Miniproject 3 – One Way ANOVAs

StudentID 100225776 Applied Statistics

Question 1

To load into R the given data on wheat growth, display the yields graphically and describe any observations of this data.

Part A – Load data

See Appendix A – R Commands – Lines 6-14.

Part B – Produce Box Plot of Data and Describe Observations

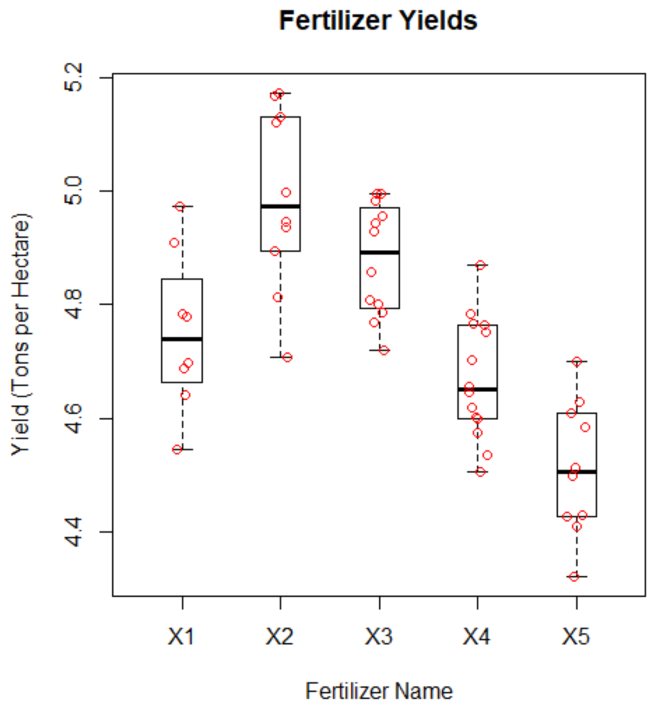


Figure 1 – Distribution of yields per fertilizer tested

Looking at Figure 1 we can make many observations on the effectiveness of each fertilizer.

The standard product (X1) has a reasonable spread with a shape indicative of a normal distribution, where the median sits evenly between the two quartiles and the whiskers extend further out than either quartile is from the mean. We can see from the data plots given that the data reflects this, with a grouping of points more towards the median.

Fertilizer X2 appears to be the most effective overall, with the highest maximum and median yield of the five tested. It also has the largest Interquartile Range (IQR) of all the fertilizers though, suggesting there is a larger distribution in the results and therefore it could be a more unreliable product. We can see from the data points that there is a group of results towards the top of the plot for X2 and trailing points at the bottom, which skews the third quartile upwards and extends the bottom whisker far from the third quartile, further showing that the product produces mixed results compared to the others in the data set.

The results for X3 show it has a very tight distribution, with the third quartile being very close to the maximum and a median that sits very close to the middle between the first and third quartiles, showing a decently even spread of yield. The minimum of X3 is approximately equal to the median of X1, showing that across the board you can expect better yields from X3 than the average yield gained from X1. Compared to X2, X3 has a lower median, Q3 and maximum with an approximately equal minimum, but the tighter spread indicates the product is more likely to deliver a yield close to its average compared to X2, but in general it is a less effective product.

Fertilizer X4 shows yields slightly worse than our standard product, with a similar distribution. We can see that the median is skewed towards the first quartile, indicating a grouping of values towards the lower end of the plot. Looking at the data points, however, we can see that there are two major groupings of points around the first and third quartiles possibly indicating a bimodal distribution, though this is hard to tell without more data. The maximum value in this set is far from the third quartile group, seeming to indicate it is an outlier or that the distribution has a tail towards the maximum. We can see that the means between groups X1 and X4 are close, indicating that there could be little to no significant difference between them, but we will have to investigate further to see if that is true. Overall, the plot for X4 indicates it performs a little worse than the standard product, with a similar spread.

