Design and analysis of one-way ANOVA

Fixed versus Random factors

Peter Tea

Department of Mathematics and Statistics
University of Ottawa

Contents

1 Analysis of one-way ANOVA

For a better understanding of the design and analysis of a single factor experiment, this chapter will explain some fundamental concepts. The information presented is largely inspired by *Montgomery (2012)*.

A one-way analysis of variance (ANOVA) is used when a researcher is interested in determining whether there exists a significant difference between means of three or more independent groups. The *one-way* which precludes the name *one-way ANOVA* implies that only one independent variable (also called a factor) is considered in the model. This factor may contain up to a different levels (i.e. the number of levels $= 1, 2, ..., a$). The focus of ANOVA is to test whether there exists any difference between any of these levels. Consider the following example:

Example 1: In an experiment which studies the effect of three different diets on cholesterol concentration, an ANOVA can be used to test whether the response variable *cholesterol concentration* vary between any of the three levels of the factor *diet*.

1.1 The effects model

In general, the observations of an experiment with a single factor containing a levels (also called *treatments*) can be modelled with the following linear statistical model:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, 3, ..., a \\ j = 1, 2, 3, ..., n \end{cases}$$

This model is called an effects model. In this model:

- y_{ij} represents the ijth observation
- μ represents the overall mean (i.e. the mean pooled across all levels)
- τ_i is a unique parameter to each treatment level and is referred to as the *treatment effect*. More importantly, τ_i represents the deviation from the overall mean resulting from the ith treatment.
- ϵ_{ij} is the random error of the experiment. The random error represents other sources of variability (eg. variability due to measurement errors or due to background noise.)

1.2 Assumptions of the effects model

In the effects model, there are some underlying assumptions made:

1. The errors are normally and independently distributed random variables

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

2. The variance σ^2 is constant for all levels of the factor. Thus,

$$y_{ij} \sim N(\mu + \tau_i, \sigma^2)$$

3. All observations are mutually independent.

1.3 Fixed vs Random Factors

Before the one-way ANOVA model is run, a decision must be made about the levels of the factor. Specifically, the treatment effects (τ_i's) can either be fixed or they can be random. The distinction between fixed and random factors affects the hypothesis being tested.

A factor is considered fixed if the treatments are specifically chosen by the researcher such that all treatment levels of interest for the study are included. In this experiment type, the inferences made can only be applied to the levels included in the model and cannot be extrapolated to other possible levels of the factor that were not included. If the study was repeated, then the exact same levels would be chosen again.

A factor is considered random if the levels of the factor are randomly sampled from a population of many other possible levels. In this experiment type, the inferences made can be extended to the population of all possible levels, even if these levels were not explicitly included in the model. If the study was repeated, then it is highly likely that different levels would be chosen.

The analysis of the data is different, depending on whether the factor is treated as fixed or as random. Consequently, inferences may be incorrect if the factor is not classified appropriately.

1.4 Fixed effects model

The effects model (introduced in 1.1) with a fixed factor is called a fixed effects model. In this type of model, the researcher is mainly interested in whether each ith treatment level produces the same response mean. Let

$$\mu_i \Big\{ i = 1, 2, ..., a$$

represent the ith treatment mean. Then, the hypotheses can be written as follows:

$$H_0: \quad \mu_1 = \mu_2 = ... = \mu_a = 0 \tag{1}$$
$$H_1: \quad \mu_i \neq \mu_j \; for \; at \; least \; one \; pair \; (i, j) \tag{2}$$

An alternative way to write the previous hypotheses is with the treatment effects (τ_i's). Consider the means model, which is an alternative way to model the data. Let μ represent the total mean of the model. Then, the means model is written as:

$$\mu = \mu_i + \tau_i, \qquad i = 1, 2, ..., a$$

If there are no treatments effects, then the response means for each level will be the same (i.e. $\mu = \mu_i$). Therefore, the hypotheses can be rewritten as follows:

$$H_0: \quad \tau_1 = \tau_2 = ... = \tau_a = 0 \tag{3}$$
$$H_1: \quad \tau_i \neq 0 \; for \; at \; least \; one \; i \tag{4}$$

These hypotheses are tested using ANOVA.

1.5 Random effects model

Inferences on random factors can be used to generalize the entire population of factor levels. For this model, there is an added assumption that the total number of factor levels are infinite in size, or large enough to be considered as infinite. From the effects model, it is assumed that:

$$\tau_i \sim N(0, \sigma_\tau^2)$$
$$\epsilon_{ij} \sim N(0, \sigma^2)$$

τ_i and ϵ_{ij} are independent.

Let

$$y_{ij} \begin{cases} i = 1, 2, ..., a \\ j = 1, 2, ..., n_i \end{cases}$$

represent the jth observation at the ith treatment level. The variance of the y_{ij} observations can then be written as:

$$V(y_{ij}) = \sigma_\tau^2 + \sigma^2$$

Unlike the fixed effects model, where all observations are mutually independent, the observations in a random effects model are only independent if they belong to different treatment levels. All observations belonging to a specific level have the same covariance.

Let y_{ij} represent the jth observation in the ith treatment level. The covariance of the y_{ij} observations can be written as follows:

$$cov(yij, yij') = \sigma_\tau^2 \quad for\ \ j \neq j'$$

$$cov(yij, yi'j') = 0 \quad for\ \ i \neq i'$$

$$\begin{cases} i = 1, 2, ..., a \\ j = 1, 2, ..., n_i \end{cases}$$

(Where n_i represents the total number of observations in the ith level.)

Hypothesis testing involving random factors focuses on the population of treatment levels. The following hypothesis is tested:

$$H_0 : \sigma_\tau^2 = 0$$

$$H_1 : \sigma_\tau^2 > 0$$

Essentially, the hypothesis testing on random factors tests whether variability exists between treatments of the factor of interest. If the null hypothesis is rejected, then we can conclude that the tested factor in general has a significant impact on the response variable.

