IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 14, Issue 4 Ver. II (Jul - Aug 2018), PP 58-67
www.iosrjournals.org

One Way Analysis of Variance: A Practical Guide with R

[1]Adeniran, A. T.,[2]Faweya, O. and[3]Balogun, K.
[1]*Department of Mathematical Sciences, Augustine University Ilara-Epe, Lagos State, Nigeria.*
[2]*Department of Mathematical Science, Ekiti State University, Ado-Ekiti, Nigeria.*
[3]*Federal School of Statistics, Ibadan, Oyo State, Nigeria*
Corresponding Author:Adeniran, A. T.

Abstract: *In this paper, we demonstrated with example the practical application of R-software to solve Analysis of Variance (ANOVA) problem. Data importation from excel, checking of ANOVA assumptions, descriptive statistics of the data by treatment groups, obtaining ANOVA model and post-hoc analysis were all explored.*

Keywords:*ANOVA, descriptive statistics, treatments, p-value, post-hoc.*

Date of Submission: 25-07-2018 Date of acceptance: 06-08-2018

I. Introduction

When experiments are designed with the analysis in mind, the researcher can before commence the experiments, identify those sources of variation that he considers important and can choose a design that will allow him to measure the extent of the contribution of these sources to the total variation. In many different types of experiments, with one or more treatments (what the experimenter want to compare), one of the most widely used statistical methods is analysis of variance or simply ANOVA. Today, ANOVA is a commonly used statistical technique in many disciplines but finds its widest application in the analysis of data derived from experiments for investigating data by comparing the means of subsets of the data. Some areas of application of ANOVA are;

- **Agriculture:**compare varieties of fertilizers on growth plant, compares breeds of animal with a particular vaccine and so on.
- **Biological and chemical sciences:** compares different level of concentration of a particular chemical on specimens e.g. different concentration of glucose on the amount of insulin released from experimental animals.
- **Education:** compare various teaching methods (treatments) on students (experimental units) academic performance (response).
- **Pharmacy:**compares effect of different malaria drugs on patients.

The simplest ANOVA is the one way or one factor or single-classification or single factor or completely randomized design which is an extension of two-sample *t* test for independent groups covering situations where there are more than two populations, or data from experiments with more than two treatments being compared. The researcher interest is to test the null hypothesis of no difference among several population means (population means ≥ 3). That is, when we are conducting an analysis of variance, the null hypothesis considered is that there is no difference in treatments mean for at-least a pair, so once rejected, the question is which pair(s) of treatment differ? then "post-hoc".

In one-way ANOVA the data is sub-divided into groups based on a single classification factor with the assumption that experimental units are homogeneous (similar in characteristics) and the standard terminology used to describe the set of factor levels is treatment even though this might not always have meaning for the particular application. There is variation in the measurements taken on the individual components of the data set and ANOVA investigates whether this variation can be explained by the grouping introduced by the classification factor. Therefore, ANOVA can be described as a technique or arithmetic procedure of partitioning the total variation exhibited or present in a set of data into several components, associated with each of these components (treatments, blocks, error) is a specific source of variation so that in the analysis it is possible to ascertain the numerical magnitude of the contribution of each of these sources to the total variation.

The model for one-way ANOVA is

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}; \quad i = 1,2,\cdots,k \quad and\ j = 1,2,\cdots,n_j \qquad (1)$$

where,

Y_{ij} =is the jth observation receiving ith treatment

μ = grand or overall mean,

τ_i =treatment effect; amount by which a group mean differs from the grand mean.$\tau_i = \mu_i - \mu$
ε_{ij} =error term; the amount by which any value differs from its group mean. That is, $\varepsilon_{ij} = y_{ij} - \mu\_i$
In this paper, the objective is neither to study mathematical theory of ANOVA nor modify the conventional Fisher's Snedecor distribution, but to show how to apply the procedure using R. To test for the assumptions, obtain the descriptive statistics, run the ANOVA model and conduct post-hoc (Fisher's Least Significant Difference, Tukey's and Bonferroni) analysis if need be using R-software.