Finally, the data from fertilizer X5 appears to have a similar spread to the standard product, however the boxplot is the lowest of the five with the maximum value being approximately equal to the median of the standard product, indicating that on average the product produces the smallest yield out of all the products tested. The data points at the minimum and maximum indicate they could be outliers but again we do not have enough data to say for certain. The data points being clustered around the median, first and third quartiles indicates that there is possibly a flat distribution towards the middle of the set with tails in the minimum and maximum quartiles.

We can see that there are no data points outside of the top and bottom whiskers of each plot, though this is because the sample size is too small to produce points in the 0-0.35% and 99.65-100% ranges that is required for these to appear. We can see from the analysis above that there are possibly still outlier values though, mostly on fertilizers X1 and X5, with possible upper outliers on X4 and lower outliers on X2.

By looking at the data, we can assume that fertilizers X1, X3, X4 and X5 have equal variances, due to the approximately equal sizes of their interquartile ranges. X2 has a noticeably larger IQR especially compared to X3 however its IQR seems to be less than twice that of X3 meaning the set likely has an equal variance assumption.

Question 2

To implement a one way ANOVA on the given data set and determine if there are any significant effects, stating any null or alternative hypotheses that apply.

Determining the Variance of a Given Data Set

In R there are two programs we can use here that can generate an ANOVA: oneway.test is used when the variances are unequal, and lm is used when variances are approximately equal. In order to work out which one we should use, we must first test the null hypothesis that the population variances are equal.

There are many tests that can be used to determine variance - including Bartlett, Hartley and Cochran - however these are known to be sensitive to departures from normality and shouldn't generally be used unless we know for certain the data is normally distributed, and even then there is no benefit to using them over other methods when we are performing the calculations in R. We can use the F-test to compare two variances, however in this case we need to compare variances over multiple groups of data so this method is unsuitable. That leaves the Levene test and the Filgner-Killeen test. A Filgner-Killeen test is beyond the scope of this project, so we will go with the Levene test as it is robust to departures from normality and easily performed in R. We import the lawstat library and perform the test on our data set with
`levene.test`. *For code see Appendix A – R Commands – lines 25-28.*
This gave us the result

```
Test Statistic = 0.81163, p-value = 0.52395934.
```

Since the P-value is far greater than the significance level of 0.05, we can accept the null hypothesis and determine that the sample variances in our data set are from random sampling of equal variances. This means I will use `lm` to calculate the ANOVA.

Implementing an ANOVA on the Set

For code see Appendix A – R Commands – lines 34+35.

The result from our use of `lm` can be expressed as

```
F(4,49) = 23.411, p < 0.0001
```

For the ANOVA, our null hypothesis H_0 is that there is no statistically significant difference between the means of the groups. Therefore, the alternative hypothesis H_1 is that the average of at least one group differs from the others. Since the P-value is far below our significance level of 0.05, we can say that there is a highly statistically significant difference between some of the means. Therefore, we reject the null hypothesis H_0 and subsequently accept H_1 to be true.

Question 3

To state any assumptions made from the ANOVA result of question 2 and investigate whether these assumptions are satisfied.

Statement of ANOVA Assumptions

There are six assumptions that we need to keep in mind when using an ANOVA test:

1. Each group within the data set is from a normally distributed population.
2. There are no significant outliers within the set.
3. There is a homogeneity (commonality) of variances.
4. The data is measured at the interval.
5. The independent variable contains at least three independent groups.
6. There is no relationship between observations within each group, and no relationship between each group.

Investigation of Assumptions 4-6

Assumptions 4-6 can be easily checked without using R. In our case, the data variable (Yield) is measured in tons per hectare, which is a measured interval and so the set passes this assumption. The set of data we are using contains five groups of data, and each group is strictly for only one type of fertilizer, so our set passes this assumption too. This also proves that there is no relationship between each group (there would be if, say, there was a fertilizer tested which was just a mix of two or more of the other fertilizers). We also know that all the data points in one group are for only one fertilizer (for example, there are no results for X2 in X3's group), so assumption 6 is also satisfied.

Investigation of Assumptions 1-3

For assumptions 1-3, we will need to use R to determine if they are satisfied.

We can test for the first assumption, that of normality in distribution of variances, in a couple different ways. One way can be by plotting the residuals against the fitted values and examining the difference by eye to see if the residuals align at 0. You can also plot the sample residuals against the theoretical quantiles on a QQ plot, in which case you will be looking for the data to fit a diagonal line. I will explore both methods below.

