Applied Statistical Methods Assignment 1

ANOVA, ANCOVA, MANOVA and Two-Way ANOVA

Fisher Group

V Sushant : 2019A7PS0045P
Chirag Wadhwa : 2019A7PS0103P
Mayank Jain : 2019A7PS0141P
Aryan Devrani : 2019A8PS0408P
Yash Bansal : 2019A7PS0484P
Atharva Chandak : 2019A7PS0062P
Suchismita Tripathy : 2019A7PS0554P
Arshdeep Taneja : 2019A4PS0469P

15 October 2021

Contents

1 ANOVA **3**
- 1.1 Theory . . . 3
- 1.2 Calculations . . . 3
- 1.3 Example . . . 4

2 ANCOVA **5**
- 2.1 Theory . . . 5
- 2.2 Purpose of ANCOVA . . . 5
- 2.3 Assumptions . . . 5
- 2.4 Example . . . 6

3 MANOVA **8**
- 3.1 Theory . . . 8
- 3.2 Methodology . . . 8
  - 3.2.1 Pillai's Trace . . . 8
  - 3.2.2 Hotelling-Lawley's trace . . . 8
  - 3.2.3 Wilk's Lambda . . . 9
  - 3.2.4 Roy's largest root . . . 9
- 3.3 MANOVA vs ANOVA . . . 9
- 3.4 Example . . . 9

4 Two - Way ANOVA **11**
- 4.1 Theory . . . 11
- 4.2 Calculation . . . 11
- 4.3 Example . . . 12
  - 4.3.1 Results of Two-way ANOVA in R: . . . 12
  - 4.3.2 Conclusion of Two-Way ANOVA: . . . 12
- 4.4 3-Way ANOVA: . . . 12

5 Conclusion **13**

6 Appendix **14**
- 6.1 Tests Used . . . 14
  - 6.1.1 Test for Normality . . . 14
  - 6.1.2 Test for multicollinearity . . . 14
  - 6.1.3 Test for homogeneity of variance . . . 14
  - 6.1.4 Test the Homogeneity of regression slopes . . . 14
  - 6.1.5 Run the ANCOVA test . . . 14
- 6.2 Two-way ANOVA Dataset . . . 14

7 References **17**

1 ANOVA

1.1 Theory

ANOVA or Analysis Of Variance is a statistical tool used to analyse the differences/lack of differences between means of two or more groups. Where a t-test can be used for the same purpose when there are 2 groups under consideration, ANOVA enables a similar analysis to be carried out for more than 2 groups. It was developed by Ronald Fisher with its first application published in 1921.
ANOVA is a method to analyse comparative experiments i.e. to determine if different *treatments* to some process affect the results significantly, or to check if there is any homogeneity between them. Here, the data involves one or more independent variables and a dependent variable. The independent variables are also known as explanatory variables and are categorical i.e. have different levels (groups). The dependent variable is also known as the response variable and are quantitative and this is what is used as a measure of the effectiveness of each level of the factor. In general, ANOVA is used to refer to One-Way ANOVA, which involves only one independent variable.

1.2 Calculations

The ANOVA process works for 3 important assumptions:

1. For each population, the dependent variable is normally distributed.
2. The variance of each population (σ^2) are the same.
3. The observations are independent.

This method takes the form of hypothesis testing for the following (in the case of 3 treatments/groups):

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_a : \text{Not all } \mu\text{s have the same value}$$

ANOVA splits the total variance into 2 components: one that is due to specific differences in the experiments and one that is due to random factors.

$$SS_{Total} = SS_{Treatments} + SS_{Error}$$

where SS : Sum of Squares

The test statistic used is an F-test, that is calculated based on the values of the mean sums of squares, each of which are ratios of sums of squares and degrees of freedom.