1.6 How it works: ANOVA

In this subchapter, a brief overview of the derivation of ANOVA is given. The following models make use of dot subscript notation. Information explaining this notation can be found in the appendix.

ANOVA is derived from the decomposition of the total variability (in the conducted experiment) into smaller components. Let SST represent the total variability in the observed data. Under the assumptions made in chapter 1.1, the total variability of observations can be defined below.

$$SS_T = \sum_{i=1}^{\alpha}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \quad n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 \quad + \quad \sum_{i=1}^{\alpha}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{.})^2$$

$$\begin{array}{ccccc} SS_T & = & SS_{Treatments} & + & SS_E \\ (N-1) & & (a-1) & & (N-a) \end{array}$$

The mean squares are obtained from the sum of squares and their respective degrees of freedom:

$$MS_{Treatments} = \frac{SS_{Treatments}}{a-1}$$

$$MS_E = \frac{SS_E}{N-a}$$

The decomposition of the total variance can be thought as being partitioned into:

1. The variation between treatment levels ($SS_{Treatments}$) and
2. The variation within each treatment level (SS_E).

$SS_{Treatments}$ is the measure of the sum of squares variation caused by the different levels and SSE is the measure of the sum of squares from random error within a level.

The F-Test

Under the null hypothesis it can be shown that, $\frac{SS_{Treatments}}{\sigma^2} \sim \chi^2_{a-1}$ and $\frac{SS_E}{\sigma^2} \sim \chi^2_{N-a}$.

With these results, the F-test statistic F_0 can be derived which is $F_{a-1,N-a}$ distributed.

$$F_0 = \frac{\frac{SS_{Treatments}}{a-1}}{\frac{SS_E}{N-a}} = \frac{MS_{Treatments}}{MS_E}$$

The null hypothesis should be rejected if

$$F_0 > F_{\alpha,a-1,N-a}$$

Interpretation of the F-Test

It can be shown that the expected values of the mean squares for fixed effects are as follows:

$$E(MSE) = E\left(\frac{SSE}{N-a}\right) = \sigma^2$$

$$E(MS_{Treatments}) = \sigma^2 + n\frac{\sum_{i=1}^{a} \tau_i^2}{a-1}$$

For fixed effects, MSE is an unbiased estimator for σ^2 whereas $MS_{Treatments}$ can also be an unbiased estimator for σ^2 if each treatment mean is the same (i.e. if the null hypothesis is true and $\tau_i = 0$). However if the treatment means are not the same, then the expected

value of $MS_{Treatments}$ increases beyond the value of σ^2 (since τ_i now is not equal to 0). Thus, by comparing $MS_{Treatments}$ to MSE, it is easy to test whether there is indeed a difference in treatment means. If the null hypothesis is true, then the F_0 ratio should be close to 1. However, if the null hypothesis is false, then the value of F_0 increases.

For the random effects model, the expected value of $MS_{Treatments}$ can be written in terms of the variance components as follows:

$$E(MS_{Treatments}) = \sigma^2 + n\sigma_\tau^2$$

Similarly, for random effects, $MS_{Treatments}$ can be compared to MSE to test whether there exists a significant σ_τ^2 that contributes to the variability of the response between treatments.

ANOVA table

Source of variation	DF	Sum of Squares	mean Square	F-value
Between treatments	$a-1$	$SS_{Treatments}$	$MS_{Treatments}$	$\frac{MS_{Treatments}}{MS_E}$
Error	$N-a$	SS_E	MS_E	
Total	$N-1$	SS_T		

Table 1: The above ANOVA table summarizes the statistics used during hypothesis testing.

2 Design of one-way ANOVA

2.1 Balanced vs Unbalanced design

A balanced design is such that the number of observations are equal in all levels of the factor. An experimental design is unbalanced if the number of observations are not equal in all levels of the factor. Even with an unbalanced design ANOVA may still be used however, some equations are modified to account for this unbalanced design. For example:

$$SS_T = \sum_{i=1}^{\alpha}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}}{N}$$

and

$$SS_{Treatments} = \sum_{i=1}^{a}\frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$$

There are advantages to choosing a balanced design over an unbalanced design:

1. The test statistic is more robust to any deviation from the assumption of equal variance between levels if the design is balanced.

2. The power is maximized in a balanced design. However, it has been suggested that for a one factor design, an unbalanced design does not pose any serious problems. The issues are more prominent for multiple factor ANOVAs.

2.2 Randomized complete block design

During the design stages of an experiment, it is oftentimes important to guard against effects from other factors (which are not explicitly considered in the model) that may unknowingly affect the variability of the experiment. These *other factors* are called *nuisance factors* and may contaminate the observations. One solution is to cluster similar experimental units together into randomized blocks. In this randomized complete block design (RCBD), it is assumed that the experimental units assigned to a specific block are essentially homogenous. Each treatment level is then randomly applied to the experimental units within each block, such that all blocks each contain all levels.

Let β_j represent the jth block and let there be b blocks. Then, the effects model can be modified to obtain a new statistical model for RCBD:

$$y_{ij} = \mu + \beta_j + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, 3, ..., a \\ j = 1, 2, 3, ..., b \end{cases}$$

In using this randomized block design, the overall MSE of the model is reduced and thus the power of the ANOVA test increases (i.e. the calculated F_0 statistic increases.) The partioning of the total sums of squares for a RBCD design can be shown:

$$SS_T = SS_{Blocks} + SS_{Treatments} + SS_E$$

Block designs are useful when experimental units are not completely homogenous throughout the entire dataset. The following example provides a unique case when block design should be considered.

