STA8190 S2 2019: Nonparametric Statistics topic.

Assignment 2 30% (100 marks)

Due: 16/10/2019 11.55pm

Justin Ross

0050040865

Email to: Taryn.Axelsen@usq.edu.au

All analyses can be performed in either SPSS or R unless otherwise indicated. Alpha=0.05 should be used for all tests. **Clearly label each part of your work.**

Please remember that STA8190 is a postgraduate level course which requires that students demonstrate an advanced level of knowledge, skills, reasoning and problem-solving. Assessment items are your opportunity to demonstrate your full understanding of the concepts and methods covered in this topic. If you do not fully explain your answers you cannot receive full marks in this assessment.

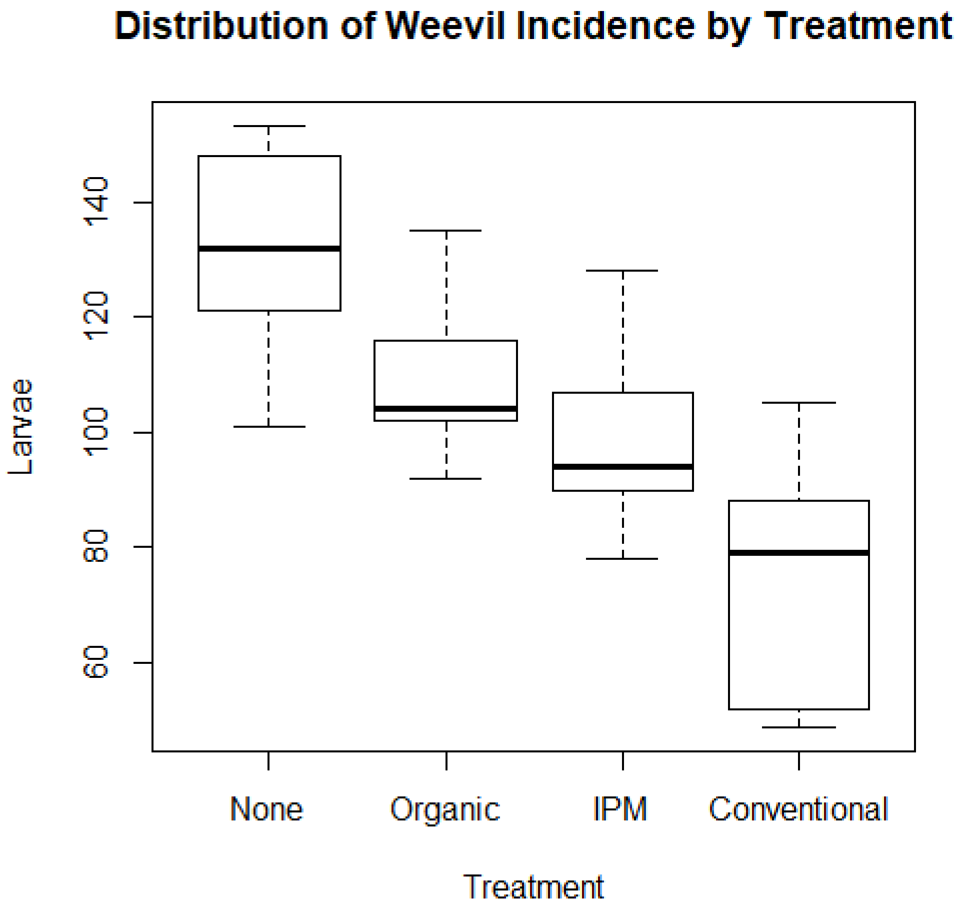
An excellent assignment
Some minor errors but
otherwise the clarity in
your explanations were
extremely well done.
Congratulations
81/100

Contents

Question 1 (30 marks):

The data file ***Weevil.txt*** contains count data of alfalfa weevil larvae incidence (Larvae) on 5 different species of alfalfa (Variety) and four different management practices (Treatment) across 20 field plots. The four Treatments were: 1=None (no proactive measures taken), 2=Organic (practices not including use of chemicals), 3=IPM (integrated pest management which includes minimal use of any available methods including chemicals) and 4=Conventional (prescribed amounts of chemicals used). The incidence of the same 5 species was recorded for each treatment.

a) Create a side-by-side box and whisker plot showing the distribution of weevil incidence in each treatment. Describe the distributions. (5 marks)



✓

Summary Statistics (B.1)

Treatment	None	Organic	IPM	Conventional
Min	101	92	78	49
1st Qu	121	102	90	52
Median	132	104	94	79
Mean	131	109.8	99.4	74.6
3rd Qu	148	116	107	88
Max	153	135	128	105
Kurtosis	-0.7693	0.6672	0.4595	-1.9811
Skewness	-0.5548	0.9265	0.7865	0.0689
Std Deviation	21.0594	16.4681	19.0473	23.9227

* Note: All values rounded down to four decimal places. All values refer to larvae incidence, but the precise way in which this unit was measured is not provided.

Description Report:

The ‘none’ treatment group has a median of 132 weevil larvae incidence and a mean of 131 weevil larvae incidence suggesting a symmetrical data set. It is slightly negatively skewed. Based on the boxplot it has quite a bit of variance compared to the other groups particularly on the lower values as the negative whisker is quite low at 101 weevil larvae incidence.

The ‘organic’ treatment group has a low median of104 weevil larvae incidence, almost reaching the first quartile value of 102 weevil larvae incidence. Its variance seems typical with its whiskers being somewhat symmetrical. The mean is comparably higher than both of these values at 109.8 weevil larvae incidence, suggesting this distribution is not as symmetrical as the former.

The ‘IPM’ treatment group has a somewhat low median of 94 weevil larvae incidence compared to its mean of 99.4 weevil larvae incidence, suggesting a lack of symmetry but otherwise seems a fairly typical distribution without notable kurtosis (-1.9811) or skewness (0.7865).

The ‘conventional’ treatment group has a median that is roughly centered in its third quartile. The presence of whiskers suggests a somewhat typical distribution, with some variance that leans positive. The most curious thing about the conventional treatment is that it the range between the first and third quartile is considerably larger than the others. This combined with kurtosis figure of -1.9811 suggests quite a strong platykurtic distribution. ✓

Wonderful!

None of the groups have any outliers, suggesting that when the treatments work they tend to work somewhat uniformly with no weevils spared.

b) Perform a Friedman test to determine if there is any difference in weevil incidence among treatments. State the hypotheses (Ho and Ha) and interpret you results. (6 marks)

Hypothesis:

The null hypothesis states that none of the treatments affect weevil incidence in the Alfalfa. The alternative hypothesis states that one or more of the treatments affect weevil incidence in the Alfalfa.

H_0: $\theta_N = \theta_O = \theta_I = \theta_C$

H_A: One or more of the treatments will affect weevil incidence in the Alfalfa.

Level of Risk:

$\alpha = 0.05$

There is a 95% chance that any observed statistical difference will be real and not due to chance.

R Output:

```
            Friedman rank sum test

data:  Larvae and Treatment and Variety

Friedman chi-squared = 15, df = 3, p-value =
0.001817
```

Interpretation:

As the p-value of 0.001817 is much lower than our critical value of $\alpha = 0.05$ we may reject the null hypothesis and accept the alternative hypothesis that one or more of the treatments will affect weevil incidence in the Alfalfa.

Since the critical value of 7.8 is less than the obtained value of 15. As such we may reject the null hypothesis that none of the treatments affect weevil incidence in the Alfalfa. The alternative hypothesis that one or more of the treatments affect weevil incidence in the Alfalfa.

Report:

$F_{r(3)} = 15$

df = 3

p-value = 0.001817

n = 5

k = 4

Critical value = 7.8

In the analysis the frequency of weevils found in different alfalfa varieties (n=5) were compared against four different treatments (k=4). Based on the analysis at least one of the treatments resulted in a difference in weevil incidence as the Friedman test was significant ($F_{R(3)} = 15$, $p < 0.05$). No follow up test was done as that was saved for part c.

Conclusion:

As the result of the Friedman test was significant it is concluded that at least one of the treatments affected the weevil incidence in the alfalfa

✓

c) If your analysis in part b) was significant use Wilcoxon signed rank tests to identify which treatments are different to each other. State the hypotheses (H0 and Ha) for the first comparison only. Interpret these results. (6 marks)

Please note that the hypothesis and other such matter are stated only for the first comparison.

Hypothesis:

The null hypothesis states that there is no difference in weevil incidence in Alfalfa's between None Treatment and Organic Treatment. The alternative hypothesis states that there is a difference in weevil incidence in Alfalfa's between None Treatment and Organic Treatment. The alternative hypothesis is a two-tailed, non-directional hypothesis because it indicates a difference but in no particular direction.

H_0: $\mu_D = 0$

H_A: $\mu_D \neq 0$

The Mann–Whitney U test / Wilcoxon rank-sum test is not the same as the Wilcoxon signed-rank test, although both are nonparametric and involve summation of ranks. The Mann–Whitney U test is applied to independent samples. The Wilcoxon signed-rank test is applied to matched or dependent samples.

−4

Level of Risk:

α = 0.05

There is a 95% chance that any observed statistical difference will be real and not due to chance.

R Output:

```
[1] 121 132 148 101 153

[1] 104 116 135  92 102

            Wilcoxon rank sum test

data:  treatmentA and treatmentB

W = 19, p-value = 0.2222

alternative hypothesis: true location shift is not
equal to 0

Sum of Positive Difference Ranks = 15

Sum of Negative Difference Ranks = 0


[1] 121 132 148 101 153

[1]  94 107 128  78  90

            Wilcoxon rank sum test
```

```
data:  treatmentA and treatmentB

W = 22, p-value = 0.05556

alternative hypothesis: true location shift is not
equal to 0

Sum of Positive Difference Ranks = 15

Sum of Negative Difference Ranks = 0


[1] 121 132 148 101 153

[1]  88  79 105  52  49

            Wilcoxon rank sum test

data:  treatmentA and treatmentB

W = 24, p-value = 0.01587

alternative hypothesis: true location shift is not
equal to 0

Sum of Positive Difference Ranks = 15

Sum of Negative Difference Ranks = 0


[1] 104 116 135  92 102

[1]  94 107 128  78  90

            Wilcoxon rank sum test

data:  treatmentA and treatmentB

W = 17, p-value = 0.4206

alternative hypothesis: true location shift is not
equal to 0

Sum of Positive Difference Ranks = 15

Sum of Negative Difference Ranks = 0


[1] 104 116 135  92 102

[1]  88  79 105  52  49

            Wilcoxon rank sum test
```

```
data:  treatmentA and treatmentB

W = 22, p-value = 0.05556

alternative hypothesis: true location shift is not
equal to 0

Sum of Positive Difference Ranks = 15

Sum of Negative Difference Ranks = 0


[1]  94 107 128  78  90

[1]  88  79 105  52  49

             Wilcoxon rank sum test

data:  treatmentA and treatmentB

W = 20, p-value = 0.1508

alternative hypothesis: true location shift is not
equal to 0

Sum of Positive Difference Ranks = 15

Sum of Negative Difference Ranks = 0
```

Interpretation:

For the first comparison between None treatment and Organic treatment the p-value was 0.2222. As it is higher than our level of risk of α = 0.05 the findings of the Wilcoxon signed rank test is not significant. We therefore must accept the null hypothesis that there is no difference in weevil incidence in Alfalfa's between None Treatment and Organic Treatment.

