STP 531

Applied Analysis of Variance

Homework 6

Nathan A. Nguyen

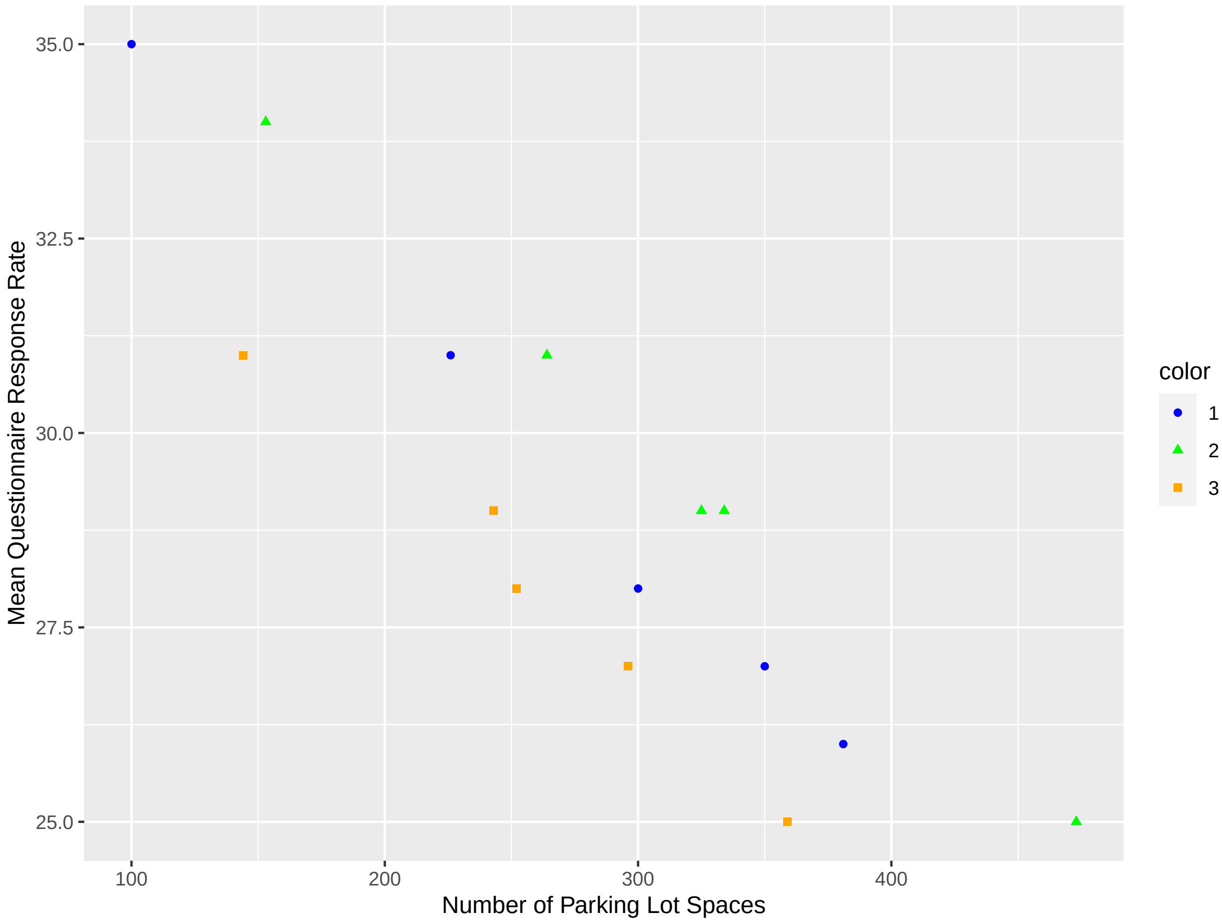
09 April 2021

Question 22.2

The analyst is right to be concerned about the nature of the relationship between the concomitant variable and the response outcome of interest. It is imperative that the analyst investigate these relationships before blindly using the ANCOVA framework, e.g., making sure that there exists a strong correlation between the concomitant and the response variable, and making sure that there doesn't exist a correlation between the concomitant and the treatment variable(s). If this is not looked at, you will have biased and misleading results. Although ANCOVA is used to make increase the precision for analysis, it will not be of any use of the concomitant is affected by the treatment variable. Furthermore, there are a lot of assumptions that the analyst would need to make about their sample if no demographic information is given. The group of people who answered the survey may be from very different groups of people, so it may be the case that the concomitant varies grossly with the groups.

Question 22.10

Part A



It does appear that there are color effects present. It looks like an orange coloured flier generally elicits lower response rates compared to blue and green coloured fliers. I wouldn't be confident enough to say whether or not a green flier performs better than a blue flier though. I'd rather rely on a formal statistical test for this.

Part C

I'm going to use R's default factor variable coding instead of using the book's regression indicator variable approach.

$$Y_{ij} = \mu_{.} + \tau_i + \gamma(X_{ij} - \bar{X}_{..}) + \epsilon_{ij} \text{ (22.3) (full model)}$$
$$Y_{ij} = \mu. + \tau_1 I_{ij1} + \tau_2 I_{ij2} + \gamma x_{ij} + \epsilon_{ij} \text{ where } x_{ij} = X_{ij} - \bar{X}_{..}, \bar{X}_{..} = 280$$