II. Methodology

2.1. Data Importation from Excel

Enter the data into excel workbook as shown in figure (1) below and save (say, *anovex1*) in csv format

```
mydata=read.csv("anovaex1.csv",header=T)
mydata
attach(mydata)
fertilizers = factor(mydata$type,labels = c("Fert1", "Fert2", "Fert3", "Fert4"))
```

2.2 Test of ANOVA Assumptions

The basic parametric assumptions of ANOVA are: Normality, Homogeneity and Independence of error components.

2.2.1 Normality of Error Assumption Using Shapiro-Wilk test and plot

We will perform Shapiro-wilk test for normality of errors term

```
normality_test1<-shapiro.test(residuals(lm(amount~type,data=mydata)))
normality_test1 or
normality_test2<-hist(residuals(lm(amount~type,data=mydata)))
```

2.2.2 Model Checking Plots

```
 par(mfrow=c(2,2))
plot(lm(amount~type, data=mydata))
```

The model residuals was plotted against the fitted values to investigate the model assumptions. First we create a data frame with the fitted values, residuals and treatment identifiers:

```
fertilizers.mod = data.frame(Fitted = fitted(anov.model), Residuals = resid(anov.model), Treatment = mydata$type)
ggplot(fertilizers.mod, aes(Fitted, Residuals, colour = Treatment)) + geom_point()
```

2.2.3 Homogeneity of Variance Assumption Using Bartlett Test and Levene's Test

```
homog.var.test1<-bartlett.test(amount~type,data=mydata) # Or
library(car)
homog.var.test2<-leveneTest(amount~type,data=mydata)
homog.var.test1
homog.var.test2
```

2.3 Descriptive Statistics

Researchers' may be interested to have the descriptive statistics (the mean, the standard deviation, the sample size n or the number of times each treatment is being replicated, coefficient of variation, confidence intervals) for all the set of treatment (fertilizers) groups.

```
library(plyr) # download this package from R package repository
fert.summary<- ddply(mydata, "type",
function(X){
data.frame( m=mean(X$amount),
std.dev=sd(X$amount),
n=length(X$amount))})
fert.summary$se<- fert.summary$s/sqrt(fert.summary$n)
fert.summary$cv<- (fert.summary$se/fert.summary$m)*100
fert.summary$ci.lower<-fert.summary$m-qt(1-0.05/2,df=fert.summary$n-1)*fert.summary$se
fert.summary$ci.upper<-fert.summary$m+qt(1-0.05/2, df=fert.summary$n-1)*fert.summary$se
fert.summary
```

2.4 Boxplot

A boxplot of the distributions of the growth rate for all the fertilizer groups (individual points, mean and CI) is created using the *ggplot* package as shown below:

```
library(ggplot2)
p<-ggplot(mydata,aes(x=type, y=amount))
p<- p+geom_hline(yintercept=mean(mydata$amount),
colour="black",linetype="dashed", size=0.3,alpha=0.5)
p<-p+geom_boxplot(size=0.75, alpha=0.5)
p<- p+geom_point(position=position_jitter(w=0.05, h=0),alpha=0.5)
p<- p+ stat_summary(fun.y =mean, geom="point", shape=18, size=6,
aes(colour=type),alpha=0.8)
p<- p+ stat_summary(fun.data="mean_cl_normal", geom="errorbar",
width=.2, aes(colour=type), alpha=0.8)
p<-p+labs(title="Combined boxplot of the data by treatment groups")+ ylab("amount of growth in cm")
print(p)
```

or simply as

```
require(ggplot2)
ggplot(mydata, aes(x =type, y = amount)) +
geom_boxplot(fill = "grey80", colour = "blue") +
scale_x_discrete() + xlab("Treatment Group (i.e types of fertilizers)") +
ylab("Growth of the maize")+
labs(title="Combined boxplot of the data by treatment groups")
```

2.5 ANOVA Model

Finally, we run ANOVA model to assess whether there are differences between pair(s) of fertilizers.

```
anov.model<- lm(amount~type, data=mydata)
anova(anov.model) #or simply as
anov.model<- aov(amount~type, data=mydata)
summary(anov.model) #or as
anov.model<- with(mydata,lm(amount~type))
summary.aov(anov.model)
```