For code see Appendix A – R Commands – line 49.

Figure 2 plots residuals against the fitted values for our data set to test whether there is a non-linear relationship between the response variable (Yield) and the predictors. We can see that although there is an upward tick towards the end of the trend line, it is still largely horizontal and therefore we can say that there is an absence of nonlinear patterns between response and predictors, as assumed.

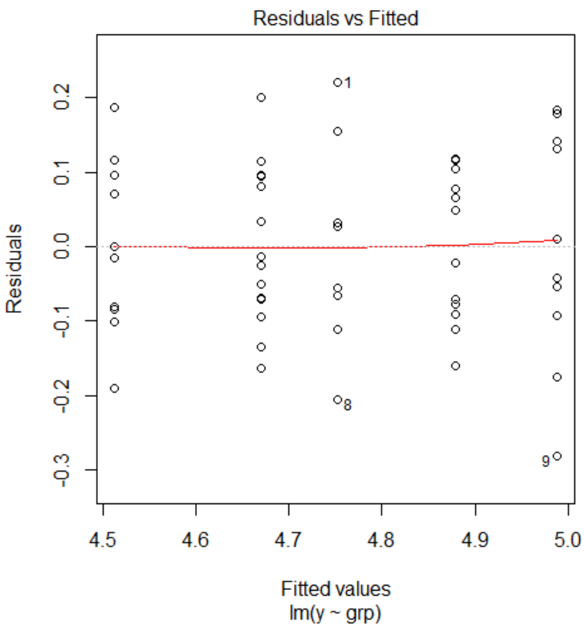


Figure 2 - Residuals vs Fitted plot

A Normal QQ plot shows the fits the standardised residuals from a data set against theoretical quartiles taken from the linear model of the data. If both sets of data are taken from a normal distribution, we expect to see the data points fitting a diagonal line. We know that the theoretical quantiles are a normal distribution due to them being a linear model, so we are checking the residuals. Residuals are standardised here to allow us to change the measured interval (from tons per hectare to, say, tonnes per square kilometre) and the graph will maintain its shape.

For code see Appendix A – R Commands – line 53.

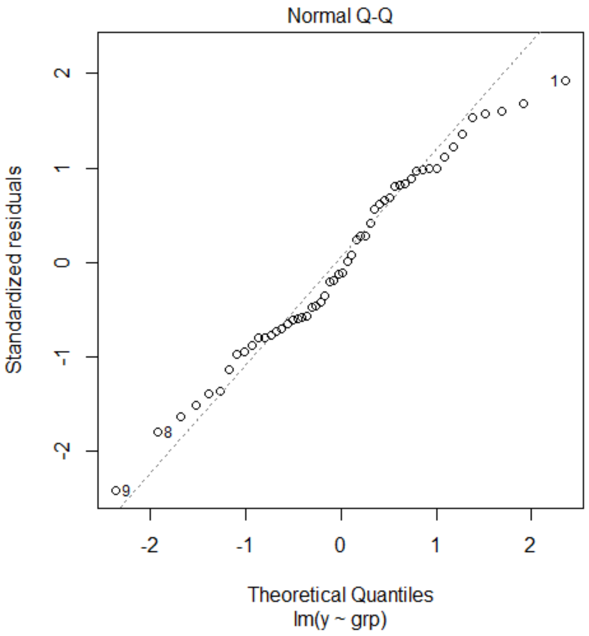


Figure 3 - QQ plot of residuals to linear model quantiles

From our data I have generated the QQ plot shown in Figure 3. I feel it is safe to say it shows an approximate normal distribution and that this assumption is satisfied. The fit is not perfect, and there appears to be outlier values, but the data does not appear to be skewed too much in either direction and the data fits the approximate direction of the line. We know that the F test used in the ANOVA is reasonably robust to non-normality.