$$Y_{ij} = \mu_{.} + \gamma x_{ij} + \epsilon_{ij} \text{ (reduced model)}$$

$$H_0 : \tau_1 = \tau_2 = 0$$
$$H_a : \text{ not both } \tau_1 \text{ and } \tau_2 = 0$$
$$F^* = \frac{SSE(R) - SSE(F)}{(n_T - 2) - [n_T - (r+1)]} \div \frac{SSE(F)}{n_T - (r+1)}$$
$$F^* = \frac{24.708 - 1.316}{(15-2) - [15 - (3+1)]} \div \frac{1.316}{15 - (3+1)} = \frac{23.392}{2} * \frac{11}{1.316} \approx 97.76$$
$$\alpha = 0.10, F_{critical} = F(0.90; 2, 11) \approx 2.86$$
$$F^* \approx 97.76 > 2.86 \approx F_{critical}, \text{ reject } H_0$$
$$p = 0^+$$

Since $F^* > F_{critcal}$ there is sufficient evidence for us to reject the null hypothesis. In other words, there is evidence to believe that there are treatment effects present. The p-value for this test is 9.8923764×10^{-8}.

Part D

Table 1: ANOVA Full Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
color	2	7.600000	3.8000000	31.75838	2.69e-05
scale(spaces, center = T, scale = F)	1	115.083812	115.0838120	961.80933	0.00e+00
Residuals	11	1.316188	0.1196535	NA	NA

Table 2: No Concomitant ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
color	2	7.6	3.8	0.3917526	0.6842074
Residuals	12	116.4	9.7	NA	NA

As we can see from the two ANOVA tables, the mean square errors for the model including the concomitant has a substantially smaller value than the model that only includes the color factor variable. I would say that using the parking lot spaces as a concomitant variable and the ANCOVA framework is justified in this case. We see a substantial reduction in error. Furthermore, there is a strong linear relationship between response rate and the parking lot spaces as well. In addition to this, the treatment regression lines all look parallel. No serious departures from this assumption is suggested based off of the plot though to be rigorous, we should perform a hypothesis test. In other words, we should take into account the lot spaces in this analysis.

