International Journal of Mathematics and Computer Research
ISSN: 2320-7167
Volume 09 Issue 10 October 2021, Page no. – 2408-2422
Index Copernicus ICV: 57.55, Impact Factor: 7.184
DOI: 10.47191/ijmcr/v9i10.04

Analysis of Variance: The Fundamental Concepts and Application with R

Adeniran, A. T.[1], Olilima, J. O.[2], Akano, R. O.[3]
[1]Department of Statistics, University of Ibadan, Ibadan, Oyo State, Nigeria.
[2]Department of Mathematical Sciences, Augustine University Ilara-Epe, Nigeria.
[3]Department of Statistics, University of Abuja, Abuja, Nigeria.

ARTICLE INFO

Published Online:
15 October 2021

Corresponding Author:
Adeniran, A. T.
at.adeniran@ui.edu.ng

ABSTRACT

The Analysis of Variance (ANOVA) test has long been an essential tool for researchers conducting studies on multiple experimental groups with or without one or more control groups. This article encapsulates the fundamentals of ANOVA for an intended benefit of the reader of scientific literature who does not possess expertise in statistics. The emphasis is on conceptually-based perspectives regarding the use and interpretation of ANOVA results, with minimal coverage of the mathematical foundations. Data entry, checking basic parametric assumptions of ANOVA, descriptive statistics of the data by treatment groups, fitting ANOVA model, statistical significance of the test based on p-value, and post-hoc analysis are all explored using R-software.

KEYWORDS: ANOVA, experimental groups, control groups, treatments, parametric assumptions, statistical significance, p-value, post-hoc analysis.
AMS subject classification: 62B15, 62J10, 65C60

INTRODUCTION

An experiment occurs when purposeful changes are made to the input variables of a process or system so that we may observe and identify the reason for changes in the output ([1], [2]). For examples: in agricultural field trials, different type of fertilizer may be applied on crop to measure their effect on growth; in industry, different methods of production using a given raw materials to investigate which method gives the best product or output etc. When experiments are designed with the analysis in mind, the researcher can before conducting the experiments, identify 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. There is variation in the measurements taken on the individual units of the data set and ANOVA investigates whether this variation can be explained by the grouping introduced by the classification factor (i.e., the identified sources of variation) by partitioning (breaking down) the total variation exhibited or present in a set of data into several recognized components in the experiment, 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 fundamental problem of ANOVA is to test the null hypothesis that all of the population means (population means ≥ 3) are the same (of no difference), i.e.,

$$H_0\colon\ \mu_1 = \mu_2 = \cdots = \mu_k = \mu \qquad (1)$$

μ unspecified, against all possible alternative hypotheses (H_1) that they are not all the same for at least a pair.

$$H_1\colon\ \mu_1 \neq \mu_2 \neq \cdots \neq \mu_k. \qquad (2)$$

The statement in (2) that the population means are not identical does not imply that each population mean is distinct. Also, from equation (1) and (2), it is easy to see that ANOVA is an extension of t-test. While t-test procedure compare between two factor/group, ANOVA/F-test is a procedure of testing whether the means of three or more normal population with unknown but common variance are equal or testing whether a set of three or more samples can be considered as being drawn from a homogeneous normal population. Today, ANOVA is a commonly used statistical technique in many disciplines but finds its widest application in the analysis of data derived from experiments. Some areas of application of

ANOVA include but not limited to:

• **Agriculture:** comparison of varieties of fertilizers on growth plant, breeds of animal after administered with dose of a particular vaccine etc.

• **Biological and chemical sciences:** comparison of 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 ([3]).

• **Medicine and Pharmacy:** comparison of effect (response) of different malaria drugs (treatments) on patients (experimental material). See [3], [4], [5] and [6].

• **Education:** comparison of various teaching methods (treatments) on students (experimental units) academic performance (response).

• **Others:** Several application of ANOVA has been found in engineering, economics, commerce, trade and industry ([7], [8], [9], [10]) etc.

The simplest ANOVA model is the "one-way" or "one-factor" or "single-classification" or "completely randomized design". In one-way ANOVA, the data is sub-divided into k groups based on a single classification factor provided the experimental units are essentially homogeneous (similar in characteristics) that the variation among them is small and grouping them in blocks would make no difference. This is the case in many types of laboratory experiments where a quantity of material is thoroughly mixed and then divided into small lots to form experimental units to which treatments are randomly assigned or in plant and animal experiments where environmental effects are much alike. 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. This can be explained by the fact that most ANOVA techniques were originally in connection with agriculture experiments where fertilizers, for example, were regarded as treatment applied to the soil. Statistical model for completely randomized design or one-way ANOVA model is:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}; \quad i = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, n_i \quad (3)$$

where,

y_{ij} = individual observation in jth plot receiving ith treatment,

μ = grand or overall mean effect,

$\tau_i = \mu_i - \mu$ = ith treatment effect; amount by which a group mean differs from the grand mean,

$\varepsilon_{ij} = y_{ij} - \mu_i$ = experimental error term; the amount by which any value differs from its group mean.

Let us consider k normal distributions with unknown means $\mu_1, \mu_2, \cdots, \mu_k$, respectively, and an unknown but common variance σ^2. One inference that we wish to consider is a test of the equality of the k means (see equation 1). To test this hypothesis, let $Y_{i1}, Y_{i2}, \cdots, Y_{iki}$ represent a random sample of size k_i from the normal distribution $N(\mu_i, \sigma^2)$, $i = 1, 2, \cdots, k$. Table 1 illustrates the randomization procedure or framework for a one-way ANOVA design.

Table 1: The data layout under CRD

	Observations				**Total**
	Treatment1	Treatment2	⋯	Treatment k	
	y_{11}	y_{21}	⋯	y_{k1}	
	y_{12}	y_{22}	⋯	y_{k2}	
	⋮	⋮	⋮	⋮	
	y_{1n_1}	y_{2n_2}	⋯	y_{kn_k}	
$\sum_{j=1}^{n_i} y_{ij}$	$Y_{1.}$	$Y_{2.}$	⋯	$Y_{k.}$	$Y_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$
$\sum_{j=1}^{n_i} y_{ij}^2$	$Y_{1.}^2$	$Y_{2.}^2$	⋯	$Y_{k.}^2$	$Y_{..}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2$

Table 1 above displays a typical data presentation of completely randomized design or one-way analysis of variance model. From Table 1,

$$\bar{Y}_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} \quad and \quad \bar{Y}_{i.} = \frac{1}{n_j}\sum_{j=1}^{n_i} n_j y_{ij}, \quad i = 1,2,\cdots,k \quad (4)$$

The dot in the notation for the means, $\bar{Y}_{..}$ and $\bar{Y}_{i.}$, indicates the index over which the average is taken. Hence, $\bar{Y}_{..}$ is an average taken over both indices while $\bar{Y}_{i.}$ is just taken over the index j. The sum of squares associated with the variance of the combined samples is

$$SS_{Total} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \mu)^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 \quad (5)$$

To determine a critical region for a test of H_0, we shall first partition (5) into two parts by adding and subtract $\bar{y}_{.i}$ to get

$$SS_{Total} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(y_{ij} - \bar{y}_{.i}) + (\bar{y}_{.i} - \bar{y}_{..})\right]^2$$

which after simple algebra, gives

$$SS_{Total} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{.i})^2 + \sum_{i=1}^{k} n_i(\bar{y}_{.i} - \bar{y}_{..})^2 \quad (6)$$

The preceding equation (6) can be summarized as

$$SS_{Total} = SS_{Error} + SS_{Treatment} \quad (7)$$

which shows that the effects are additive. These sum of squares (total, treatment and error) can be re-written as:

$$SS_{Total} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^2}{N}, \quad (8)$$

$$SS_{Treatment} = \sum_{i=1}^{k}\left[\frac{\left(\sum_{j=1}^{n_i} y_{ij}\right)^2}{n_i}\right] - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^2}{N} \quad (9)$$

and from (7),

$$SS_{Error} = SS_{Total} - SS_{Treatment} \quad (10)$$

which are more convenient for computational purpose. Analysis of variance procedures rely on a distribution called the F-distribution, named in honor of Sir Ronald Fisher [1]. The F-statistic is

$$F_{statistic} = \frac{\frac{SS_{Treatment}}{k-1}}{\frac{SS_{Error}}{N-k}} = \frac{MS_{Treatment}}{MS_{Error}} \quad (11)$$

After all the necessary computations has been done, for brevity, the computations leading to the F-statistic are usually put in a tabular form and (Table 2) is called the **analysis of variance** table for one-way ANOVA model.

