STA 207: Homework 1

Eshita Nandini

ID: 916693682

Rouyan Yin

ID: 916666619

(Proof of problem 16.41, 17.34 are attached at the end of the homework)

Problem 16.5

a)

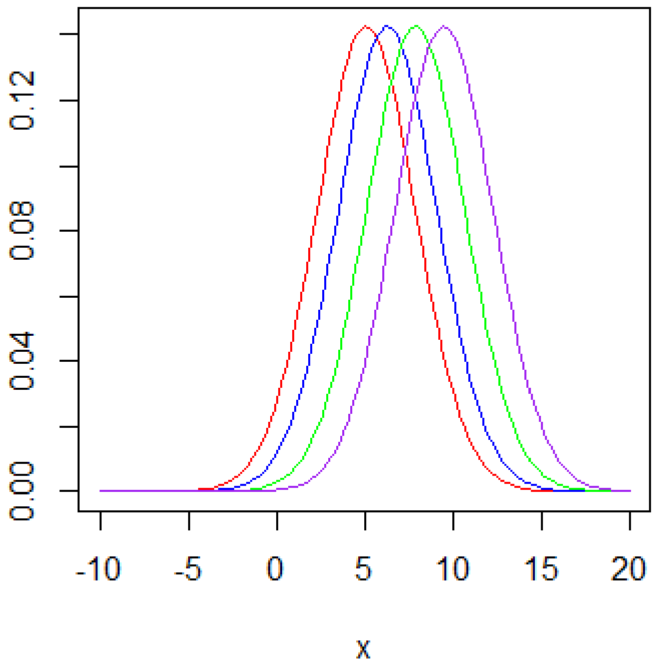


Figure 1.1

b)

E[MSTR] = 2+ni(i-)2/(r-1) = 374.5067

E[MSE]=2=7.84

The expectation of MSTR is substantially larger than that of MSE. From the result, it is reasonable to infer that the means under different levels are not all the same.

c)

The updated expectation of MSTR will be changed to:

523.1733.

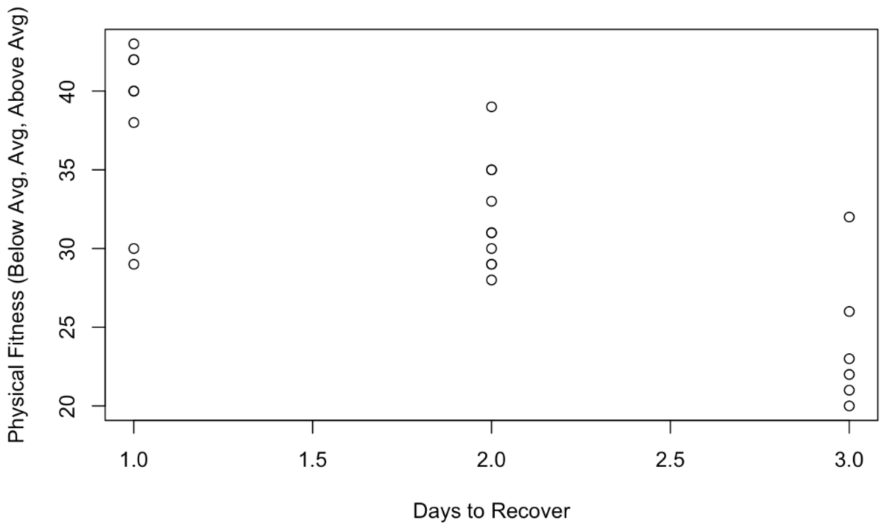
Even the range stays the same, but the data are more widely spread than it was before, which means that the means are even more obviously different than in part b. Under this case, even the range of the data does not change, E[MSTR] gets larger

Problem 16.9
R CODE

nine = read.table("CH16PR09.txt", col.names=c("days to recover", "prior physical fitness", "sample size for group"))

plot(nine$prior.physical.fitness, nine$days.to.recover, xlab = 'Days to Recover', ylab = 'Physical Fitness (Below Avg, Avg, Above Avg)')

a)



The factor level means do appear to differ, but it is hard to tell if there is variability in the observations of each factor level, which means there is probably equal variance across all factor levels.