For subsequent comparisons the table below compares the p-value, the level of risk or alpha and determines if the findings are significant:

Treatment Comparison Table (B.2)			
Treatment Comparison	**p-value**	**alpha (α)**	**Significant**
None-Organic	0.2222	0.05	No
None-IPM	0.05556	0.05	No
None-Conventional	0.01587	0.05	Yes
Organic-IPM	0.4206	0.05	No
Organic-Conventional	0.05556	0.05	No
IPM-Conventional	0.1508	0.05	No

As the above table shows the only results of significance is the comparison between 'None' treatment and 'Conventional' treatment. As its p-value was 0.0158 (rounded down) compared to a level of risk of α = 0.05, we must accept the alternative hypothesis that there is a difference in weevil incidence in Alfalfa's between 'None' Treatment and 'Conventional' treatment.

Report:

T-statistic = 0

n = 5

P-value = 0.2222

ΣR+ = 15

ΣR- = 0

Critical Value = 0

The Wilcoxon signed rank test (T = 0, n = 5, $p > 0.05$) indicated that the alfalfa weevil larvae incidence was not significantly different between use of 'None' treatment and 'Organic' treatment. The sum of the positive difference ranks ($\Sigma R_+ = 15$) was larger than the sum of the negative difference ranks ($\Sigma R_- = 0$), showing a positive result in the 'None' treatment. In other words, there were more weevil incidence in the 'None' treatment than the 'Organic' treatment, but this was not enough to reject the null hypothesis that there is no difference in weevil incidence in Alfalfa's between None Treatment and Organic Treatment. Therefore, the analysis shows that in the comparison between no treatment and organic treatment there is no significant difference in weevil incidence.

However, while this trend extends to most of the other comparisons as shown in the above table (B.2) it does not extend to the comparison between 'None' treatment and ‘Conventional’ treatment. In that case the findings of the Wilcoxon signed rank test ($T = 0$, $n = 5$, $p < 0.05$) were significant, with a p-value of 0.0158 (rounded down) compared to the level of risk of $\alpha = 0.05$. To add further the sum of the positive difference ranks being exclusively positive shows that in all varieties of alfalfa using no treatment resulted in a greater incidence of weevil than using the conventional treatment.

Conclusion:

The analysis concludes that there is no significant difference in weevil incidence in alfalfa in any of the treatment comparisons with the exception of the comparison between no treatment and conventional treatment. In the case of the comparison between no treatment and conventional treatment, there were significantly lower levels of weevil incidence in the alfalfa when treated using the conventional treatment.

d) Show the Bonferroni correction calculation by hand. How does this correction change your results and interpretation in part c)? How does this correction affect the probability of a type-I error? (3 marks)

$$\alpha_B = \frac{\alpha}{k}$$

Where,
α_B is the adjusted level of risk
α is the original level of risk, 0.05
k is the number of comparisons, 6

$$\alpha_B = \frac{0.05}{6}$$

✓

$$\alpha_B = 0.0083\ (rounded\ down)$$

Treatment Comparison Table (B.3)

Treatment Comparison	p-value	alpha (α_B)	Significant
None-Organic	0.2222	0.0083	No
None-IPM	0.05556	0.0083	No
None-Conventional	0.01587	0.0083	No
Organic-IPM	0.4206	0.0083	No
Organic-Conventional	0.05556	0.0083	No
IPM-Conventional	0.1508	0.0083	No

As shown above when we apply the new, corrected alpha none of the results are significant. Even the previously significant comparison between 'None' and 'Conventional' treatments is no longer significant. As such the null hypothesis that there is no difference between the treatments must be accepted in all cases, contradicting previous conclusions.

This correction technique results in a lower probability of a Type 1 error, or a lower chance in rejecting the null hypothesis when it is true. Without the correction in place, at α = 0.05, the Type I error rate would equal:

$1-(1-0.05)^6$
1-0.7350
= 0.265

However, with the correction in place:

$1-(1-(0.05/6))^6$
1-0.7350
= 0.0489

It is noteworthy that the probability with the correction in place (0.0489) is roughly equal to our originally desired probability (0.05). This is the intention of the correction, to have the same 0.05 level of risk but for multiple comparisons. To achieve this for 6 comparisons an adjusted level of risk of 0.0083 (rounded down) needs to be used for reach individual comparison.

✓

e) Calculate the Friedman test by hand using equation 5.2 from the text book. Clearly define *Ri*, *n*, *k* and *C*F as part of your answer. Does this match your result in part a)? (10 marks)

Treatment Type				
Variety	**None**	**Organic**	**IPM**	**Conventional**
1	121	104	94	88
2	132	116	107	79
3	148	135	128	105
4	101	92	78	52
5	153	102	90	49

Ranks of Treatment Type				
Variety	**None**	**Organic**	**IPM**	**Conventional**
1	4	3	2	1
2	4	3	2	1
3	4	3	2	1
4	4	3	2	1
5	4	3	2	1

For treatment 1 (none),

R_N = 4 + 4 + 4 + 4 + 4 = 20

For treatment 2 (organic),

R_O = 3 + 3 + 3 + 3 + 3 = 15

For treatment 3 (IPM),

R_I = 2 + 2 + 2 + 2 + 2 = 10

For treatment 4 (Conventional),

R_C = 1 + 1 + 1 + 1 + 1 = 5

First we need to calculate C_F:

$$C_F = \left(\frac{1}{4}\right) nk\ (k+1)^2$$

Where k is the number of groups, in this case 4.
n is the number of rows, in this case 5.

$$C_F = \left(\frac{1}{4}\right) (5)(4)\ (k+1)^2$$

$$C_F = \left(\frac{1}{4}\right) (20)\ (4+1)^2$$

$$C_F = \left(\frac{1}{4}\right) (20)\ (4+1)^2$$

$$C_F = (5)\ (4+1)^2$$

$$C_F = (5)\ (5)^2$$

$$C_F = (5)(25)$$
$$C_F = 125$$ ✓

Then we need to calculate Σr^2_{ij}:

Squared Ranks of Treatment Type				
Variety	**None**	**Organic**	**IPM**	**Conventional**
1	16	9	4	1
2	16	9	4	1
3	16	9	4	1
4	16	9	4	1
5	16	9	4	1

For treatment 1 (none treatment),

R_N = 16 + 16 + 16 + 16 + 16 = 80

For treatment 2 (organic),

R_O = 9 + 9 + 9 + 9 + 9 = 45

For treatment 3 (IPM),

R_I = 4 + 4 + 4 + 4 + 4 = 20

For treatment 4 (Conventional),

R_C = 1 + 1 + 1 + 1 + 1 = 5

$$\sum r_{ij}^2 = 80 + 45 + 20 + 5$$

$$\sum r_{ij}^2 = 150$$

Now we find F_r:

$$F_r = \frac{n(k-1)\left[\sum_{i=1}^{k} \frac{R_i^2}{n} - C_F\right]}{\sum r_{ij}^2 - C_F}$$

Where,
C_F = 125
Σr^2_{ij} = 150
k = 4
n = 5
Critical Value: 7.8

$$F_r = \frac{n(k-1)\left[\sum_{i=1}^{k} \frac{R_i^2}{n} - C_F\right]}{\sum r_{ij}^2 - C_F}$$

$$F_r = \frac{5(4-1)\left[\frac{20^2}{5} + \frac{15^2}{5} + \frac{10^2}{5} + \frac{5^2}{5} - 125\right]}{150 - 125}$$

$$F_r = \frac{5(4-1)\left[\frac{400}{5} + \frac{225}{5} + \frac{100}{5} + \frac{25}{5} - 125\right]}{150 - 125}$$

$$F_r = \frac{5(4-1)[80 + 45 + 20 + 5 - 125]}{150 - 125}$$

$$F_r = \frac{5(4-1)[150 - 125]}{150 - 125}$$

$$F_r = \frac{5(4-1)[25]}{150 - 125}$$

$$F_r = \frac{5(3)[25]}{150 - 125}$$

$$F_r = \frac{375}{150 - 125}$$

$$F_r = \frac{375}{25}$$

$$F_r = \frac{375}{25}$$ ✓

06/30

$$F_r = 15$$

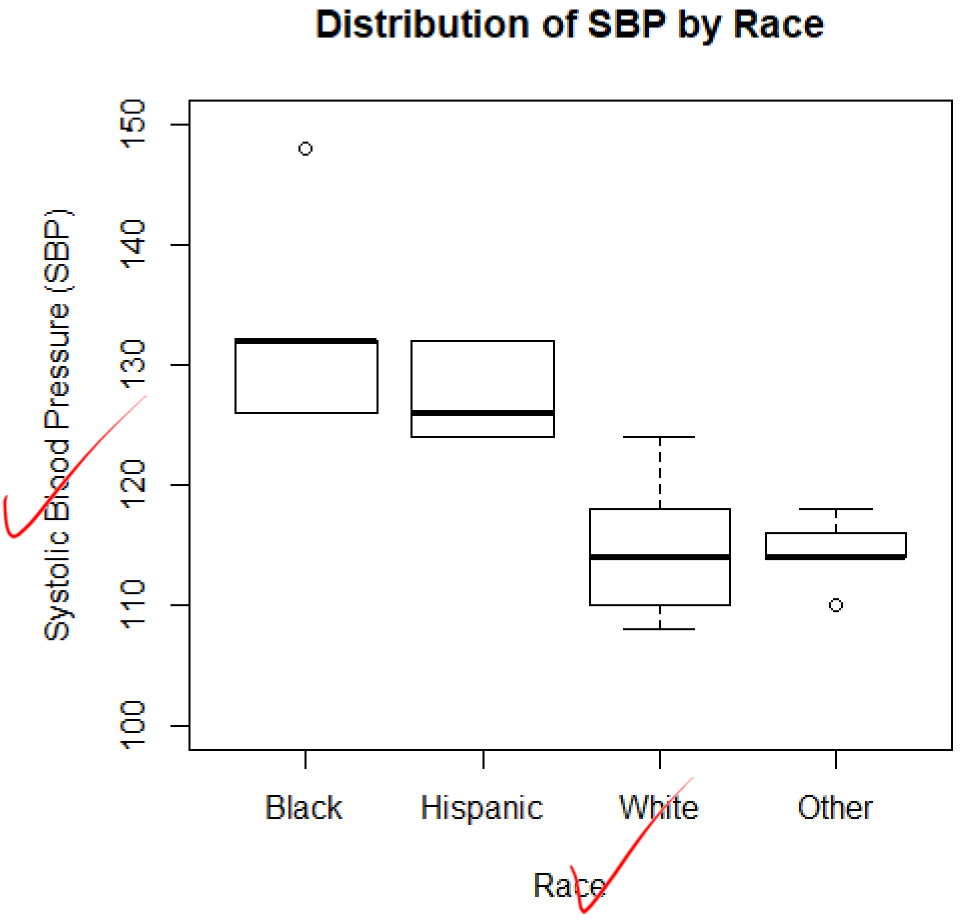
Critical Value: 7.8

The result in part A is 15 and the result here is 15, so Fr matches in both cases. The answer is Fr = 15. ✓

Question 2 (20 marks):

The data file ***systolic.txt*** contains data on systolic blood pressure (SBP) measured on 20 Americans across four different Race groups (Black=1, Hispanic=2, White=3, Other=4). Sex of the patients (Male=1 and Female=2) is also included in the data set.

a) Create a side-by-side box and whisker plot showing the distribution of SBP for each Race. Describe these distributions. (5 marks)



✓

Summary Statistics (B.4)

Race	Black	Hispanic	White	Other
Min	126.0	124.0	108.0	110.0
1st Qu	126	124	110	114
Median	132	126	114	114
Mean	132.8	127.6	114.8	114.4
3rd Qu	132	132	118	116
Max	148	132	124	118
Kurtosis	2.9306	-3.1632	-0.6814	0.8677
Skewness	1.6411	0.4414	0.608	-0.5516
Std Deviation	9.0111	4.0987	6.4187	2.9664

* Note: All values rounded down to four decimal places. All values refer to Systolic Blood Pressure (SBP), but the unit of measure was not provided in either the question or data set.