The function *confint*is used to calculate confidence intervals on the treatment parameters, by default 95% confidence intervals:

```
CI<-confint(anov.model)
CI
```

2.6 Post-Hoc Analysis

The ANOVA F-test checks whether all the population means are equal. Post-Hoc tests are often used as a follow-up to a significant ANOVA F-test to determine which population means are different. This study will discuss Fisher's Least Significance Difference (FLSD), Tukey's and Bonferroni's test for comparing all pairs of means.

2.6.1 Fisher's Least Significance Difference (FLSD)

One way to get Fisher comparisons in R uses *pairwise.t.test()* with *p.adjust.method*. The resulting summary of the multiple comparisons is in terms of p-values for all pairwise two-sample t-tests using the pooled standard deviation from the ANOVA, *pool.sd=TRUE*. This output can be used to generate groupings. An easy way to compare all pairs of treatments is to order the sample by their sample means. The sample can then be grouped easily, noting that two treatments are in the same group if the absolute difference between their sample means is smaller than the FLSD. The pairwise comparisons using t-tests with pooled standard deviation is

$$FLSD = t_{\frac{\alpha}{2}} \times S_{pooled} \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \qquad (2)$$

If all sample sizes are equal,$(n_A = n_B = \cdots, n_k = n)$ then equation (2) reduces to (3) which is the same for each comparison.

$$FLSD = t_{\frac{\alpha}{2}} \times S_{pooled} \sqrt{\frac{2}{n}} \qquad (3)$$

The minimum absolute difference between$\overline{y_A} - \overline{y_B}$ needed to reject H_o is the FLSD. That is, there is a significant difference between pair A and B if

$$|\overline{y_A} - \overline{y_B}| \geq t_{\frac{\alpha}{2}} \times S_{pooled} \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = FLSD \quad (4)$$

The treatment group means is obtained using
comb_mean<-tapply(mydata$amount, mydata$type, mean)
comb_mean

Therefore, multiple comparisons for all pairwise comparisons among levels of fertilizers using FLSD in *R* is
LSD_posthoc<-pairwise.t.test(mydata$amount, mydata$type, pool.sd=TRUE, p.adjust.method="none")
LSD_posthoc

2.6.2 Post-Hoc Analysis Using Tukey Test

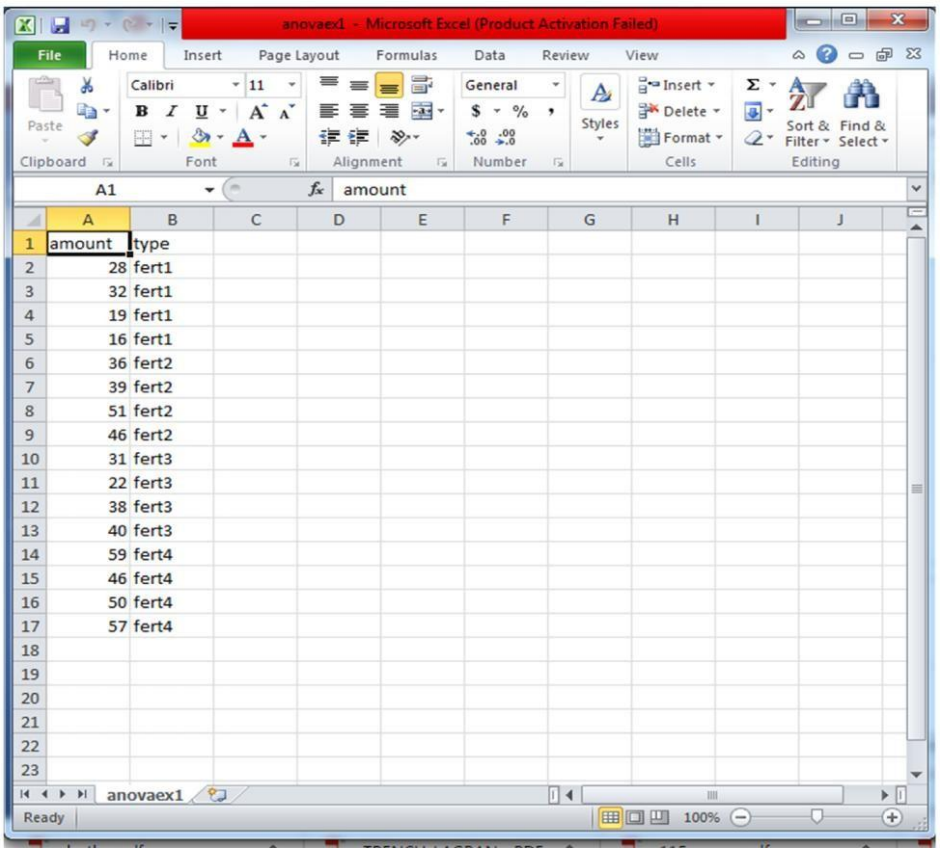
anov.mod<- aov(amount~type)
Tukey_posthoc<- TukeyHSD(x=anov.mod, 'type', conf.level=0.95)
Tukey_posthoc## or simply as
TukeyHSD(anov.mod)