MS	SS	Degrees of Freedom
$MS_{Treatments}$	$SS_{Treatments}$	k - 1
MS_{Error}	SS_{Error}	$n_T - k$

where k : number of treatments i.e. levels of the independent/explanatory variable n_T : total number of readings (= n x k where n is the number of readings per treatment, if the total is split equally)

$$F_{(k-1),(n_T-k)} = \frac{MS_{Treatments}}{MS_{Error}}$$

1.3 Example

A pharmaceutical company conducts an experiment to test the effect of a new cholesterol medication. The company selects 15 subjects randomly from a larger population. Each subject is randomly assigned to one of three treatment groups. Within each treatment group, subjects receive a different dose of the new medication. In Group 1, subjects receive 0 mg/day; in Group 2, 50 mg/day; and in Group 3, 100 mg/day.

Group 1	Group 2	Group 3
210	210	180
240	240	210
270	240	210
270	270	210
300	270	240

Data Summary :

Groups	Count	Sum	Average	Variance
Group 1	5	1290	258	1170
Group 2	5	1230	246	360
Group 3	5	1050	210	450

The independence assumption is satisfied by the design of the study, which features random selection of subjects and random assignment to treatment groups.The tests for normality and Homogeneity of Variance are specified in Appendix section 6.1 **ANOVA**(performed using excel) :

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	6240	2	3120	4.16	0.04241751818	3.885293835
Within Groups	9000	12	750			
Total	15240	14				

We accept the null hypothesis when the P-value is greater than the significance level; when it is lower, we reject it. For significance level 0.05 we reject the null hypothesis because the P-value (0.04) is lesser.

2 ANCOVA

2.1 Theory

1. ANCOVA or Analysis of Covariance is an extension of ANOVA that removes the impact of one or more metric-scaled undesired variables from the original dependent variable before undertaking research. It is a way of controlling for initial individual differences that could not be randomized.

2. The focus is to determine the effects of the independent variable on the dependent variable adjusted for the presence of Covariates in the model.

3. An independent and continuous control variable, called a Covariate, is used to study the differences among different sample groups and the factor being observed, while controlling for the effects of covariate on the associated dependent variable.

4. ANCOVA can explain within-group variance. It takes the unexplained variances from the ANOVA test and tries to explain them with a covariate. There can be multiple possible covariates. However, the more the covariates, the fewer degrees of freedom we have. A strong covariate increases the power of the test and a weak covariate reduces it.

2.2 Purpose of ANCOVA

1. To increase the power and sensitivity of the test (F-test) by reducing the error term (SSE); the error term is adjusted for, and hopefully reduced by, the relationship between the dependent variable and the Covariate(s). Covariates are used to assess the "noise" where noise is the undesirable variance in the dependent variable that is estimated by scores on the covariate.

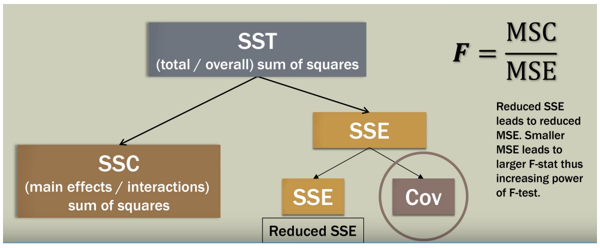

Figure 1: Distribution of Sum of Squares when Covariate is taken into consideration

2. To adjust the means on the levels of dependent variable itself to what they would be if all subjects scored equally on the covariate(s). Differences between subjects on covariate(s) are removed so that presumably, the only differences that remain are related to the effects of the grouping independent variables. The covariate(s) enhance prediction of the dependent variable, but there is no implication of causality.

2.3 Assumptions

1. The dependent variable(s) and covariate should be continuous and their regression relationship must be linear for each group.

2. There should be homogeneity in the error variances, i.e, equal variances for the different categorical groups. The regression slopes of different groups must be equivalent.