Part E

From page 931:

$$
\begin{aligned}
&\mu_{.} + \tau_1 \rightarrow \hat{\mu}_{.} + \hat{\tau}_1 \text{ with variance} \\
&\sigma^2\{\hat{\mu}_{.}\} + \sigma^2\{\hat{\tau}_1\} + 2COV\{\hat{\mu}_{.}, \hat{\tau}_1\} \\
&= 0.024 \\
s\{\hat{Y}\} &\approx 0.15 \\
29.14 &= \mu_{.} + \tau_1 \pm 1.80(0.15) \\
t_{critical} &= t(0.95; 11) \approx 1.80
\end{aligned}
$$

In our case since colour 1 (blue is the reference class, it is the intercept value 29.14

Table 3: Confidence Estiamte

	Estimate
Lower Limit	28.87062
Upper Limit	29.40938

Question 22.9

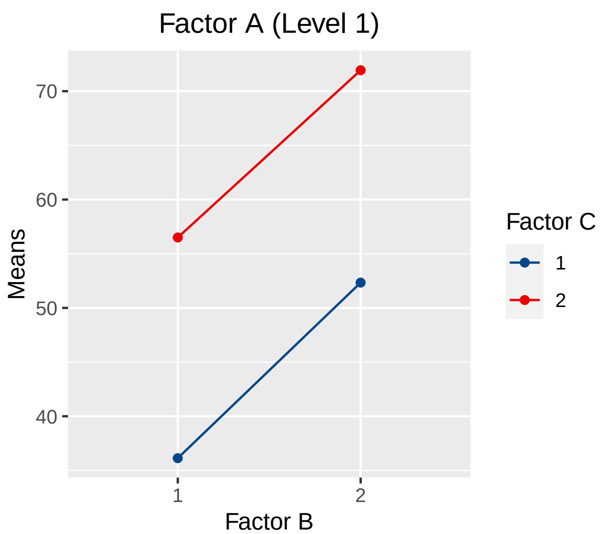
Part D

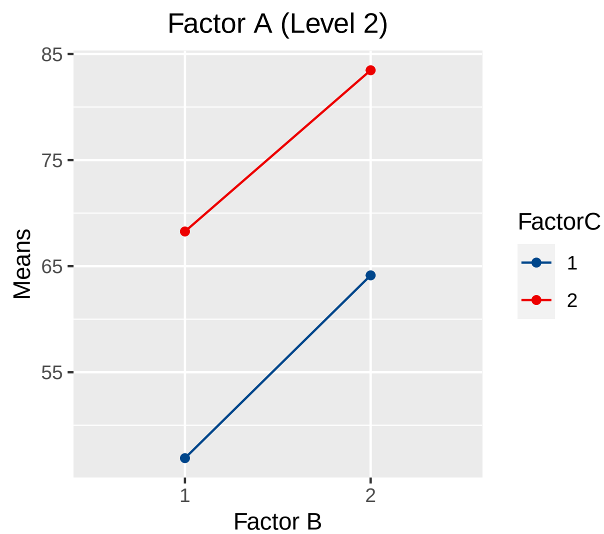
We would not be able to conduct a lack of fit test because we don't have any replicates of the response for the respective X's.