Description Report:

The black group has a high median of 132 SBP that is curiously equal with its third quartile of 132 SBP. Its median would also be its maximum if not for the affect of an outlier with a high SBP of 148. Interestingly, the presence of the outlier seems to be compensated for by the presence of two people with an SBP of 126 resulting in the first quartile value. This has resulted in a mean of 132.8 SBP that is surprisingly close to the median of 132 SBP given the outlier and small number of people in the sample (n=5). This has, however, resulted in a skewness toward the positive with a value of 1.6411 SBP and a heavy kurtosis of 2.9306 SBP suggesting a leptokurtic distribution. The lack of any whiskers in the Box Plot means that, with the exception of outliers, there is not much variance in this group.

The Hispanic group in many ways resembles the black group, with a mean and median that are quite close, 126 and 127.7 SBP respectively. The negative kurtosis in this case is quite extreme being -3.1632 SBP, suggesting a very flat distribution. This is further reinforced by the lack of any whiskers or outliers in the Box Plot. There is not much variance in this group with a quite low standard deviation of 4.0987 SBP.

The white group in many ways has a more normal distribution with a kurtosis (-0.6814 SBP) quite close to 0 compared to the others and a skewness that is not so extreme at 0.608 SBP. It is lower than the previous three overall, with a median (114 SBP) and mean (114.8 SBP) hovering in the middle and quite close to each other. It has no outliers but the whiskers suggest more variance, though that variance seems skewed higher rather than lower.

The other group is the oddest distribution of the four, with a low outlier (110 SBP), a low median (114 SBP). The outlier in this group seems to have largely been caused by how closely together the dataset is clustered, with a low standard deviation of 2.9664 SBP. This is the lowest standard deviation value of the four suggesting low variance in this group.

✓

Lovely

b) Perform a Kruskal-Wallis H-test to determine if there is any difference in SBP among the four races. State the hypotheses (Ho and Ha) and interpret you results (Do not include pairwise-comparison analysis). (7 marks)

Hypothesis:

The null hypothesis, shown below, states that there is no tendency for systolic Blood Pressure (SBP) to rank systematically higher or lower for any of the four races (Black, Hispanic, White or Other). The research hypothesis states that there is a tendency for SBP to rank systematically higher or lower for at least one of the races.

H_0: $\theta_B = \theta_H = \theta_W = \theta_O$

H_A: There is a tendency for Systolic Blood Pressure (SBP to rank systematically higher or lower for at least one race when compared to the other races.

Level of Risk:

$\alpha = 0.05$

There is a 95% chance that any observed statistical difference will be real and not due to chance.

R Output:

```
          Kruskal-Wallis rank sum test

data:  SBP by Race

Kruskal-Wallis chi-squared = 14.391, df = 3, p-value
= 0.002418
```

Interpretation:

As the p-value of 0.002418 is lower than our critical value of α = 0.05 we may reject the null hypothesis and accept the alternative hypothesis that there is a tendency for Systolic Blood Pressure (SBP to rank systematically higher or lower for at least one race when compared to the other races.

Report:

H = 14.391

df = 3

p-value = 0.002418

n = 20

$n_B = 5$

$n_H = 5$

$n_W = 5$

$n_O = 5$

A group of 20 Americans ($n = 20$) were split into four different racial groups, Black ($n_B = 5$), Hispanic ($n_H = 5$), White ($n_W = 5$) and Other ($n_O = 5$), and then had their Systolic Blood Pressure (SBP) measured. The obtained p-value of 0.002418 was less than the critical value of $\alpha = 0.05$, therefore we must reject the null hypothesis. According to the data the results from the Kruskal-Wallis rank sum test indicated that the SBP ✓ of at least one of the four races was significantly different ($H_{(3)} = 14.391$, $p < 0.05$).

Conclusion:

The analysis concludes that at least one of the racial groups has a significant tendency to rank systematically higher or lower for Systolic Blood Pressure (SBP) compared to the other groups.

✓

c) Perform a correlation analysis between SBP and sex and interpret your results. Include an appropriate plot showing the correlation between the variables, the correlation coefficient and p-value in your answer. (8 marks)

Hypothesis:

The null hypothesis states that there is no correlation between Systolic Blood Pressure (SBP) and sex. The alternative hypothesis is that there is a correlation between Systolic Blood Pressure (SBP) and sex.

H_0: $p_{pb} = 0$ ✓

H_A: $p_{pb} \neq 0$

Level of Risk:

$\alpha = 0.05$

There is a 95% chance that any observed statistical difference will be real and not due to chance. ✓

R Output:

```
r = -0.2233604

p-value = 0.3438
```

✓

Interpretation:

The critical value for rejecting the null hypothesis is 0.444 and the obtained value is $|r_{pb}|$= 0.2233 (rounded down). To add further the p-value of 0.3438 is greater than the level of risk of $\alpha = 0.05$. As the critical value is greater than the obtained value and the p-value is greater than the level of risk the null hypothesis cannot be rejected. ✓

Report:

r_{pb} = -0.2233 (rounded down)

$|r_{pb}|$= 0.2233 (rounded down)

Critical value = 0.444

n = 20

df = n -2 = 20 - 2 = 18

p-value = 0.3438

$\bar{x}_M$ = 124.1667

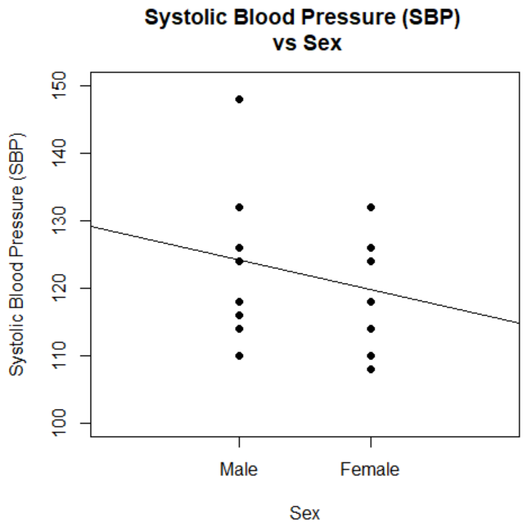
$\bar{x}_F$ = 119.75

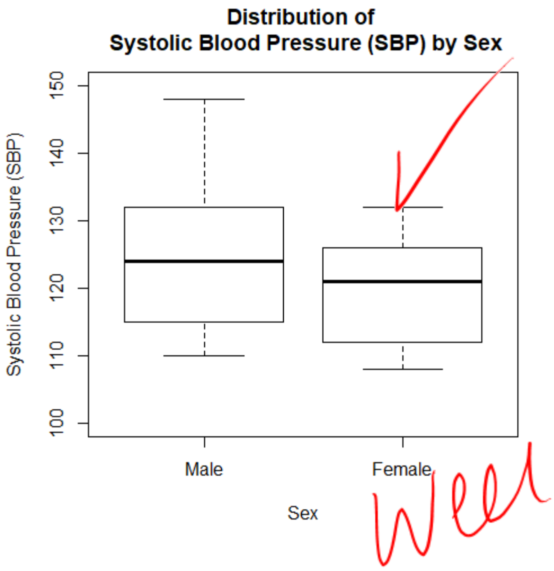
20 people (n = 20) were split into two groups based on gender, 12 males (n_M = 12) and 8 females (n_F = 8). Systolic blood pressure (SBP) was measured for each person. A point biserial correlation did not produce significant results ($R_{pb(18)}$ = 0.637, p > 0.05). The p-value of 0.3438 further demonstrated this, being greater than the level of risk α = 0.05. This suggests that there is a not a significant relationship between gender and SBP, though there is a slight indirection correlation between gender and SBP. The mean scores are quite close, with the males ($\bar{x}_M$ = 124.1667) having a slightly higher SBP ($\bar{x}_F$ = 119.75) than the females. The end result of the analysis is that the null hypothesis that there is no correlation between Systolic Blood Pressure (SBP) and sex must be accepted.

20/20

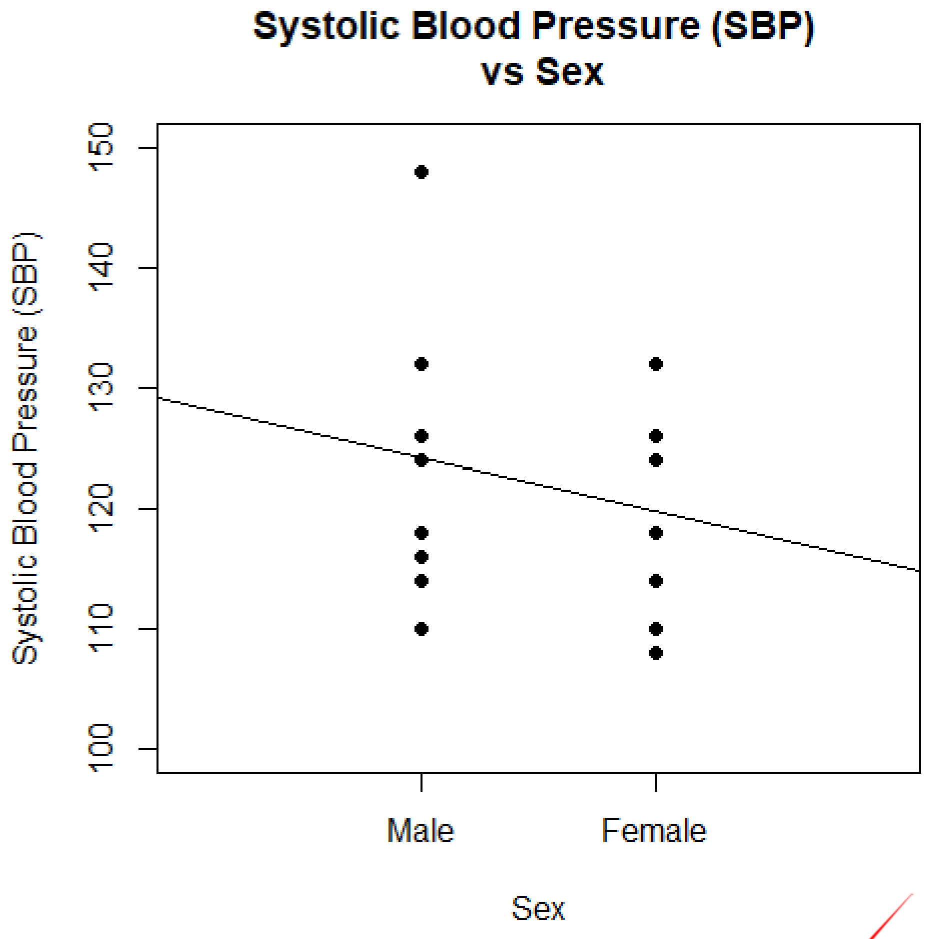
Conclusion:

The analysis concluded that there was no significant correlation between Systolic Blood Pressure (SBP) and sex.

