Assignment 1

Ayato Tanemura (atan524)

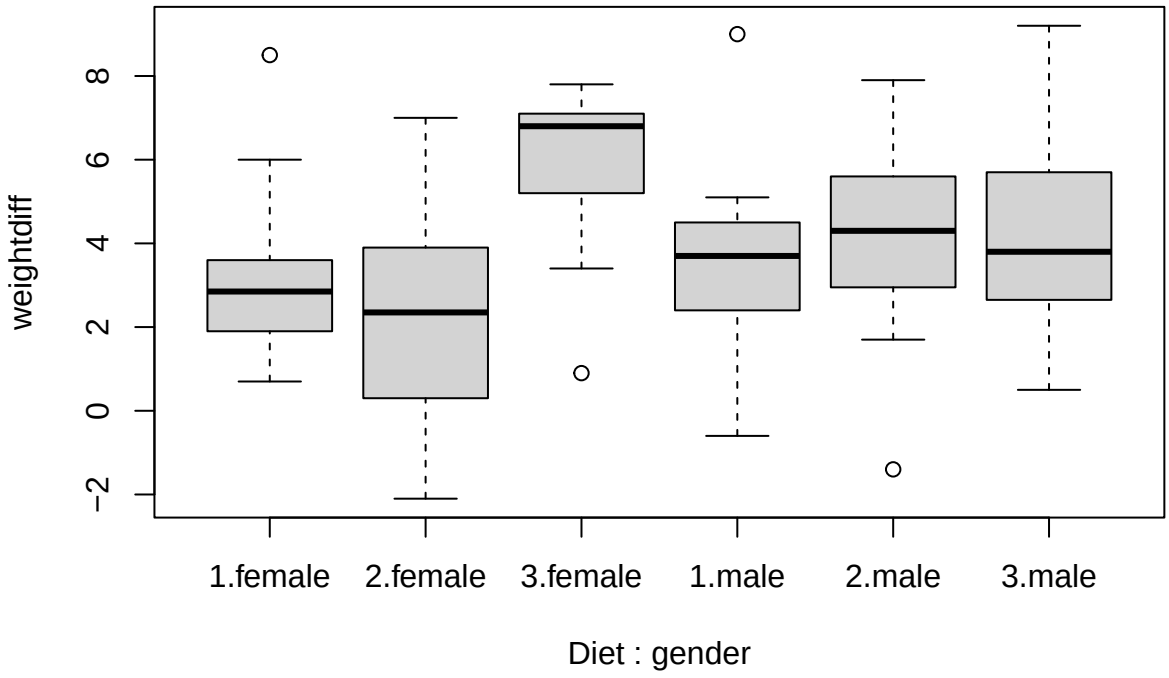
Contents

Question 1

(a) Essential assumption check

Normality Distribution

Normality plot:



- weightdiff shows the difference between before and after six weeks.
- The formula is `weightdiff = (weight before diet) - (weight 6 weeks after)`

Outliers check: 4 outliers are identified via this plot.

There are 1 (Diet1, Male), 1 (Diet1, Female), 1 (Diet2, Male), and 1 (Diet3, Female) observations should be omitted.

Therefore, 4 observations will be deleted as outliers.

SW test:

```
>
>   Shapiro-Wilk normality test
>
> data:  weightdiff.NoOut
> W = 0.99234, p-value = 0.9383
```

We want to see non-significant result.

P-value for weight difference test of SW is 0.938 which is more than 0.05. H0 is not rejected and we conclude that the assumption for normality is satisfied.