Table 2: A typical example of ANOVA table

Source of Variation	**degree of freedom (df)**	**Sum of squares (SS)**	**Mean Square (MS)**	**F-statistic**
Factor	$k-1$	$SS_{Treatment}$	$\frac{SS_{Treatment}}{k-1}$	$F = \frac{MS_{Treatment}}{MS_{Error}}$
Error	$N-k$	SS_{Error}	$\frac{SS_{Error}}{N-k}$	
Total	$N-1$	SS_{Total}		

Given a pre-assigned significance level (α), the critical value F_α with $df = (k-1, N-k)$ is:

$$F_{critical} = F(\alpha, df_{treatment}, df_{error}) = F(\alpha, k-1, N-k) \quad (12)$$

To make relevant inference, the rule is: reject H_0 if the value of the test statistic falls in the rejection region (i.e., F-statistic>F-critical); otherwise, do not reject H_0. When $k-1=1$, in other words, when $k=2$, the test-statistic equals the student t-statistic.

After conducting an analysis of variance test, we might conclude that there is sufficient evidence to reject a claim of equal population means ($H_0: \mu_i = \mu_j$ for all $i \neq j$), but we cannot conclude from ANOVA that any particular mean is different from the others. **Post-hoc** (which in Latin means "after this") tests explore differences between multiple group means while controlling the experiment-wise error rate ([1], [2]). There are several procedures for identifying which means differ from the others. Yet, no consensus on which test is best, but some of the common tests are: Least Significant Difference (LSD) or Fisher Least Significant Difference (FLSD) test, Dunn-Bonferroni test, Tukey test (or Tukey honestly significant difference test), Duncan test, Student-NewmanKeuls test (or SNK test), Scheffé test, Dunnett test etc. To use FLSD procedure, the theory compare the observed difference between each pair of mean to the corresponding LSD ([11], [12], [13]). The quantity LSD is given by

$$LSD = t_{\left(\frac{\alpha}{2}, N-k\right)}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \quad (13)$$

where, N = total number of observed values, k = number of distinct group/treatment, MSE = Mean square error from ANOVA, and n_i, n_j are the number of observation in ith group and jth group, respectively. In case where the groups sample sizes are equal, that is, $n_1 = n_2 = \cdots = n_k = n$,

$$LSD = t_{\left(\frac{\alpha}{2}, N-k\right)}\sqrt{\frac{2MSE}{n}}. \qquad (14)$$

If $|\bar{y}_i - \bar{y}_j| > LSD$, we conclude that the pairs of means are significantly different. Otherwise, it is not.

John Tukey's honest significant difference method is to reject the equality of a pair of means based, not on the t-distribution, but the studentized range distribution ([11], [12], [14]). To implement Tukey's method with a FER of α, reject $H_0: \mu_i = \mu_j$ if

$$\left|\bar{Y}_i - \bar{Y}_j\right| \geq \frac{q_{critical}}{\sqrt{2}}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \qquad (15)$$

where $q_{critical}$ is the α level critical value of the studentized range distribution.

In this study, the objective is neither to study mathematical theory of ANOVA nor modify the conventional Fisher's Snedecor distribution, but to demonstrate computation of necessary statistic and fitting of ANOVA model using R-software.

MANUAL COMPUTATION

For the sake of clarity and ease of calculation, a data set with an inappropriately small sample size (an unbalanced design, i.e., unequal sample sizes in each group) is used as illustration to achieve the aforementioned objective. In addition, it is very important to know that **R**, as well as many other software programs, are not substitute for your brain. Therefore, researchers are strongly advised to figure out from time-to-time whether **R-commands** do what they want. Hence, this justify the exploration of mathematical computation of ANOVA in this study.

Example 2.1: The drug *Scopolamine* is often used as a sedative to induce sleep in patients. Medical researchers, [15] examined *Scopolamine's* effect on memory for word-pair associates. A total of 28 human subjects were randomly divided into three treatment groups. Group 1 subjects were administered an injection of *Scopolamine*, group 2 subjects were given an injection of *Glycopyrrolate* (an active placebo), and group 3 subjects were not given any drug. Four hours later, subjects were tested on how many of the associated word pairs they could recall. The data on number of pairs recalled (based on summary information provided in the research article) are presented in Table (3) below.

Table 3: Amount of word recalled

Scopolamine	5	8	8	6	6	6	6	8	6	4	5	6
Glycopyrrolate (Placebo)	8	10	12	10	9	7	9	10				
No drug	8	9	11	12	11	10	12	12				

Conduct the ANOVA F-test on the data. Is there sufficient evidence (at $\alpha = 0.05$) to conclude that the mean number of word pairs recalled differs among the three treatment groups?

Solution: The model is a completely randomized design since classification is based on single factor (drug). The treatment is drug and response variable is the number of word pairs recalled. The hypotheses to be tested are:

$$H_0: \mu_1 = \mu_2 = \mu_3 \quad versus \quad H_1: \mu_1 \neq \mu_2 \neq \mu_3 \quad for\ at\ least\ a\ pair.$$

Table 4 presents a summary statistics of Table 3 indicating the number of times each treatment was replicated, sum and sum of squares of observations in each group.

Table 4: Summary statistics of amount of word recalled

	Drugs			
Statistic	**Scopolamine**	**Active placebo**	**No drug**	**Total**
n_i	12	8	8	28
$\sum_{j=1}^{n_i} y_{ij}$	74	75	85	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = 234$
$\sum_{j=1}^{n_i} y_{ij}^2$	474	719	919	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 = 2112$

First and foremost, the correction factor is

$$Correction\ factor = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\right)^2}{N} = \frac{(234)^2}{28} = \frac{54756}{28} = 1955.571$$

Sum of squares are:

$$SS_{Total} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - Correction\ factor$$
$$= 2112 - 1955.571 = 156.429$$

$$SS_{Treatment} = \sum_{i=1}^{k}\left[\frac{\left(\sum_{j=1}^{n_i} y_{ij}\right)^2}{n_i}\right] - Correction\ factor$$
$$= \frac{(74)^2}{12} + \frac{(75)^2}{8} + \frac{(85)^2}{8} - 1955.571 = 107.0123$$

$$SS_{Error} = SS_{Total} - SS_{Treatment} = 156.429 - 107.0123 = 49.4167$$

More often than not, we summarize the computation in tabular form as shown in Table 5.