3. The errors must be independent and uncorrelated. The residuals should be normally distributed.

2.4 Example

Let us look at the data below showing study skill assessment of university students in different years of study. A typical one-way ANOVA would run considering the year as a fixed factor and respective scores as the dependent variable. Here, the GPAs for different years serve as a Covariate, which will help us to study the difference between the sample groups (Years) and the dependent variable being observed, i.e Scores.

Year 1 Scores	Year 1 GPA	Year 2 Scores	Year 2 GPA	Year 3 Scores	Year 3 GPA
53	2.46	62	2.42	56	2.49
61	2.44	66	2.32	64	2.64
69	2.39	71	3.01	69	2.79
70	3.02	71	3.51	72	3.25
74	2.9	78	3.14	73	3.32
82	3.55	85	3.4	78	3.34
93	3.22	94	3.49	87	3.57

Tests of Between-Subjects Effects

Dependent Variable: Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	67.524[a]	2	33.762	.256	.777
Intercept	111180.190	1	111180.190	841.465	<.001
Year	67.524	2	33.762	.256	.777
Error	2378.286	18	132.127		
Total	113626.000	21			
Corrected Total	2445.810	20			

a. R Squared = .028 (Adjusted R Squared = -.080)

Figure 2: ANOVA results obtained on IBM SPSS

Tests of Between-Subjects Effects (ANCOVA)

Dependent Variable: Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1637.244[a]	3	545.748	11.474	<.001
Intercept	49.550	1	49.550	1.042	.322
GPA	1569.720	1	1569.720	33.003	<.001
Year	92.109	2	46.055	.968	.400
Error	808.566	17	47.563		
Total	113626.000	21			
Corrected Total	2445.810	20			

a. R Squared = .669 (Adjusted R Squared = .611)

Figure 3: ANCOVA results obtained on IBM SPSS

In order to justify the assumptions, we perform 3 tests:

(a) Test for Multicollinearity (Ref.[6.1.1])

(b) Test for Homogeneity of variances (Levene's test) (Ref.[6.1.2])

(c) Test for homogeneity of regression slopes (Ref.[6.1.3])

The difference in results obtained before and after considering the Covariate are as follows:

1. Sum of Squares accounted for by the model increased from 67.524 to 1637.244, thus improving the coefficient of determination (r^2) from -0.8 to 0.611

2. The Sum of Squares due to error (SSE) reduced from 2378.286 to 808.566

3. The F-statistic improved from a mere 0.256 to 11.474

4. The Model (F-test) significance reduced drastically from 0.777 to 0.000236, which drastically reduces the chances of concluding that a difference exists when there is actually no difference.

   The dataset and .sav file for SPSS can be found on the github repository here.

3 MANOVA

3.1 Theory

Multivariate Analysis Of Variance (MANOVA) is the multivariate equivalent of univariate ANOVA, which uses at least two dependent variables to examine changes in the independent variable between many groups. MANOVA compares the vectors containing the group mean of each dependent variable. MANOVA uses multivariate tests like Wilk's Lambda, Hotelling-Lawley's test, Pillai's Trace, and Roy's Largest Root. Pillai's Trace has the highest statistical power and is the most robust. Based on the best linear combinations of the many response variables, MANOVA maximises discrimination between groups rather than within groups.

Assumptions of MANOVA:

1. The groups being compared are independent.
2. The independent variables should be categorical.
3. Multivariate normality of the population distribution.
4. Homogeneity of Covariance matrices.
5. Non collinearity between different dependent variables i.e. low correlation between them.
6. Linear relationship between the dependent variables for each group

After rejecting NULL hypothesis of MANOVA, we can run separate ANOVAs on individual

3.2 Methodology

Unlike univariate ANOVA, which only provides one test statistic (the F ratio), MANOVA includes numerous alternative test statistics that can be utilised to confirm the findings and establish a conclusive conclusion. The Hypothesis matrix H and the Error matrix E are two matrices that describe these statistical tests. These matrices are created using the sum of squares and cross-product methods and are based on the dependent and independent variables generated for each degree of freedom in the model.