Homogeneity of Variance

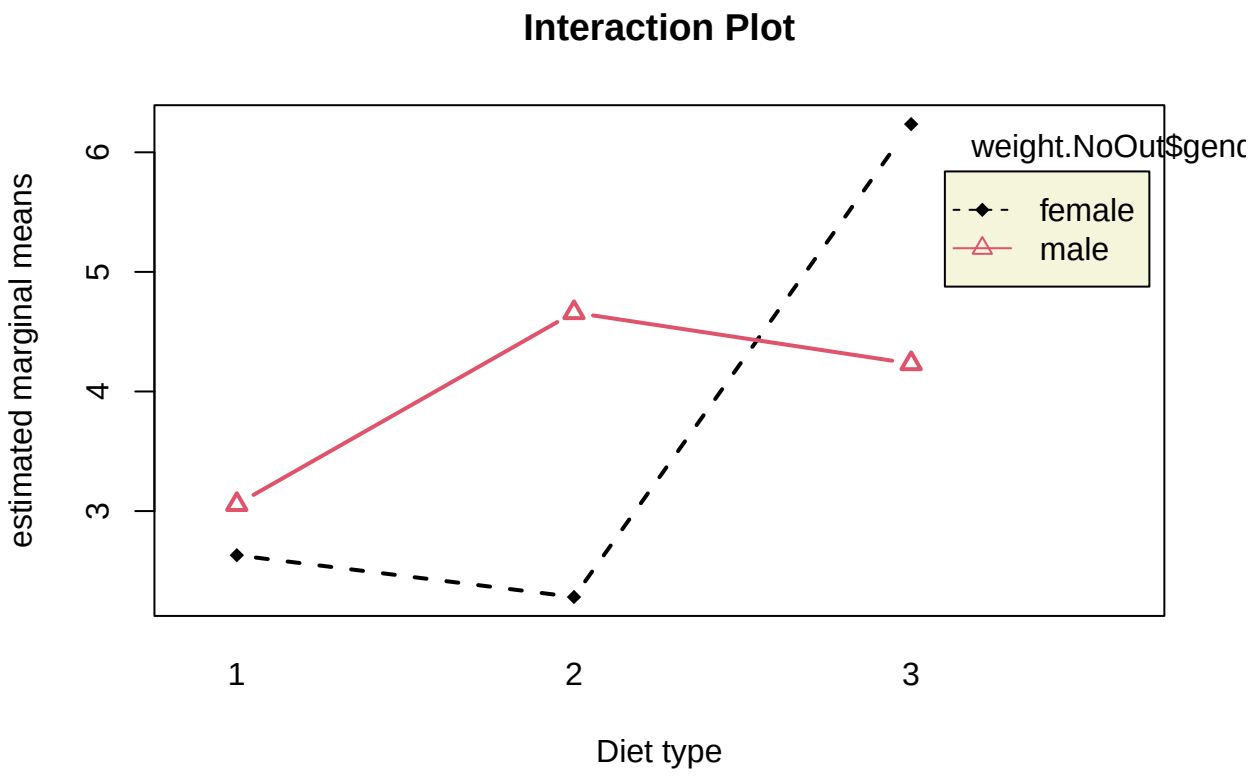
Levenne's test

```
> Levene's Test for Homogeneity of Variance (center = median)
>       Df F value Pr(>F)
> group  5  1.5479 0.1867
>       68
```

The p-value is more than 0.05 (level of significance), which means the result is non-significant.

Therefore, Homogeneity is met.

(b) Interaction Plot



(c) Difference between Male Diet 1 and 2

ANOVA summary:

```
>             Df Sum Sq Mean Sq F value   Pr(>F)
```

```
> gender       1   2.27    2.27   0.582 0.448008
> Diet         2  89.25   44.63  11.474 5.09e-05 ***
> gender:Diet  2  60.81   30.41   7.818 0.000879 ***
> Residuals   68 264.48    3.89
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey test:

```
>
>   Pairwise comparisons using Tamhane's T2-test for unequal variances

> data: weightdiff.NoOut and as.factor(weight.NoOut$gender):as.factor(weight.NoOut$Diet)

> alternative hypothesis: two.sided

> P value adjustment method: T2 (Sidak)

> H0

>                          t value   Pr(>|t|)
> female:2 - female:1 == 0  -0.503  0.9999995
> female:3 - female:1 == 0   6.826 6.0304e-06 ***
> male:1 - female:1 == 0     0.593  0.9999958
> male:2 - female:1 == 0     2.905  0.1418846
> male:3 - female:1 == 0     1.832  0.7380165
> female:3 - female:2 == 0   5.818 7.3056e-05 ***
> male:1 - female:2 == 0     0.928  0.9988759
> male:2 - female:2 == 0     2.904  0.1151982
> male:3 - female:2 == 0     2.005  0.5901037
> male:1 - female:3 == 0    -4.539  0.0073939  **
> male:2 - female:3 == 0    -2.308  0.4148369
> male:3 - female:3 == 0    -2.322  0.4064584
> male:2 - male:1 == 0       1.918  0.6752379
> male:3 - male:1 == 0       1.192  0.9861627
> male:3 - male:2 == 0      -0.437  0.9999999

> ---

> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As you can see the Tamhane test result above, the p-value for male:1 - male:2 == 0 (7th row in gender:Diet table) is 0.9988759 which is larger than 0.05 significance level. The hull hypothesis is not rejected. Null hypothesis in this case is Male Diet 1 and Male Diet 2 has no difference.

Therefore, there is no significant difference in the weight loss of males between diet 1 and diet 2.

(d) Highest Efficiency

As you can see from the interaction plot, Female Diet 3 has the largest impact on weight difference. It is regarding both gender and diet type.