Table 5: ANOVA table for effect of drug on amount of word recalled

Source of Variation	degree of freedom (df)	Sum of squares (SS)	Mean Square (MS)	F-statistic
Drug	k-1=3-1=2	107.0123	53.50615	27.06886
Error	N-k=28-3=25	49.4167	1.976668	
Total	N-1=28-1=27	156.429		

Due to the abridge nature of the adopted statistical table, the critical-value $F(0.05, df_{drug}, df_{error}) = F_{0.05,2,25}$ is not directly obtained. Thus, we use two-points Langrangian interpolation formula to obtain the approximated value as follows: From standard statistical table of F-distribution, $F(0.05,2,24) = 3.40$ and $F(0.05,2,30) = 3.32$. By Langrangian interpolation,

$$f(x) \cong \frac{x-x_1}{x_0-x_1}f(x_0) + \frac{x-x_0}{x_1-x_0}f(x_1) = \frac{(x-30)}{24-30}(3.40) + \frac{(x-24)}{30-24}(3.32)$$

Putting $x = 25$ in the preceding equation gives

$$f(25) \cong \frac{1}{6}[-3.40(25-30) + 3.32(25-24)] = 3.386667$$

Therefore,

$$F_{critical} = F(\alpha, df_{drug}, df_{error}) = F(0.05,2,25) = 3.386667$$

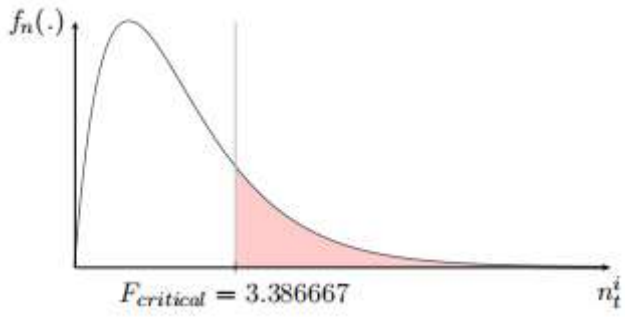

Figure 1: F-curve for $\nu_1 = 2$ and $\nu_2 = 25$ at $\alpha = 5\%$

Decision rule: Since $F_{statistic} = 27.06886 > F_{critical} = 3.386667$ and consequently fall in the critical region (see Figure 1), then the study reject H_0. Hence, data suggest that the population mean word-pair recalled differ across drug groups for at least a pair.

At the first step, we reject the hypothesis that the population mean are equal. At the second step, we compare all pairs of drugs at the 5% level to determine which of the group means differ from each other. Using FLSD as a basis for comparison, there is a significant difference between pair i and j if

$$|\bar{y}_i - \bar{y}_j| \geq t_{\left(\frac{\alpha}{2},N-k\right)}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = FLSD \qquad (16)$$

Otherwise, it is not. The t-critical-value for a two-sided test based on 25 df (df of Error) is:

$$t_{critical} = t_{\left(\frac{\alpha}{2},N-k\right)} = t_{\left(\frac{0.05}{2},28-3\right)} = t_{(0.025,25)} = 2.060$$

So, the FLSD for No-drug and *Scopolamine* comparison for example is:

$$FLSD = t_{\left(\frac{\alpha}{2},N-k\right)}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$
$$= 2.060\sqrt{1.976668\left(\frac{1}{8}+\frac{1}{12}\right)}$$
$$= 1.321945$$

Any two sample means that differ by at least $FLSD_{ij}$ (LSD of pair i and j) in magnitude are significantly different at the 5% level. An easy way to compare all pairs of drugs is to order the samples by their sample means. The samples can then be grouped easily, noting that two drugs are in the same group (not significantly different) if the absolute difference

between their sample means is smaller than the FLSD/LSD. The treatment group means is:

Table 6: The mean for each drug group

Drug groups	Mean
No drug	10.625000
Placebo	9.375000
Scopolamine	6.166667

There are $\binom{k}{2} = \binom{3}{2} = 3$ possible comparisons of two drugs. From this table, you can visually assess which sample means differ by at least their corresponding LSD, and which ones do not. Therefore, multiple comparisons for all pairwise comparisons among levels of drugs using FLSD is succinctly illustrates in the Table 7 below.

Table 7: Comparison of absolute difference in means with LSD

Comparison	$\lvert\bar{y}_i - \bar{y}_j\rvert$	LSD	Absolute difference in means exceeds LSD?
No drug and Placebo	1.25	1.448118	No
No-drug and *Scopolamine*	4.458333	1.321945	Yes
Placebo and *Scopolamine*	3.208333	1.321945	Yes

The groupings imply that we have sufficient evidence to conclude that population means No drug and *Scopolamine*, *Glycopyrrolate* (active placebo) and *Scopolamine* are significantly different while No drug and *Glycopyrrolate* (active placebo) are not significantly different.

COMPUTATION WITH R-SOFTWARE

What is, and Why R?

One may ask a question: **What is, and Why R?** Answer to this question is simple and direct. **R** is a computer language initially written by **Ross Ihaka** and **Robert Gentleman** in the mid-1990s specifically for statistical computing [16]. **R** is excellent software to use while first learning statistics, it provides a coherent, flexible system for data analysis that can be extended as needed. The open-source nature of **R** ensures its availability. The **R** home page (http://www.r-project.org/) contains more information about **R** and instructions for downloading a copy. A large and growing fraction of the world's quantitative methodologists and statisticians are moving to **R**, and the base of programs available for **R** is quickly surpassing all alternatives ([17], [18]). In addition to built-in functions, **R** is a complete programming language, which allows you to design new functions to suit your needs. Finally, despite its reputation, **R** is as suitable for students learning statistics as it is for researchers using statistics. In this study, we use in-built R functions in conjuction with extended R-packages "vioplot", "nortest", "car", "plyr", "ggplot2" developed by [13], [17], [19], [20], [21], [22], respectively.

Data Entry to R-environment

The study used information from Table 3 to illuminate how ANOVA can be carried out with R-software. First and foremost, reading the data into **R-environment** using **list-wise** approach as shown below:

```
#### Data entry using "listing format"
y1 <- c(5,8,8,6,6,6,6,8,6,4,5,6) # Group 1: Scopolamine
y2 <- c(8,10,12,10,9,7,9,10) #Group 2: Glycopyrrolate (active placebo)
y3 <- c(8,9,11,12,11,10,12,12) # Group 3: No drug
amount <- c(y1, y2, y3) # combined data
drugs <-c(rep("Scopolamine",length(y1)), rep("Placebo",length(y2)), rep("No drug", length(y3)))
drugs.long <- data.frame(amount, drugs) View(drugs.long)
```

It is pertinent to understand that in real work with data using **R**, one would generally not import data into **R** by explicit listings in an R-script file as done here. This only works for very small data set. The more realistic approach is to import the data from somewhere else, e.g. from a spread sheet program such as Microsoft Excel. **Note:** In **R**, # is for comment.

Descriptive Statistics

Researchers' may be interested to have the summary statistics (the mean, the standard deviation, the sample size n or the number of times each factor/treatment is being replicated, standard error, coefficient of variation, confidence intervals) for all the set of factor or groups (drug-type, in this case).

```
library(plyr)## download this from R package repository
drug.summary <- ddply(drugs.long, "drugs",
 function(X){
data.frame(m = mean(X$amount),
s = sd(X$amount),
n = length(X$amount) )})
drug.summary$se <- drug.summary$s/sqrt(drug.summary$n)# standard errors
drug.summary$cv<- (drug.summary$se/drug.summary$m)*100
drug.summary$ci.l <- drug.summary$m - qt(1-.05/2, df=drug.summary$n-1) * drug.summary$se
drug.summary$ci.u <- drug.summary$m + qt(1-.05/2, df=drug.summary$n-1) * drug.summary$se
drug.summary
```

This helps obtain descriptive statistics as shown in the Table 8 below.

