STAT 6338: ADVANCED STATISTICAL METHODS II

Homework 2

1 Question 16.4

a) Below is a representation of the model:

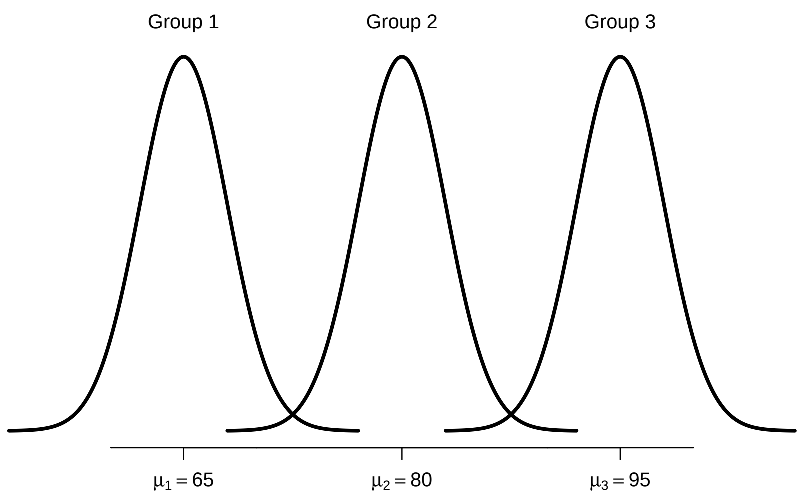


Figure 1: *Analysis of Variance Model Representation*

b)

$$
\begin{aligned}
\mu. &= \frac{\sum \mu_i}{r} = \frac{65+80+95}{3} = 80 \\
E\{MSTR\} &= \sigma^2 + \frac{\sum n_i(\mu_i - \mu.)^2}{r-1} \\
&= (3)^2 + \frac{25(65-80)^2 + 25(80-80)^2 + 25(95-80)^2}{3-1} = 5634 \\
E\{MSE\} &= \sigma^2 = (3)^2 = 9
\end{aligned}
$$

$E\{MSTR\}$ is substantially larger than $E\{MSE\}$, which implies that the factor level means μ_i are not equal.

2 Question 16.5

a) Figure 2 is a representation of the model:

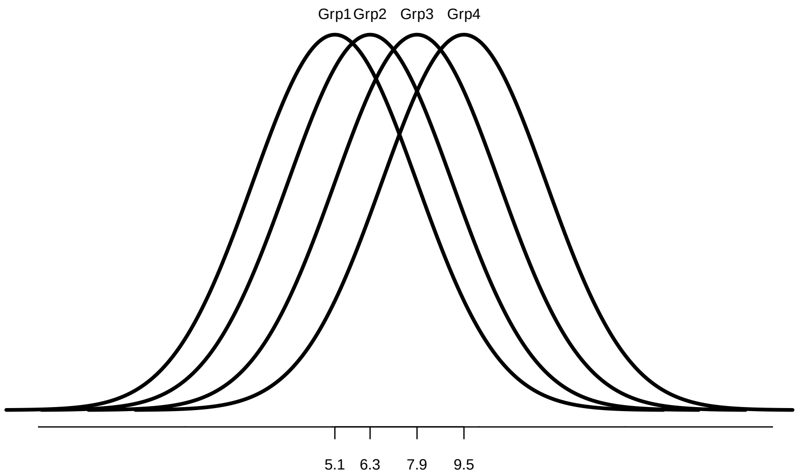


Figure 2: *Analysis of Variance Model Representation*

b)

$$\begin{aligned}
\mu. &= \frac{\sum \mu_i}{r} = \frac{5.1 + 6.3 + 7.9 + 9.5}{4} = 7.2 \\
E\{MSTR\} &= \sigma^2 + \frac{\sum n_i(\mu_i - \mu.)^2}{r-1} \\
&= (2.8)^2 + \frac{100\left[(5.1-7.2)^2 + (6.3-7.2)^2 + (7.9-7.2)^2 + (9.5-7.2)^2\right]}{4-1} \\
&= 374.5067 \\
E\{MSE\} &= \sigma^2 = (2.8)^2 = 7.84
\end{aligned}$$