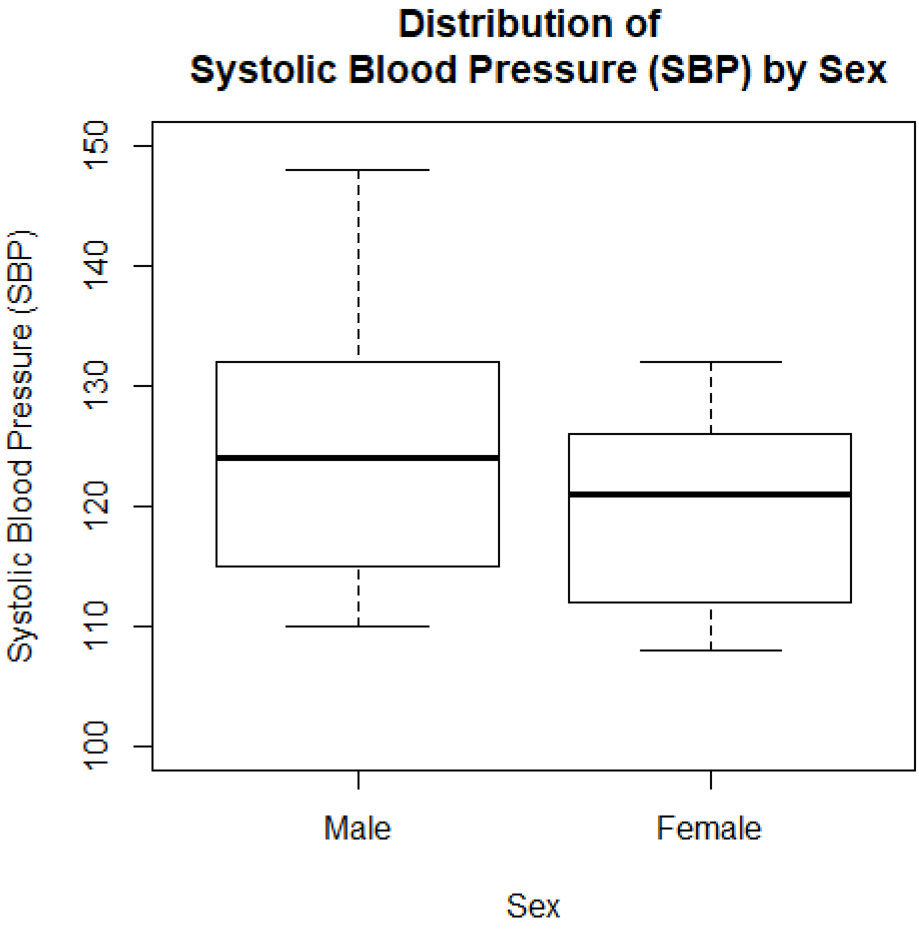




Well Done


Systolic Blood Pressure (SBP)
vs Sex
Systolic Blood Pressure (SBP)
150
140
130
120
110
100
Male
Female
Sex

✓


Distribution of
Systolic Blood Pressure (SBP) by Sex
Systolic Blood Pressure (SBP)
150
140
130
120
110
100
Male
Female
Sex

✓

Question 3 (20 marks):

A biologist is recording the sex and the presence/absence of a clearly expressed trait in a sample of 60 animals. The trait of interest is expected to occur in 20% of the population. In the sample of 60 the trait was observed in 20 animals.

a) Perform a Chi-square Goodness of Fit test. State the hypotheses (Ho and Ha) as part of your answer and interpret your results. You do not need to import a data file into SPSS or R to perform this analysis. (10 marks)

Hypothesis:

The null hypothesis states that there is no difference between the expected distribution that the trait will occur in 20% of the population of animals and the observed distribution. The alternative hypothesis states that there is a difference between the expected distribution that the trait will occur in 20% of animals and the observed distribution.

H_0: There is no difference between the expected distribution that the trait will occur in 20% of the population of animals and the observed distribution. ✓

H_A: There is a difference between the expected distribution that the trait will occur in 20% of animals and the observed distribution. ✓

Level of Risk:

α = 0.05

There is a 95% chance that any observed statistical difference will be real and not due to chance.

Calculation:

n = 60
Expected trait = 20%
Observed trait = 20

	No Trait	Trait
Expected Frequencies	48	12
Observed Frequencies	40	20

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$f_o = 20$

$f_e = P_i n = 0.20 \text{ x } 60 = 12$

$$X^2 = \frac{(40-48)^2}{48} + \frac{(20- \ 12)^2}{12}$$

$$X^2 = \frac{64}{48} + \frac{64}{12}$$

$$X^2 = \frac{64}{48} + \frac{64}{12}$$

$$X^2 = \ 1.3333 \ (rounded\ down) + \ 5.3333(rounded\ down)$$

✓

$$X^2 = \ 6.6666$$

Df = C - 1

C = 2

Df = 2 - 1 ✓

Df = 1

✓

Critical value is 3.84

Interpretation:

The critical value for rejecting the null hypothesis is 3.84 and the obtained value X^2 = 6.6666. As the obtained value of 6.6666 is greater than the critical value of 3.84 the results are significant at α = 0.05. Subsequently, we must reject the null hypothesis that there is no difference between the expected distribution that the trait will occur in 20% of the population of animals and the observed distribution and accept the alternative hypothesis that there is a difference between the expected distribution that the trait will occur in 20% of animals and the observed distribution. ✓

Report:

$x^2_{(1)}$ = 6.6666

n= 60

df = 1

Critical value = 3.84

A dataset of 60 animals were examined for the presence or absence of a trait. The 60 animals were split into female and male, though this was not directly relevant for the analysis conducted. Each of the 60 animals sampled were examined for the trait, which has not been defined other than that it is either not present or present. It was expected that 20% of the animals would exhibit the trait. An observation was conducted and the observed distribution of animals with the trait was compared to the expected distribution of animals with the trait. The chi-square test was significant ($X^2_{(1)} = 6.6666$, $p < 0.05$), suggesting that there was a difference between the expected distribution and the observed distribution. As such we must accept the alternative hypothesis that there is a difference between the expected distribution that the trait will occur in 20% of animals and the observed distribution. Based on these results the trait appears significantly more than expected in the population of animals.

Conclusion:

The analysis concludes that there is a difference between the expected distribution that the trait will occur in 20% of animals and the observed distribution. The analysis suggests that the trait appears more often in the population than expected.

The data file ***GenderTrait.txt*** contains the data from the 60 animals sampled showing each animal classified as either female or male (Gender = 1 or 2 respectively) and the absence or presence of the trait (Trait = 1 or 2 respectively).

b) Perform a Chi-square Test of Independence. State the hypotheses (Ho and Ha) and interpret your results. (10 marks)

Hypothesis:

The null hypothesis states that there is no association between the presence of the trait and the animal's gender. The alternative hypothesis is that there is an association between the presence of the trait and the animal's gender.

H_0: In the general population there is no association between the presence of the trait in an animal and the animal's gender.

H_A: In the general population there is an association between the presence of the trait in an animal and the animal's gender.

Level of Risk:

α = 0.05

There is a 95% chance that any observed statistical difference will be real and not due to chance.

-.5

Should have included Yates continuity correction

Because the chi-square distribution is continuous and if we examine only two groups, in a large series of experiments in which the null hypothesis is known to be true, the values obtained cause us to reject the null hypothesis more than the expected number of times for any critical value of the chi-square (type I error). To reduce the error, Yates' correction for continuity is often advised, especially if the actual numbers are small

R Output:

```
Pearson's Chi-squared test

data:  data1

X-squared = 2.0109, df = 1, p-value = 0.1562, phi =
0.183071

    Presence

Gender   No Trait    Trait

  Female 28.66667 14.333333

  Male   11.33333  5.666667
```

Interpretation:

As the p-value of 0.1562 is greater than α = 0.05 we must accept the null hypothesis that the proportions of the trait do not differ between genders.

The chi-square test of independence was not significant ($x^2_{(1)}$ = 2.0109, $p > 0.05$). Based on these results it is suggested that the trait does not appear more in one gender than the other.

Report:

$x^2_{(1)}$ = 2.0109

n= 60

df = 1

p-value = 0.1562

Effect Size (Φ) = 0.183 (rounded down)

A dataset of 60 animals were examined for the presence or absence of a trait. The 60 animals were split into female and male. Each of the 60 animals sampled were examined for the trait, which has not been defined other than that it is either absent (‘1’) or present (‘2’). The chi-square test was not significant ($x^2_{(1)}$ = 2.0109, $p > 0.05$). The effect size, using phi, was 0.183 (rounded down). Based on these results there is no tendency for the presence of the trait in animals to differ between genders.

Conclusion:

The analysis concludes that there is no association between the presence of the trait in an animal and the animal’s gender in the general population. ✓

15/20

Lovely reporting

Question 4 (30 marks):

For the following three scenarios select the most appropriate nonparametric test from those studied in this course. Give clear reasons for your selection and state the hypotheses (Ho and Ha).

Scenario 1:

In a study of the comparative strengths of tape-closed or sutured wounds 12 rats were tested. Each rat had a tape-closed and a sutured wound at the same time and each was tested 40 days later for strength. The researchers would like to know if taped wounds are stronger than sutured wounds. (10 marks)

Wilcoxon Signed Ranks Test ✓

As the same rat was used in both tests the data is paired. The response variable, the strength of the wound closure, is presumed to be numerical rather than categorical. The non-parametric tests for numerical paired data are the Wilcoxon signed ranks test and the sign test. The Wilcoxon signed rank test has been chosen as it is more powerful than the sign test. In this case the test is one-tailed as the only topic of interest is whether the strength of taped wounds is greater than the strength sutured wounds.

Although the hypothesis is focused on the median, what we are really interested in after the alternative hypothesis has been confirmed is whether the sum of the ranks strength of the taped wounds (ΣR_+) is greater than the sum of the ranks of the strength of the sutured wounds (ΣR_-).