The four test statistics based on the Eigenvalues of the H(E+H)-1 , HE-1 , and E(E+H)-1 matrices are as follows -

3.2.1 Pillai's Trace

This is considered it to be the most powerful and most robust of the four statistics. The formula is

$$Pillai'sTrace = trace[H(H+E)^{-1}] = \sum_{i=1}^{q} \mathrm{\lambda_i}/(1+\mathrm{\lambda_i}) \tag{1}$$

3.2.2 Hotelling-Lawley's trace

$$Hotelling - Lawley'strace = trace[A] = trace[HE_{-1}] = \sum_{i=1}^{q} \mathrm{\lambda_i} \tag{2}$$

3.2.3 Wilk's Lambda

Wilk's Λ was the first MANOVA test statistic developed and is very important for several multivariate procedures in addition to MANOVA.

$$\Lambda = det(E)/det(H + E) \tag{3}$$

3.2.4 Roy's largest root

This gives an upper bound for the F statistic.

$$Roy'sLargestroot = max(\lambda_i) \tag{4}$$

or the maximum eigenvalue of A.

In some cases, the four tests will produce F identical statistics and identical probabilities. They will differ in others. Pillai's trace is frequently used when they differ since it is the most powerful and robust. Roy's largest root is an upper bound on F, hence it provides a lower bound estimate of F's probability. As a result, when Roy's largest root is significant but the others are not, it is often disregarded.

3.3 MANOVA vs ANOVA

1. Many dependent variables are accounted for in a single experiment.
2. Which one of the independent variables is truly important can be identified.
3. It is possible to avoid the type 1 error that happens when multiple ANOVAs are conducted. Each dependent variable is subjected to a series of ANOVA tests. By performing a series of ANOVAs separately for each dependent variable, we increase the Type I error. This leads to a more significant outcome, even if there is no difference between the groups in reality.
4. Multiple ANOVA does not possibly show significant effect on dependent variables whereas taking in combination may show significant effect (interaction).
5. MANOVA can capture the relationship between outcome variables by considering all dependent variables in the same analysis. In regard to this point, ANOVA can only tell us if groups differ along a single dimension, whereas MANOVA can tell us if groups differ along multiple dimensions.

3.4 Example

Consider a plant dataset of 4 varieties with their associated phenotypic properties of canopy volume and plant height. We can intuitively understand that the two properties, height and canopy volume are dependent on each other. Here, we wish to see if these properties are associated with different plant varieties using MANOVA. The dataset and the analysis code can be found on the github repository here.

plant_var	Value	Num DF	Den DF	F Value	Pr > F
Wilks' lambda	0.0797	6.0000	70.0000	29.6513	0.0000
Pillai's trace	1.0365	6.0000	72.0000	12.9093	0.0000
Hotelling-Lawley trace	10.0847	6.0000	44.9320	58.0496	0.0000
Roy's greatest root	9.9380	3.0000	36.0000	119.2558	0.0000

Figure 4: Results of MANOVA on the plants dataset

After performing the preliminary tests for verifying the fulfillment of assumptions like multivariate normality (ensuring univariate normality), Homogeneity of the variance-covariance matrices (Box's M test), Multicollinearity and Linearity, we perform One-way MANOVA on the data using Python3 language. The figure below shows results as obtained by us (the code is linked in the appendix).

From all the above test statistics, since Pillani's trace is considered the most robust, we make conclusions based on it. Since it is statistically significant, with a value of 1.0365 and $F(6, 72) = 12.90$ and p ¡ 0.01, we can safely conclude that plant varieties have a statistically significant relation with both canopy volume and plant height.

4 Two - Way ANOVA

4.1 Theory

The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two.

ANOVA tests for significance using the F-test. It compares the variance in each group mean to the overall variance in the dependent variable. If the variance within groups is smaller than the variance between groups, the F-test will find a higher F-value, and therefore a higher likelihood that the difference observed is real and not due to chance.