$E\{MSTR\}$ is substantially larger than $E\{MSE\}$, which implies that the factor level means μ_i are not equal.

c)

$$\begin{aligned}
\mu. &= \frac{\sum \mu_i}{r} = \frac{5.1 + 5.6 + 9.0 + 9.5}{4} = 7.3 \\
E\{MSTR\} &= \sigma^2 + \frac{\sum n_i(\mu_i - \mu.)^2}{r-1} \\
&= (2.8)^2 + \frac{100\left[(5.1-7.3)^2 + (5.6-7.3)^2 + (9.0-7.3)^2 + (9.5-7.3)^2\right]}{4-1} \\
&= 523.1733
\end{aligned}$$

$E\{MSTR\}$ is substantially larger here than in (b) because now the factor level means μ_i are further away from the grand mean ($\mu.$).

3 Question 16.7

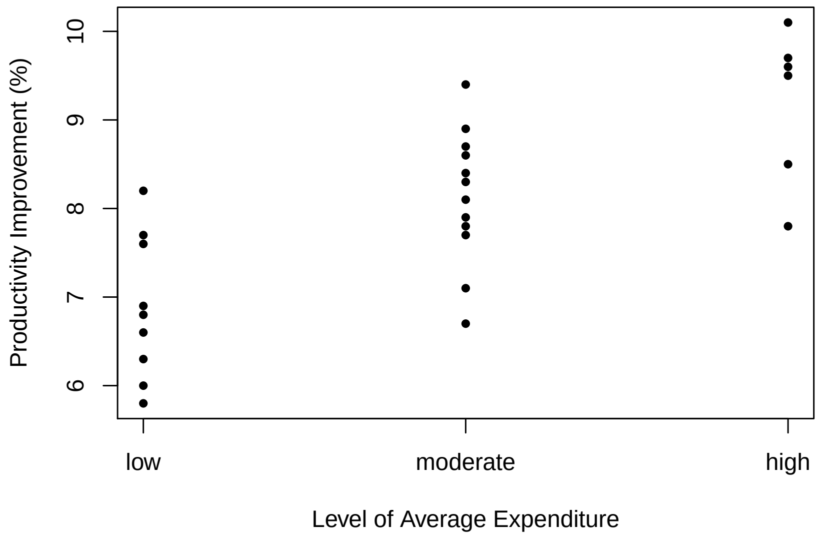


Figure 3: *Aligned dot plot of productivity improvement of firms against level of their average expenditures for research and development*

a) The level means appear to differ but the variability of the observations within each factor level appear to be approximately the same.

b) The fitted values are:

$$\hat{Y}_{1i} = \bar{Y}_{1\cdot} = 6.8778, \qquad \hat{Y}_{2i} = \bar{Y}_{2\cdot} = 8.1333, \qquad \hat{Y}_{3i} = \bar{Y}_{3\cdot} = 9.2000.$$

c) Residuals e_{ij}:

	Firm (j)											
i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0.722	1.322	-0.078	-1.078	0.022	-0.278	-0.578	0.822	-0.878			
2	-1.433	-0.033	1.267	0.467	-0.333	-0.433	0.767	-0.233	0.167	0.567	-1.033	0.267
3	-0.700	0.500	0.900	-1.400	0.400	0.300						

Yes, they sum to zero.

d) Analysis of Variance Table:

Source of Variation	**SS**	**df**	**MS**	***F***
Between Levels	$\sum n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = 20.12$	$r-1=2$	10.06	15.72
Errors	$\sum\sum(Y_{ij} - \bar{Y}_{i\cdot})^2 = 15.36$	$n_T - r = 24$	0.64	
Total	$\sum\sum(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = 35.48$	$n_T - 1 = 26$		

e) <u>Hypothesis</u>:
$H_0 : \mu_1 = \mu_2 = \mu_3$
H_a : at least one μ_i is different
at 5% significance level.
<u>Test Statistics</u>:

$$F^* = \frac{MSTR}{MSE} = \frac{10.08}{0.64} = 15.72$$

<u>Decision Rule:</u> If $F^* \geq F_{0.95,2,24} = 3.4028$, reject H_0, otherwise fail to reject H_0.
<u>Conclusion</u>: Reject H_0 and conclude that the mean productivity improvement differs according to the level of research and development expenditures at a significant level of 5%.

f) The $P-$value is $P(F > 15.72) = 4.33 \times 10^{-5}$.

g) It appears that the higher the research and development expenditure, the higher the expected productivity improvement.

4 Question 16.8

a) The level means appear to be the same but the variability of the observations within each factor level appear to differ.

b) The fitted values are:

$$\hat{Y}_{1i} = \bar{Y}_{1\cdot} = 29.4, \qquad \hat{Y}_{2i} = \bar{Y}_{2\cdot} = 29.6, \qquad \hat{Y}_{3i} = \bar{Y}_{3\cdot} = 28.$$

c) Residuals e_{ij}:

	j					
i	**1**	**2**	**3**	**4**	**5**	Total
Blue	-1.4	-3.4	1.6	-2.4	5.6	0
Green	4.4	-0.6	-4.6	1.4	-0.6	0
Orange	3	-3	-1	1	0	0

d) Analysis of Variance Table

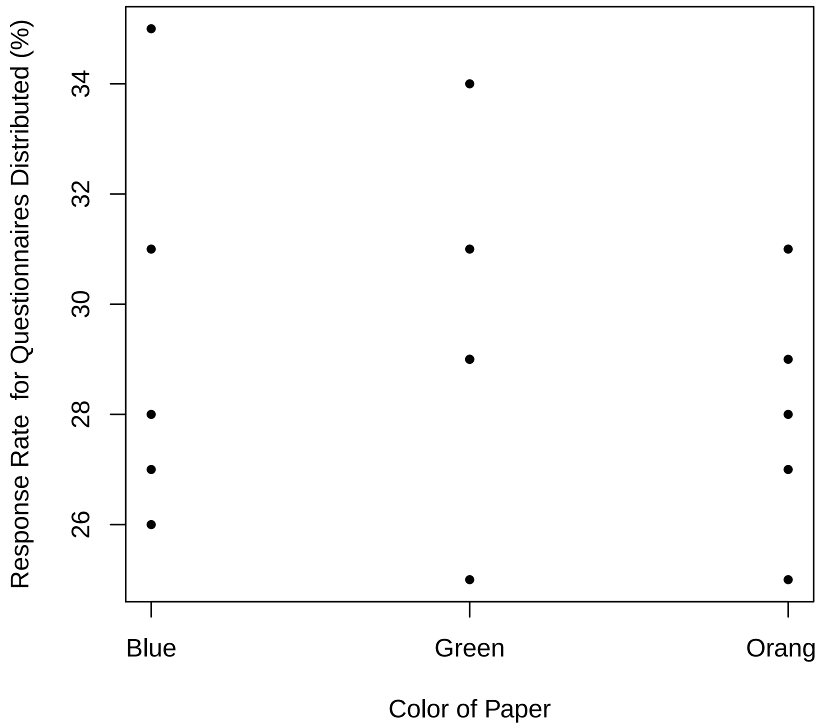


Figure 4: *Aligned dot plot of response rates for questionnaires distributed by the "windshield method" in supermarket parking lots.*

Source of Variation	**SS**	**df**	**MS**	***F***
Between Levels	7.6	2	3.8	0.392
Errors	116.4	12	9.7	
Total	124	14		

e) Hypothesis:
$H_0 : \mu_1 = \mu_2 = \mu_3$
H_a : at least one μ_i is different, $i = 1, 2, 3$
at 10% significance level.
Test Statistics:

$$F^* = \frac{MSTR}{MSE} = \frac{3.8}{9.7} = 0.392$$

Decision Rule: If $F^* \geq F_{0.9,2,12} = 2.8068$, reject H_0, otherwise fail to reject H_0.
Conclusion: Fail to reject H_0 and conclude that the mean response rates are not different at a significant level of 10%.