Hypothesis:

The null hypothesis states that there was no difference in strength between tape-closed and sutured wounds in the rats. The alternative hypothesis states that there was a difference in strength between tape-closed and sutured wounds in the rats. Our research hypothesis is a one-tailed, directional hypothesis because it indicates a positive difference in the direction of the strength of the taped wounds. ✓

H_0: $\mu_D = 0$

H_A: $\mu_D > 0$ ✓

Scenario 2:

Twenty pigs are matched into 5 sets of 4 pigs with each set based on similarity of initial weight. In each set a pig is fed one of 4 diets (A, B, C or D) for 3 months. The farmer would like to know if there is a difference in weight gain among the diets. (10 marks)

Kruskal-Wallis H-Test

they are related by weight

Although the pigs do have a similarity of initial weight within their sets, they are unrelated as we do not try the different diets on the same pig nor is there any relation whatsoever between pigs of different sets. In other words, there is no dependence between set 1's weight on diet A and set 2's weight on diet A. As a result the data is considered unrelated and thus independent. The data, however, contains more than two samples. As it meets the criteria of both having independent samples and having more than two samples it is concluded that the Kruskal-Wallis H-Test is the correct test to apply.

Hypothesis:

The null hypothesis states that there is no tendency for one of the diets to rank systematically higher or lower for weight gain in the pigs. The research hypothesis states that there is a tendency for one of the diets to rank systematically higher or lower for weight gain in the pigs.

H_0: $\theta_A = \theta_B = \theta_C = \theta_D$

H_A: There is a tendency for one of the diets to rank systematically higher or lower for weight gain in the pigs.

-5

For this analysis, there are four diets to be analyses, diet A, diet B, diet C, and diet D.

Furthermore, the pigs were grouped based on the similarity of initial weight. Therefore, the groups are related.

This means that Friedman test is the most suitable nonparametric test to be used for this analysis as it is suitable for analysing three or more related samples.

Scenario 3:

Fifty patients who recently experienced some type of medical procedure were asked to rate their pain from 1 to 10 approximately 24 hours after the procedure. The rating scale was: 1 (No pain), 5 (moderate pain) and 10 (worst possible pain). For each of the 50 patients a self-rating of pain while in bed (Bed), while standing (Standing) and while walking (Walking) was recorded. Of interest is whether there is a relationship between a self-rating of pain while standing and while walking. (10 marks)

Point-Biserial Correlation

In this particular case we are interested in the relationship between two variables, which means that some form of correlation analysis applies. We are interested in the relationship between two variables, self-rating of pain while standing and self-rating of pain while walking. Self rating of pain is an interval scale variable and 'while standing' and 'while walking' can be considered discrete dichotomous in this case, though would require some feature engineering to achieve in practice in order to exclude 'while in bed' from consideration.

However, it starts to get tricky as it is unclear whether there is a sharp distinction between being in bed, standing and walking. In many ways these states merge into each other somewhat making it difficult. The result is some confusion as to whether this is a continuous or a discrete dichotomous variable. Given that we are only interested in whether the patient is standing or walking, and these two states are intuitively distinctive enough to be considered discrete the Point-Biserial Correlation has been chosen. It was chosen as it is the test for correlation between an interval scale variable and a discrete dichotomous variable.

Hypothesis:

The null hypothesis states that there is no correlation between self-rating of pain while standing and self-rating of pain while walking. The alternative hypothesis states that there is a correlation between self-rating of pain while standing and self-rating of pain while walking.

H_0: $p_{pb} = 0$

H_A: $p_{pb} \neq 0$

Activity level is ordinal not continuous and thus Spearman's correlation is the test to do here.

The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data. Spearman correlation is often used to evaluate relationships involving ordinal variables

-5

Appendices

R Analysis

Input

```
#Open Libraries

#library(e1071)

library(dplyr)
library(car)

library(lattice)
library(BSDA)
library(psych)
library(PerformanceAnalytics)
library(tidyr)
library(ltm)

#skew(data$scores, na.rm = TRUE,type=2) #this is closes to SPSS
#kurtosi(data$scores, na.rm = TRUE,type=2)#this is closes to SPSS

#Declaring variables Section

weevil<- read.table(file="C:/Users/Justin/Desktop/STA8190 Non-Parametrics/Assignment 2/Data Files/Weevil.txt", header=TRUE, sep="")
weevilDf<-data.frame(weevil)
weevil_variety_split <- split(weevilDf, weevilDf$Variety)
weevil_treatment_split <- split(weevilDf, weevilDf$Treatment)
#weevil_treatment_split <- split(weevil_variety_split, weevil_variety_split$Treatment)


weevilNoneSplit <- weevil_treatment_split$`1`
weevilOrganicSplit <- weevil_treatment_split$`2`
weevilIPMSplit <- weevil_treatment_split$`3`
weevilConventionalSplit <- weevil_treatment_split$`4`


weevilNone <- weevilNoneSplit %>% pull(Larvae)
weevilOrganic <- weevilOrganicSplit %>% pull(Larvae)
weevilIPM <- weevilIPMSplit %>% pull(Larvae)
weevilConventional <- weevilConventionalSplit %>% pull(Larvae)


#weevilNoneVar=weevil_treatment_split %>% pull(Control)
```

```
#print(weevilNone)



systolic<-read.table(file="C:/Users/Justin/Desktop/STA8190 Non-Parametrics/Assignment 2/Data Files/systolic.txt", header=TRUE, sep="")
systolicSexSplit <- split(systolic, systolic$Sex)
systolicRaceSplit <- split(systolic, systolic$Race)


systolicMales <- systolicSexSplit$`1`
systolicFemales <- systolicSexSplit$`2`

systolicMales <- systolicMales[ -c(1) ]
systolicFemales <- systolicFemales[ -c(1) ]


systolicBlackSplit <- systolicRaceSplit$`1`
systolicHispanicSplit <- systolicRaceSplit$`2`
systolicWhiteSplit <- systolicRaceSplit$`3`
systolicOtherSplit <- systolicRaceSplit$`4`


systolicBlack <- systolicBlackSplit %>% pull(SBP)
systolicHispanic <- systolicHispanicSplit %>% pull(SBP)
systolicWhite <- systolicWhiteSplit %>% pull(SBP)
systolicOther <- systolicOtherSplit %>% pull(SBP)




gender<-read.table(file="C:/Users/Justin/Desktop/STA8190 Non-Parametrics/Assignment 2/Data Files/GenderTrait.txt", header=TRUE, sep="")
#print(gender)

#twin<- read.table(file="C:/Users/Justin/Desktop/STA8190 Non-Parametrics/Assignments to Students 2019/Data Files/twin.txt", header=TRUE, sep="")
#twinDf<-data.frame(twin)
#controlVar=twinDf %>% pull(Control)
#treatmentVar=twinDf %>% pull(Treatment)

#goatDf<-data.frame(goats)
#goat_split <- split(goatDf, goatDf$Treatment)
#goatControl <- goat_split$`1` %>% pull(Judgement)
```

```
#goatTreat <-goat_split$`2` %>% pull(Judgement)


# Display output Section

print('######################')
print('##### Question 1 #####')
print('######################')

print('a)Create a side-by-side box and whisker plot showing the distribution of weevil
incidence in each treatment.
Describe the distributions.')

boxplot(Larvae ~ Treatment, data = weevilDf,
        xlab = "Treatment", ylab = "Larvae",
        main = "Distribution of Weevil Incidence by Treatment",
        xaxt = "n"
)

raceTicks <- c("None", "Organic", "IPM", "Conventional")

axis(1, at=1:4, labels=raceTicks)






weevilNone <- weevilNoneSplit %>% pull(Larvae)
weevilOrganic <- weevilOrganicSplit %>% pull(Larvae)
weevilIPM <- weevilIPMSplit %>% pull(Larvae)
weevilConventional <- weevilConventionalSplit %>% pull(Larvae)



print('*** None Group ***')

print(summary(weevilNone))
print(paste0("Standard Deviation = ",sd(weevilNone)))
print(kurtosi(weevilNone,na.rm = TRUE,type=2))
print(skew(weevilNone,na.rm = TRUE,type=2))


print('*** Organic Group ***')
print(summary(weevilOrganic))
print(paste0("Standard Deviation = ",sd(weevilOrganic)))
print(kurtosi(weevilOrganic,na.rm = TRUE,type=2))
print(skew(weevilOrganic,na.rm = TRUE,type=2))


print('*** IPM Group ***')

print(summary(weevilIPM))
print(paste0("Standard Deviation = ",sd(weevilIPM)))
print(kurtosi(weevilIPM,na.rm = TRUE,type=2))
print(skew(weevilIPM,na.rm = TRUE,type=2))


print('*** Conventional Group ***')

print(summary(weevilConventional))
print(paste0("Standard Deviation = ",sd(weevilConventional)))
print(kurtosi(weevilConventional,na.rm = TRUE,type=2))
print(skew(weevilConventional,na.rm = TRUE,type=2))



print('b)Perform a Friedman test to determine if there is any difference in weevil incidence among treatments. State the hypotheses (Ho and Ha) and interpret you results.')

#friedman.test(data1$late, data1$month, data1$Employee)

#friedman.test(Likert ~ Instructor | Rater, data = Data)

print(friedman.test(Larvae ~ Treatment | Variety, data = weevilDf)) # Larvae is the dependant variable, Treatment is the Independent, Variety is blocking


#friedman.test(weevil_treatment_split$late, weevil_treatment_split$month, data1$Employee)

print('c)If your analysis in part b) was significant use Wilcoxon signed rank tests to identify which treatments are different to each other. State the hypotheses (Ho and Ha) for the first comparison only. Interpret these results.')

#weevil_treatment_split

treatment1 <- weevil_treatment_split$`1` %>% pull(Larvae)
```

```
treatment2 <- weevil_treatment_split$`2` %>% pull(Larvae)
treatment3 <- weevil_treatment_split$`3` %>% pull(Larvae)
treatment4 <- weevil_treatment_split$`4` %>% pull(Larvae)

#print(treatment1)
#print(treatment2)
#print(treatment3)
#print(treatment4)

#print(paste('treatment','1',sep=""))


#wilcox.test(Larvae ~ Treatment, data=weevilDf)




for (i in 1:4) {
  treatmentA <- c(get(paste('treatment',i,sep="")))
  for (j in 1:4) {
    if (j > i){
      treatmentB <- c(get(paste('treatment',j,sep="")))
      print(treatmentA)
      print(treatmentB)
      print(wilcox.test(treatmentA,treatmentB))
      #print(wilcox.test(treatmentA,treatmentB,paired=TRUE, alternative = "two.sided", mu
= 0.0,
      #                exact = TRUE, correct = FALSE, conf.int = TRUE, conf.level = 0.95))
#paired=TRUE results in wilcox signed rank test for paired data, defaults to two tailed.
      #print(wilcox.test(treatmentA, treatmentB, paired = TRUE, alternative = "two.sided",
mu = 0.0,
      #                exact = TRUE, correct = TRUE, conf.int = TRUE, conf.level = 0.95))
      #print(pairwise.wilcox.test(treatmentA, treatmentB, alternative = c("two.sided")))

      diff <- c(treatmentA - treatmentB) #calculating the vector containing the differences
      diff <- diff[ diff!=0 ] #delete all differences equal to zero
      diff.rank <- rank(abs(diff)) #check the ranks of the differences, taken in absolute
      diff.rank.sign <- diff.rank * sign(diff) #check the sign to the ranks, recalling the signs of
the values of the differences
      ranks.pos <- sum(diff.rank.sign[diff.rank.sign > 0]) #calculating the sum of ranks
assigned to the differences as a positive, ie greater than zero
      ranks.neg <- -sum(diff.rank.sign[diff.rank.sign < 0]) #calculating the sum of ranks
assigned to the differences as a negative, ie less than zero
      print(ranks.pos) #it is the value V of the wilcoxon signed rank test # Sum of Positive
Difference Ranks
```

```
      print(ranks.neg) # Sum of Negative Difference Ranks
    }
  }
}


print('#### new tests ####!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!')


A <- c(121, 132, 148, 101, 153)

B <- c(88, 79, 105, 52, 49)


print(wilcox.test(A, B, paired = TRUE, alternative = "two.sided", mu = 0.0,
            exact = TRUE, correct = TRUE, conf.int = TRUE, conf.level = 0.95))


print(pairwise.wilcox.test(weevilDf$Larvae, weevilDf$Treatment, p.adjust.method =
"none"))

print(pairwise.wilcox.test(weevilDf$Larvae, weevilDf$Treatment,p.adjust.method =
"bonferroni"))



#print(pairwise.wilcox.test(Larvae ~ Treatment, data = weevilDf, paired=TRUE, alternative
= c("two.sided")))

#print(friedman.test(Larvae ~ Treatment | Variety, data = weevilDf))

#print(wilcox.test(Larvae ~ Treatment, data = weevilDf, paired = TRUE, alternative = c(
"two.sided"))) #agrees with SPSS


print('d)Show the Bonferroni correction calculation by hand. How does this correction
change your results and interpretation in part c)? How does this correction affect the
probability of a type-I error?')
```

```
print('e)Calculate the Friedman test by hand using equation 5.2 from the text book. Clearly
define Ri, n, k and CF as part of your answer. Does this match your result in part a)? (10
marks)')





print('######################')
print('##### Question 2 #####')
print('######################')

print('a)Create a side-by-side box and whisker plot showing the distribution of SBP for
each Race. Describe these distributions.')



boxplot(SBP ~ Race, data = systolic,
        xlab = "Race", ylab = "Systolic Blood Pressure (SBP)",
        main = "Distribution of \n Systolic Blood Pressure (SBP) by Race",
        xaxt = "n",
        ylim=c(100, 150)
)

raceTicks <- c("Black", "Hispanic", "White", "Other")

axis(1, at=1:4, labels=raceTicks)



print('*** Black Group ***')

print(summary(systolicBlack))
print(paste0("Standard Deviation = ",sd(systolicBlack)))
print(kurtosi(systolicBlack,na.rm = TRUE,type=2))
print(skew(systolicBlack,na.rm = TRUE,type=2))


print('*** Hispanic Group ***')

print(summary(systolicHispanic))
print(paste0("Standard Deviation = ",sd(systolicHispanic)))
print(kurtosi(systolicHispanic,na.rm = TRUE,type=2))
print(skew(systolicHispanic,na.rm = TRUE,type=2))
print('*** White Group ***')

print(summary(systolicWhite))
print(paste0("Standard Deviation = ",sd(systolicWhite)))
print(kurtosi(systolicWhite,na.rm = TRUE,type=2))
print(skew(systolicWhite,na.rm = TRUE,type=2))


print('*** Other Group ***')

print(summary(systolicOther))
print(paste0("Standard Deviation = ",sd(systolicOther)))
print(kurtosi(systolicOther,na.rm = TRUE,type=2))
print(skew(systolicOther,na.rm = TRUE,type=2))




print('b)Perform a Kruskal-Wallis H-test to determine if there is any difference in SBP among the four races. State the hypotheses (Ho and Ha) and interpret you results (Do not include pairwise-comparison analysis).')

#data1$treat<-as.factor(data1$treat) #must convert to factor or test will not work
#kruskal.test(data1$strength, data1$treat) #agrees with SPSS

#kruskal.test(systolic$Race, systolic$SBP) #agrees with SPSS

print(kruskal.test(SBP ~ Race, data = systolic))


#(data1<-systolic %>% gather(Race, SBP, 1:3))
#data1$group<-as.factor(data1$Race) #must convert to factor or test will not work
#kruskal.test(data1$Race, data1$SBP) #agrees with SPSS




print('c)Perform a correlation analysis between SBP and sex and interpret your results. Include an appropriate plot showing the correlation between the variables, the correlation coefficient and p-value in your answer. ')

# https://www.statmethods.net/advgraphs/axes.html
```