Example 2: Suppose an experiment is conducted to test the therapeutic effects of three different drugs on a rare disease. Given that the disease is rare, it may be impossible to recruit enough patients all at once. It may be possible to instead administer the different drug options to groups of patients recruited at different time intervals. Patients can be grouped into blocks corresponding to the starting time they were administered their given drug. In doing so, the variability of the response variable due to the nuisance factor *time* may be controlled.

It should be noted that a RCBD is not completely random since the blocks are not randomly chosen. Only the assignment of the a treatment levels within a block is random. As such, it has been proposed to add another restriction error component in the statistical model which accounts for the restriction of randomization in the experiment.

The procedures for ANOVA F-tests concerning fixed or random effects block cases are still identical to the ones previously seen, which did not consider block design.

An example in R:
In the following experiment, 4 different machines (A, B, C and D) were tested on the number

of units they can each produce on 5 different days. The researchers would like to test if the machines are producing different amounts of units, while controlling for the nuisance factor day. The data was obtained from *Xu (2014)*.

The data is analyzed with the following R output:

```
machine.data <- read.csv("Blocking_example.csv")
machine.data
   Machine Day Output
1        A   1    293
2        A   2    298
3        A   3    280
4        A   4    288
5        A   5    260
6        B   1    308
7        B   2    353
8        B   3    323
9        B   4    358
10       B   5    343
11       C   1    323
12       C   2    343
13       C   3    350
14       C   4    365
15       C   5    340
16       D   1    333
17       D   2    363
18       D   3    368
19       D   4    345
20       D   5    330

> class(machine.data$Day)
#Determine the class of the variable "Day".
[1] "integer"

> machine.data$Day <- as.factor(machine.data$Day)
#Convert the class of variable "Day" into a factor.

> class(machine.data$Day)
#Verify that the conversion was made.
[1] "factor"

> anova.machine <-aov(Output ~ Machine + Day , data = machine.data)
#Run Anova with the factor and the blocking effect.
```

```
> anova(anova.machine)
#Display Anova results

Analysis of Variance Table

Response: Output
          Df  Sum Sq Mean Sq F value    Pr(>F)
Machine    3 13444.8  4481.6 20.4780 5.178e-05 ***
Day        4  2146.2   536.6  2.4517    0.1027
Residuals 12  2626.2   218.8
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

From the R output, we see that the ANOVA test gave significance for the factor *Machine* at the significance level of 0.05 (p-value = 5.178e-05). For demonstration purposes, we can compare these ANOVA results to an ANOVA model run without the blocking effect to visualize the importance of randomized blocking:

```
> anova.machine2 <-aov(Output ~ Machine, data = machine.data)
#Run Anova without blocking effects

> anova(anova.machine2)
Analysis of Variance Table

Response: Output
          Df  Sum Sq Mean Sq F value    Pr(>F)
Machine    3 13444.8  4481.6  15.025 6.486e-05 ***
Residuals 16  4772.4   298.3
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

With the blocking effect, the F-value increases from 15.025 to 20.478 and the obtained p-value decreases from $6.486e-05$ to $5.178e-05$. Thus, there is more power when the nuisance variable is accounted for. Also, the mean square error for the residuals decreases from 298.3 to 218.8 when the blocking factor is included in the model.

Balanced Incomplete Block Design

Many times in an experiment, it is not possible to allocate every treatment level of interest in each block. If only a subset of treatment levels k (where $k < a$), is present in each block, then it is referred to as an incomplete block design. The treatments are still randomly allocated in each block and all pairs of treatments appear together in the same number of blocks.

Since each treatment is not present within each block, the calculations for $SS_{Treatments}$ must be adjusted. Extra precautions must be taken to separate the effects caused by the

	Treatment			
Block	A	B	C	D
1	X	X	X	
2	X		X	X
3	X	X		X
4		X	X	X

Figure 1: An example of a balanced incomplete block design. Here, there are 4 levels and only a subset of 3 levels are assigned to each block. Image was obtained from *Borkowski(2015)*

.

treatments and the effects caused by the blocks.

$$SS_T = SS_{Blocks} + SS_{Treatments(Adjusted)} + SS_E$$

2.3 Latin square design

Latin square design can be used to control the variability of two extraneous nuisance factors in an experiment. Let p be the number of levels in each of the 2 nuisance factors. The two nuisance factors are assigned to the rows and the columns respectively, of a p x p latin square where each treatment level appears only once in each row and once in each column.

		Column			
		1	2	3	4
Rows	1	A	C	B	D
	2	D	A	C	B
	3	B	D	A	C
	4	C	B	D	A

Figure 2: A 4X4 latin square. The letters A,B,C and D represent the 4 different treatment levels. Each row is a level of the row nuisance factor and each column is a level of the column nuisance factor. The number of levels of each nuisance factor must equal the number of levels of the treatment. Image was obtained from *Borkowski(2015)*

.

Much like in the RCBD, in executing a latin square design the overall MSE will be reduced thus increasing the power of the F-Test. Through ANOVA, it can be shown that the partitioning of the sums of squares for a latin square design is:

$$SS_T = SS_{Rows} + SS_{Columns} + SS_{Treatments} + SS_E$$

Replication of Latin Squares

For latin squares, the degrees of freedom for SS_E is $(p-2)(p-1)$. If the levels of a factor are small, then the researcher may be concerned with a small corresponding degrees of freedom for SS_E. One solution to this problem is replicating the latin square. There are three unique ways to replicate a latin square design:

1. Use the same levels of both the rows and columns nuisance factors
2. Use different levels of the rows, but the same levels of the columns of the nuisance factors and vice versa.
3. Use different levels of both the rows and columns nuisance factors.