2.6.3 Post-Hoc Analysis Using Bonferroni Test

Assuming all comparisons are of interest, you can implement the Bonferroni adjustment in R by specifying *p.adjust.method="bonf"*. A by-product of the Bonferroni adjustment is that we have at least$100(1-\alpha)\%$ confidence that all pairwise t-test statement holds simultaneously. Bonferroni 95%individual p-values for all pairwise comparisons among levels of fertilizersis obtained by
Bonferroni_posthoc<-pairwise.t.test(mydata$amount, mydata$type,
pool.sd=TRUE, p.adjust.method="bonf")
Bonferroni_posthoc

III. Results And Discussion

3.1 Data Importation from Excel



	A	B
1	amount	type
2	28	fert1
3	32	fert1
4	19	fert1
5	16	fert1
6	36	fert2
7	39	fert2
8	51	fert2
9	46	fert2
10	31	fert3
11	22	fert3
12	38	fert3
13	40	fert3
14	59	fert4
15	46	fert4
16	50	fert4
17	57	fert4

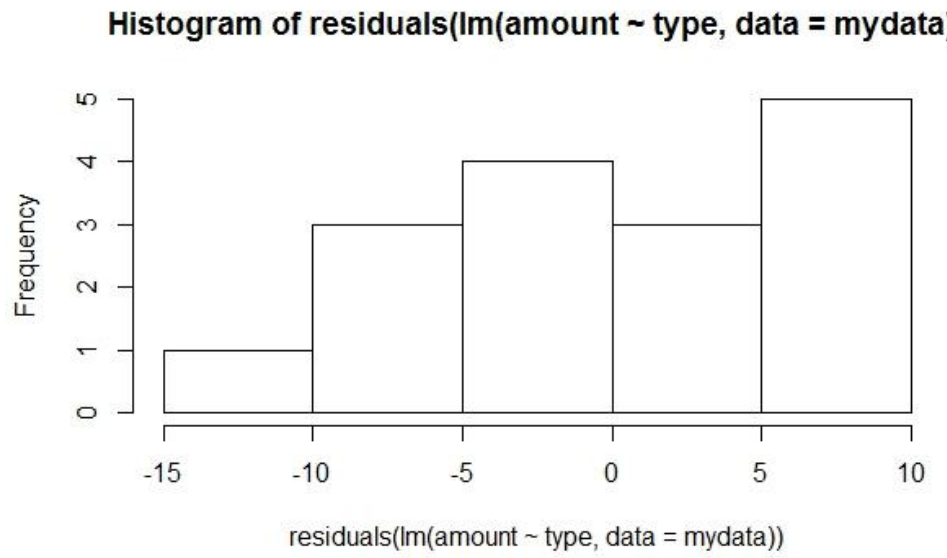
3.2 Test of ANOVA Assumptions

3.2.1 Normality of Error Assumption Results

We will perform Shapiro-wilk test for normality of errors term and visual inspection of normality through histogram

Shapiro-Wilk normality test
data: residuals(lm(amount ~ type, data = mydata))
W = 0.91155, p-value = 0.1233

Interpretation: The result here is Shapiro-Wilk normality test of the errors component W=0.91155, p-value=0.1233 indicating the test is not significant. That is, we do not reject the null hypothesis (H_o) which state H_o: errors are distributed normally Vs H_1 error terms are not normally distributed.