As can bee seen the difference between before and after six weeks, the larger difference of weight for Male Diet 1 and 2 while it has larger impact on weight difference for Female Diet 3.

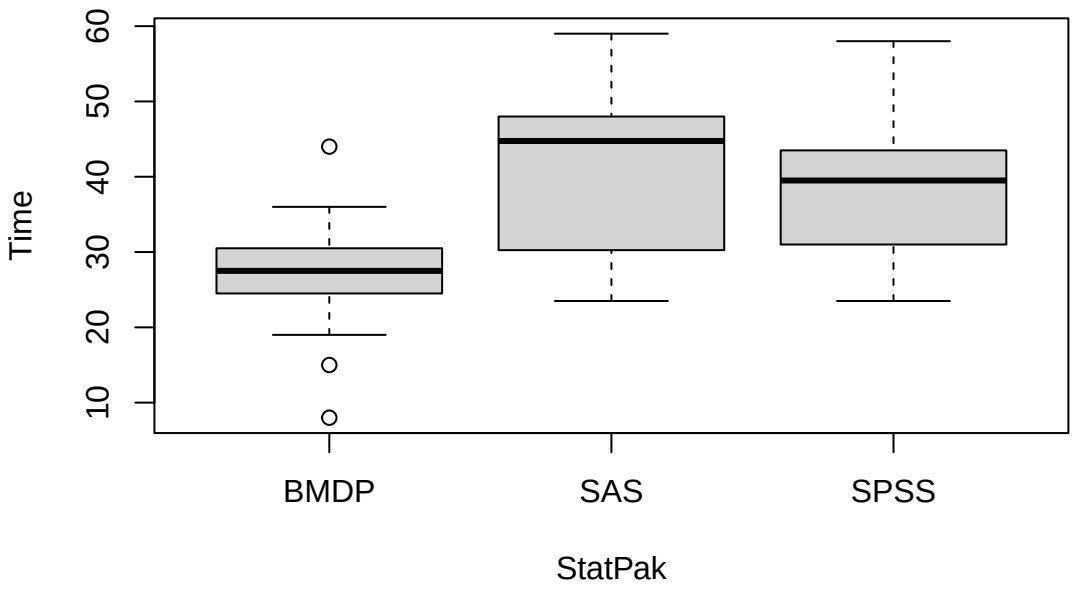
Question 2

```
df.stat <- read.csv("Data/STATPAK.csv") %>%
  mutate(StatPak = factor(StatPak))
```