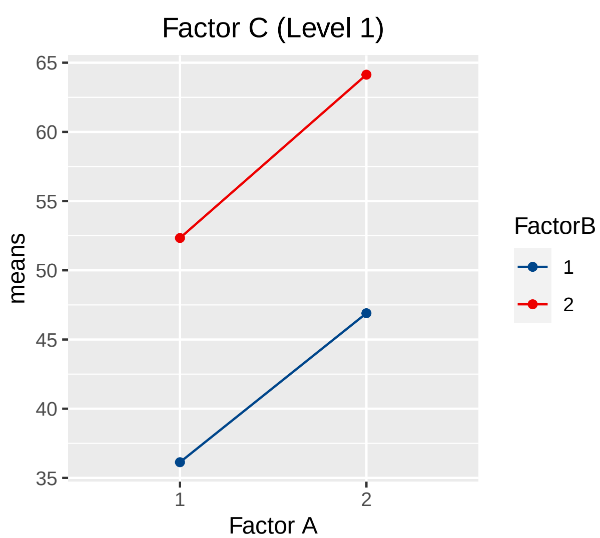
Question 24.7

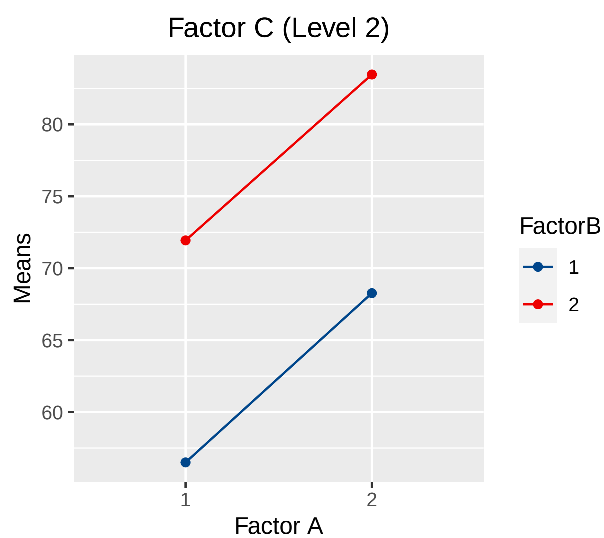
$$a = 2, \text{ factor A: alloy molten state}$$
$$b = 2, \text{ factor B: temperature hardening process}$$
$$c = 2, \text{ factor C: time length of hardening process}$$
$$2 \times 2 \times 2 \text{ design}, \; n_i = 3$$
$$Y_{ijkm} = \mu_{...} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkm} \; (24.14)$$

Part A









There do not appear to be any interaction among the factors but there are main effects present.

Part B

Table 4: Three Way ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FactorA	1	788.9066667	788.9066667	234.8810321	0.0000000
FactorB	1	1539.2016667	1539.2016667	458.2662201	0.0000000
FactorC	1	2440.1666667	2440.1666667	726.5103585	0.0000000
FactorA:FactorB	1	0.2400000	0.2400000	0.0714552	0.7926448
FactorA:FactorC	1	0.2016667	0.2016667	0.0600422	0.8095443
FactorB:FactorC	1	2.9400000	2.9400000	0.8753256	0.3633923
FactorA:FactorB:FactorC	1	0.6016667	0.6016667	0.1791341	0.6777527
Residuals	16	53.7400000	3.3587500	NA	NA

Part C

$$\begin{aligned}
\alpha &= 0.025 \\
H_0 &: \text{ all } (\alpha\beta\gamma)_{ijk} = 0 \\
H_a &: \text{ not all } (\alpha\beta\gamma)_{ijk} = 0 \\
F^* &= \frac{MSABC}{MSE} = \frac{0.60}{3.36} \approx 0.179 \\
F_{critical} &= F(0.975; 1, 16) \approx 6.12 \\
F^* &< F_{critical}, \text{ conclude } H_0 \\
p &= 0.68 > 0.025 = \alpha
\end{aligned}$$

Since $F^* > F_{critical}$ and $p = 0.68 > 0.025 = \alpha$ there is not evidence evidence to reject the null hypothesis. In other words, there is insufficient evidence to conclude that interactions are present.

Part D

$\alpha = 0.025$

$$\begin{aligned}
&\text{Test for AB interactions}\\
H_0 &: \text{ all } (\alpha\beta)_{ij} = 0\\
H_a &: \text{ not all } (\alpha\beta)_{ij} = 0\\
F^* &= \frac{MSAB}{MSE} = \frac{0.24}{3.36} \approx 0.071\\
F_{critical} &= F(0.975; 1, 16) \approx 6.12\\
F^* &< F_{critical} \text{ fail to reject null}\\
p &\approx 0.79
\end{aligned}$$