The figure above (an alternative approach for checking normality) help us to obtain a visual inspection of the nature of normality which conforms with the Shapiro-Wilk result.

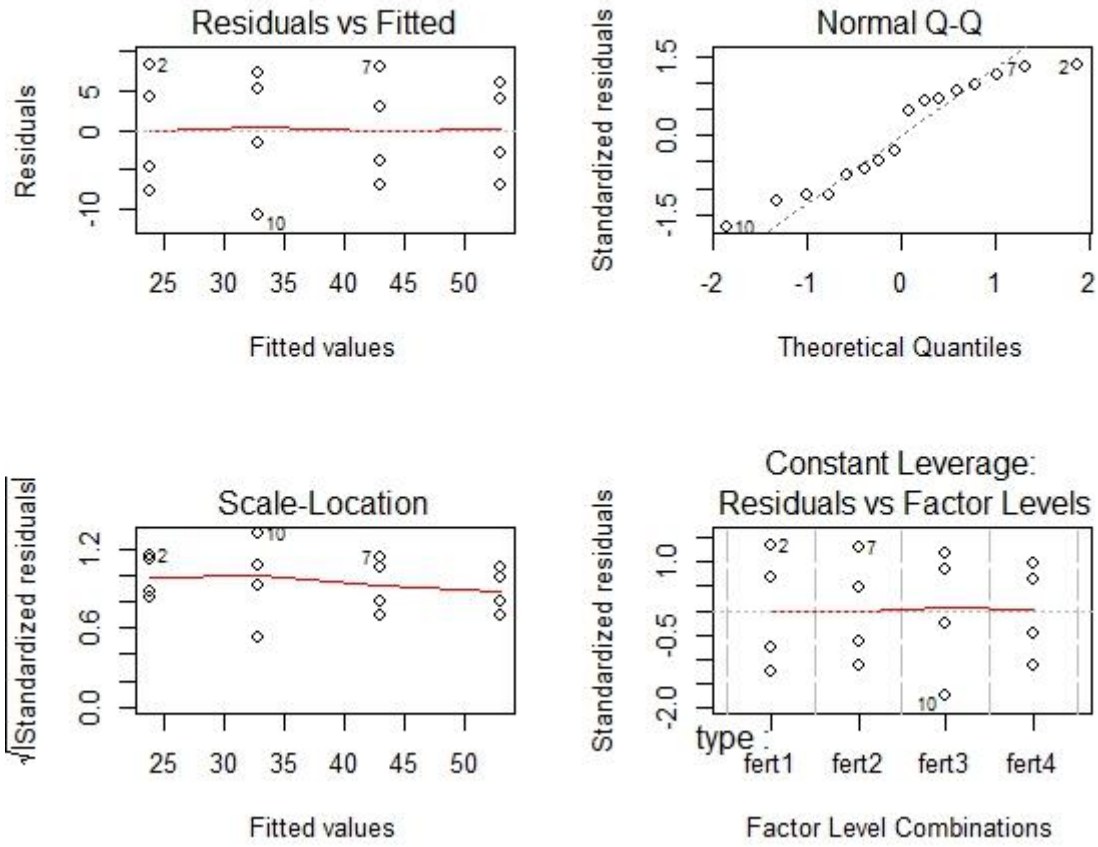
3.2.2 Homogeneity of Variance Assumption Using Bartlett Test and Levene's Test

Bartlett test of homogeneity of variances
data: amount by type
Bartlett's K-squared = 0.25383, df = 3, p-value = 0.9685
Or
Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	3	0.1865	0.9035
	12		