The $p-$value of the test is 68.4%.

f) No, this conclusion does not follow from the findings of the study because, plain white paper wasn't included in the study.

5 Question 16.18

a)

$$\mathbf{Y} = \begin{bmatrix} 7.6 \\ 8.2 \\ 6.8 \\ 5.8 \\ 6.9 \\ 6.6 \\ 6.3 \\ 7.7 \\ 6.0 \\ 6.7 \\ 8.1 \\ 9.4 \\ 8.6 \\ 7.8 \\ 7.7 \\ 8.9 \\ 7.9 \\ 8.3 \\ 8.7 \\ 7.1 \\ 8.4 \\ 8.5 \\ 9.7 \\ 10.1 \\ 7.8 \\ 9.6 \\ 9.5 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \mu. \\ \tau_1 \\ \tau_2 \end{bmatrix}$$

b)

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mu. + \tau_1 \\ \mu. + \tau_1 \\ \mu. + \tau_1 \\ \mu. + \tau_1 \\ \mu. + \tau_1 \\ \mu. + \tau_1 \\ \mu. + \tau_1 \\ \mu. + \tau_1 \\ \mu. + \tau_1 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. - \tau_1 - \tau_2 \\ \mu. - \tau_1 - \tau_2 \\ \mu. - \tau_1 - \tau_2 \\ \mu. - \tau_1 - \tau_2 \\ \mu. - \tau_1 - \tau_2 \\ \mu. - \tau_1 - \tau_2 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_3 \\ \mu_3 \\ \mu_3 \\ \mu_3 \\ \mu_3 \\ \mu_3 \end{bmatrix}$$

c) The fitted regression function is

$$\hat{Y} = 8.07037 - 1.19259X_1 + 0.06296X_2$$

The intercept term is estimate of $\mu.$, the unweighted average of all factor level means.

d) Analysis of Variance Table

Source of Variation	**SS**	**df**	**MS**	***F***
Regression	20.125	2	10.0626	15.72
Errors	15.362	24	0.6401	
Total	35.487	26		

e) Hypothesis:
$H_0 : \tau_1 = \tau_2 = 0$
$H_a : \tau_1 \neq 0$ or $\tau_2 \neq 0$
at 5% significance level.
Test Statistics:

$$F^* = \frac{MSR}{MSE} = \frac{10.0626}{0.6401} = 15.72$$

According to Decision Rule: Since $F^* \geq F_{0.95,2,24} = 3.4028$, reject H_0.

6 Question 16.34

a) $\Delta/\sigma = 0.15/0.15 = 1$, $\alpha = 0.05$, $1 - \beta = 0.7$, $r = 6$.
From the table, the required sample sizes is 22.

b)

$$\mu. = \frac{\Sigma\mu_i}{r} = \frac{0.09 + 0.18 + 0.30 + 0.20 + 0.10 + 0.20}{6} = 0.0034$$

$$\phi = \frac{1}{\sigma}\sqrt{\frac{n}{r}\sum(\mu_i - \mu.)^2} = \frac{1}{0.15}\sqrt{\frac{22}{6}(0.02968)} = 2.1994$$

The power of the test is

$$P\{F^* > F(0.95; 5, 126)|\phi = 2.1994\} \geq 0.97$$

c)

$$\frac{\lambda\sqrt{n}}{\sigma} = \frac{0.10\sqrt{n}}{0.15} = 3.1591 \quad \Longrightarrow \quad n \geq \left(\frac{3.1591 \times 0.15}{0.1}\right)^2 = 22.45$$

Take $n = 23$

7 Question 17.2

The trainee is "data snooping". One should not test H_o that are supported by data, because this may violate the standard assumptions.