Next, we check for outliers in the data set. As we can see from the QQ plot in Figure 3, there appears to be two outlier values in the set. These are labelled '1' and '9' and appear to exist outside of (-1.98, 1.98) on either axis. As outliers tend to increase the estimate of sample variance (and therefore decrease the F statistic for an ANOVA), they lower the chance of us rejecting the null hypothesis. Outliers mean we could perform a nonparametric test such as a Kruskal-Wallis test as the ANOVA, however this is both beyond the scope of this project and not required, as we know the F statistic found in question 2 was sufficiently high to reject our null hypotheses we know that these two outlier values were not enough to increase the estimate of sample variance by enough to have an impact on the results of our ANOVA therefore we can safely say this assumption is satisfied.
As for where these outliers could have come from, it could be that they are from recording errors during the harvest of the crop, or from the samples not being from entirely from the same population. The latter could be true if the field that those samples come from was previously treated with another fertilizer or a herbicide, and those chemicals remained in the soil when our experiment was carried out.

Finally, we must investigate the commonality of variances. Fortunately, this was previously covered under question 2 where we sort to discover which ANOVA program to use on our data. We used the Lavene Test and determined that with a p-value << 0.05, the assumption of commonality of variances is satisfied.

Through investigation, we have analysed six different assumptions made in the ANOVA and determined that all six have been satisfied.

Question 4

Compare each new fertilizer (X2-X5) against the standard product (X1). This should include comparisons on significance levels, parameter estimates and confidence levels.

Method of Comparison

By using the `summary()` function in R on the linear model previously generated, we can return a comparison of the standard product group to the other groups in our data set. This comparison includes Standard Deviation, Error, t value and p value. This can also be found using `pairwise.t.test()`, however this compares every group to every other group in the data set rather than just X1 against the others, and also provides us with less information.

Significance Level or α (alpha) refers to the probability of rejecting the null hypothesis if the null hypothesis is true. For most experiments, an alpha of 0.05 is low enough to safely reject our null hypothesis should the results conform to it and will be the value we will use for all our comparisons.

Parameter estimates or sample statistics are estimates of the population parameters we are looking to model, as it is generally impossible to measure an entire population. For our data, we are looking to model several parameters. The first being $\hat{\mu}$ (mu-hat) which is equal to $\bar{x}_1$ (x-bar-sub-one), which is the estimation of the population mean equal to the mean of the first sample set (X1). $\bar{x}_I$ (x-bar-sub-i) can be found in R using the `tapply()` function across the y, and grp variables, using the mean function. We also look for $\hat{\beta}$ (beta-hat), the sample estimate of the population parameter β (beta) for a given group, which is equal to $\bar{x}_i$- $\bar{x}_1$ for group *i*. This can be easily found in R using the program `confint()`.

Confidence level is determined by our significance level α, where CL = 100 – α. Therefore, in our results we are looking for a CL of 0.95. Related to this is the confidence interval. The CI is the range in which we are [CL]% likely to find the mean. To find this for each group (*i*) compared with group 1, we first find the critical value for the t distribution with the degrees of freedom of our data set. This is then multiplied by the Residual Standard Error of the data set as well as by the square root of $\left(\frac{1}{n_1}+\frac{1}{n_i}\right)$ where X_1 and X_i are the numbers of entries in each set being compared, found using `tapply()` on y and grp variables, using the length function. The CI range is the result of that calculation ± the estimate of the coefficient for group *i* against group 1.

Comparison Results

First, I will find the values needed for calculating the comparison for each group as listed above.

For code see Appendix A – R Commands – lines 65-93

Running our code, we are left with the following results:
Group coefficients from `summary()`:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.206e+00	1.953e-01	16.421	< 2e-16
X2	1.318e-02	3.241e-03	4.067	0.000182
X3	4.787e-02	1.252e-02	3.823	0.028940
X4	-1.923e-03	7.851e-05	-24.496	0.138103
X5	4.426e-01	4.738e-02	9.341	0.000151

Residual standard error from `summary()`: `0.123 on 49 degrees of freedom`.

Group means and counts from `tapply()`:

Group	*X1*	*X2*	*X3*	*X4*	*X5*
Mean	4.752125	4.988373	4.878447	4.669961	4.512522
Length	8	10	12	14	10

Therefore μ̂ = 4.752125
Critical value from `qt(.975, 49) = 2.009575.`

With these values, we can find the parameter estimates and confidence intervals when comparing groups X2 to X5 with X1. For this we can use `confint()`, supplying it with the linear model and requesting a 0.95 confidence level.