Table 8: Descriptive statistics of the data by treatment group

	drugs	m	s	n	se	cv	ci. l	ci. u
1	No drug	10.625	1.505941	8	0.5324304	5.011110	9.3660	11.883998
2	Placebo	9.375	1.505941	8	0.5324304	5.679258	8.1160	10.633998
3	Scopolamine	6.167	1.267304	12	0.3658393	5.932529	5.3615	6.971874

In addition, boxplot of the distributions of the amount of word-pair recalled for each of the drug classification/type (individual points, mean and CI) is created using the **ggplot, Hmisc** packages as shown below:

```
library(ggplot2)## download both from R package repository
p <- ggplot(drugs.long, aes(x = drugs, y = amount))
p <- p + geom_hline(yintercept = mean(drugs.long$amount), colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)# boxplot,
p <- p + geom_boxplot(size = 0.75, alpha = 0.5)# points for observed data
p <- p + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.5)
p <- p + stat_summary(fun = mean, geom = "point", shape = 18, size = 6,
aes(colour = drugs), alpha = 0.8)# confidence limits based on normal distribution
p <- p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar", width = .2, aes(colour=drugs), alpha = 0.8)# confidence limits based on normal distribution
p <- p + labs(title = "Word paired-associate memory task") + ylab("amount of word recalled")
print(p)
```

which gives

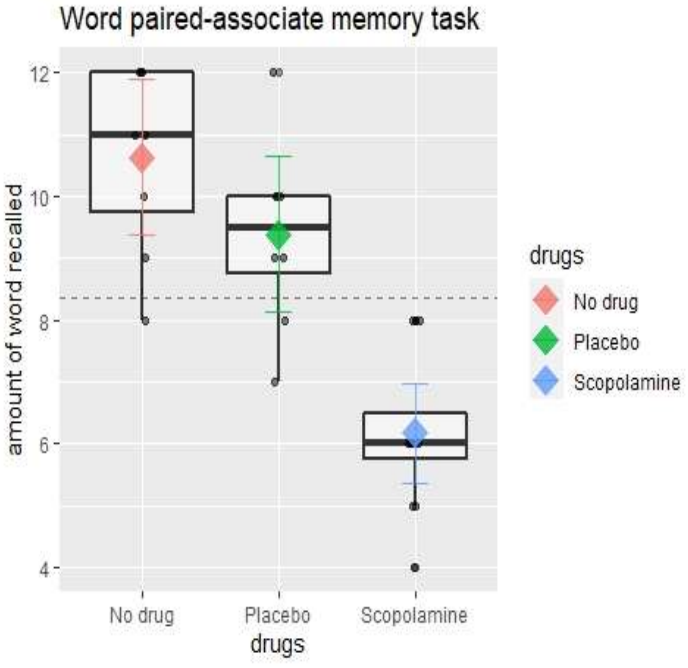

Figure 2: Combined boxplot of the data by drug-type (a)

A simpler alternative code for the boxplot is:

```
library(ggplot2)## download both from R package repository
ggplot(drugs.long, aes(x =drugs, y = amount)) +
geom_boxplot(fill = "grey80", colour = "blue") +
scale_x_discrete()+xlab("Treatment Group (i.e varieties of drug)") +
ylab("amount of word recalled")+
labs(title="Combined boxplot of the data by drug-type")
```

to produce

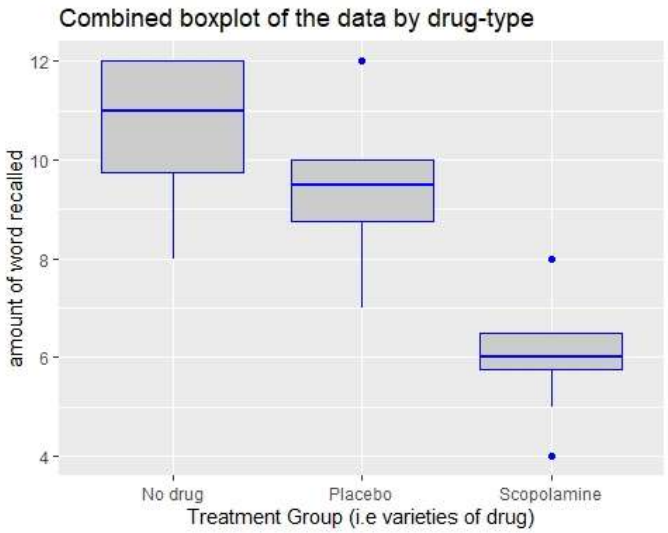

Figure 3: Combined boxplot of the data by drug-type (b)

Interpretation: From Figures 2-3, initial inspection of the data suggests that there are differences in the recalling rate for the two groups *Scopolamine* and No drug but it is not so clear to conclude for No drug and *Glycopyrrolate*(active placebo). Except for the mild outlier in the *Glycopyrrolate* and *Scopolamine* sample, the observed distributions are fairly symmetric, with similar spreads. The small deviations we are seeing here are not likely to impact our conclusions. We expect the standard ANOVA to perform well.

Prior to fitting ANOVA model, the entire body of classical statistical inference techniques is based on fairly specific assumptions regarding the nature of the underlying population distribution: usually its form and some parameter values must be stated. Given the right set of assumptions,

certain test statistics can be developed using mathematics which is frequently elegant and beautiful. The derived distribution theory is qualified by certain prerequisite conditions, and therefore all conclusions reached using these techniques are exactly valid only so as the assumptions themselves can be substantiated ([11], [14], [23]). Similarly, first, the basic parametric assumptions of ANOVA is that the effects are additive. Secondly, the experimental errors are randomly, normally and independently distributed about zero mean and a common variance, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$. Mathematically, $\varepsilon_{ij} \sim NID(0, \sigma^2)$. This assumption of common variance is popularly known in applied statistics as **homoscedasticity** assumption.

Test of Normality of Error Assumption

There is more than one way to skin a cat, as the saying goes, and more than one way to test the normality of error assumption. In this study, we shall embrace two distinct approaches: Exploratory Data Analysis (EDA) or graphical approach and Inference-based techniques (Shapiro-Wilk [24], Anderson-Darling, and Cramer-von Mises normality test). First and foremost, for data visualization approach, use the R-code below:

```
library(vioplot) ## download this package from R package repository
library(car) ## download this package from R package repository
fit.d <- aov(amount ~ drugs, data = drugs.long)
win.graph(width=6, height=7, pointsize=6) # this is optional as it is used to re-size the graph
par(mfrow=c(2,2))
hist(fit.d$residuals, freq = FALSE, breaks = 20)# Histogram with kernel density curve points(density(fit.d$residuals), drugs = "l")
rug(fit.d$residuals)
vioplot(fit.d$residuals, horizontal=TRUE, col="gray")# violin plot boxplot(fit.d$residuals, horizontal=TRUE)# boxplot
qqPlot(fit.d$residuals, las = 1, lwd = 1, main="QQ Plot")# QQ plot
```

This produce the following figures:

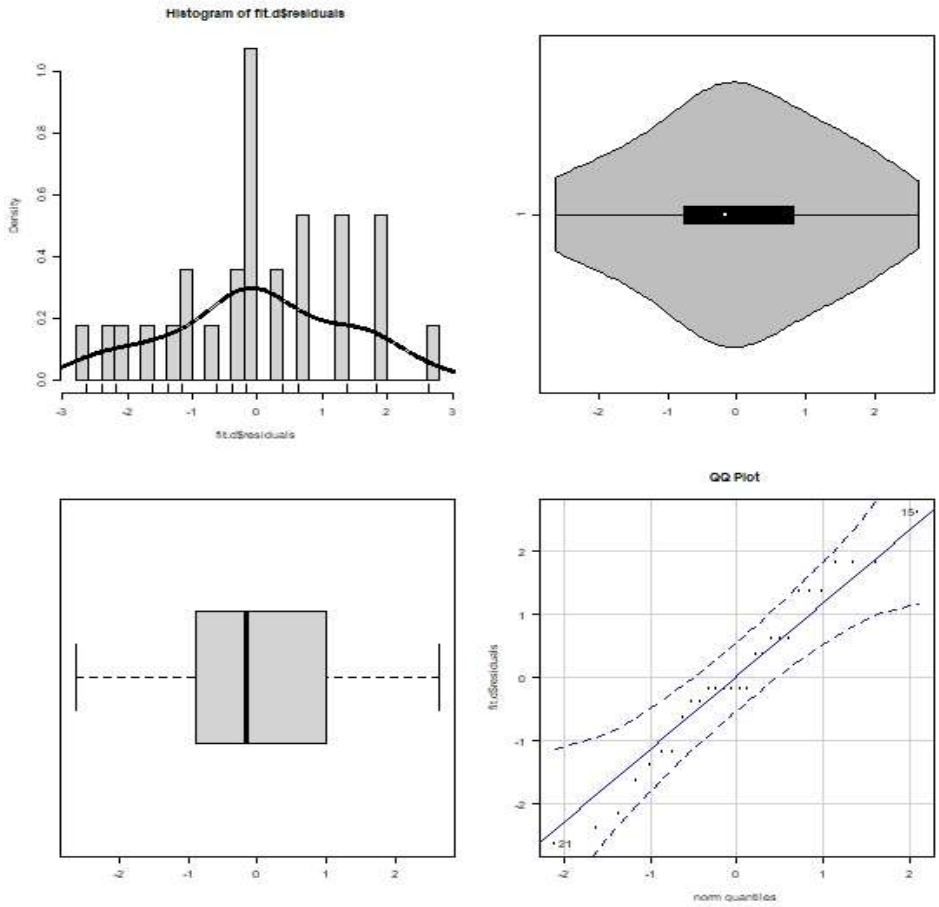

Figure 4: Inspection of basic parametric assumptions of ANOVA

To elucidate more on the first graph in the matrix of Figure (4) above, we write:

```
win.graph(width=7, height=6, pointsize=5) # optional: is used to re-size the graph hist(residuals(fit.d), freq=FALSE,
```

```
col="grey", main="normal curve over histogram", ylim=c(0.00,0.35))
curve(dnorm(x, mean=mean(residuals(fit.d)), sd=sd(residuals(fit.d))), add=TRUE, col="red")
lines(density(residuals(fit.d)), col="blue", lwd=2)
legend("topright", legend=c("normal density","empirical density"), fill=c("red","blue"), col=c("red", "blue"))
```

to get

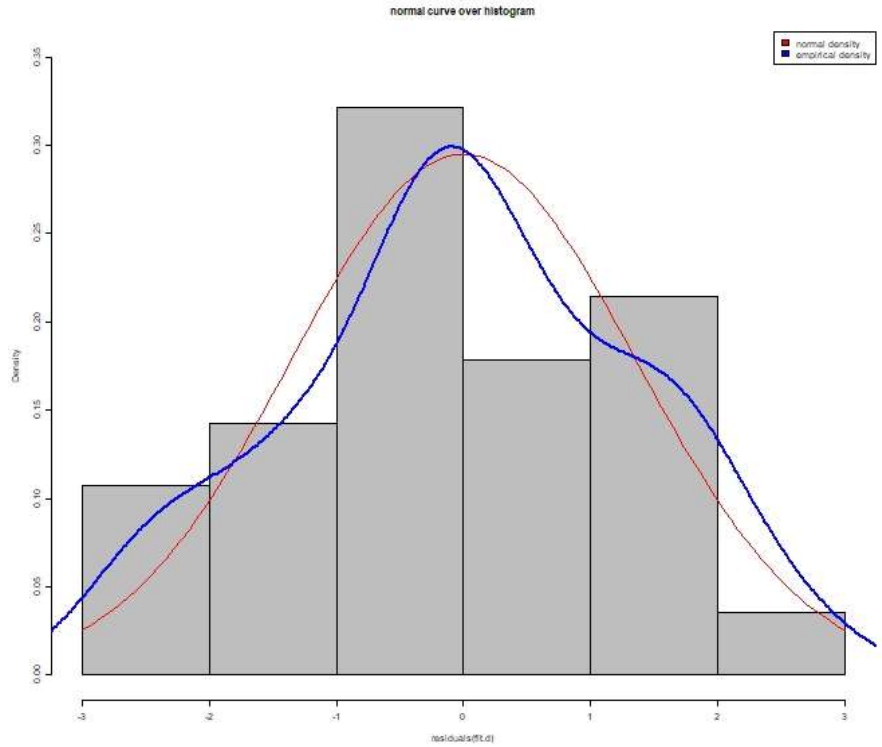

Figure 5: Visualization of normality of error assumption with histogram

The Figure 4 and 5 show a visual inspection of the nature of normality. The histogram, violin plot, box plot and QQ plot are all look pretty normal which indicate that there is no strong evidence against normality. In short, any distribution that resembles a bell-shaped curve will be "normal enough" to pass normality tests, especially if the sample size is adequate. There is no fixed definition of "large enough", but a rule of thumb is $N \geq 30$ ([1], [2], [25]).
For the inferential statistical techniques, we set hypothesis $H_0: \varepsilon_{ij} \sim N(0, \sigma^2)$ against H_1: H_0 is false and employ Shapiro-wilk test for normality using the R-code:

```
shapiro.test(fit.d$residuals) ## Shapiro-Wilk normality test
```

and we have this result:

Shapiro-Wilk normality test

data: fit.f$residuals
W = 0.97096, p-value = 0.6067

Interpretation: Result of Figure 4–5 is supported by (or conform with) the results of Shapiro-Wilk test with test-statistic $W = 0.97096$ and $p - value = 0.6067$ indicating the test is not significant. That is, no indication that normality is violated. Hence, we do not reject the null hypothesis (H_o) which state that H_o : errors are approximately normally distributed against alternative hypothesis H_1: error terms are not normally distributed. Other alternative tests are Anderson-Darling and Cramer-von Mises normality test presented as follows:

```
library(nortest) ## download this package from R package repository
ad.test(fit.d$residuals) ## Anderson-Darling Normality Test
```

which gives

Anderson-Darling normality test

data: fit.f$residuals A = 0.33208, p-value = 0.4932

and

```
cvm.test(fit.d$residuals)## Cramer-von Mises normality test
```

with results output as

Cramer-von Mises normality test

data: fit.f$residuals
W = 0.057227, p-value = 0.399

Interpretation of Anderson-Darling normality and Cramer-von Mises normality test results follow same as in Shapiro-Wilk normality test. Suppose there is violations of normality, in the context of one-way ANOVA, the easiest solution is probably to switch to a non-parametric test (i.e., one that doesn't rely on any particular assumption about the kind of distribution involved) known as Kruskal-Wallis rank sum test.

