STAT30040 Homework 5

Sarah Adilijiang

Problem 1

(a)

Analysis of Variance

Response: Survival

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Poison	1	93.161	93.161	20.813	P(F(1,44)>20.813) = 4.033e-05
Method	1	8.251	8.251	1.843	P(F(1,44)>1.843) = 0.182
Poison:Method	1	0.012	0.012	0.003	P(F(1,44)>0.003) = 0.957
Residuals	44	196.960	4.476		
Total	47	298.384			

(b)

An additive model is a model without interaction terms. Since the total degrees of freedoms and the total sum of squares does not change, thus the degree of freedoms and the sum of squares of the interaction term get added to those of the residuals: $df_{Residuals} = 44 + 1 = 45, \quad SS_{Residuals} = 196.960 + 0.012 = 196.972$

So the ANOVA table will be:

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Poison	1	93.161	93.161	21.284	P(F(1,45)>21.284) = 3.292e-05
Method	1	8.251	8.251	1.885	P(F(1,45)>1.885) = 0.177
Residuals	45	196.972	4.377		
Total	47	298.384			

Answer:

The p-values are different from the corresponding p-values in the model (a), but the differences are small.

This is because the sum of squares of interaction term "Poison:Method" is very small relative to sum of squares of the two main effect predictors and the residuals, thus when removing the interaction term, there is only a little change in the $SS_{residuals}$, hence the F values do not change much.

And the degree of the freedom of the interaction term is also small relative to the degrees of the freedom of the residuals, thus the degrees of freedom of the F distribution dose not change much.

As a result, the similar F values and F distributions with similar degrees of freedoms end up with similar p-values.

Problem 2

(a)

The four Plays are the different treatments in this case, so $I = 4$, and there are three observations for each treatment Player, so $J = 3$, and $N = I \times J = 12$

One-Way ANOVA Table:

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Player	$I-1=3$	436.2500	145.4167	6.3686	P(F(3,8)>6.3686) = 0.0163
Residuals	$N-I=8$	182.6667	22.8333		
Total	$N-1=11$	618.9167			

Since the p-value = 0.0163 < 0.05, so we reject the null hypothesis that the Player does not have an effect on the number of Home Runs. Therefore, there is a significant Player effect at 5% significance level.

```
# data
Player1=c(23,12,18); Player2=c(21,8,15); Player3=c(31,25,30); Player4=c(15,10,15)
baseball = matrix(c(Player1, Player2, Player3, Player4),3,4)
row.names(baseball) = c("Zone1", "Zone2", "Zone3")

# Sum of squares
SST = sum( (baseball-mean(baseball))^2 );    SST
```

```
## [1] 618.9167
```

```
SS_Player = 3 * sum( (colMeans(baseball)-mean(baseball))^2 );   SS_Player
```

```
## [1] 436.25
```

```
SS_resids = SST - SS_Player;    SS_resids
```

```
## [1] 182.6667
```

```
# Mean Sq
MS_Player = SS_Player/3;  MS_Player
```

```
## [1] 145.4167
```

```
MS_resids = SS_resids/8;  MS_resids
```

```
## [1] 22.83333
```

```
# F value
F_val = MS_Player/MS_resids;  F_val
```

```
## [1] 6.368613
```

```
# p value
1 - pf(F_val, 3, 8)
```

```
## [1] 0.01631226
```

(b)

The three Zones are the different treatments in this case, so $I=3$, and there are four observations for each treatment Zone, so $J=4$, and $N=I\times J=12$

One-Way ANOVA Table:

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Zone	$I-1=2$	158.1667	79.0833	1.5448	P(F(2,9)>1.5448) = 0.2650
Residuals	$N-I=9$	460.7500	51.1944		
Total	$N-1=11$	618.9167			

Since the p-value = 0.2650 > 0.05, so we do not reject the null hypothesis that the Zone does not have an effect on the number of Home Runs. Therefore, there is not a significant Zone effect at 5% significance level.

```r
# Sum of squares
SS_Zone = 4 * sum( (rowMeans(baseball)-mean(baseball))^2 );   SS_Zone
```

```
## [1] 158.1667
```

```r
SS_resids = SST - SS_Zone;    SS_resids
```

```
## [1] 460.75
```

```r
# Mean Sq
MS_Zone = SS_Zone/2;  MS_Zone
```

```
## [1] 79.08333
```

```r
MS_resids = SS_resids/9;  MS_resids
```

```
## [1] 51.19444
```

```r
# F value
F_val = MS_Zone/MS_resids;  F_val
```

```
## [1] 1.544764
```

```r
# p value
1 - pf(F_val, 2, 9)
```

```
## [1] 0.2650021
```

(c)