8 Question 17.3

a) (i) $c_1 = 1$, $c_2 = 3$, $c_3 = -4$ and $\sum c_i = 0$, therefore it is a contrast.
(ii) $c_1 = 0.3$, $c_2 = 0.5$, $c_3 = 0.1$, $c_4 = 0.1$ and $\sum c_i = 1 \neq 0$, therefore it is not a contrast.
(iii) $c_1 = \frac{1}{3}$, $c_2 = \frac{1}{3}$, $c_3 = \frac{1}{3}$, $c_4 = -1$ and $\sum c_i = 0$, therefore it is a contrast.

b) (i) $\hat{L} = \bar{Y}_1 + 3\bar{Y}_2 - 4\bar{Y}_3$ and $s^2\{\hat{L}\} = MSE\left(\frac{1^2+3^2+(-4)^2}{n}\right) = MSE\left(\frac{26}{n}\right)$

(ii) $\hat{L} = 0.3\bar{Y}_1 + 0.5\bar{Y}_2 + 0.1\bar{Y}_3 + 0.1\bar{Y}_4$ and $s^2\{\hat{L}\} = MSE\left(\frac{0.3^2+0.5^2+0.1^2+0.1^2}{n}\right) = MSE\left(\frac{0.36}{n}\right)$

(iii) $\hat{L} = \frac{1}{3}\bar{Y}_1 + \frac{1}{3}\bar{Y}_2 + \frac{1}{3}\bar{Y}_3 - \bar{Y}_4$ and $s^2\{\hat{L}\} = MSE\left(\frac{1/3^2+1/3^2+1/3^2+(-1)^2}{n}\right) = MSE\left(\frac{4}{3n}\right)$

9 Question 17.4

a) $r = 6$, $n_i \equiv 10$

$$T = \frac{1}{\sqrt{2}}q(1-\alpha; r, n_T - r) = \frac{1}{\sqrt{2}}q(0.9; 6, 54) = \frac{1}{\sqrt{2}}(3.7652) = 2.6624$$

$$S = \sqrt{(r-1)F_{(1-\alpha; r-1, n_T-r)}} = \sqrt{5F_{0.9;5,54}} = \sqrt{5(1.9570)} = 3.1281$$

$$B = t_{1-\frac{\alpha}{2g}, n_T-r} = t_{1-\frac{0.1}{2g}, 54}$$

g	T	S	B
2	2.6624	3.1281	2.0049
5	2.6624	3.1281	2.3974
15	2.6624	3.1281	2.8220

In this problem, for small g, Bonferroni multiplier is the smallest. Scheffé multiplier is generally large. We therefore may prefer a B multiplier when we do multiple contrast when g is reasonably small. T and S don't depend on the # of comparisons. They are designed for all pairwise comparison or contrasts. B increases in g, then each estimation is less accurate.

b)

g	S	B
2	3.1281	2.0049
5	3.1281	2.3974
15	3.1281	2.8220

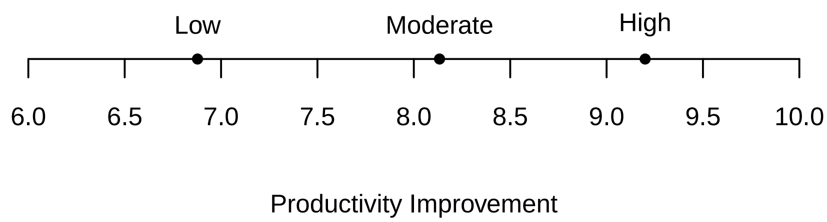
We cannot apply Tukey for general contrast. For B and S, the same finding.