$$\begin{aligned}
&\text{Test for AC interactions}\\
H_0 &: \text{ all } (\alpha\gamma)_{ik} = 0\\
H_a &: \text{ not all } (\alpha\gamma)_{ik} = 0\\
F^* &= \frac{MSAC}{MSE} = \frac{0.20}{3.36} \approx 0.06\\
F_{critical} &= F(0.975; 1, 16) \approx 6.12\\
F^* &< F_{critical} \text{ fail to reject null}\\
p &\approx 0.81
\end{aligned}$$

$$\begin{aligned}
&\text{Test for BC interactions}\\
H_0 &: \text{ all } (\beta\gamma)_{jk} = 0\\
H_a &: \text{ not all } (\beta\gamma)_{jk} = 0\\
F^* &= \frac{MSBC}{MSE} = \frac{0.24}{3.36} \approx 0.88\\
F_{critical} &= F(0.975; 1, 16) \approx 6.12\\
F^* &< F_{critical} \text{ fail to reject null}\\
p &\approx 0.36
\end{aligned}$$

All $p > 0.025$ so we failed to reject all three null hypothesis. There no two-way interactions present.

Part E

$\alpha = 0.025$ and all have common critical F-value ≈ 6.12

$$\begin{aligned}
&\text{Test for A main effects}\\
H_0 &: \text{ all } (\alpha)_i = 0\\
H_a &: \text{ not all } (\alpha)_i = 0\\
F^* &= \frac{MSA}{MSE} = \frac{788.91}{3.36} \approx 234.79\\
F_{critical} &= F(0.975; 1, 16) \approx 6.12\\
F^* &> F_{critical},\ p \approx 5.54 \times 10^{-11} < \alpha\\
&\text{conclude the alternative } (H_a)
\end{aligned}$$

$$\begin{aligned}
&\text{Test for B main effects}\\
H_0 &: \text{ all } (\beta)_j = 0\\
H_a &: \text{ not all } (\beta)_j = 0\\
F^* &= \frac{MSB}{MSE} = \frac{1539.20}{3.36} \approx 458.10\\
F_{critical} &= F(0.975; 1, 16) \approx 6.12\\
F^* &> F_{critical},\ p \approx 3.36 \times 10^{-13} < \alpha\\
&\text{conclude the alternative } (H_a)
\end{aligned}$$

$$\begin{aligned}
&\text{Test for C main effects}\\
H_0 &: \text{ all } (\gamma)_k = 0\\
H_a &: \text{ not all } (\gamma)_k = 0\\
F^* &= \frac{MSC}{MSE} = \frac{2440.17}{3.36} \approx 726.24\\
F_{critical} &= F(0.975; 1, 16) \approx 6.12\\
F^* &> F_{critical},\ p \approx 9.24 \times 10^{-15} < \alpha\\
&\text{conclude the alternative } (H_a)
\end{aligned}$$

All $p << 0.025 = \alpha$, so there is sufficient evidence to reject all three null hypothesis. In other words, there is reason to believe that are main effects present for all three factor variables.

Part F

$$\alpha < 1 - (1 - 0.025)^7 \approx 0.16$$

1. There are no three-way interactions present ($p > 0.025$)
2. There are no two-way interactions present ($p > 0.025$)
3. There are main effects present ($p < 0.025$)

Part G

Yes they do.

Question 24.8

$$B = t\left[1 - \frac{\alpha}{2g}; (n-1)abc\right] = t\left[1 - \frac{0.05}{6}; 16\right] \approx 2.673$$

Part A

$$\begin{aligned}
\hat{D}_1 &= \hat{\mu}_{2..} - \hat{\mu}_{1..} = 65.7 - 54.2 = 11.5 \\
\hat{D}_2 &= \hat{\mu}_{.2.} - \hat{\mu}_{.1.} = 68 - 52 = 16 \\
\hat{D}_3 &= \hat{\mu}_{..2} - \hat{\mu}_{..1} = 70 - 49.9 = 20.1 \\
MSE &= 3.36, s\{\hat{D}\} \approx 0.74 \\
\hat{D}_i &\pm Bs\{\hat{D}\}
\end{aligned}$$