In the additive model(without interactions), there are two factors: the three Zones, i.e. $I = 3$, and the four Players, i.e. $J = 4$, and $N = I \times J = 12$

Two-Way ANOVA Table:

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Zone	$I - 1 = 2$	158.1667	79.0833	19.3674	P(F(2,6)>19.3674) = 0.0024
Player	$J - 1 = 3$	436.2500	145.4167	35.6122	P(F(3,6)>35.6122) = 0.0003
Residuals	$(I - 1)(J - 1) = 6$	24.5000	4.08333		
Total	$N - 1 = 11$	618.9167			

Since both p-values are smaller than 0.05, therefore, there is a significant Player effect and also a significant Zone effect at 5% significance level.

```r
# Sum of squares
SS_resids = SST - SS_Player - SS_Zone;    SS_resids
```

```
## [1] 24.5
```

```r
# Mean Sq
MS_resids = SS_resids/6;  MS_resids
```

```
## [1] 4.083333
```

```r
# F values
F_Zone   = MS_Zone/MS_resids;  F_Zone
```

```
## [1] 19.36735
```

```r
F_Player = MS_Player/MS_resids;  F_Player
```

```
## [1] 35.61224
```

```
# p values
1 - pf(F_Zone, 2, 6)
```

```
## [1] 0.002412795
```

```
1 - pf(F_Player, 3, 6)
```

```
## [1] 0.0003222602
```

(d)

In the two one-way ANOVA analyses, the Player effect is significant but the Zone effect is not significant both at 5% significance level. However, in the two-way ANOVA analysis, both the Player effect and the Zone effect are significant at 5% significance level.

This is because in one-way ANOVA analyses, $SS_{Residuals} = SS_{Total} - SS_{Player}$ (or $SS_{Residuals} = SS_{Total} - SS_{Zone}$). However, in the two-way ANOVA analysis, $SS_{Residuals} = SS_{Total} - SS_{Player} - SS_{Zone}$, which is much smaller than the $SS_{Residuals}$ in one-way ANOVA analyses. Though the $df_{Residuals}$ in two-way ANOVA is also smaller than that in one-way ANOVA, the difference is much smaller than the difference in $SS_{Residuals}$, thus the $MS_{Residuals}$ is much smaller in the two-way ANOVA, resulting in much larger F values and smaller p-values. Therefore, the Zone effect becomes significant in the two-way ANOVA analysis.

(e)

We cannot use a two-way analysis of variance to test the interaction effect of Player and Zone, because there is no replications in each combination of Player and Zone, i.e. $K = 1$.

When there is more than one observations in each combination, the df of the interaction term will be $(I-1)(J-1)$, and the df of the residuals will be $IJ(K-1)$. When there is no replications, the df of the residuals will become zero, which is used up by the interaction term, thus the model will be saturated. Therefore, we will not be able to estimate the interactions between factors without replications in each combination.

Problem 3

(a)

Set the height of a mothers is $X > 0$, and the height of a daughter is $Y > 0$, which are both in centimeters (cm).

Thus we have: $[\log(X)\ \ \log(Y)]^T \sim N(\mu_X = \log(160),\ \mu_Y = \log(165),\ \sigma_X = \sigma_Y = 0.05,\ \rho = 0.5)$

$\Rightarrow$
$Cov(\log(X), \log(Y)) = \rho * \sigma_X * \sigma_Y = 0.5 \times 0.05 \times 0.05 = 0.00125$
$E[\log(X) - \log(Y)] = \log(160) - \log(165) = \log(\frac{160}{165})$
$Var[\log(X) - \log(Y)] = \sigma_X^2 + \sigma_Y^2 - 2Cov(\log(X), \log(Y)) = 0.05^2 + 0.05^2 - 2 \times 0.00125 = 0.0025 = 0.05^2$

$\Rightarrow \quad \log(X) - \log(Y) \sim N(\log(\frac{160}{165}),\ 0.05^2)$

$\Rightarrow \quad P(X > Y) = P(\log(X) - \log(Y) > 0) = P(Z > \frac{0 - \log(\frac{160}{165})}{0.05}) \approx P(Z > 0.6154) = 1 - \Phi(0.6154) \approx 0.2691$

(b)

$P(X \geq 0.9Y) = P(\log(X) - \log(Y) \geq \log(0.9)) = P(Z \geq \frac{\log(0.9) - \log(\frac{160}{165})}{0.05}) \approx P(Z \geq -1.4918) = 1 - \Phi(-1.4918) \approx 0.9321$

(c)

Set $X^* = X/2.54$ and $Y^* = Y/2.54$, thus $\log(X^*) = log(X) - \log(2.54)$ and $\log(Y^*) = log(Y) - \log(2.54)$