Homogeneity of Variance Assumption

Three commonly invoked tests of homogeneity of variance assumption: Bartlett, Fligner-Killeen and Levene [25] are adopted in the study.

```
bartlett.test(amount ~ drugs, data = drugs.long)## Bartlett test of homogeneity of variances
```

and this gives:

Bartlett test of homogeneity of variances data: amount by drugs

Bartlett's K-squared = 0.34085, df = 2, p-value =0.8433

Alternatively, we use Fligner-Killeen test of homogeneity of variances with the aid of R-code presented as:

```
fligner.test(amount ~ drugs, data = drugs.long)
```

which give result as:

Fligner-Killeen test of homogeneity of variances

data: amount by drugs

Fligner-Killeen:med chi-squared = 1.086, df = 2, p-value = 0.581

Interpretation: Formal tests of equal population variances are far from significant. The p-values for Bartlett's test and Fligner-Killeen homogeneity of variances test are greater than 0.05. In case of Bartlett's test, the $p-value = 0.8433 > 0.05$. So, we fail to reject the null hypothesis that the population variances are equal. This result is not surprising given how close the sample standard deviations (evidenced from Table 8) are to each other. Thus, the standard ANOVA appears to be appropriate here. For Levene-test, the R-code is as follows:

```
library(car) ## download this package from R package repository
leveneTest(amount ~ drugs, data = drugs.long)
```

The result is

Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	2	0.3253	0.7253

Interpretation: The result in this case is Levene's Test. $Test_{statistic} = 0.3253$, $df = 2$, $p-value = 0.7253 > 0.05$, indicating that the test is not significant. That is, there is no sufficient evidence to reject the null hypothesis which states that H_o: variances assumed equal versus H_1: variances assumed not equal. Thus, Levene's test conforms to Bartlett's and Flinger's test.

Test of Independence of error components

This assumption is not only limited to ANOVA, rather, a general assumption of parametric analysis is that the value of each observation for each subject is independent of (i.e., not related to or influenced by) the value of any other observation. For independent groups designs, this issue is addressed with random sampling, random assignment to groups, and experimental control of extraneous variables. This assumption is an inherent concern for repeated measures designs, in which an assumption of **sphericity** comes into play. When subjects are exposed to all levels of an independent variable (e.g., all treatments), it is conceivable that the effects of a treatment can persist and affect the response to subsequent treatments. For example, if a treatment effect for one level has a long half-time (analogous to a drug effect) and there is inadequate "wash out" time between exposures to different levels (treatments), there will be a carryover effect. A well designed and executed cross-over experimental design can mitigate carryover effects. Mauchly's test of sphericity is commonly employed to test the assumption of independence in repeated measures designs. If the Mauchly test is statistically significant, corrections to the F-statistic calculation are warranted. The two most commonly used correction methods, the Greenhouse-Geisser and Huynh-Feldt, are not discussed here.

```
drug.mod = data.frame(Fitted = fitted(lm(amount~drugs, data=drugs.long)),
Residuals = resid(lm(amount~drugs, data=drugs.long)), Treatment = drugs.long$drugs)
library(ggplot2) # download this package from R package repository
ggplot(drug.mod, aes(Fitted, Residuals, colour=Treatment))+geom_point()
```

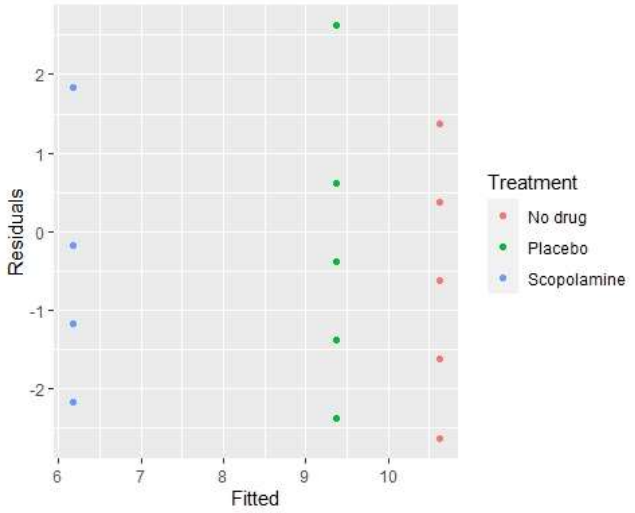

Figure 6: Diagnostic plot for linearity and independence of error components

Interpretation: From Figure 6, we can see that there is no major problem with the diagnostic plot but some evidence of different variability in the spread of the residuals for the three treatment groups. Moreover, it seems that this increase occurs in a linear fashion.

In a nutshell, all the assumptions can be investigated in a single computation by plotting the model residuals against the fitted values. First, create a data frame with the fitted values, residuals and treatment identifiers:

```
win.graph(width=7, height=7, pointsize =6) ## optional: used to re-size the graph par(mfrow=c(2,2)) ## to put the figure in matrix form of dimension 2 by 2 plot(lm(amount~drugs, data=drugs.long))
```

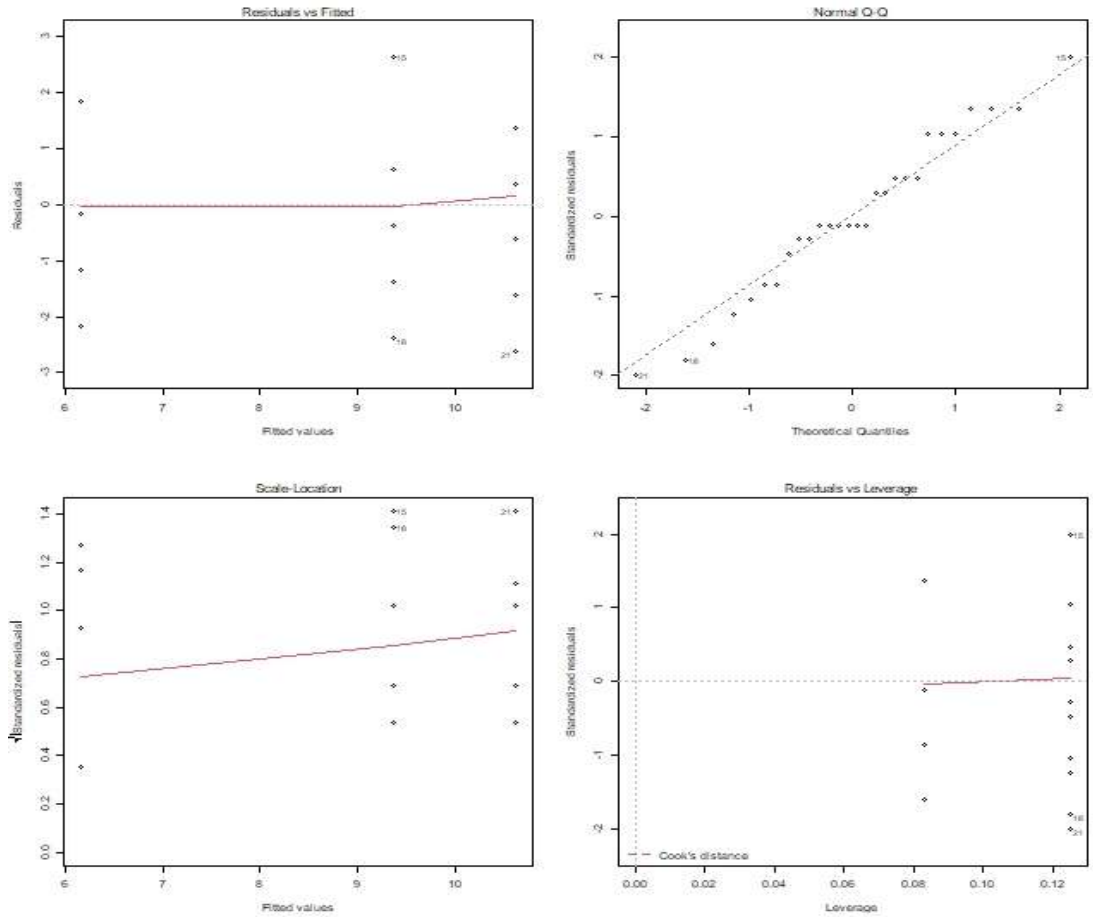

Figure 7: Checking model assumption and adequacy

Interpretation of the Plots in Figure 7:

• **Residuals versus 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, 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 6 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. The result which corroborates Shapiro-Wilk

normality 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 *Scopolamine*. How far is the points to the centre of the treatment factor. From Figure 7d, to what extent are the points 15, 18, 21 actually influence the ANOVA model.