A two-way ANOVA tests three null hypotheses at the same time:

1. There is no difference in the dependent variable by the first independent variable.
2. There is no difference in the dependent variable by the second independent variable.
3. There is no interaction effect between both the independent variables on the dependent variable.

4.2 Calculation

Two-way ANOVA makes all the following assumptions:

1. The variation around the mean for each group being compared should be similar among all groups.
2. The independent variables should not be dependent on one another and the dependent variable should represent unique observations.
3. The dependent variable should follow Normal distribution i.e. it should follow a Bell curve.

Let these independent variables be called factors. Each of these factors have certain number of levels. Let factor A have a levels and factor B have b levels. For factor,

$$df_A = a - 1$$
$$df_B = b - 1$$

Now,

$$SS_{Total} = SS_A + SS_B + SS_{Interaction} + SS_{Error}$$

Compiling all the required metrics below:

Source	**df**	**SS**	**MS**	**F**
Factor A	a-1	$SS_A = \sum_{i=1}^{a} n_i(y_i - \overline{y})^2$	$\frac{SS_A}{a-1}$	$\frac{MS_A}{MS_E}$
Factor B	b-1	$SS_B = \sum_{j=1}^{b} n_j(y_j - \overline{y})^2$	$\frac{SS_B}{b-1}$	$\frac{MS_B}{MS_E}$
Error within groups	(a-1)(b-1)	$SS_{Error} = SS_{Total} - (SS_A + SS_B)$	$MS_{Error} = \frac{SS_{Error}}{(a-1)(b-1)}$	
Total	N-1	$SS_{Total} = \sum_{i=1}^{a} \sum_{j=1}^{b}(y_{ij} - \overline{y})^2$		

$$SS_{Interaction} = SS_{Total} - (SS_A + SS_B + SS_{Error})$$

We can then proceed to find the corresponding p-values for the F-Statistic found and then compare it against the chosen significance level to conclude whether to reject H_o or not.

4.3 Example

A botanist is researching the optimum conditions for crop growth. She obtains a dataset containing crop yield data from a region, corresponding to a particular crop and weather conditions. Farmers in that region have used 3 fertilizers (Type 1, Type 2, and Type 3) and have planted their crops in two planting densities (1 = Low density and 2 = High density). At harvest time, the measure of crop yield is in Bushels per Acre.

Let's look at the summary of data:

Data	Type	Possible Values
Fertilizer	Categorical	1,2 and 3
Planting Density	Categorical	1 and 2
Crop Yield	Continuous (float)	Range

To get a sense of the data, refer to the graphs in 6.2.
We'll be dealing simultaneously with **3 Null Hypotheses** here:

1. H_{o1}: Type of Fertilizer used has no effect on Crop Yield.
2. H_{o2}: Planting density has no effect on Crop Yield.
3. H_{o3}: The interaction effect between Fertilizer used and Planting Density has no effect on Crop Yield.

4.3.1 Results of Two-way ANOVA in R:

After running two-way ANOVA in R, we get the following:

	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F-value	p-value
Density	1	5.1217	5.1217	15.2297	0.0001811
Fertilizer	1	5.7432	5.7432	17.0779	7.902e-05
Interaction between Density and Fertilizer	1	0.1502	0.1502	0.4466	0.5056303
Residuals	92	30.9391	0.3363		

4.3.2 Conclusion of Two-Way ANOVA:

After looking at F-statistic and p-values above, we conclude the following:

1. We can reject the first null hypothesis H_{o1} (*"Type of Fertiliser has no effect on Crop Yield"*) since F-statistic = 17.0779 and p-value < 0.001.
2. We can reject the second null hypothesis H_{o2} (*"Planting density has no effect on Crop Yield"*) since F-statistic = 15.2297 and p-value < 0.001.
3. However, we cannot reject the third null hypothesis H_{o3} (*"Interaction effect between Fertilizer used and Planting Density has no effect on Crop Yield"*) since F-value = 0.4466 and p-value = 0.505.