2.4 Repeated Measures Design

Many studies such as pharmaceutical or behavioural studies require that the experimental units be humans or animals. These types of studies may be problematic since the experimental units vary (eg. in terms of physical characteristics or health history), and thus many different factors may affect the variability in the response variable. The subject-to-subject variability can be controlled by instead designing an experiment such that each subject receives all a treatments. In this design, average responses will be free from subject-to-subject variability since each subject serves as their own control.

One issue with this experimental design is the possibility of carry over effects where there is interaction between treatment levels. In the study, it is crucial to prevent carry over effects (eg. by allowing enough time to pass between treatments). The order of treatment administered must also be completely randomized.

The total sums of squares for this design can be partitioned as follows:

$$SS_T = SS_{Betweensubjects} + SS_{Treatments} + SS_E$$

3 Post-Hoc analysis

It should be noted that the ANOVA only provides an overall omnibus test for the model. This implies that if statistical significance is obtained through ANOVA testing, the only conclusion that can be made is that at least one pair of levels differed from one another. Further post-hoc analysis is needed to determine which specific levels differed from one another. In this chapter, some multiple comparison methods are introduced which are used to make simultaneous comparisons among levels.

3.1 Multiple comparisons

It can be noted that an ANOVA F-test is very much similar to a two-sample t-test, in the sense that both tests can be run to test whether there is indeed a difference in mean between

two independent groups. In fact, it can be shown that for 2 levels (a = 2) the ANOVA test statistic is equal to the two-sample t-test test statistic (F = t^2). However, despite this similarity, it turns out that using multiple comparison t-tests is not a good solution once the number of levels is increased beyond two. The reasons are as follows:

1. Firstly, it may take a long time to use solely the t-test for all possible combination of comparisons of the levels under study.

2. Secondly, the type I error rate (α) becomes inflated when multiple pairwise comparisons are conducted on the same data using the two sample t-test. In other words, if the hypothesis testing is conducted at a significance level α, then α applies to only one individual test and not for a set of tests. *Borkowski (2015).*

To visualize this problem, consider the following example.

Example 3: An experiment is conducted to test the equality of means for four different levels. Given a significance level of $\alpha = 0.05$, if the null hypothesis is true then there is a $(1-\alpha) = 0.95$ chance of obtaining the correct decision for one pairwise comparison. However, given that there are 6 total possible unique comparisons that can be made on the same data, the probability of obtaining the correct decision in *all* comparisons made is $(1-\alpha)^6 = 0.74$, assuming that all groups are mutually independent. This implies that the type I error rate is inflated to: 1 - 0.74 = 0.26. The probability of making at least one error in all comparisons is 0.26. This is referred to as the experimentwise error rate (α_e.)

Multiple comparison procedures allow the analyst to make more than one comparison among two or more treatment levels. Many different procedures exist that vary in statistical power and in their approach to controlling the experimentwise error rate.

3.2 Contrasts

Comparisons between levels can be made in the form of contrasts.

Let

$$c_i \Big\{ i = 1, 2, ..., a$$

represent the contrast coefficients and μ_i represent the ith treatment mean.

A contrast is a linear combination of parameters of the form:

$$\Gamma = \sum_{i=1}^{a} c_i \mu_i$$

such that $\sum_{i=1}^{a} c_i = 0$.

Example 4: Suppose that in an experiment of a = 4 levels, the researcher would like to test whether the average of levels 1 and 2 differed from the average of levels 3 and 4. The

null and alternative hypotheses would be:

$$H_0 : \mu_1 + \mu_2 = \mu_3 + \mu_4$$

$$H_1 : \mu_1 + \mu_2 \neq \mu_3 + \mu_4$$

or

$$H_0 : \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$$

$$H_1 : \mu_1 + \mu_2 - \mu_3 - \mu_4 \neq 0$$

In this case, the contrast constants are $c_1 = c_2 = 1$ and $c_3 = c_4 = -1$. The example was chosen to illustrate that with contrasts, elaborate comparisons can be made. I.e. The comparisons can involve more than just two groups (unlike pairwise comparisons) which allows for more creative and versatile comparisons to be made.

With contrasts, a t-test can be used for hypothesis testing. The following t-test statistic is derived:

$$t_0 = \frac{\sum_{i=1}^{a} c_i \bar{y}_{i.}}{\sqrt{\frac{MS_E}{n} \sum_{i=1}^{a} c_i^2}}$$

The null hypothesis should be rejected if

$$|t_0| \geq t_{\frac{\alpha}{2}, N-a}$$

In conducting multiple comparisons, the analyst should take precautions and specify the contrasts before the experiment is conducted. If the contrasts are selected after preliminary data analysis, then it is likely that the analyst will choose observations with the highest observed differences that will likely be statistically significant. This is referred to as data snooping and may inflate the type 1 error rate.

Before data is collected, the researcher will usually have in mind the preplanned comparisons to be made if the ANOVA test turns out to be significant. These preplanned comparisons are usually the comparisons that are the most meaningful and biologically interesting. By only testing the comparisons that the researchers are interested in making, they can simplify their post-hoc analysis and reduce type 1 error rate. While there are some procedures such as Tukey's test that is able to conduct all possible pairwise comparisons while correcting for the experiment wise type 1 error rate, such tests may come at a cost of power (*Montgomery 2012).*

It should be noted that for random effects, the interest is focused on variance components and not on means. Therefore, multiple comparison analysis should only be conducted for fixed effects.