Remarks: 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 ([2], [24], [27]). This is what make statistics doable because no real data set 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).
According to [26] and [28], ANOVA is robust even when the homogeneity assumption is not fulfilled, as long as the sample sizes are roughly equal or the deviation is only of a moderate level. As a rule of thumb, if the largest std.dev $<$ 2 $\times$ the smallest std.dev) then we need not to be concerned about this assumption.

ANOVA Model
Finally, we run ANOVA model to assess whether there are differences between pair(s) of drugs using the R-code:

```
fit.d <- aov(amount ~ drugs, data = drugs.long)
summary(fit.d)
```

The following results are generated

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Drugs	2	107.01	53.51	27.07	5.55e-07 ***
Residuals	25		49.42	1.98	

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

Interpretation: Theoretically, the rule is: reject H_0 if $P-value \leq \alpha$; otherwise, do not reject. From the model summary output, the $P-value = 5.55e-07 < \alpha = 0.05$. We therefore, reject H_0 at any of the usual test levels (such as, 0.05 or 0.01). The data suggest that there is sufficient evidence to conclude that the population mean of word recalled differ across drugs in some way. This is desirable since it is expected that the *Scopolamine* affect memory.

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

```
confint(fit.d, level=0.95)
```

and we have the results as

	2.5%	97.5%
(Intercept)	9.601255	11.6487447
drugsPlacebo	-2.697794	0.1977936
drugsScopolamine	-5.779982	-3.1366847

Post-Hoc Analysis
We shall use Fisher's Least Significance Difference (FLSD), Bonferroni and Tukey test to identify between which pair the difference is significant.

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. The treatment group means is obtained using

```
combined_mean<-tapply(drugs.long$amount,
drugs.long$drugs, mean) combined_mean
```

to get

No drug	Placebo	Scopolamine
10.625000	9.375000	6.166667

Therefore, multiple comparisons for all pairwise comparisons among levels of drugs using FLSD in **R** is:

```
pairwise.t.test(drugs.long$amount, drugs.long$drugs,
pool.sd = TRUE, p.adjust.method = "none")
```

This generates the following results:

Pairwise comparisons using t tests with pooled SD

data: drugs.long$amount and drugs.long$drugs

	No drug	Placebo
Placebo	0.088	-
Scopolamine	2.8e-07	3.7e-05

P value adjustment method: none

Interpretation: The output above indicate that there is no significance difference between pair No drug-placebo since

absolute difference in means does not exceed FLSD and $p-value > 0.05$ but there is significance difference between pairs: No drug-*scopolamine* and placebo-*scopolamine* since their absolute difference in means exceed FLSD and $p-values < 0.05$. We judge the significant based on their p-value being $>$ or $<$ α (preassigned significance level).

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 hold simultaneously. Bonferroni 95% individual p-values for all pairwise comparisons among drugs is obtained by

```
pairwise.t.test(drugs.long$amount, drugs.long$drugs, pool.sd = TRUE, p.adjust.method = "bonf")
```

This produce

Pairwise comparisons using t tests with pooled SD

data: drugs.long$amount and drugs.long$drugs

	No drug	Placebo
Placebo	0.26263	-
Scopolamine	8.3e-07	0.00011

P value adjustment method: bonferroni

Interpretation: There is significance difference between No drug-*scopolamine* and placebo-*scopolamine* while the pair No.drug-placebo is not significant. The criterion is the same as in Fisher' least significance difference.

Tukey Test

A commonly-used alternative is Tukey's honest significant difference method (HSD). Procedure for post-hoc analysis using Tukey test is as follow:

```
fit.d<-aov(amount~drugs, data=drugs.long)# this has been previously defined TukeyHSD(x=fit.d, 'drugs', conf.level=0.95)
```

or in a more simpler form as

```
TukeyHSD(fit.d)
```

In either case, results of the **Tukey test** is:

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = amount ~ drugs, data = drugs.long)

$drugs

	diff	lwr	upr	p adj
Placebo-No drug	-1.250000	-3.000979	0.500986	0.1973818
Scopolamine-No drug	-4.458333	-6.056751	-2.859916	0.0000008
Scopolamine-Placebo	-3.208333	-4.806751	-1.609916	0.0001073

Interpretation: John Tukey's honest significant difference method is to reject the equality of a pair of means based, not on the t-distribution, but the studentized range distribution. This output indicates that the differences between *Scopolamine*-No drug and *Scopolamine*-Placebo are significant, while Placebo-No drug 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 *Scopolamine*-No drug differ significantly. How? because the interval does not contain 0. The confidence intervals for Placebo-No drug contain 0. Thus, it appears that those pairs do not differ among themselves. For the drug data, the groupings based on Fisher's LSD, Bonferroni and Tukey comparisons are identical. This is coincident and not conventional.

Readers are encourage to consult [14] to learn about **False Discovery Rate (FDR):** expected proportion of false discoveries amongst the rejected hypotheses. The method FDR by Benjamini, Hochberg, and Yekutieli is a less popular, less stringent but a more statistically powerful test. By statistical power, we mean, ability of an inferential test to detect a difference that actually exists, i.e., a true positive.

CONCLUSION

In this study, concept of ANOVA is expounded. A real-life survey data was analysed using manual (theoretical formulas) and R-software, and results from the two approaches are well-agree in terms of magnitude and interpretation. Hence, this study is able to achieve its set objective. Therefore, if all the necessary details (as elucidated in the methodological framework of this paper) are put into consideration, future researchers should worry-less on theoretical or manual method and employed the statistical tools (R-codes) explicated in this study for sake of computational advantage.

SIGNIFICANCE STATEMENT

This material should be of pedagogical interest to researchers whose data layout follows analysis of variance and intends to use **R**. In addition, it can serve as an excellent teaching reference in computing classes where understanding of some elementary rudiment of statistical inference, introduction to R-environment and basic R-code are the only background requirements. The procedures and results discussions are straightforward, comprehensive and lucidly presented.

ACKNOWLEDGMENT

The authors are eternally grateful to International Statistical

Institute (ISI) and International Association for Statistical Computing (IASC) for given us the opportunity to present the manuscript in her 2020 workshop titled **"Capacity Building for Statistician on Modeling Financial and Agricultural Data Using R"** (https://iasc-isi.org/2020/09/12/iasc-webinar-capacity-building-for-statistician-on-modeling-financial-and-agricultural-data-using-r/). Many thanks also go to the UI-LISA (University of Ibadan-Laboratory for Interdisciplinary Statistical Analysis) for the collegiality enjoyed. In addition, authors thank the Chief-editor of *International Journal of Mathematics and Computer Research*, and anonymous reviewers for their careful reading, some kindly comments and suggestions on improving the presentation of this paper.