Interpretation: The result in this case is Bartlett's K-squared = 0.25383, df=3, p-value = 0.9685. p-value>0.05, the test is not significant, i.e. there is no sufficient evidence to reject the null hypothesis which state that H_o variances assumed equal Vs H_1 Variances assumed not equal. Thesame interpretation goes for Levene's test for homogeneity of variances

3.2.3 Model Checking Plots



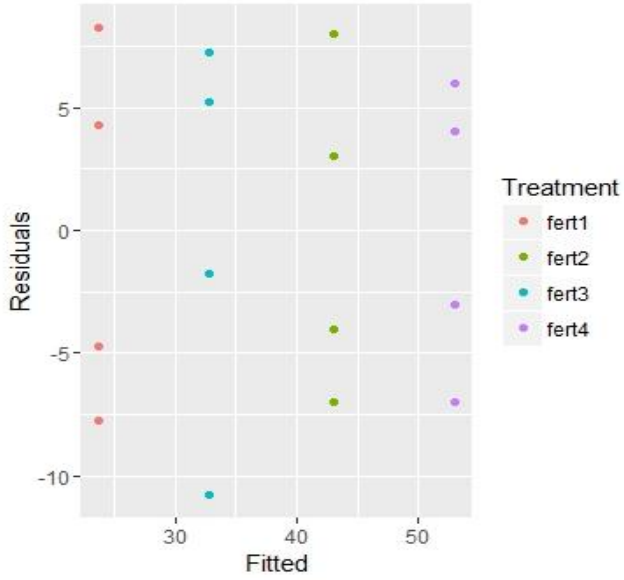
Interpretation of the Plot:

Residuals vs Fitted: This checks for a pattern in the residuals, and ideally should show similar scatter for each condition. Here, no worrying effect, there is *homoscedasticity*. There is a worrying effect if there are larger residuals for larger fitted values. This is called *heteroscedasticity* meaning that not only is variance in the response not equal across groups(types of fertilizers), but that the variance has some specific relationship with the size of the response. In fact, you could see this in the original boxplots. This is also separately illustrated in the Diagnostic plot in figure (4) below.

Normal QQ: This looks for normality of the residuals assumption. If they are not normal, the normality assumption of ANOVA is potentially violated. Here normality is achieved in corroboration with Shapiro-Wilk test and Histogram plot.

Scale-Location: This is like the first plot, but now to specifically test if the residuals increase with the fitted values, which they do not. Hence, no worrying effect.

Constant Leverage: This gives an idea of which levels of the factor(treatments) are best fitted. Here, is fertilizer3.



Interpretation: We can see that there is no major problem with the diagnostic plot but some evidence of different variabilities in the spread of the residuals for the four treatment groups.

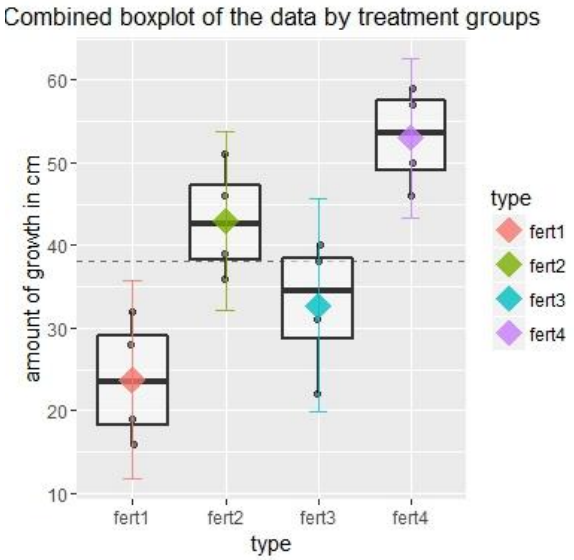
Remark:

Theoretically speaking, whenever any of these assumptions is not met, the ANOVA technique cannot be employed to yield valid inferences. However, in some situations, departure from one of these assumptions does not markedly affect conclusions based on F-test. For example, looking for exact normality is a bit of a red herring because, we also have the Central Limit Theorem (CLT) that says that if the errors are not normal but still identically and independently distributed then the distribution of the coefficients will approach normality as the sample size increases. This is what make statistics doable because no real dataset entered into the computer is perfectly normal. The more important question is, are the residuals "normal enough"? for which there is no a definitive test (experience and plots help).