b)
Fitted values
below_av= subset(nine, prior.physical.fitness=='1')$days.to.recover
avg = subset(nine, prior.physical.fitness=='2')$days.to.recover
abov_av = subset(nine, prior.physical.fitness=='3')$days.to.recover

fit_1 = mean(below_av)
fit_2 = mean(avg)
fit_3 = mean(abov_av)

fit_1 = 38

fit_2 = 32
fit_3 = 24

c.)
Residuals
resid_1 = below_av - fit_1
resid_2 = avg - fit_2
resid_3 = abov_av - fit_3

sum(resid_1) f
sum(resid_2)
sum(resid_3)

All of these residuals come out to zero.

d.)
aov.out = aov(days.to.recover ~ factor(prior.physical.fitness), data=nine)
summary(aov.out)

ANOVA Table

```
                              Df Sum Sq Mean Sq F value   Pr(>F)    
factor(prior.physical.fitness)  2    672   336.0   16.96 4.13e-05 ***
Residuals                      21    416    19.8                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e.) F-test
H0:$\mu 1 = \mu 2 = u3$
Ha: Not all μi's are the same
At level $\alpha = 0.01$
F* = 16.96 (from the ANOVA table)
F(1-0.01; 3-1, 24-3) = F(0.99; 2,21) = 5.78
F* > 5.78 so we reject the null hypothesis.
Conclusion: Mean number of days required for successful rehab is not the same for all 3 fitness groups.

f.) Obtain p-value
pf(16.96, 2, 21)
We get a p-value of 0.999

g.)

It appears that more active on fitness status, less time will the physical therapy take.

16.26

Based on the assumption:

$$\mu_1 = 7.0, \mu_2 = 8.0, \mu_3 = 9.0$$

$$\sigma^2 = 20.25$$

$$n_1 = 8, n_2 = 10, n_3 = 6$$

$$n_t = n_1 + n_2 + n_3 = 24$$

$$\mu = \sum n_i/n_t * \mu_i = 33.92$$

$$\text{MSTR} = \left(\sum n_i * (\mu_i - \mu)^2\right)/(r - 1) = 148.92$$

$$\text{MSE} = \left(\sum\sum (Y_i j - \mu_i)^2\right) = 29.04$$

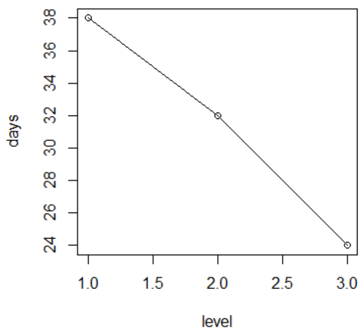
$$F^* = \text{MSTR}/\text{MSE} = 5.13$$

$$F(0.99, 2, 21) = 5.78$$

So, under this case we cannot reject the null hypothesis. We are 99% confident to say the means are the same.

17.10

a)



From the plot, we can infer that the days required for the therapy decreases with higher fitness level.

b)

The interval is:

$$\overline{Y_2}.\pm t(0.995; n_t - r) * s_2$$

$$s_2 = 2.267787$$

So, the interval should be:

$$[31.57908{,}44.42092]$$

c)

$$\boldsymbol{D_1} = \boldsymbol{\mu_2} - \boldsymbol{\mu_3}$$

$$\boldsymbol{D_2} = \boldsymbol{\mu_1} - \boldsymbol{\mu_2}$$

$$\mathbf{MSE} = \mathbf{19.80952}$$

$$\mathbf{B} = \mathbf{2.413845}$$

$$\widehat{\boldsymbol{D_1}} = \mathbf{8}$$

$$\widehat{\boldsymbol{D_2}} = \mathbf{6}$$

So, for D1, the interval is:

$$[\ 0.9039033\ 11.0960967]$$

For D2, the interval is:

[0.9039033, 11.0960967]

According to the intervals, all of them contain zero, so we cannot reject H_0, which means that $\boldsymbol{\mu_2}$ and $\boldsymbol{\mu_3}$ may equal. $\boldsymbol{\mu_1}$ may also equal to $\boldsymbol{\mu_2}$.

d)

The multiplier obtained by Tukey-Kramer method is 2.520568 which is larger than using Bonferroni method. Under this case, it is more efficient to use Bonferroni method.