An example in R:
In the following dataset, measurements on the sizes of sturgeon (a type of fish) collected

from a river in Saskatchewan were studied. Measurements of 118 sturgeons were taken during the years: 1954, 1958, 1965 and 1966. The measurements taken were on the fork lengths (fklngth) of the sturgeons, which measures from the tip of the snout to the end of the middle caudal fin ray. It has been proposed that in 1960-1962, a dam was built in the Saskatchewan river which caused the bigger sturgeons to die while the smaller sturgeons were able to survive. Therefore, researchers hypothesized that the average size of the sturgeon is greater before 1960 and smaller after 1960. Data was obtained from *Findlay, S. et al (2016)* The researchers would like to test the following hypothesis:

$$H_0 : \mu_{1954} + \mu_{1958} - \mu_{1965} - \mu_{1966} = 0$$

$$H_1 : \mu_{1954} + \mu_{1958} - \mu_{1965} - \mu_{1966} > 0$$

Where μ_i represents the mean fork length at the specified year i. A boxplot plot is provided to help visualize the data:

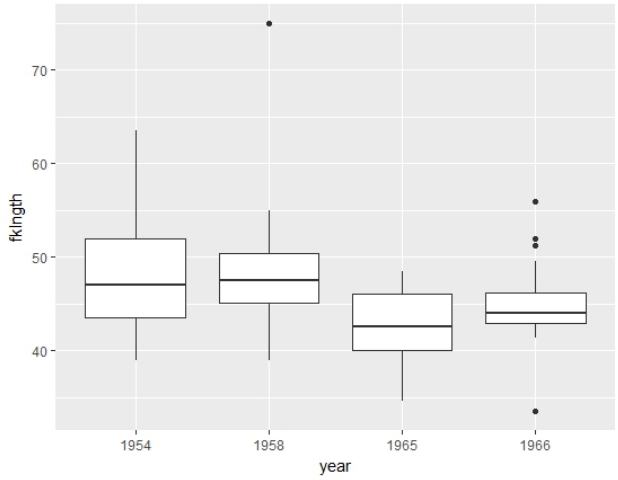

Figure 3: Boxplot for the fork lengths of sturgeon measured at 4 different years. The black lines in each box represents the median. The vertical length of the boxes represent the interquartile range. The "whiskers" which extend from the boxes represents the spread of the extreme points.

.

The following R output is as follows:

```
> Dam.data <-read.csv("Dam_data.csv", header = TRUE, sep = ",")

> class(Dam.data$year) #Determine class of variable "year"
[1] "integer"
```

```
> Dam.data$year <- as.factor(Dam.data$year)
# Convert the column "year" into categorical type.

> class(Dam.data$year) #Verify conversion was made correctly.
[1] "factor"

> model <- aov(fklngth ~ year, data = Dam.data)
#Set up the One-Way ANOVA

> summary(model)
             Df Sum Sq Mean Sq F value Pr(>F)
year          1    383   383.2    13.9  3e-04 ***
Residuals   116   3197    27.6
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

From the ANOVA table, the p-value is less than 0.05 (p-value = 3e-04), meaning that there is significance. We can now move on and test the pre-planned contrast in the post-hoc analysis. From the earlier stated null hypothesis, the contrast coefficients in order of ascending years are: $c_1 = c_2 = 1$ and $c_3 = c_4 =$ -1.

```
> contrast.coef.matrix <- matrix(c(1,1,-1,-1),nrow=1)
#Construct the contrast coefficient matrix

> contrast.coef.matrix #Verify matrix was made correctly
     [,1] [,2] [,3] [,4]
[1,]    1    1   -1   -1

> contrast.test <- glht(model, linfct = mcp(year=contrast.coef.matrix),
 alternative = "greater")
# Perform the one sided contrast comparison test.

> summary(contrast.test)

 Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: User-defined Contrasts
Fit: aov(formula = fklngth ~ year, data = Dam.data)

Linear Hypotheses:
       Estimate Std. Error t value   Pr(>t)
1 <= 0    9.008      2.162   4.167 3.01e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
(Adjusted p values reported -- single-step method)

> confint(contrast.test, level = 0.95)
# Compute a 95% confidence interval for the comparison made.

 Simultaneous Confidence Intervals
Multiple Comparisons of Means: User-defined Contrasts
Fit: aov(formula = fklngth ~ year, data = Dam.data)
Quantile = -1.6583
95% family-wise confidence level

Linear Hypotheses:
       Estimate lwr    upr
1 <= 0 9.0076   5.4230    Inf
```

From the R output, the p-value is smaller than 0.05 (p-value = 3.01e-05), meaning that the pooled average fork length during the years 1954 and 1958 is greater than the pooled average fork length during the years 1965 and 1966. The confidence interval for this estimated difference is $(5.423, \infty)$.
As an aside, it may have been worthwhile to first test if it is even appropriate to pool the average fork lengths before 1960-1962 and also to pool the average fork lengths after 1960-1962. I.e. we may wish to test whether the sizes of sturgeon were similar during the years 1954 and 1958 and also similar during the years 1965 and 1966. The hypotheses can be written as:

$$H_0 : \mu_{1954} - \mu_{1958} = 0 \qquad vs. \qquad H_1 : \mu_{1954} - \mu_{1958} \neq 0$$

and

$$H_0 : \mu_{1965} - \mu_{1966} = 0 \qquad vs. \qquad H_1 : \mu_{1965} - \mu_{1966} \neq 0$$

```
> contrast2 <- rbind("50's" = c(1,-1,0,0),
                     "60's" = c(0,0,1,-1))
#Construct a contrast coefficient matrix containing the contrast coefficients with the
hypotheses being tested as noted above.