10 Question 17.8

a)

b) $\hat{\mu}_3 = \bar{Y}_3 = 9.2$, $s^2\{\bar{Y}_3\} = \frac{MSE}{n_3} = \frac{0.64}{6}$

$$\bar{Y}_3 \pm t_{1-\frac{\alpha}{2}, n_T-r}\sqrt{\frac{MSE}{n_3}} \quad \Longrightarrow \quad 9.2 \pm 2.0639\sqrt{\frac{0.64}{6}} \quad \Longrightarrow \quad 8.5259 \le \mu_3 \le 9.8741$$



c) $\hat{D} = \bar{Y}_2 - \bar{Y}_1 = 8.1333 - 6.8778 = 1.2556$

$$\hat{D} \pm t_{1-\frac{\alpha}{2}, n_T - r}\sqrt{MSE\left(\frac{1}{n_2} + \frac{1}{n_1}\right)} \implies 1.2556 \pm 2.0639\sqrt{0.64\left(\frac{1}{12} + \frac{1}{9}\right)} \implies -0.1805 \leq D \leq 2.6917$$

The interval contains 0. Hence, the difference is not significant.

d) $T = \frac{1}{\sqrt{2}q(1-\alpha; r, n_T - r)} = \frac{1}{\sqrt{2}}q(0.9, 3, 24) = \frac{1}{\sqrt{\sqrt{2}}}(3.0471)$

$$\hat{D}_1 = \bar{Y}_3 - \bar{Y}_1 = 9.2000 - 6.8777 = 2.3223$$

$$\hat{D}_2 = \bar{Y}_2 - \bar{Y}_1 = 8.1333 - 6.8777 = 1.2556$$

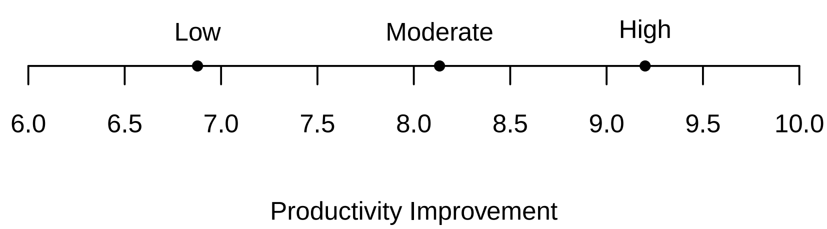
$$\hat{D}_3 = \bar{Y}_3 - \bar{Y}_2 = 9.2000 - 8.1333 = 1.0667$$

$$\hat{D}_1 \pm T\sqrt{MSE\left(\tfrac{1}{n_3} + \tfrac{1}{n_1}\right)} \implies 2.322 \pm \tfrac{1}{\sqrt{2}}(3.0471)\sqrt{0.64\left(\tfrac{1}{6} + \tfrac{1}{9}\right)} \implies 1.4137 \leq D_1 \leq 3.2310$$

$$\hat{D}_2 \pm T\sqrt{MSE\left(\tfrac{1}{n_2} + \tfrac{1}{n_1}\right)} \implies 1.2556 \pm \tfrac{1}{\sqrt{2}}(3.0471)\sqrt{0.64\left(\tfrac{1}{12} + \tfrac{1}{9}\right)} \implies 0.4955 \leq D_2 \leq 2.0156$$

$$\hat{D}_3 \pm T\sqrt{MSE\left(\tfrac{1}{n_3} + \tfrac{1}{n_2}\right)} \implies 1.066 \pm \tfrac{1}{\sqrt{2}}(3.0471)\sqrt{0.64\left(\tfrac{1}{6} + \tfrac{1}{12}\right)} \implies 0.2048 \leq D_3 \leq 1.9285$$

The difference between any two group means is significantly different at 90% family confidence level.



e) Yes, Tukey procedure is the most efficient because $t_{(0.9833,24)} = 2.257$, $F_{(0.9,2,24)} = 2.54$ and $S = 2.25$.