For code see Appendix A – R Commands – line 99.

Confidence intervals for our data as a result from using `confint()`:

	2.5 % CL	97.5% CL
(Intercept)	4.6647515	4.83949830
X2	0.1190247	0.35347203
X3	0.0135231	0.23912018
X4	-0.1916920	0.02736466
X5	-0.3568269	-0.12237952

Group X2 – X1:

$$\bar{x}_2 = 4.988373$$

$$\hat{\beta} = \bar{x}_2 - 4.752125 = 0.236248$$

$$\text{CI} = 0.2362 \pm 2.0096 \times 0.123 \times \left(\frac{1}{10} + \frac{1}{8}\right)^{1/2} = 0.2362 \pm 0.1172$$

$$\text{CI} = (0.119, 0.3535)$$

With 95% confidence, we can say that the difference in yield between fertilisers X2 and X1 is between 0.119 and 0.3535 units. Our estimate of the difference is 0.2362 units. The standard error of the difference is 0.123 and the margin of error is 0.1172. The p-value for this comparison is 0.0002 which is highly statistically significant. This all shows that the means differ significantly in this comparison as the interval range does not contain 0.

Group X3 – X1:

$$\bar{x}_3 = 4.878447$$

$$\hat{\beta} = \bar{x}_3 - 4.752125 = 0.126322$$

$$\text{CI} = 0.1263 \pm 2.0096 \times 0.123 \times \left(\frac{1}{12} + \frac{1}{8}\right)^{1/2} = 0.1263 \pm 0.1128$$

$$\text{CI} = (0.014, 0.2391)$$

With 95% confidence, we can say that the difference in yield between fertilisers X3 and X1 is between 0.014 and 0.2391 units. Our estimate of the difference is 0.1263 units. The standard error of the difference is 0.123 and the margin of error is 0.1128. The p-value for this comparison is 0.0289 which is just statistically significant to pass our test. This all shows that the means differ significantly in this comparison as the interval range does not contain 0.

Group X4 – X1:

$$\bar{x}_4 = 4.669961$$
$$\hat{\beta} = \bar{x}_4 - 4.752125 = -0.082164$$
$$\text{CI} = -0.0822 \pm 2.0096 \times 0.123 \times \left(\frac{1}{14} + \frac{1}{8}\right)^{1/2} = -0.0822 \pm 0.1096$$
$$\text{CI} = (-0.1917, 0.0274)$$

With 95% confidence, we can say that the difference in yield between fertilisers X4 and X1 is between -0.1917 and 0.0274 units. Our estimate of the difference is -0.0822 units. The standard error of the difference is 0.123 and the margin of error is 0.1096. The p-value for this comparison is 0.1381 which is not statistically significant as it is above our α of 0.05. This all shows that the means do not differ significantly enough as the interval range contains the number 0.

Group X5 – X1:

$$\bar{x}_5 = 4.512522$$
$$\hat{\beta} = \bar{x}_5 - 4.752125 = -0.239603$$
$$\text{CI} = \text{-}0.2396 \pm 2.0096 \times 0.123 \times \left(\frac{1}{10} + \frac{1}{8}\right)^{1/2} = \text{-}0.2396 \pm 0.1172$$
$$\text{CI} = (-0.3568, \text{-}0.1224)$$

With 95% confidence, we can say that the difference in yield between fertilisers X5 and X1 is between -0.3568 and -0.1224 units. Our estimate of the difference is -0.2396 units. The standard error of the difference is 0.123 and the margin of error is 0.1172. The p-value for this comparison is 0.0002 which is highly statistically significant. This all shows that the means are significantly different for this comparison.

In conclusion, all the comparisons against the standard control product showed significantly different means aside from fertilizer X4, which did not. We therefore conclude that there is no statistically significant difference between the crop yield on fertilizer X4 compared to our standard control, but there is significant difference in all other products tested.