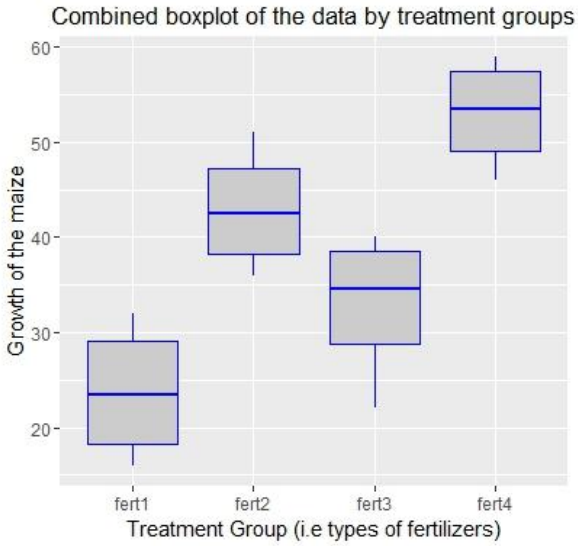
3.3 Descriptive Statistics

	type	m	std.dev	n	se	cv	ci.lower	ci.upper
1	fert1	23.75	7.500000	4	3.750000	15.789474	11.81583	35.68417
2	fert2	43.00	6.782330	4	3.391165	7.886430	32.20780	53.79220
3	fert3	32.75	8.139410	4	4.069705	12.426581	19.79838	45.70162
4	fert4	53.00	6.055301	4	3.027650	5.712548	43.36467	62.63533

3.4 Box plot



or simply as



Interpretation: Initial inspection of the data suggests that there are differences in the growth rate for the two treatments fert1 and fert3 but it is not so clear to conclude, hence, we run ANOVA model.

3.5 ANOVA Model

Call:
lm(formula = amount ~ type, data = mydata)
Residuals:

Min	*1Q*	*Median*	*3Q*	*Max*
-10.750	*-5.312*	*0.625*	*5.438*	*8.250*

Coefficients:

	Estimate	*Std. Error*	*t value*	*Pr(>|t|)*
(Intercept)	*23.750*	*3.581*	*6.632*	*2.42e-05 ****
typefert2	*19.250*	*5.064*	*3.801*	*0.00252 ***
typefert3	*9.000*	*5.064*	*1.777*	*0.10087*
typefert4	*29.250*	*5.064*	*5.776*	*8.80e-05 ****

---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Residual standard error: 7.162 on 12 degrees of freedom
Multiple R-squared: 0.7575, Adjusted R-squared: 0.6968
F-statistic: 12.49 on 3 and 12 DF, p-value: 0.0005312

Analysis of Variance Table

Response: amount

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	3	1922.2	640.75	12.492	0.0005312 ***
Residuals	12	615.5	51.29		

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

Interpretation: This table confirms that there are differences between the groups (type of fertilizers) which were highlighted in the model summary. From the output we see that the p-value is 0.0005312 for fertilizers indicating that the fertilizers have significant effect on the response. This is desirable since it is expected that the fertilizers can affect growth at different rate. Also in the table we see that the ANOVA p-value for the type of fertilizers is highly significant, indicating the difference between them.

The function *confint* is used to calculate confidence intervals on the treatment parameters, by default 95% confidence intervals:

	2.5 %	**97.5 %**
(Intercept)	15.947867	31.55213
typefert2	8.216118	30.28388
typefert3	-2.033882	20.03388
typefert4	18.216118	40.28388

3.6 Post-Hoc Analysis

3.6.1 Fisher's Least Significance Difference Results

Treatment Combined Mean

Treatments:	fert1	fert2	fert3	fert4
Means:	23.75	43.00	32.75	53.00

Pairwise comparisons using t tests with pooled SD

data: example$amount and example$type

	fert1	*fert2*	*fert3*
fert2	*0.0025*	-	-
fert3	*0.1009*	*0.0658*	-
fert4	*8.8e-05*	*0.0718*	*0.0018*

P value adjustment method: none

Interpretation: The output above indicate that there is no significance difference between pairs fert1-fert3, fert2-fert3 and fert2-fert4 since their r$p - value > 0.05$ but there is significance difference between pairs fert1-fert2, fert1-fert4 and fert3-fert4 since their$p - value < 0.05$. We judge the significant based on their p-value being$> or < \alpha$ (significance level).