$\Rightarrow$
$\mu_{X^*} = E(\log(X^*)) = \log(160) - \log(2.54) \approx 4.1430, \quad \mu_{Y^*} = E(\log(Y^*)) = \log(165) - \log(2.54) \approx 4.1738$
$\sigma_{X^*} = \sqrt{Var(\log(X^*))} = \sqrt{Var(\log(X))} = 0.05, \quad \sigma_{Y^*} = \sqrt{Var(\log(Y^*))} = \sqrt{Var(\log(Y))} = 0.05$
$\rho^* = \frac{Cov(\log(X^*),\log(Y^*))}{\sigma_{X^*}\sigma_{Y^*}} = \frac{Cov(\log(X),\log(Y))}{\sigma_X\sigma_Y} = \rho = 0.5$

$\Rightarrow \quad [\log(X^*) \ \log(Y^*)]^T \sim N(\mu_{X^*} = 4.1430, \mu_{Y^*} = 4.1738, \sigma_{X^*} = \sigma_{Y^*} = 0.05, \rho^* = 0.5)$

$\Rightarrow \quad f(\log(x^*), \log(y^*))$

$$= \frac{1}{2\pi\sigma_{X^*}\sigma_{Y^*}\sqrt{1-\rho^{*2}}} \exp\{-\frac{1}{2(1-\rho^{*2})}[(\frac{\log(x^*)-\mu_{X^*}}{\sigma_{X^*}})^2+(\frac{\log(y^*)-\mu_{Y^*}}{\sigma_{Y^*}})^2-2\rho^*(\frac{\log(x^*)-\mu_{X^*}}{\sigma_{X^*}})(\frac{\log(y^*)-\mu_{Y^*}}{\sigma_{Y^*}})]\}$$

$$\approx \frac{1}{2\pi 0.05^2\sqrt{1-0.5^2}} \exp\{-\frac{1}{2(1-0.5^2)}[(\frac{\log(x^*)-4.1430}{0.05})^2+(\frac{\log(y^*)-4.1738}{0.05})^2-2\times 0.5(\frac{\log(x^*)-4.1430}{0.05})(\frac{\log(y^*)-4.1738}{0.05}$$

$$\approx \frac{1}{0.0043\pi} \exp\{-\frac{1}{1.5}[(\frac{\log(x^*)-4.1430}{0.05})^2 + (\frac{\log(y^*)-4.1738}{0.05})^2 - (\frac{\log(x^*)-4.1430}{0.05})(\frac{\log(y^*)-4.1738}{0.05})]\}$$

(d)

Set $Z = X/Y > 0$ and $U = \log(Z)$

So we have: $U = \log(Z) = \log(X) - \log(Y) \sim N(\mu_U = \log(\frac{160}{165}), \sigma_U^2 = 0.05^2)$

$$\Rightarrow \quad f_Z(z) = f_U(\log(z)) \ |\frac{d}{dz}\log(z)| = \frac{1}{\sqrt{2\pi\sigma_U^2}} \exp\{-\frac{(\log(z)-\mu_U)^2}{2\sigma_U^2}\} \ |\frac{1}{z}|$$

$$= \frac{1}{z\,\sigma_U\sqrt{2\pi}} \exp\{-\frac{(\log(z)-\mu_U)^2}{2\sigma_U^2}\} = \frac{1}{0.05z\sqrt{2\pi}} \exp\{-\frac{1}{2}(\frac{\log(z)-\log(\frac{160}{165})}{0.05})^2\} \quad (z>0)$$

Problem 4

Here we want to get the probability: $P(Y \geq X + 10) = P(\log(Y) \geq \log(X + 10))$, for which it is not possible to get an exact answer in terms of any standard special funcitons.

Therefore, I used bootstrapping method to simulate the data from the bivariate normal distributions of the logarithms of the mother's and daughter's heights (in centimeters): $[\log(X) \ \log(Y)]^T \sim N(\mu_X = \log(160), \mu_Y = \log(165), \sigma_X = \sigma_Y = 0.05, \rho = 0.5)$, and then get the probabily that $P(Y \geq X + 10)$ in these simulations.