> contrast2 #View the constructed contrast coefficient matrix.
    [,1] [,2] [,3] [,4]
50's    1   -1    0    0
60's    0    0    1   -1

> contrast.test2 <- glht(model, linfct = mcp(year = contrast2))

> summary(contrast.test2)

 Simultaneous Tests for General Linear Hypotheses
```

```
Multiple Comparisons of Means: User-defined Contrasts
Fit: aov(formula = fklngth ~ year, data = Dam.data)

Linear Hypotheses:
         Estimate Std. Error t value Pr(>|t|)
50s == 0  -0.1872     1.3335   -0.14    0.988
60s == 0  -2.1950     1.7012   -1.29    0.358
(Adjusted p values reported -- single-step method)

> confint(contrast.test2)
#Obtain 95% confidence intervals

 Simultaneous Confidence Intervals
Multiple Comparisons of Means: User-defined Contrasts
Fit: aov(formula = fklngth ~ year, data = Dam.data)
Quantile = 2.2654
95% family-wise confidence level


Linear Hypotheses:
         Estimate lwr     upr
50s == 0 -0.1872  -3.2081  2.8337
60s == 0 -2.1950  -6.0489  1.6590
```

From the output, we observe that the p-values obtained from the 2 comparisons are both greater than 0.05 (0.988 and 0.358 respectively). Also, both confidence intervals produced contained the value 0 ((-3.2081, 2.8337) and (-6.0489, 1.6590).) Therefore, there are no significant differences among the closely related years and it is indeed appropriate to pool the closely related years together.

3.3 Fishers Least Significant Difference

If the F-test null hypothesis is rejected, then it is possible to perform individual t-tests comparing any pair of treatment means. Specifically, this procedure tests:

$$H_0 : \mu_i = \mu_j \qquad vs. \qquad H_1 : \mu_i \neq \mu_j$$

for all i and j pairs.

The least significant difference quantity is defined as:

$$LSD = t_{\frac{\alpha}{2},N-a}\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}$$

Where:

- $t_{\frac{\alpha}{2},N-a}$ is the critical t-statistic
- N represents the total number of observations
- n represents the number of observations in the ith or jth treatment level
- $i = 1, 2, ..., a$, $i \neq j$

The null hypothesis is rejected if:

$$|\bar{y}_{i.} - \bar{y}_{j.}| > LSD$$

where $\bar{y}_{i.}$ is a point estimate for μ_i.

This procedure does not guard against the inflation of type 1 error rate. Therefore, if many comparisons are made, there is an increased probability of obtaining erroneous conclusions.

An example in R: From the previous sturgeon data set, Fisher's LSD test can be used to make all pairwise comparisons for each level of the factor *year*.

```
> LSD.test(model, "year", console = TRUE)
#Perform Fisher's LSD test.

Study: model ~ "year"
LSD t Test for fklngth
Mean Square Error:  27.15171

year,  means and individual ( 95 %) CI

      fklngth      std  r      LCL      UCL  Min  Max
1954 48.02432 5.864365 37 46.32733 49.72132 39.0 63.5
1958 48.21154 6.757474 26 46.18715 50.23593 39.0 75.0
1965 42.51667 4.123289 12 39.53684 45.49649 34.6 48.5
1966 44.71163 3.547716 43 43.13747 46.28578 33.5 56.0

alpha: 0.05 ; Df Error: 114
Critical Value of t: 1.980992

t-Student: 1.980992
Alpha    : 0.05
Minimum difference changes for each comparison

Means with the same letter are not significantly different
Groups, Treatments and means
```

```
a   1958   48.21154
a   1954   48.02432
b   1966   44.71163
b   1965   42.51667
```

From the R output, the only pairs that did not differ significantly from each other (i.e. pairs whose absolute difference did not exceed Fisher's LSD) are the pairs: 1954 - 1958 and 1965 - 1966. These results are the same as the results seen previously in the post-hoc test using preplanned contrasts.

3.4 Tukeys Honest Significant Difference Test

In many cases, all of the pairwise comparisons of each treatment level are of interest. To determine which treatment means differ, it is indeed possible to conduct tests which compare all possible pairs of treatment means. That is, all possible pairwise comparisons are made between all a treatment levels. In this case, the null and alternative hypotheses that are tested are as follows: For all $i \neq j$.

$$H_0 : \mu_i = \mu_j$$

$$H_1 : \mu_i \neq \mu_j$$

Tukey's approach of testing hypotheses for all possible pairwise comparisons is unique in that this approach can keep the overall significance level (the experiment wise error rate) at α. The type I error rate remains the same, irrespective of the number of comparisons made. Tukey's test is based on the studentized range distribution and works as follows: If the absolute value of a pair of means (i and j) is greater than the critical value, then the null hypothesis should be rejected.

$$|\mu_i - \mu_j| \geq T_\alpha = q_\alpha(a, f)\sqrt{\frac{MS_E}{n}}$$

Where $q_\alpha(a, f)$ is the critical value from a studentized range distribution.

An example in R:
The following dataset comes from an experiment which tested the effects of three different virus treatments (denoted *cc, ff and fc*) and a control treatment with no virus (denoted *oo*) on the yield of sweet potatoes. Each treatment was made three times. The data was obtained from the R package *agricolae*.

```
> library("agricolae")
> data("sweetpotato") #Load the data
> sweetpotato #Display the data

   virus yield
1     cc  28.5
```

```
2     cc  21.7
3     cc  23.0
4     fc  14.9
5     fc  10.6
6     fc  13.1
7     ff  41.8
8     ff  39.2
9     ff  28.0
10    oo  38.2
11    oo  40.4
12    oo  32.1
```

To better visualize the data, the means of each treatment level can be computed as follows:

```
> attach(sweetpotato)
> tapply(yield, virus ,mean)
#Loop function that computes the mean at each virus level.
      cc       fc       ff       oo
24.40000 12.86667 36.33333 36.90000
```

These means can be further visualized with a gg plot.