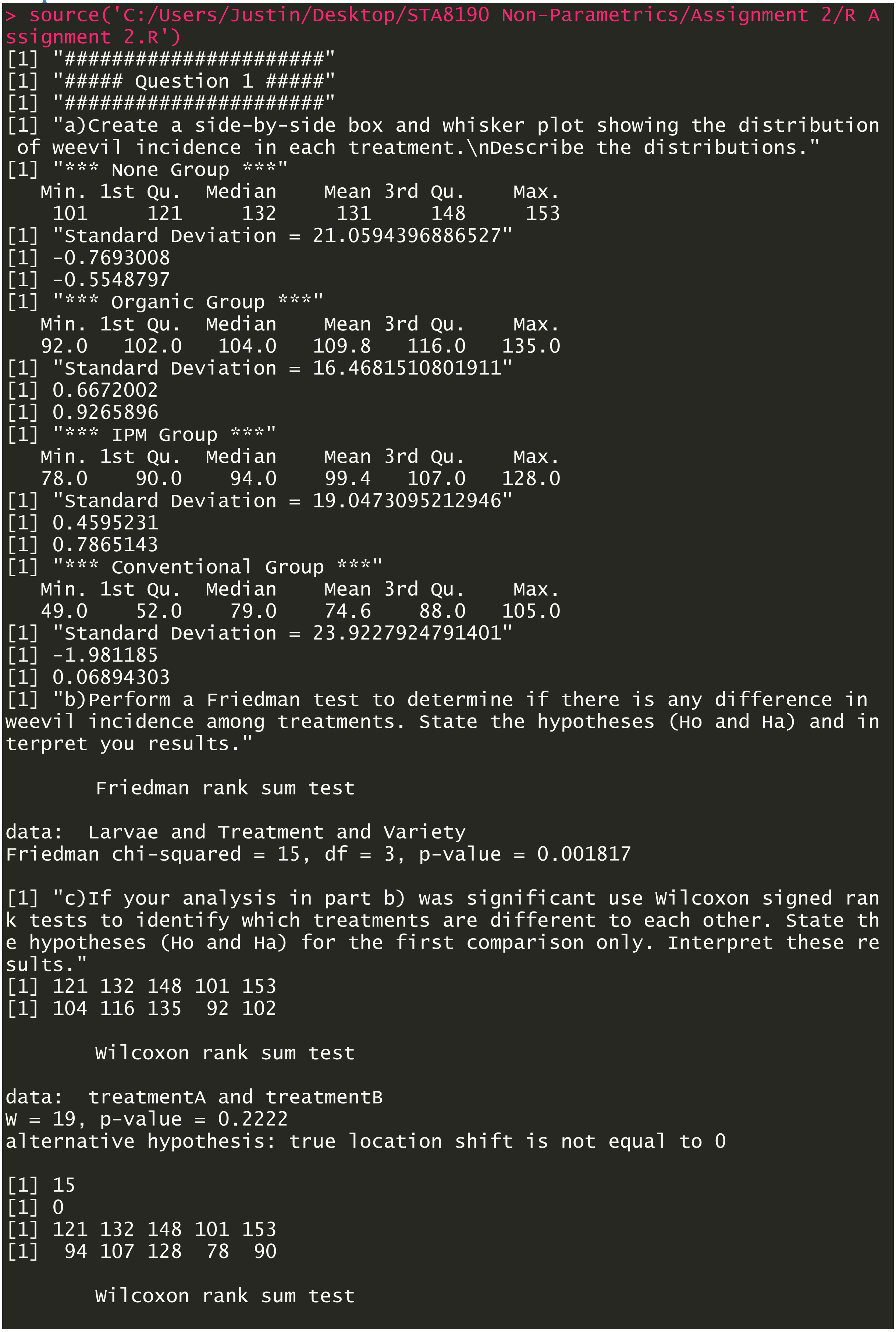
```
#corr.test(data[,2:3], use = "pairwise", method = "spearman", adjust = "none")


#sexVar<- systolic$Sex
#SBPVar<- systolic$SBP

# print(corr.test(x = sexVar, y = SBPVar, use = "pairwise", method = "spearman", adjust =
"none"))

# print(systolic)


#(systolic)


#axis(1, xaxp=c(1, 2, 5), las=0)

###axis(1, at=c(1,2,3,4), las=1)

###abline(lm(systolic$SBP~systolic$Sex), col="blue") # regression line (y~x)

#legend(location, title, legend, ...)

#lines(lowess(systolic$Race, systolic$SBP), col = "blue")


#print(corr.test(systcoli$Race, systolic$SBP, use = "pairwise", method = "spearman",
adjust = "none"))


print(biserial.cor(systolic$SBP,systolic$Sex, level = 2))

print(cor.test(systolic$SBP, systolic$Sex))

print(mean(systolicMales$SBP))
print(mean(systolicFemales$SBP))

par(xpd=FALSE)

plot(systolic$Sex, systolic$SBP, main = "Scatter Plot of \n Systolic Blood Pressure (SBP) vs
Sex",
    xlab = "Sex", ylab = "Systolic Blood Pressure (SBP)",
    pch = 19, frame = TRUE,
    xaxt = "n",
    ylim=c(100, 150),
    xlim=c(0, 3)
)

sexTicks <- c("Male", "Female")

axis(1, at=1:2, labels=sexTicks)


par(xpd=FALSE)
abline(lm(systolic$SBP~systolic$Sex), col="black") # regression line (y~x)

#abline(lm(systolic$SBP~systolic$Sex), col="black") # regression line (y~x)

boxplot(SBP ~ Sex, data = systolic,
        xlab = "Sex", ylab = "Systolic Blood Pressure (SBP)",
        main = "Box Plot of Distribution of \n Systolic Blood Pressure (SBP) by Sex",
        xaxt = "n",
        ylim=c(100, 150)
)

sexTicks <- c("Male", "Female")

axis(1, at=1:2, labels=sexTicks)

plot(systolic$Sex,systolic$SBP, xlab = "Sex", ylab = "SBP", main = "Distribution of SBP by
Sex")
abline(lm(systolic$SBP~systolic$Sex), col="blue") # regression line (y~x)

#plot(x = sexVar, y = SBPVar)



#plot(systolic$Race,systolic$SBP)
#plot(systolic$Race,systolicMales$SBP)
#points(systolicMales$SBP, col=2)
#print(systolicMales)
#print(systolicFemales)

plot(systolicMales$Race,systolicMales$SBP, col="Blue", xlab = "Sex", ylab = "SBP", main =
"Distribution of SBP by Sex")
points(systolicFemales$SBP, col="Pink")


abline(lm(systolicMales$SBP~systolicMales$Race), col="blue") # regression line (y~x)
abline(lm(systolicFemales$SBP~systolicFemales$Race), col="Pink") # regression line (y~x)
```

```
barplot(systolicMales$SBP,systolicMales$Race)
barplot(systolicFemales$SBP,systolicFemales$Race)

#margin.table(systolic,1)

#barplot(systolic,
#       main = "Survival of Each Class",
#       xlab = "Class",
#       col = c("blue","pink")
#)
#legend("topleft",
#      c("Not survived","Survived"),
#      fill = c("red","green")
#)


print('######################')
print('##### Question 3 #####')
print('######################')

print('a)Perform a Chi-square Goodness of Fit test. State the hypotheses (Ho and Ha) as
part of your answer and interpret your results. You do not need to import a data file into
SPSS or R to perform this analysis.')

#goodness of fit
#chisq.test(x = count, p = expected, rescale.p = TRUE) #rescale turns expected counts into
proportions
#?chisq.test



# printing the p-value
#chisq$p.value
# printing the mean
#chisq$estimate


print('b)Perform a Chi-square Test of Independence. State the hypotheses (Ho and Ha)
and interpret your results.')

# Not by Gender

#$genderMorphA <- xtabs(Trait~Gender, data=gender)
#chisqVarA <- chisq.test(genderMorphA)
#print(chisqVarA)
```