11 Question 17.16

a) $\hat{\mu_1} = \bar{Y}_{1\cdot} = 21.5$, $\hat{\mu_2} = \bar{Y}_{2\cdot} = 27.75$, $\hat{\mu_3} = \bar{Y}_{3\cdot} = 21.4167$, $t_{0.995,33} = 2.7333$

$$\hat{L} = (\bar{Y}_3 - \bar{Y}_2) - (\bar{Y}_2 - \bar{Y}_1) = \bar{Y}_3 - 2\bar{Y}_2 + \bar{Y}_1 = 21.4167 - 2(27.75) + 21.5 = -12.5833$$

$$s^2\{\hat{L}\} = MSE\sum_{i=1}^{r}\frac{c_i^2}{n_i} = 2.49\left(\frac{1^2 + (-2)^2 + 1^2}{12}\right) = 1.245 \quad\Longrightarrow\quad s\{\hat{L}\} = 1.1158$$

The 99% confidence interval will be

$$\hat{L} \pm t_{(1-\alpha/2;n_T-r)}s\{\hat{L}\}$$

$$-12.5833 \pm 2.7333(1.1158)$$

$$-15.6331 \leq L \leq -9.5335$$

The interval doesn't contain 0. Hence, the difference is significant.

b) For multiple comparison, according to our textbook, we may choose one from Scheffé, Bonferroni, and Tukey. However, Tukey is designed for pairwise comparison only, so it cannot be applied in this problem since the last contrast is not pairwise. In many cases, Bonferroni provides a narrower multiplier than Scheffé's. However, for this problem, Scheffé multiplier is smaller.

$$\hat{D}_1 = \bar{Y}_{2\cdot} - \bar{Y}_{1\cdot} = 27.75 - 21.5 = 6.25 \qquad \hat{D}_2 = \bar{Y}_{3\cdot} - \bar{Y}_{2\cdot} = 21.41667 - 27.5 = -6.3333$$

$$\hat{D}_3 = \bar{Y}_{3\cdot} - \bar{Y}_{1\cdot} = 21.41667 - 21.5 = -0.0833 \qquad \hat{L}_1 = \bar{D}_2 - \bar{D}_1 = 6.25 - (-6.333) = -12.5833$$

$$s^2\{\hat{D}_i\} = MSE\left(\frac{2}{n}\right) = 2.49\left(\frac{2}{12}\right) = 0.415 \quad\Longrightarrow\quad s\{\hat{D}_i\} = 0.6442,\ i = 1,2,3.$$

$$s^2\{\hat{L}_1\} = MSE\left(\frac{4}{n}\right) = 2.49\left(\frac{6}{12}\right) = 1.245 \quad\Longrightarrow\quad s\{\hat{L}_1\} = 1.1158$$

$$F_{(1-\alpha,r-1,n_T-r)} = F_{(0.9,2,33)} = 2.471 \quad\Longrightarrow\quad S^2 = (r-1)F = 2(2.471) = 4.942 \quad\Longrightarrow\quad S = 2.223$$

$$\begin{array}{lclcl}
\hat{D}_1 \pm Ss\{\hat{D}_1\} & \Longrightarrow & 6.25 \pm 2.223(0.6442) & \Longrightarrow & 4.8179 \leq D \leq 7.6821 \\
\hat{D}_2 \pm Ss\{\hat{D}_2\} & \Longrightarrow & -6.333 \pm 2.223(0.6442) & \Longrightarrow & -7.765 \leq D \leq -4.901 \\
\hat{D}_3 \pm Ss\{\hat{D}_3\} & \Longrightarrow & -0.0833 \pm 2.223(0.6442) & \Longrightarrow & -1.515 \leq D \leq 1.349 \\
\hat{L}_1 \pm Ss\{\hat{L}_1\} & \Longrightarrow & -12.5833 \pm 2.223(1.1158) & \Longrightarrow & -15.064 \leq D \leq -10.103
\end{array}$$

With 90% confidence coefficient, the third one contains 0, therefore the difference is significant. The other contrasts are not significant.