```
> ggplot(data = sweetpotato, aes(x = virus, y = yield)) + geom_boxplot(notch = FALSE)
```

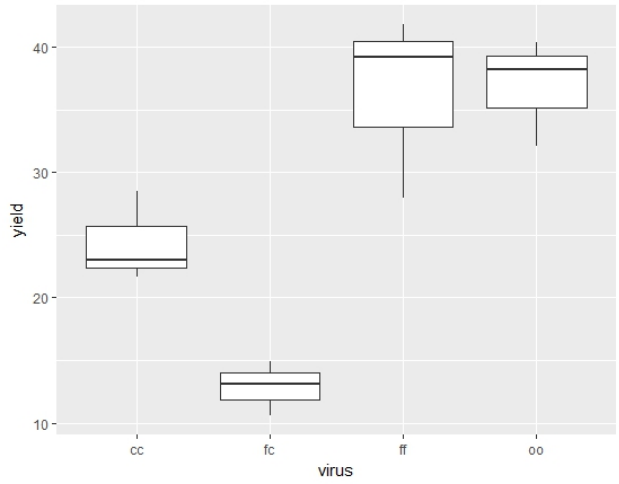

Figure 4: Boxplot representing the yield of potatoes for the different levels of *virus*.

Continuing with ANOVA testing, we get the following results:

```
> model2 <- aov(yield ~ virus , data = sweetpotato)
> anova(model2)
Analysis of Variance Table

Response: yield
          Df  Sum Sq Mean Sq F value    Pr(>F)
virus      3 1170.21  390.07  17.345 0.0007334 ***
Residuals  8  179.91   22.49
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Since the p-value obtained is 0.0007334, the null hypothesis can be rejected at the level of $\alpha = 0.05$. Thus, we may wish to perform post-hoc analysis to determine which specific treatment levels differ from each other.

```
> model <- aov(yield ~ virus , data = sweetpotato)

> posthoc.tukey <- glht(model, linfct = mcp(virus = "Tukey"))
#Perform Tukey's multiple comparison test.

> summary(posthoc.tukey)

 Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = yield ~ virus, data = sweetpotato)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
cc - oo == 0 -12.5000     3.8721  -3.228  0.04807 *
fc - oo == 0 -24.0333     3.8721  -6.207  0.00132 **
ff - oo == 0  -0.5667     3.8721  -0.146  0.99879
fc - cc == 0 -11.5333     3.8721  -2.979  0.06831 .
ff - cc == 0  11.9333     3.8721   3.082  0.05916 .
ff - fc == 0  23.4667     3.8721   6.061  0.00136 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Adjusted p values reported -- single-step method)

> confint(posthoc.tukey)
#Obtain 95% confidence intervals for each pairwise comparison

 Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = yield ~ virus, data = sweetpotato)
```

```
Quantile = 3.2034
95% family-wise confidence level


Linear Hypotheses:
             Estimate  lwr       upr
cc - oo == 0 -12.50000 -24.90356  -0.09644
fc - oo == 0 -24.03333 -36.43689 -11.62977
ff - oo == 0  -0.56667 -12.97023  11.83689
fc - cc == 0 -11.53333 -23.93689   0.87023
ff - cc == 0  11.93333  -0.47023  24.33689
ff - fc == 0  23.46667  11.06311  35.87023

> plot(posthoc.tukey)
#Plot the confidence intervals for each pairwise comparison.
```

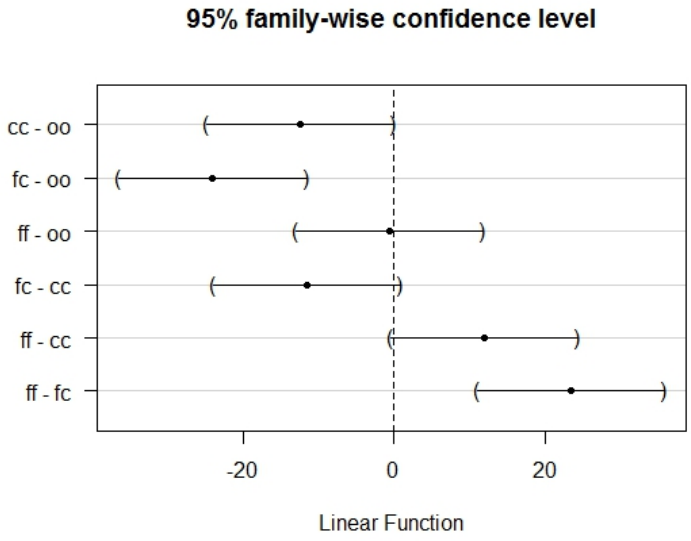

Figure 5: Confidence intervals for all pairwise comparisons of the yield of potatoes for the factor *virus*

.

From the R output, we see that only the comparisons of the levels cc - oo, fc - oo, and ff - fc were significant at the significance level of $\alpha = 0.05$. For these comparisons, their respective p-values are below 0.05 (0.04807, 0.00132 and 0.00136 respectively.) Additionally, the confidence intervals for these pairwise comparisons do not include the value 0.

3.5 Dunnetts Test

Dunnetts test can also be used as a follow up test to ANOVA if the experiment conducted included a control level. For Dunnetts procedure, the analyst is only interested in comparing each treatment level to the control level such that the total number of comparisons made is exactly $(a-1)$. This procedure also controls the experiment wise type 1 error rate like in Tukeys test, however the Dunnetts test makes fewer comparisons which results in greater statistical power and narrower confidence intervals for the difference of two treatment levels. The null hypothesis for the Dunnetts test can be written as follows:

$$H_0 : \mu_i = \mu_a$$

$$H_1 : \mu_i \neq \mu_a$$

Where μ_a is the control mean and $i = (1, 2, ..., a-1)$

The null hypothesis should be rejected if

$$|\bar{y}_{i.} - \bar{y}_{a.}| \geq d_\alpha(a-1, f)\sqrt{MSE(\frac{1}{n_1} + \frac{1}{n_a})}$$

where $d_\alpha(a-1, f)$ is a critical value calculated for Dunnett's test.