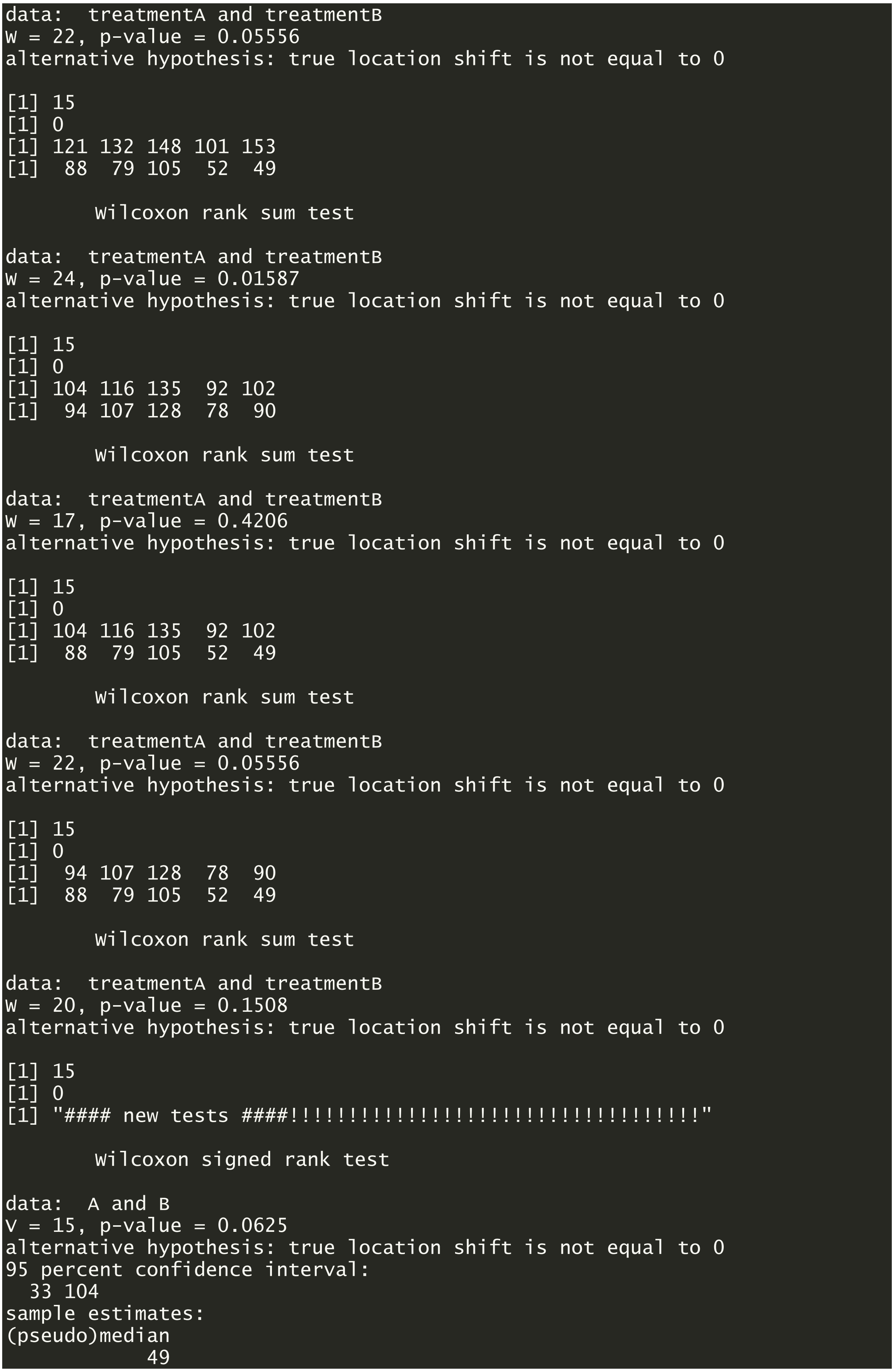
```
#print(chisqVarA$expected)





#print(gender)

y = count(gender, 'Trait')
print(y)

z <- gender %>% group_by(Gender) %>% count(Trait)

z[1:2,1:1] <- 'Female'
z[3:4,1:1] <- 'Male'

z[1:1,2:2] <- 'No Trait'
z[2:2,2:2] <- 'Trait'
z[3:3,2:2] <- 'No Trait'
z[4:4,2:2] <- 'Trait'

z <-data.frame(z)

colnames(z) <- c("Gender", "Presence","Count")

print(z)

##test of independence
tblGender = table(gender$Trait, gender$Gender) # trait is row, gender is column
print(tblGender)

chiVar <- chisq.test(tblGender,correct=F)

print(chiVar)
print(chiVar$expected)


##test of independence 2
(data <- z)
data1<-xtabs(Count~Gender + Presence, data=data)
(a<-chisq.test(data1,correct=F))
print(a)
print(a$expected)
```

```
#genderMorph <- xtabs(Trait~Gender, data=gender)
#genderMorphTest<-xtabs(Trait~Gender+Subject, data=gender) # Unsure, page 178 of textbook do more reading.

#print(genderMorph)
#print(genderMorphTest) # Unsure

#chisqVar <- chisq.test(genderMorph)

#print(chisqVar)

#print(chisqVar$expected)

effectSize <- sqrt(2.0109/60)

print(effectSize)




#cramers <- apply(newdat, 2, cramersV)



##test of independence
#(data <- read.table("table8.10.txt",header=TRUE))
#data1<-xtabs(Count~Type + Behaviour, data=data)
#(a<-chisq.test(data1,correct=F))
#a$expected



print('######################')
print('##### Question 4 #####')
print('######################')
```