e)

The multiplier will be changed if still use Bonferroni method. That is because the multiplier are related to the number of the contrasts we want to do. But it is not the case by using Tukey-Kramer method because Tukey-Kramer method is based on do all the contrasts in one time. So when using this method, we already do this estimate.

f)

By using Tukey-Kramer method, we get this table:

	diff	lwr	upr	p adj
2-1	-6	-11.32141	-0.6785856	0.0253639
3-1	-14	-20.05870	-7.9413032	0.0000254
3-2	-8	-13.79322	-2.2067778	0.0060547

From the table, we find that no interval contain zero. So it is reasonable to conclud e that all the means are not equal.

17.15

a)

$$L = \mu_1 - 2 * \mu_2 + \mu_3$$
$$s^2(L) = 13.70159$$
$$t(0.995; 21) = 2.83136$$

So the interval should be:

[-12.480463 8.480463]

This interval contain zero, which means the factor effects between adjacent level of fitness are reasonably the same.

b)

B = t(0.998, 21)= 2.731632

Parameter	Estimate	SE	Interval
D1	6	2.111195	[0.2329921 11.7670079]

D2	14	2.403701	[7.433973 20.566027]
D3	8	2.298378	[1.721677 14.278323]
L1	-2	3.701566	[-12.111317 8.111317]

From the table, we can conclude by using Bonferroni method under at least 99 percent confidence, μ_1, μ_2, μ_3 are not equal to each other. But the difference between the adjacent level is same.

c)

Since in this case, m is 4 which is considerably small. So Scheffe should not work as well as Bonferroni method.

16.41.

$$\sum_i (\mu_i - \mu_\cdot)^2 = \sum (\mu_i^2 - 2\mu_i \cdot \mu_\cdot + \mu_\cdot)^2$$
$$= \sum \mu_i^2 - 2\mu_\cdot \cdot 3\mu_\cdot + 3\mu_\cdot^2$$
$$= \sum \mu_i^2 - 3\mu_\cdot^2$$
$$= 1 + \mu_2^2 - 3 \cdot \left(\frac{1+\mu_2}{3}\right)^2 = 1 + \mu_2^2 - \frac{(1+\mu_2)^2}{3}$$
$$= \frac{2}{3}\mu_2^2 - \frac{2}{3}\mu_2 + \frac{2}{3}$$
$$= \frac{2}{3}(\mu_2^2 - \mu_2 + 1) = f(\mu_2).$$

$\therefore f'(\mu_2) = \frac{2}{3}(2\mu_2 - 1) = 0 \Rightarrow \mu_2 = \frac{1}{2}.$

$\therefore \sum_i (\mu_i - \mu_\cdot)^2$ is minimized when $\mu_2 = 0.5$.

17.34.

$$Var(\hat{L}_1) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2 = \left(\frac{1}{n} + \frac{1}{kn}\right)\sigma^2.$$
$$Var(\hat{L}_2) = \left(\frac{1}{n_3} + \frac{1}{n_2}\right)\sigma^2 = \left(\frac{1}{n} + \frac{1}{kn}\right)\sigma^2.$$

$\Rightarrow Var(\hat{L}_1) = Var(\hat{L}_2)$

$\Rightarrow$ The k minimize $Var(\hat{L}_1)$ will also minimize $Var(\hat{L}_2)$.

So just focus on $var(\hat{L}_1)$:

$$n_1 + n_2 + n_3 = (k+2)n = n_T \Rightarrow n = \frac{n_T}{k+2}.$$

$$\therefore Var(\hat{L}_1) = \left(\frac{k+2}{n_T} + \frac{k+2}{k \cdot n_T}\right)\sigma^2 = \left(\frac{k}{n_T} + \frac{2}{n_T} \cdot \frac{1}{k} + \frac{3}{n_T}\right)\sigma^2. =$$

Let $f(k) = \frac{k}{n_T} + \frac{2}{n_T} \cdot \frac{1}{k}$

$\therefore f'(k) = \frac{1}{n_T} - \frac{2}{n_T} \cdot \frac{1}{k^2} = 0 \Rightarrow k = \sqrt{2}.$

$\therefore$ The variance will be minimized when $k = \sqrt{2}$