3.6.2 Tukey Test Results

posthoc

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = amount ~ type)

$type

	diff	*lwr*	*upr*	*p adj*
fert2-fert1	*19.25*	*4.214975*	*34.285025*	*0.0116806*
fert3-fert1	*9.00*	*-6.035025*	*24.035025*	*0.3298460*
fert4-fert1	*29.25*	*14.214975*	*44.285025*	*0.0004417*
fert3-fert2	*-10.25*	*-25.285025*	*4.785025*	*0.2327916*
fert4-fert2	*10.00*	*-5.035025*	*25.035025*	*0.2502188*
fert4-fert	*20.25*	*5.214975*	*35.285025*	*0.0082805*

Interpretation: This output indicates that the differences fert1-fert2, fert1-fert4 and fert3-fert4 are significant, while fert1-fert3, fert2-fert3 and fert2-fert4 is not significant. An easier way to interpret this output is visualizing the confidence intervals for the mean differences. That is, one can see that fert1-fert2, fert1-fert4 as well as fert3-fert4 differ significantly. How? because the interval does not contain 0. The confidence intervals for fert1-fert3, fert2-fert3 and fert2-fert4 contain 0. Thus, it appears that the pairs do not differ among themselves.

3.6.3 Bonferroni Test Result

Pairwise comparisons using t tests with pooled SD

data: mydata$amount and mydata$type

	fert1	*fert2*	*fert3*
fert2	*0.01515*	-	-
fert3	*0.60521*	*0.39490*	-
fert4	*0.00053*	*0.43063*	*0.01060*

P value adjustment method: bonferroni

Interpretation: Thesame as Fisher'sLeast significance difference.

IV. Conclusion

This study is able to achieve its set objective. The material should be of pedagogical interest to researchers whose data layout follows analysis of variance and intended to use R. In addition, it can serve as an excellent teaching reference in computing classes where only introduction to R-environment and basic R-code are the only background requirements. The procedures and results discussions are straightforward and require only understanding of some element of statistical inference.

Acknowledgement

The authors thank the editor and reviewers for their careful reading and some kindly comments and suggestions on improving the presentation of this paper.

References

[1]. Adebowale, S. A. (2006). Statistics for Engineers, Managers and Scientists, Alfredo Graphics Limited, ISBN-978-06043-2-4.
[2]. Brian, S. E. and Torsten, H. (2010). *A Handbook of statistical Analysis Using R*, Taylor & Francis Group, LLC, ISBN 978-1-4200-7933-3.
[3]. Christophe, L. (2005). *R Companion to Montgomery's Design and Analysis of Experiments*.
[4]. Erik, B. E., Edward, J. B. and Ronald, M. S. (2013). *Advanced Data Analysis 1 (ADA1), Stat 427/527*, Fall 2013, University of New Mexico
[5]. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*, Springer, ISSN 1431-875X, ISBN 978-1-4614-7137-0, ISBN 978-1-4614-7138-7 (eBook), DOI 10.1007/978-1-4614-7138-7.
[6]. Gupta, S. P. (2008).*Statistical Methods*, Sultan Chand & Sons Educational publishers, New Delhi,36th edition.
[7]. Mood, A. M. and Graybill, F. A. (1963). *Introduction to the Theory of Statistics*, McGraw Hill Book Company, 2nd edition.

Adeniran, A. T."One Way Analysis of Variance: A Practical Guide with R" IOSR Journal of Mathematics (IOSR-JM) 14.4 (2018) PP: 58-67.