Table 5: Bonerroni 95%

	D1	D2	D3
Lower Limit	9.499686	13.99969	18.09969
Upper Limit	13.500314	18.00031	22.10031

Part B

$$\begin{aligned}
\bar{Y}_{222} &= 83.5 \\
s\{\bar{Y}_{222}\} &\approx 1.06 \\
\bar{Y}_{222} &\pm t(0.975, 16)s\{\hat{Y}\} \\
83.5 &\pm 2.12(1.06)
\end{aligned}$$

Table 6: 95% Estimate

	Y.222
Lower Limit	-1.123497
Upper Limit	3.363497

Appendix

Question 22.10

Part A

```
# getdata <- function(...){
#   e = new.env()
#   name = data(..., envir = e)[1]
#   e[[name]]
# }
#
# data <- getdata("QuestionnaireColor")

data <- read.table("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData
                   col.names = c("response", "color", "index", "spaces"))

data$color <- as.factor(data$color)

p1 <- data %>%
  ggplot(aes(x = spaces, y = response, color = color)) +
  geom_point() + geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Number of Parking Lot Spaces",
       y = "Mean Questionnaire Response Rate") +
  scale_color_manual(values = c("blue", "green", "orange"))
p1
```

Part B

Part C

Part E

```
# regression coefficient variance-covariance matrix
# mod.F <- lm(response ~ relevel())
# summary(mod.f)
blah <- vcov(mod.F)

# Xmat <- model.matrix(mod.F)
term <- 29.14
t.value <- qt(0.95,11)
s <- 0.15
LL <- term - t.value*s; UL <- term + t.value*s
moo <- data.frame(
  Estimate = c(LL, UL)
)

row.names(moo) <- c("Lower Limit", "Upper Limit")
kable(moo[], caption = "Confidence Estiamte", format = "markdown") %>%
  kable_styling(position = "center")
```

Question 22.9

Part D

Question 24.7

Part A

```
data <- read.table("https://people.stat.sc.edu/hitchcock/casehardening.txt",
                   header = F, col.names = c("Hardness", "FactorA", "FactorB", "FactorC", "Index"))

data$FactorA <- as.factor(data$FactorA)
data$FactorB <- as.factor(data$FactorB)
data$FactorC <- as.factor(data$FactorC)

means <- data %>%
  group_by(FactorA, FactorB, FactorC) %>%
  summarise(means = mean(Hardness))

FA.1 <- means %>%
  filter(FactorA == 1)

p2 <- FA.1 %>%
  ggplot(aes(x = FactorB, y = means, colour = FactorC)) +
  geom_point() + geom_line(aes(group = FactorC)) +
  scale_color_lancet() +
  labs(x = "Factor B",
       y = "Means",
       color = "Factor C",
       title = "Factor A (Level 1)") +
  theme(plot.title = element_text(hjust = 0.5))

FA.2 <- means %>%
  filter(FactorA == 2)

p3 <- FA.2 %>%
  ggplot(aes(x = FactorB, y = means, colour = FactorC)) +
  geom_point() + geom_line(aes(group = FactorC)) +
  scale_color_lancet() +
  labs(x = "Factor B",
       y = "Means",
       title = "Factor A (Level 2)") +
  theme(plot.title = element_text(hjust = 0.5))

FC.1 <- means %>%
  filter(FactorC == 1)

p4 <- FC.1 %>%
  ggplot(aes(x = FactorA, y = means, colour = FactorB)) +
  geom_point() + geom_line(aes(group = FactorB)) +
  scale_color_lancet() +
  labs(x = "Factor A",
       y = "means",
       title = "Factor C (Level 1)") +
  theme(plot.title = element_text(hjust = 0.5))

FC.2 <- means %>%
  filter(FactorC == 2)
```

```r
p5 <- FC.2 %>%
  ggplot(aes(x = FactorA, y = means, colour = FactorB)) +
  geom_point() + geom_line(aes(group = FactorB)) +
  scale_color_lancet() +
  labs(x = "Factor A",
       y = "Means",
       title = "Factor C (Level 2)") +
  theme(plot.title = element_text(hjust = 0.5))

g1 <- ggarrange(p2, p3, p4, p5, nrow = 2, ncol = 2)
g1
```