4.4 3-Way ANOVA:

Just like 2-way ANOVA, 3-way ANOVA deals with 3 independent variables or factors. So, for the aforementioned example, if the dataset had another feature, say "Rainfall", then 3-way ANOVA could have been used to find the effect of these 3 factors, along with pair-wise interaction effects and the cumulative effect of all three factors on the Crop Yield. 3-Way ANOVA has widely known applications in Social Science and Finance.

5 Conclusion

This report throws light on ANOVA and its different variants useful for various situations.
ANOVA is a collection of statistical models and estimate techniques for analysing differences between group means in a sample. It looks at how two or more groups interact quantitatively with one another. It's used to see if there are any statistically significant differences between two or more independent groups' means. We utilise variance-like quantities in ANOVA to investigate whether population means are equal or unequal.
We have also gone through all the different specific variants of ANOVA, along with detailed descriptions on assumptions, methodology, examples, etc. We have seen how ANCOVA seeks to identify and removes the effects of certain variables. Although the approach is similar to that of ANOVA, the purpose for which it is used is somewhat different. The difference between 2 way ANOVA and ANOVA is that we study two factors at the same time instead of one. In 3-way ANOVA, the number of independent variables or factors in focus increments to 3. In case of MANOVA, it has more than one dependent variables, which is different from single variable case of ANOVA.

We conclude that each approach is unique in its own way and may be used in a variety of real-world situations.

6 Appendix

6.1 Tests Used

6.1.1 Test for Normality

For small sample sizes, normality can be tested for using the following descriptive statistics:

Central tendency: The mean and median are considered a summary measure of central tendency. In a normal distribution the mean is equal to the median

Skewness: Measure of asymmetry of a distribution. The normal distribution has a skewness value of zero. As a rule of thumb, skewness between -2 and +2 is consistent with a normal distribution.

Kurtosis: Kurtosis is a measure of whether observations cluster around the mean of the distribution or in the tails of the distribution. The normal distribution has a kurtosis value of zero. As a rule of thumb, kurtosis between -2 and +2 is consistent with a normal distribution.

6.1.2 Test for multicollinearity

If a covariate is highly related to another covariate (which occurs at a correlation of 0.5 or more), then it will not adjust the dependent variable over and above the other covariate. One of them should be removed since they are statistically redundant. So finally we need to run linear regressions between each pair of possible pairs of CVs, and remove one from each pair that has a good correlation.

6.1.3 Test for homogeneity of variance

Tested by Levene's test of equality of error variances. If the Levene test is positive ($P < 0.05$) then the variances in the groups are different (the groups are not homogeneous), and therefore the assumptions for ANCOVA are not met. To explain this, let us first look at residuals in Linear Regression. A residual is basically the vertical distance of a point from the regression line. What ANCOVA does internally, is that it runs an ANOVA test on the residuals from the Linear Regression of CV with DV and CV*IV with DV. Hence, to satisfy the ANOVA assumption of equal variance, this test needs to be performed.

6.1.4 Test the Homogeneity of regression slopes

We run an ANCOVA analysis to check if the CV affects our IV, by including the CV*IV and the IV terms. The Minitab software uses mediation analysis to check if the CV affects the IV's effect on the DV.

6.1.5 Run the ANCOVA test

You need to now simply run a Linear Regression Model with the IV as the Independent Variable and the DV as the dependent Variable, along with another model for the DV vs. CV. You need to note the R-squared values and see if the Explained-Variance (i.e. the contribution of each of IV and CV) is significant.

6.2 Two-way ANOVA Dataset

The data was sourced from Rebecca Bevans (PhD, Soil ecology) Scribbr Page. Here's the link to the dataset: Dataset Zip File

The graphs below show the distribution of Crop Yield with respect to the planting density, fertilizer used and the interaction effect between the two variables.