Question 5

We will implement Holm and Bonferroni correction for multiple testing on all p values analysed in question 4 and comment on any changes to the conclusions.

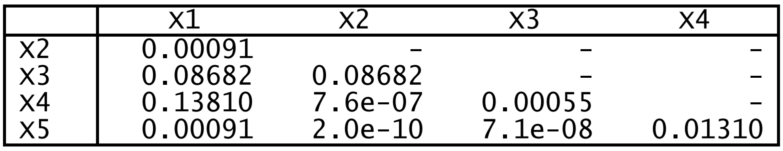
(i) Holm Correction of Results

With our previous test in question 4, we were testing each fertilizer (X2-X5) against our standard control product to see if each one was significantly different, which formed our null hypothesis. We found that for three of the products that there was a high likelihood we could reject the null hypothesis based on the results we gathered. There is a possibility, however, that we identified a product that was not different from the control when it is, or that we identified a product as being significantly different from the control when it is not. These are known as type I (false positive) and type II (false negative) errors. Holm correction is an approach designed to control the type I errors in our findings by adjusting the rejection criteria.

We can run a Holm correction on our p-values in R with the `pairwise.t.test()` program. This compares each group in the data set together and gives adjusted p-values for each.

For code see Appendix A – R Commands – line 115.

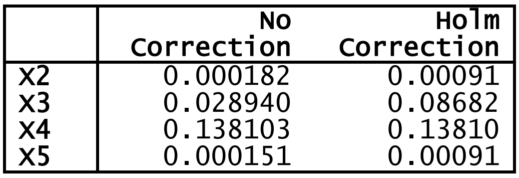
T test p-values with Holm correction:



	X1	X2	X3	X4
X2	0.00091	-	-	-
X3	0.08682	0.08682	-	-
X4	0.13810	7.6e-07	0.00055	-
X5	0.00091	2.0e-10	7.1e-08	0.01310

As we have been testing groups X2-X5 against only X1, we are only interested in the first column. Looking carefully, you will see an interesting result here. I will compare these p-values to the p-values without correction to make it clearer.

T tests against X1 with and without correction:



	No Correction	Holm Correction
X2	0.000182	0.00091
X3	0.028940	0.08682
X4	0.138103	0.13810
X5	0.000151	0.00091

The p-value for group X4, which we assumed to not have a significant difference in means, did not change. This is expected since if this were an incorrect assumption, it would be classified as a type II error which Holm does not correct for. We can see that the result for group X3, however, has changed from our initial finding of $p = 0.0289$ to $p = 0.0868$. This would be above our α of 0.05 and therefore, under Holm correction, we can say that fertilizer X3 does not have a significant difference in means.

We can see that the p-values for groups X2 and X5 have changed slightly but still show high statistical significance.

(ii) Bonferroni Correction of Results

Bonferroni correction is a similar correction method to Holm in that it aims to reduce type I errors in results but is generally considered inferior to Holm as it is too conservative in its application. It is commonly regarded that Bonferroni correction was better suited to times before we had computers with languages like R to calculate corrections, as it is much easier than Holm to perform by hand.

To test for Bonferroni correction of p-values, we can again use `pairwise.t.test()` program, specifying Bonferroni instead of Holm correction.

For code see Appendix A – R Commands – line 125.

T test p-values with Bonferroni correction:

	X1	X2	X3	X4
X2	0.00182	-	-	-
X3	0.28940	0.42049	-	-
X4	1.00000	9.5e-07	0.00079	-
X5	0.00151	2.0e-10	7.9e-08	0.03276

To make this difference more apparent I will again compare the p-values to ones we have found in the past.

T tests against X1 with and without correction:

	No	Holm	Bonferroni
	Correction	Correction	Correction
X2	0.000182	0.00091	0.00182
X3	0.028940	0.08682	0.28940
X4	0.138103	0.13810	1.00000
X5	0.000151	0.00091	0.00151

We can see that the Bonferroni correction has done nothing to our p-values for groups X2, X3 and X5, which we should be testing for, and has decided that the probability that X4 is not statistically significant is certain, which seems unusual but is no concern to us since X4 was again not a type I error candidate.
We know that smaller sample sizes like ours are subject to increases in type I error when there are moderate differences in group variance. In cases such as this, we can use the Welch t test to compare means. In R, this is carried out by repeating the previous command but using the `pool.sd = FALSE` flag. Let us see if that has any effect on our results.