REFERENCES

1. Johnson, R. A. and Bhattacharyya, G. K. (2010). Statistics: Principles and Methods, *John Wiley & Sons*, Inc., 6th edition, ISBN 978-0-470-40927-5, https://book4you.org/book/1162381/485dfd.
2. Weiss, N. A. (2012). Introductory Statistics, Addison-Wesley: Pearson Education, Inc., 9th edition, ISBN-13: 978-0-321-69122-4, https://book4you.org/book/1226780/d58353.
3. Zhang J. T., Cheng, M. Y., Wu, H. T. and Zhou, B. (2018). A new test for functional one-way ANOVA with applications to ischemic heart screening, *Computational Statistics and Data Analysis*, DOI:10.1016/j.csda.2018.05.004, https://doi.org/10.1016/j.csda.2018.05.004.
4. Cook, C. (2008). Clinimetrics corner: Use of effect sizes in describing data, *The Journal of Manual & Manipulative Therapy*, **16**(3), 54-55, DOI:10.1179/106698108790818398, https://dx.doi.org/10.1179/106698108790818398.
5. Shrefe, A. M. A., Alacrouk, S. A., Fadia, T. M., Oshkondali, S. T. Aburas, K.M., Ahmed, A. B., Alqamoudy, H., Almunir, N. and Elshlli, M. M. (2019). Using ANOVA One-way test for determination of suitable dose of Alloxan for induction of type II diabetes in Mice, *Scholars Journal of Physics, Mathematics and Statistics*, **6**(9), 160-162, ISSN 2393-8056 (Print)|ISSN 2393-8064 (Online), DOI: 10.36347/sjpms.2019.v06i09.003, saspjournals.com/sjpms-69 .
6. Cabral, H. J. (2008). Multiple comparisons procedures, *Circulation*, **117**(5), 698-701, DOI: 10.1161/CIRCULATIONAHA.107.700971, https://doi.org/10.1161/CIRCULATIONAHA.107.700971 .
7. Ostertagova, E. and Ostertag, O. (2013). Methodology and application of one-way ANOVA, *American Journal of Mechanical Engineerin*g, **1**(7), 256-261, DOI: 10.12691/ajme-1-7-21, https://pubs.sciepub.com/ajme/1/7/21.
8. Bloomfield, R., O'Hara, M. and Saar, G. (2009). How noise trading affects markets: An experimental analysis,*The Review of Financial Studies*, 22(6), 2275-2302, DOI:10/hhhn102,http://ssrn.com/abstract=1408433, or http://dx.doi.org/hhn102.
9. [Cianci, A. M. and Bierstaker, L. J. (2009). The effect of client importance and performance feedback on auditors' technical and ethical judgments, *Managerial Auditing Journal*, **24**(5), 455-474, DOI: 10.1108/02686900910956810, http://dx.doi.org/10.1108/02686900910956810
10. Todua, N. and Dotchviri, T. (2015). Anova in marketing research of consumer behaviour of different categories in Georgian Market, *Economy Series*, **1**(1), ISSN 2344-3685/ISSN-L 1844-7007, https://www.utgjiu.ro/revista/ec/pdf/2015-01.Volumul%201/26_TODUA,%20%20Dotchviri.pdf
11. Montgomery, D. C. (2017). Design and analysis of experiments, *John Wiley & Sons Incorporated*, 9th Edition, ISBN 9781119113478 (PBK), ISBN 9781119299455 (EVALC), https://book4you.org/book/3591248/4df2ac.
12. McHugh, M. L. (2011). Multiple comparison analysis testing in ANOVA, *Biochemia Medica*, **21**(3), 203-209, DOI: 10.11613/BM.2011.029, https://www.biochemia-medica.com/en/journal/21/3/10.11613/BM,2011.029
13. Assaad, H. I. Lan Zhou, Carroll, R. J. and Wu, G. (2014). Rapid publication-ready MS-Word tables for one-way ANOVA, *SpringerPlus*, **3**(1), 474, 1-8, DOI:10.1186/2193-1801-3-474, http://www.springerplus.com/content/3/1/474
14. Everitt, B. S. and Hothorn, T. (2010). *A handbook of statistical analysis using R*, Taylor & Francis Group, LLC, ISBN 978-1-4200-7933-3, https://book4you.org/book/716450/d9f17e.
15. Atri, A., Sherman, S., Norman, K. A., Kirchhoff, B. A., Nicolas, M. M., Greicius, M. D., Cramer, S. C., Breiter, H. C., Hasselmo, M. E., Stern, C. E. (2004). Blockade of central cholinergic receptors impairs new learning and increases proactive interference in a word paired-associate memory task, *Behavioral Neuroscience*, **118**(1), 223-236, DOI:10.1037/0735-7044.118.1.223, https://doi.org/10.1037/0735-7044.118.1.223.
16. Li, K. Yan, E. and Feng, Y. (2017). How is R cited in research outputs? Impacts, and citation standard, *Journal of Informetrics*, **11**(2017), 1-15, DOI:10.1016/j.joi.2017.08.003, https://dx.doi.org/10.1016/j.joi.2017.08.003.
17. Fox, J. and Weisberg, H. S. (2019). An R companion to applied regression, Third Edition, *Thousand Oaks*

CA: Sage, ISBN-13: 978-1544336473, ISBN-10: 1544336470, https://socialsciences.mcmaster.ca/jfox/Books/Companion/

18. Górecki, T. and Smaga, K. (2019). fdANOVA: an R software package for analysis of variance for univariate and multivariate functional data, *Computational Statistics*, **34**, 571–597, DOI:10.1007/s00180-018-0842-7,https://doi.org/10.1007/s00180-018-0842-7.
19. Adler, D. and Kelly, S. T. (2020). vioplot: violin plot, R package version 0.3.6, https://github.com/TomKellyGenetics/vioplot
20. Gross, J. and Ligges, U. 2015. nortest: Tests for Normality, https://CRAN.R-project.org/package=nortest
21. Wickham, H. (2011). The split-apply-combine strategy for data analysis, *Journal of Statistical Software*, **40**(1),1-29, http://www.jstatsoft.org/v40/i01/, http://www.amstat,org/
22. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis, *Springer-Verlag New York*, ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org
23. Gibbons, J. D. and Chakraborti, S. (2003). Nonparametric Statistical Inference, Fourth edition, *Library of Congress Cataloging-in-Publication Data*, ISBN: 0-8247-4052-1, https://book4you.org/book/905501/78c4c3
24. Ghasemi, A. and Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians, *International Journal of Endocrinology Metabolism*, **10**(2), 486-489, DOI: 10.5812/ijem.3505.
25. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*, *Springer Texts in Statistics*, ISSN 1431-875X, ISBN 978-1-4614-7137-0, ISBN 978-1-4614-7138-7 (eBook), DOI: 10.1007/978-1-4614-7138-7, https://doi.org/10.1007/978-1-4614-7138-7.
26. Gastwirth, J. L., Gel, Y. R. and Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice, *Statistical Science*, **24**(3), 343-360, DOI: 10.1214/09-STS301.
27. Sawyer, S. F. (2017). Analysis of variance: The fundamental concepts, *The Journal of Manual & Manipulative Therapy*, **17**(2), 27E-38E, DOI: 10.1179/jmt.2009.17.2.27E, https://doi.org/10.1179/jmt.2009.17.2.27E .
28. Blanca, M. J., Alarcon, R., Arnau, J., Bono, R. and Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit?, *Behavior Research Methods*, **50**(2018), 937-962, DOI:10.3758/s13428-017-0918-2, https://link.springer.com/content/pdf/10.3758/s13428-017-0918-2.pdf .