.8154045	$\mu_4 < \mu_1$
-1.9731334	$\mu_5 - \mu_1$	-1.2718666	$\mu_5 < \mu_1$
-2.5165524	$\mu_3 - \mu_2$	-1.8034476	$\mu_3 < \mu_2$
-1.9037892	$\mu_4 - \mu_2$	-1.1962108	$\mu_4 < \mu_2$
-0.3370488	$\mu_5 - \mu_2$	0.3470488	$\mu_5 = \mu_2$
0.2836093	$\mu_4 - \mu_3$	0.9363907	$\mu_4 > \mu_3$
1.8513736	$\mu_5 - \mu_3$	2.4786264	$\mu_5 > \mu_3$
1.2445186	$\mu_5 - \mu_4$	1.8654814	$\mu_5 > \mu_4$

From the table above, we can conclude that

$$\mu_1 > \mu_2 = \mu_5 > \mu_4 > \mu_3$$

We know that μ_i is the mean days for agent i to copmlete a transaction. Based on the data, agent 1 has produces the highest mean days and agent 3 and 4 produces the first and second lowest mean days correspondignly. The conclusion suggests that agents 3 and 4 should handle a higher percentage of the premium distributions and agent 1 should have alower the percantage of premium distridubtions. Since, agent 3 and 4 distribute cash-value coupons only, there is evidence to suggest that cash value coupon only method is the best for distributing premiums.