Part B

```r
three_way.mod <- lm(Hardness ~ FactorA*FactorB*FactorC, data = data)
three_way_table <- anova(three_way.mod)
kable(three_way_table[], caption = "Three Way ANOVA", format = "markdown") %>%
  kable_styling(position = "center")
# alt.3 <- aov(Hardness ~ FactorA*FactorB*FactorC, data = data)
```

Part C

```r
F.crit <- qf(1-0.025,1,16)
p <- pf(F.star,1,16,lower.tail = FALSE)
```

Part D

```r
# AB
MSE <- 3.36
F.star <- 0.24/MSE
p <- pf(F.star,1,16,lower.tail=FALSE)

# AC
F.star <- 0.20/MSE
p <- pf(F.star,1,16,lower.tail=FALSE)

# BC
F.star <- 2.94/MSE
p <- pf(F.star,1,16,lower.tail=FALSE)
```

Part E

```r
# all have a common F critical
blah <- (3-1)*2*2*2
F.crit <- qf(0.975,1,16)

# A Main
F.star <- 788.91/MSE
p <- pf(F.star,1,16,lower.tail=FALSE)

# B main
```

```r
F.star <- 1539.20/MSE
p <- pf(F.star,1,16,lower.tail=FALSE)

# C Main
F.star <- 2440.17/MSE
p <- pf(F.star,1,16,lower.tail=FALSE)
```

Part F

```
alpha <- 0.025
right <- 1 - (1-alpha)^7
```

Part G

Question 24.8

Part A

```r
B <- qt(1 - (0.05/6), 16)
MSE <- MSE

# Factor A
FA.mean <- means %>%
  group_by(FactorA) %>%
  summarise(means = mean(means))

mu2.. <- 65.7
mu1.. <- 54.2
D1 <- mu2.. - mu1..

# Factor B
FB.mean <- means %>%
  group_by(FactorB) %>%
  summarise(means = mean(means))

mu.2. <- 68
mu.1. <- 52
D2 <- mu.2. - mu.1.

# Factor C
FC.mean <- means %>%
  group_by(FactorC) %>%
  summarise(means = mean(means))

mu..2 <- 70
mu..1 <- 49.9
D3 <- mu..2 - mu..1

S2 <- MSE/(3*2)
s <- sqrt(S2)

D1.LL <- D1 - B*s; D1.UL <- D1 + B*s
D2.LL <- D2 - B*s; D2.UL <- D2 + B*s
D3.LL <- D3 - B*s; D3.UL <- D3 + B*s

blah <- data.frame(
  D1 = c(D1.LL, D1.UL),
  D2 = c(D2.LL, D2.UL),
  D3 = c(D3.LL, D3.UL)
)

rownames(blah) <- c("Lower Limit", "Upper Limit")
kable(blah[], caption = "Bonerroni 95%", format = "markdown") %>%
  kable_styling(position = "center")
```

Part B

```r
t.value <- qt(0.975,16)
S2.Y.bar <- MSE/(3)
s <- sqrt(S2.Y.bar)
Y.LL <- S2.Y.bar - t.value*s; Y.UL <- S2.Y.bar + t.value*s
moo <- data.frame(
  Y.222 = c(Y.LL, Y.UL)
)

row.names(moo) <- c("Lower Limit", "Upper Limit")
kable(moo[], caption = "95% Estimate", format = "markdown") %>%
  kable_styling(position = "center")
```