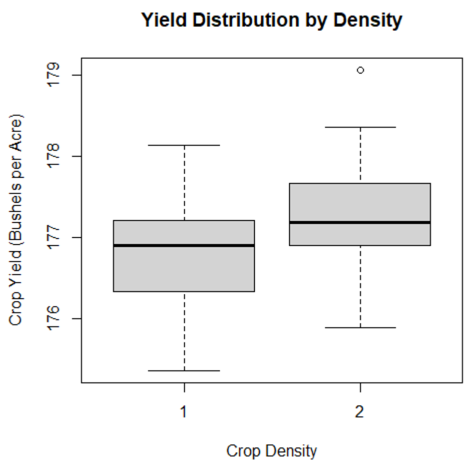

Figure 5: Box and Whisker plot for Crop Yield distribution with respect to Planting Density; generated using R

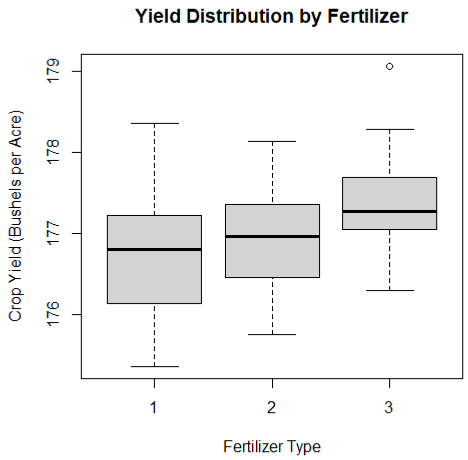

Figure 6: Box and Whisker plot for Crop Yield distribution with respect to type of Fertilizer; generated using R

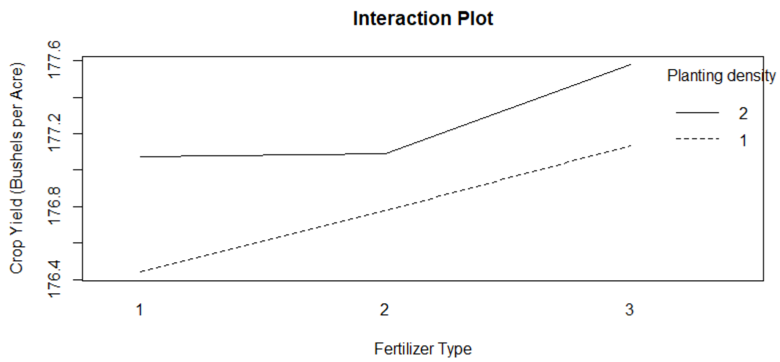

Figure 7: Line Graph for Interaction Effects between Fertilizer and Planting Density on Crop Yield; generated using R

7 References

[1] D.R. Anderson, D.J. Sweeney, T.A. Williams, J.D. Camm, and J.J. Cochran. *Statistics for Business & Economics, Revised.* Cengage Learning, 2017.

[2] Renesh Bedre. Manova using R. https://www.reneshbedre.com/blog/manova.html. Accessed 2021-10-08.

[3] Rebecca Bevans. Two-way anova. https://www.scribbr.com/author/beccabevans/. Accessed: 2021-10-07.

[4] Brandon Foltz. Pl13 anova (analysis of variance). https://www.bcfoltz.com/blog/video-companion-guides/pl13-anova-analysis-of-variance/. Accessed: 2021-10-05.

[5] Lisa Sullivan. Hypothesis testing - analysis of variance (anova). https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_hypothesistesting-anova/bs704_hypothesistesting-anova_print.html. Accessed: 2021-10-05.

[6] Lisa Sullivan. One-way analysis of variance: Example. https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_hypothesistesting-anova/bs704_hypothesistesting-anova_print.html. Accessed: 2021-10-05.

[7] Stat Trek. One-way analysis of variance: Example. https://stattrek.com/anova/completely-randomized/one-way-example.aspx. Accessed 2021-10-07.