For code see Appendix A – R Commands – line 135.

Welch t test p-values with Bonferroni correction:

	X1	X2	X3	X4
X2	0.04090	-	-	-
X3	0.48030	0.76817	-	-
X4	1.00000	0.00063	0.00028	-
X5	0.01741	8.0e-06	3.7e-06	0.03291

We can now see that the p-values have wildly changed from before, once again eliminating X3. However, the correction has been so extreme that both X2 and X5 are now considered barely statistically significant. Although this has given us the same conclusion as the Holm correction above, I feel this test could be more likely to start introducing type II errors into our findings via overcorrection.

Question 6

We will examine the given fertilizer costs for each fertilizer to determine which fertilizer should be recommended when the average profit per ton of wheat is £120 and explain why. We will also explain what additional information I would request for better predictions and how that information would be used.

Fertilizer Profit Recommendation

We must work out the potential profits of the five fertilizers given. If we can assume that each ton of product sells for exactly £120, we can find how much profit each fertilizer generates.

For code see Appendix A – R Commands – lines 150-155.

Results for $\sum \bar{x}_i \times 120 - c$ with accompanying data.

	X1	*X2*	*X3*	*X4*	*X5*
Mean Yield (tons per hectare)	4.752125	4.988373	4.878447	4.669961	4.512522
Cost (£)	120	125	100	112	80
Profit per hectare (yield × 120 – cost)	450.26	473.60	485.41	448.40	461.50

Going by these values, X3 would be the fertilizer to choose as it gives the largest profit per hectare of wheat out of all the fertilisers tested. However, this does not consider the range of results for each fertilizer and whether that would have any effect on our result.

Further Information Recommendation

To more accurately simulate what fertilizer to choose, it might be good to have a list of prices we can sell the wheat at over time, to see if it's worth using a more or less expensive fertilizer when the sales price for our product varies.

There is the possibility that the wheat being harvested is of a lower quality using some fertilisers than others and could therefore sell for a lower price, which would be another data point to consider in choosing which fertilizer to use.