Finally, I get the approximate answer that: $P(Y \geq X + 10) \approx 26.97\%$

```
set.seed(123)
library(MASS)
prob = NULL
for (i in 1:1000) {
    mu = c(log(160), log(165))
    cov_matrix = matrix(c(0.05^2,0.5*0.05^2,0.5*0.05^2,0.05^2),2,2)
    bvn = mvrnorm(n=100, mu, cov_matrix)
    log_x = bvn[,1]
    log_y = bvn[,2]
    x = exp(log_x)
    y = exp(log_y)
    prob[i] = mean( y >= (x+10) )
}
mean(prob)
```

```
## [1] 0.26967
```

Problem 5

(a)

Since the four treatment groups are independent and have equal variance σ^2, we should carry out pooled two-sample t-tests with equal variances.

```
mydata = read.table("./data/multitest-Stat2.txt")
data = as.matrix(mydata)

# pooled two-sample t-tests (equal variances)
p_vals = matrix(NA,3,4)
for (i in 1:3) {
    for (j in (i+1):4) {
        p_vals[i,j] = t.test(data[,i], data[,j], mu=0, paired=FALSE, var.equal=TRUE)$p.value
    }
}
p_vals
```

```
##      [,1]      [,2]       [,3]        [,4]
## [1,]   NA 0.2709039 0.06382406 0.001472481
## [2,]   NA        NA 0.41130054 0.026832192
## [3,]   NA        NA         NA 0.170659222
```

Answer:

There are three p-values smaller than 0.1: p_{13}, p_{14}, p_{24}, so the following three hypotheses are rejected at test level $\alpha = 0.1$:

$$H_{13} : \mu_1 = \mu_3 \quad , \quad H_{14} : \mu_1 = \mu_4 \quad , \quad H_{24} : \mu_2 = \mu_4$$

(b)

Using the Bonferroni-adjusted test level: $\alpha/\binom{4}{2} = 0.1/6 \approx 0.0167$, there will be only one p-value in question (a) smaller than 0.0167: p_{14}, so its corresponding hypothesis $H_{14} : \mu_1 = \mu_4$ is rejected at Bonferroni-adjusted test level 0.0167.

(c)

```
# test statistics Tij's
t_stats = matrix(NA,3,4)
for (i in 1:3) {
    for (j in (i+1):4) {
        t_stats[i,j] = t.test(data[,i], data[,j], mu=0, paired=FALSE, var.equal=TRUE)$statistic
    }
}
t_stats
```

```
##      [,1]      [,2]       [,3]      [,4]
## [1,]   NA -1.117245 -1.9090849 -3.428873
## [2,]   NA        NA -0.8307567 -2.303223
## [3,]   NA        NA         NA -1.396530
```

```
# max|Tij|
abs(t_stats[which.max(abs(t_stats))])
```

```
## [1] 3.428873
```

Answer:

In question (a), for a hypothesis $H_{ij} : \mu_i = \mu_j$ $(1 \le i < j \le 4)$, the test statistics under H_{ij} is :

$$T_{ij} = \frac{(\bar{X}_i - \bar{X}_j) - (\mu_{X_i} - \mu_{X_j})}{s_{p_{ij}}\sqrt{\frac{1}{n} + \frac{1}{n}}} = \frac{\bar{X}_i - \bar{X}_j}{s_{p_{ij}}\sqrt{\frac{2}{n}}} \sim t(2n-2) = t(38)$$

where $n = 20$ is the sample size in each treatment group, and $s^2_{p_{ij}} = \frac{(n-1)s^2_{X_i} + (n-1)s^2_{X_j}}{2n-2} = \frac{s^2_{X_i} + s^2_{X_j}}{2}$ is the pooled sample variance of the two groups.

And in this question, we get that the observed test statistic for our data is: $T_{observed} = \max |T_{ij}| = 3.428873$

Thus the p-value is:

$$p = P(\max |T_{ij}| > 3.428873 | H_0) \le \sum_{1 \le i < j \le 4} P(|T_{ij}| > 3.428873 | H_0) = 6 \times 2 \times P(t(38) > 3.428873) \approx 0.0088$$

```
6 * 2 * pt(3.428873, 38, lower.tail=FALSE)
```

```
## [1] 0.008834885
```

So a simple upper bound on the p-value of the test based on T is 0.0088.

(d)

The upper bound in question (c) is 0.0088, which is smaller than 0.1, so we reject the null hypothesis H_0 at the test level $\alpha = 0.1$.

(e)

```
# construct data
x = as.factor(rep(1:4, c(20,20,20,20)))
y = as.vector(data)

# One-way ANOVA F-test
anova(lm(y~x))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x          3 11.261  3.7536  4.0547 0.009947 **
## Residuals 76 70.357  0.9257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since $I = 4$, $J = 20$, $N = I \times J = 80$

So the One-Way ANOVA Table is:

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	$I - 1 = 3$	11.261	3.7536	4.0547	P(F(3,76)>4.0547) = 0.009947
Residuals	$N - I = 76$	70.357	0.9257		
Total	$N - 1 = 79$	81.618			

Since the p-value $= 0.009947 < 0.1$, so we reject the null hypothesis H_0 at the test level $\alpha = 0.1$.