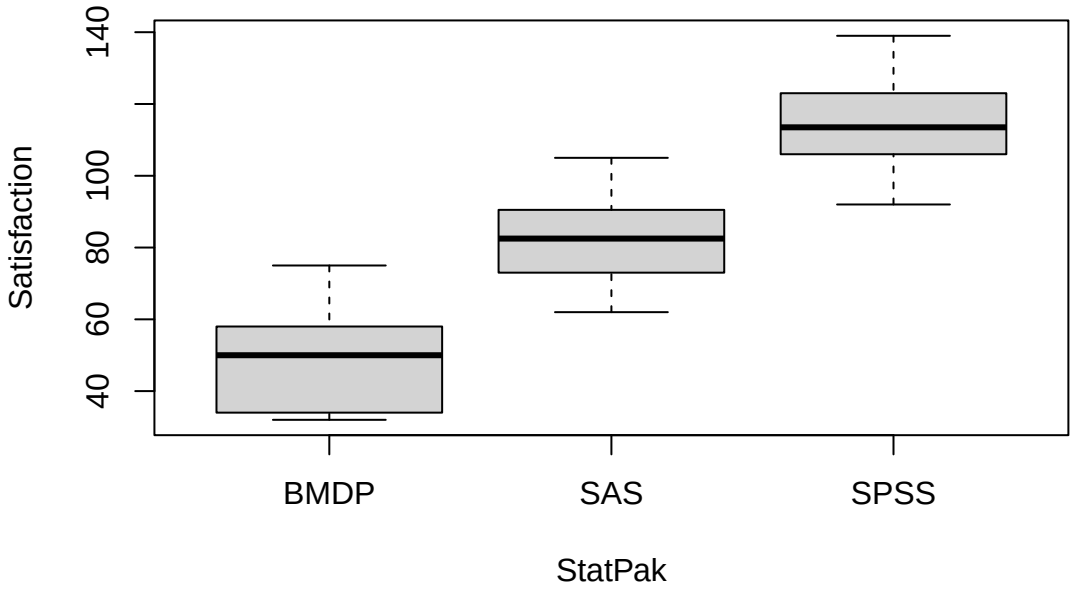
(a) Outliers & Homogeneity

Outliers

```
# Outliers Detestion
olTime <- boxplot(Time ~ StatPak, data = df.stat)$out
```




```
olSatisfaction <- boxplot(Satisfaction ~ StatPak, data = df.stat)$out
```




```
olTime
```

```
> [1] 15 44  8
```

```
olSatisfaction
```

```
> numeric(0)
```

```
out <- df.stat[df.stat$StatPak == "BMDP" & (df.stat$Time == 15 | df.stat$Time == 44 | df.stat$Time == 8
out
```

```
>    No StatPak Platform Experience Comp Time Satisfaction
> 35 35    BMDP  Windows         12   29   15           58
> 36 36    BMDP      Mac          0   88   44           44
> 65 65    BMDP  Windows         10   38    8           58
```

```
# Remove Outliers
df.NoOutlier <- df.stat[-which(df.stat$No %in% out$No),] %>%
  mutate(Satisfaction = as.numeric(Satisfaction))
```

SW test

```
>
>   Shapiro-Wilk normality test
>
> data:  df.NoOutlier$Time
> W = 0.95591, p-value = 0.008155
```

P-value for Time of SW is 0.008 which is less than 0.01 significance level. H0 is rejected and the assumption of multivariate normality is not satisfied.

```
>
>   Shapiro-Wilk normality test
>
> data:  df.NoOutlier$Satisfaction
> W = 0.96576, p-value = 0.03225
```

P-value for Satisfaction of SW is 0.03 which is more than 0.01 significance level. H0 is not rejected and the assumption of multivariate normality is satisfied.

Homogeneity of Variance and covariance

Levene's test for Time

```
> Levene's Test for Homogeneity of Variance (center = median)
>       Df F value   Pr(>F)
> group  2  5.8648 0.004279 **
>       76
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value of Levene's for Time is $0.004 < 0.05$ -> assumption is not satisfied.

Levene's test for Satisfaction

```
> Levene's Test for Homogeneity of Variance (center = median)
>       Df F value Pr(>F)
> group  2  0.4613 0.6322
>       76
```

P-value of Levene;s for Satisfaction is $0.6322 < 0.05$ -> assumption is satisfied.

Box's M test

```
>
>   Box's M-test for Homogeneity of Covariance Matrices
>
> data:  df.NoOutlier[, 6:7]
> Chi-Sq (approx.) = 17.523, df = 6, p-value = 0.007541
```

P-value of Box test is $0.007 < 0.05$ -> Homogeneity of cov matrices assumption not met. We should use Pillai's Trace for interpretation of MANOVA results.

(b) Main effects

```
>           Df Pillai approx F num Df den Df    Pr(>F)
> StatPak    2 1.0073   38.557      4    152 < 2.2e-16 ***
> Residuals 76
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, Pillai trace is used for multivariate test. Based on the Pillai's Trce result, StatPak has a statistically significant effect on Time and Satisfaction as a group (F=38.6, p=0.000<0.01)

```
>  Response 1 :
>             Df Sum Sq Mean Sq F value    Pr(>F)
> StatPak      2 2466.7 1233.35  17.605 5.218e-07 ***
> Residuals   76 5324.3   70.06
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
>  Response 2 :
>             Df Sum Sq Mean Sq F value    Pr(>F)
> StatPak      2  54757 27378.7  183.23 < 2.2e-16 ***
> Residuals   76  11356   149.4
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Response 1 -> Time

Response 2 -> Satisfaction

There are significant relationships between both StatPak & Time and StatPak & Satisfaction. This means different stat packages has a significant effect on the results of Time and Satisfactions as an individual.

(c)?? Independent Group

Tamhane test for Time

```
>
>   Pairwise comparisons using Tamhane's T2-test for unequal variances

> data: df.NoOutlier$Time and df.NoOutlier$StatPak

> alternative hypothesis: two.sided

> P value adjustment method: T2 (Sidak)

> H0

>                  t value   Pr(>|t|)
> SAS - BMDP == 0    6.185 8.6602e-07 ***
> SPSS - BMDP == 0   5.646 3.6583e-06 ***
> SPSS - SAS == 0   -1.078    0.63613

> ---

> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tukey test for satisfaction

```
>   Tukey multiple comparisons of means
>     95% family-wise confidence level
>
> Fit: aov(formula = Satisfaction ~ StatPak, data = df.NoOutlier)
>
> $StatPak
>               diff      lwr      upr p adj
> SAS-BMDP  33.75466 25.53156 41.97776     0
> SPSS-BMDP 65.79037 57.56727 74.01347     0
> SPSS-SAS  32.03571 24.22610 39.84532     0
```