Appendix A - R Commands

```
## Question 1 #################################################################
# We will load the data into R and create graphs for the different
# fertilizers, as well as explain any observations on the data.
#
# Part A - Load data:
library(readr)
yields <- read_csv("WheatGrowth.csv", col_names = FALSE)
colnames(yields) <- c('Fertilizer', 'Yield')
# Part B - Produce Box Plot on this data:
boxplot(Yield~Fertilizer, data = yields, main = "Fertilizer Yields",
        xlab = "Fertilizer Name", ylab = "Yield (Tons per Hectare)",
        boxwex = 0.40)
stripchart(Yield~Fertilizer, data = yields, add = TRUE, vertical = TRUE,
           method = "jitter", pch = 1, col = "red")
# Saved as Fig1.png
# See report for analysis.
###############################################################################

## Question 2 #################################################################
# We will implement a one way ANOVA on the given data set and determine if
# there are any significant effects, stating any hypotheses that apply.
#
# To ensure we're using the correct ANOVA program in r, we should first test
# the equality of the variances. We have determined to use the Levene test.
y <- yields$Yield
grp <- factor(yields$Fertilizer)
require(lawstat)
levene.test(y, grp, location = "median")
# RESULTS:
# data:  y
# Test Statistic = 0.81163, p-value = 0.5239
#
# These results varify the null hypothesis therefore I will use lm.
yields.lm <- lm(y ~ grp)
anova(yields.lm)

oneway.test(y ~ grp, var.equal = TRUE)
# RESULTS:
# Response: y
#           Df  Sum Sq Mean Sq F value    Pr(>F)
# grp        4 1.41619 0.35405  23.411 7.297e-11 ***
# Residuals 49 0.74103 0.01512
# See report for analysis
###############################################################################

## Question 3 #################################################################
# We will state any assumptions made from the ANOVA and investigate whether
# these assumptions are satisfied.
#
# We start with plotting residuals verses fitted values.
plot (yields.lm, which = 1)
# Saved as Fig2.png.
# See report for analysis.
# Now we create a QQ plot of Standardised Residuals vs Theoretical Quartiles
plot (yields.lm, which = 2)
```

```
# Saved as Fig3.png.
# See report for analysis.
###############################################################################

## Question 4 #################################################################
# Compare each new fertilizer (X2-X5) against the standard product (X1).
# Include significance levels, parameter estimates and confidence levels.
#
# We start by finding the comparison of group X1 with against the rest of
# the data set.
summary(yields.lm)
# RESULTS:
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  4.75212    0.04348 109.298  < 2e-16
# X2           0.23625    0.05833   4.050 0.000182
# X3           0.12632    0.05613   2.250 0.028940
# X4          -0.08216    0.05450  -1.508 0.138103
# X5          -0.23960    0.05833  -4.108 0.000151
# Residual standard error: 0.123 on 49 degrees of freedom
# Multiple R-squared:  0.6565, Adjusted R-squared:  0.6284
# F-statistic: 23.41 on 4 and 49 DF,  p-value: 7.297e-11
#
# We then find the means and counts of each group:
tapply(y, grp, mean)
tapply(y, grp, length)
tapply(y, grp, sd)
# RESULTS:
# Group means:
#       X1       X2       X3       X4       X5
# 4.752125 4.988373 4.878447 4.669961 4.512522
# Group counts:
# X1 X2 X3 X4 X5
#  8 10 12 14 10
#
# Then we find the critial value for this set:
qt(.975, 49)
# RESULTS:
# [1] 2.009575
#
# From here we can finally calculate the CI for each group using a
# two-sample t test:
confint(yields.lm, level = 0.95)
# RESULTS:
#                  2.5 %      97.5 %
# (Intercept)  4.6647515  4.83949830
# X2           0.1190247  0.35347203
# X3           0.0135231  0.23912018
# X4          -0.1916920  0.02736466
# X5          -0.3568269 -0.12237952
# See report for analyses.
###############################################################################

## Question 5 #################################################################
# We will implement Holm and Bonferroni correction for multiple testing on all
# p values analysed in question 4 and comment on any changes.
#
```

```
# Holm Correction:
pairwise.t.test(y, grp, p.adjust.method = "holm")
# RESULTS:
#    X1      X2      X3      X4
# X2 0.00091 -       -       -
# X3 0.08682 0.08682 -       -
# X4 0.13810 7.6e-07 0.00055 -
# X5 0.00091 2.0e-10 7.1e-08 0.01310
# See report for analysis.
#
# Bonferroni Correction:
pairwise.t.test(y, grp, p.adj = "bonferroni")
# RESULTS:
#    X1      X2      X3      X4
# X2 0.00182 -       -       -
# X3 0.28940 0.42049 -       -
# X4 1.00000 9.5e-07 0.00079 -
# X5 0.00151 2.0e-10 7.9e-08 0.03276
# See report for analysis.
#
# Bonferroni Correction on Welch t test:
pairwise.t.test(y, grp, p.adj = "bonferroni", pool.sd = FALSE)
# RESULTS:
#    X1      X2      X3      X4
# X2 0.04090 -       -       -
# X3 0.48030 0.76817 -       -
# X4 1.00000 0.00063 0.00028 -
# X5 0.01741 8.0e-06 3.7e-06 0.03291
# See report for analysis.
################################################################################

## Question 5 ##################################################################
# We will examine the given fertilizer costs for each fertilizer to determine
# which fertilizer should be recommended when the average profit is £120.
#
# First, read in the data:
costs <- read_csv("Costs.csv", col_names = TRUE)
colnames(costs) <- c('Fertilizer', 'Cost')
# Then calculate the profit per hectare
yield.mean <- tapply(y, grp, mean)
salePrice <- 120
yield.mean * salePrice - costs$Cost
# RESULTS:
#       X1       X2       X3       X4       X5
# 450.2550 473.6048 485.4136 448.3954 461.5026
# See report for summary.
################################################################################
```