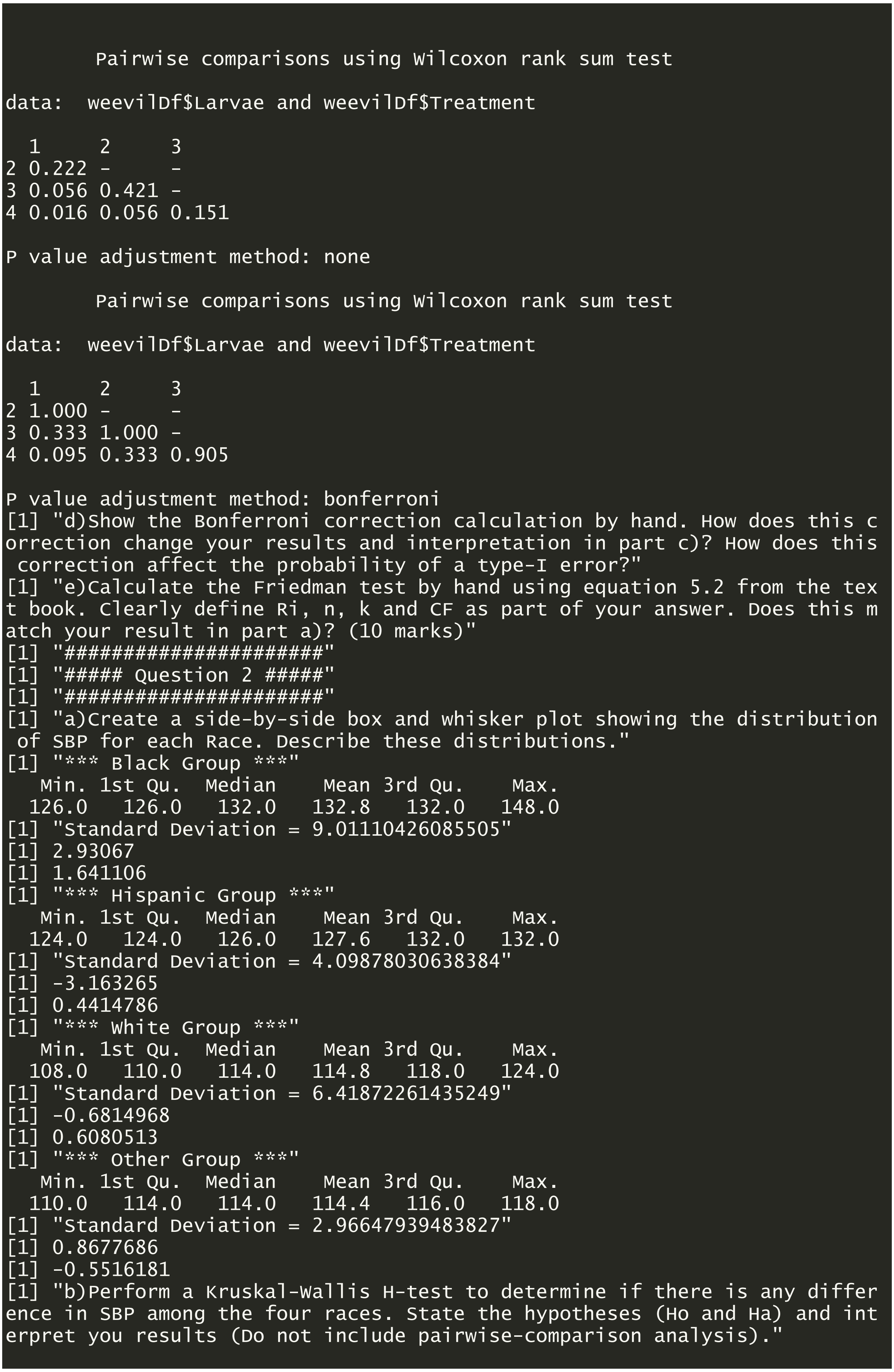
Output



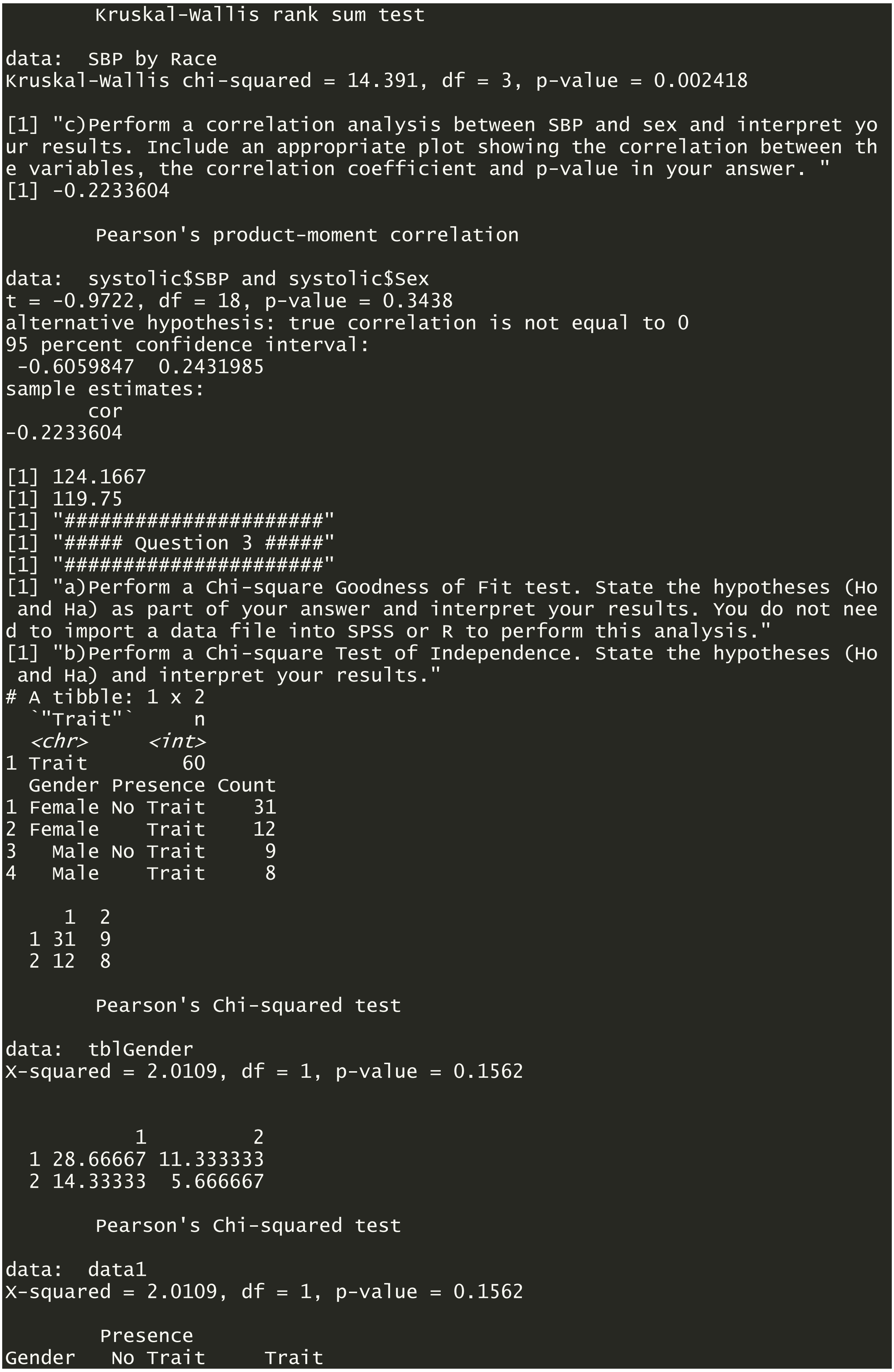
```
> source('C:/Users/Justin/Desktop/STA8190 Non-Parametrics/Assignment 2/R A
ssignment 2.R')
[1] "######################"
[1] "##### Question 1 #####"
[1] "######################"
[1] "a)Create a side-by-side box and whisker plot showing the distribution
 of weevil incidence in each treatment.\nDescribe the distributions."
[1] "*** None Group ***"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    101     121     132     131     148     153 
[1] "Standard Deviation = 21.0594396886527"
[1] -0.7693008
[1] -0.5548797
[1] "*** Organic Group ***"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   92.0   102.0   104.0   109.8   116.0   135.0 
[1] "Standard Deviation = 16.4681510801911"
[1] 0.6672002
[1] 0.9265896
[1] "*** IPM Group ***"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   78.0    90.0    94.0    99.4   107.0   128.0 
[1] "Standard Deviation = 19.0473095212946"
[1] 0.4595231
[1] 0.7865143
[1] "*** Conventional Group ***"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   49.0    52.0    79.0    74.6    88.0   105.0 
[1] "Standard Deviation = 23.9227924791401"
[1] -1.981185
[1] 0.06894303
[1] "b)Perform a Friedman test to determine if there is any difference in
weevil incidence among treatments. State the hypotheses (Ho and Ha) and in
terpret you results."

        Friedman rank sum test

data:  Larvae and Treatment and Variety
Friedman chi-squared = 15, df = 3, p-value = 0.001817

[1] "c)If your analysis in part b) was significant use Wilcoxon signed ran
k tests to identify which treatments are different to each other. State th
e hypotheses (Ho and Ha) for the first comparison only. Interpret these re
sults."
[1] 121 132 148 101 153
[1] 104 116 135  92 102

        Wilcoxon rank sum test

data:  treatmentA and treatmentB
W = 19, p-value = 0.2222
alternative hypothesis: true location shift is not equal to 0

[1] 15
[1] 0
[1] 121 132 148 101 153
[1]  94 107 128  78  90

        Wilcoxon rank sum test
```



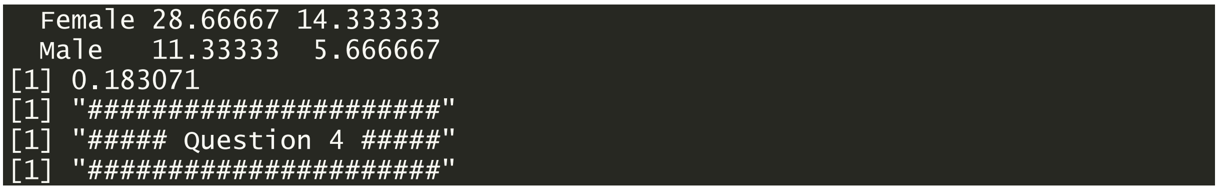
```
data:  treatmentA and treatmentB
W = 22, p-value = 0.05556
alternative hypothesis: true location shift is not equal to 0

[1] 15
[1] 0
[1] 121 132 148 101 153
[1]  88  79 105  52  49

        Wilcoxon rank sum test

data:  treatmentA and treatmentB
W = 24, p-value = 0.01587
alternative hypothesis: true location shift is not equal to 0

[1] 15
[1] 0
[1] 104 116 135  92 102
[1]  94 107 128  78  90

        Wilcoxon rank sum test

data:  treatmentA and treatmentB
W = 17, p-value = 0.4206
alternative hypothesis: true location shift is not equal to 0

[1] 15
[1] 0
[1] 104 116 135  92 102
[1]  88  79 105  52  49

        Wilcoxon rank sum test

data:  treatmentA and treatmentB
W = 22, p-value = 0.05556
alternative hypothesis: true location shift is not equal to 0

[1] 15
[1] 0
[1]  94 107 128  78  90
[1]  88  79 105  52  49

        Wilcoxon rank sum test

data:  treatmentA and treatmentB
W = 20, p-value = 0.1508
alternative hypothesis: true location shift is not equal to 0

[1] 15
[1] 0
[1] "#### new tests ####!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!"

        Wilcoxon signed rank test

data:  A and B
V = 15, p-value = 0.0625
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  33 104
sample estimates:
(pseudo)median
            49
```



```

        Pairwise comparisons using Wilcoxon rank sum test

data:  weevilDf$Larvae and weevilDf$Treatment

  1     2     3
2 0.222 -     -
3 0.056 0.421 -
4 0.016 0.056 0.151

P value adjustment method: none

        Pairwise comparisons using Wilcoxon rank sum test

data:  weevilDf$Larvae and weevilDf$Treatment

  1     2     3
2 1.000 -     -
3 0.333 1.000 -
4 0.095 0.333 0.905

P value adjustment method: bonferroni
[1] "d)Show the Bonferroni correction calculation by hand. How does this c
orrection change your results and interpretation in part c)? How does this
 correction affect the probability of a type-I error?"
[1] "e)Calculate the Friedman test by hand using equation 5.2 from the tex
t book. Clearly define Ri, n, k and CF as part of your answer. Does this m
atch your result in part a)? (10 marks)"
[1] "######################"
[1] "##### Question 2 #####"
[1] "######################"
[1] "a)Create a side-by-side box and whisker plot showing the distribution
 of SBP for each Race. Describe these distributions."
[1] "*** Black Group ***"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  126.0   126.0   132.0   132.8   132.0   148.0 
[1] "Standard Deviation = 9.01110426085505"
[1] 2.93067
[1] 1.641106
[1] "*** Hispanic Group ***"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  124.0   124.0   126.0   127.6   132.0   132.0 
[1] "Standard Deviation = 4.09878030638384"
[1] -3.163265
[1] 0.4414786
[1] "*** White Group ***"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  108.0   110.0   114.0   114.8   118.0   124.0 
[1] "Standard Deviation = 6.41872261435249"
[1] -0.6814968
[1] 0.6080513
[1] "*** Other Group ***"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  110.0   114.0   114.0   114.4   116.0   118.0 
[1] "Standard Deviation = 2.96647939483827"
[1] 0.8677686
[1] -0.5516181
[1] "b)Perform a Kruskal-Wallis H-test to determine if there is any differ
ence in SBP among the four races. State the hypotheses (Ho and Ha) and int
erpret you results (Do not include pairwise-comparison analysis)."
```



```
        Kruskal-Wallis rank sum test

data:  SBP by Race
Kruskal-Wallis chi-squared = 14.391, df = 3, p-value = 0.002418

[1] "c)Perform a correlation analysis between SBP and sex and interpret yo
ur results. Include an appropriate plot showing the correlation between th
e variables, the correlation coefficient and p-value in your answer. "
[1] -0.2233604

        Pearson's product-moment correlation

data:  systolic$SBP and systolic$Sex
t = -0.9722, df = 18, p-value = 0.3438
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6059847  0.2431985
sample estimates:
       cor
-0.2233604

[1] 124.1667
[1] 119.75
[1] "######################"
[1] "##### Question 3 #####"
[1] "######################"
[1] "a)Perform a Chi-square Goodness of Fit test. State the hypotheses (Ho
 and Ha) as part of your answer and interpret your results. You do not nee
d to import a data file into SPSS or R to perform this analysis."
[1] "b)Perform a Chi-square Test of Independence. State the hypotheses (Ho
 and Ha) and interpret your results."
# A tibble: 1 x 2
  `"Trait"`     n
  <chr>     <int>
1 Trait        60
  Gender Presence Count
1 Female No Trait    31
2 Female    Trait    12
3   Male No Trait     9
4   Male    Trait     8

     1  2
  1 31  9
  2 12  8

        Pearson's Chi-squared test

data:  tblGender
X-squared = 2.0109, df = 1, p-value = 0.1562


           1         2
  1 28.66667 11.333333
  2 14.33333  5.666667

        Pearson's Chi-squared test

data:  data1
X-squared = 2.0109, df = 1, p-value = 0.1562

        Presence
Gender   No Trait     Trait
```



```
  Female 28.66667 14.333333
  Male   11.33333  5.666667
[1] 0.183071
[1] "######################"
[1] "##### Question 4 #####"
[1] "######################"
```