An example in R:
In the previous dataset, the experiment involving the study of yield of sweet potatoes with varying levels of virus treatments can be further analysed. Since one of the treatment levels included in the experiment was a control level, we may be interested in *only* comparing each treatment level to the baseline control level. Thus, Dunnett's test may be an appropriate test to consult in the post-hoc analysis.

```
##Dunnett's test in R assumes that the first level of the factor is the control level.
#However, by default the levels of a factor are ordered in alphabetical order.

> levels(sweetpotato$virus)
#Check the levels of the factor "virus"
[1] "oo" "cc" "fc" "ff"

> sweetpotato$virus <- factor(sweetpotato$virus, levels = c("oo", "cc", "fc", "ff"))
#Re-order the levels of the virus, such that the control is the first level.

> levels(sweetpotato$virus)
#Check that the levels of the factor "virus" were appropriately changed.
[1] "oo" "cc" "fc" "ff"

> model <- aov(yield ~ virus , data = sweetpotato)
```

```
> potato.dunnetts = glht(model, linfct = mcp(virus = "Dunnett"))
#Run Dunnett's test

> summary(potato.dunnetts)

 Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Dunnett Contrasts
Fit: aov(formula = yield ~ virus, data = sweetpotato)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
cc - oo == 0 -12.5000     3.8721  -3.228   0.0298 *
fc - oo == 0 -24.0333     3.8721  -6.207   <0.001 ***
ff - oo == 0  -0.5667     3.8721  -0.146   0.9976
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Adjusted p values reported -- single-step method)
```

At the level of $\alpha = 0.05$, the output for Dunnett's test indicates that there is a statistically significant different between the levels *cc* and *fc* to the control level (p-values are 0.0298 and < 0.001 respectively.)

4 Model adequacy

Recall that assumptions about the random errors (ϵ_{ij}) were made for the one-way ANOVA model. Before interpreting the results from the ANOVA test, it is essential to verify the validity of the underlying assumptions; otherwise, interpreting the results of the test are not appropriate. The model adequacy can be verified through residual analysis.

If the assumptions of normality and of equal variances are not sufficiently met, then extra precautions can be taken to analyse the data. For example, if the assumption of constant variance is not valid, then one solution could be to transform the data. Data transformation can stabilize the variance. ANOVA is robust to deviations from the normality assumption, given that the sample size is large enough. For a small sample size that is not normally distributed, other approaches like a randomization approach can be used instead.

An example in R:
From the *sweetpotato* dataset, the model adequacy can be checked with the following residual plot:

From the QQ plot, it appears most of the points lie on the line or are closely situated to the line. Therefore, the assumption of normality is met. In the residual vs fitted values plot, there appears to be an equal spread above and below the horizontal line with some slight deviation near the right side of the plot. To further test the homogeneity of variance assumption, Levenes test can be run:

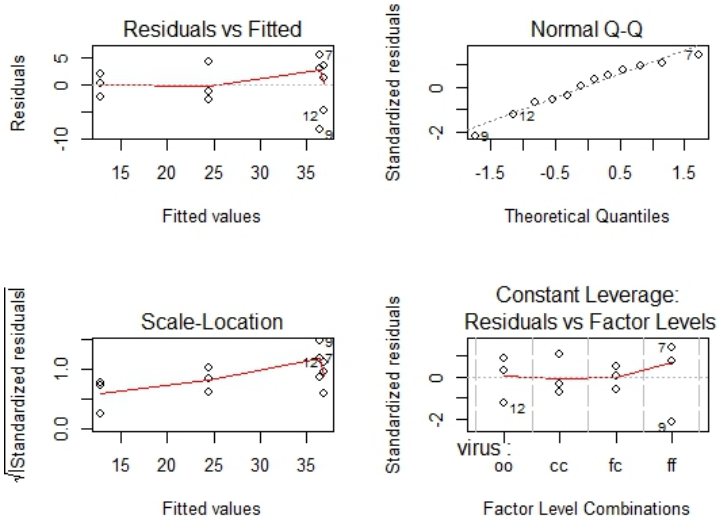

Figure 6: Residual diagnostic plot for the ANOVA model run on the dataset *sweetpotato*

.

```
> leveneTest(model)
#HOV test. H_0: There is HOV, H_1: There is no HOV.

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.4004 0.7566
       8
```

Since the p-value output from Levenes test is greater than 0.05, there is not enough sufficient evidence to reject the null hypothesis of homogeneity of variance.

4.1 Non- Parametric approach: Kruskal Wallis

An alternative to the F-test ANOVA when the normality assumption is not justified is the Kruskal Wallis test. This test is a non-parametric approach which is used to test the same hypotheses seen for a fixed factor (equality of means). The approach works first by ordering the observations y_{ij}'s in ascending order and assigning the new order of observations with ranks (R_{ij}'s). The smallest observation begins with an assigned rank 1. If observations have the same value, then the average rank is assigned to the same tied observations. The test statistic is then calculated as follows:

$$H = \frac{1}{S^2}\left(\sum_{i=1}^{a}\frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4}\right)$$

Under the null hypothesis: $H \sim \chi^2_{a-1}$. The null hypothesis should be rejected if $H > \chi^2_{\alpha,a-1}$

An example in R:
For illustration purposes, we can run a non-parametric Kruskal-wallis test (which would normally be run if the assumptions of a parametric ANOVA are not met) and compare its results to a parametric ANOVA test on the ANOVA model with the *sweetpotato* dataset.

```
>kruskal.test(yield ~ virus, data = sweetpotato)

Kruskal-Wallis rank sum test

data:  yield by virus
Kruskal-Wallis chi-squared = 8.6923, df = 3, p-value = 0.03367
```

The p-value obtained from the Kruskal-Wallis test is significant, however it is still smaller than the p-value obtained from the parametric ANOVA (0.03367 vs 0.000734). The assumptions of a parametric ANOVA are met and therefore the use of a parametric ANOVA has more statistical power than a non-parametric test.

5 Appendix

5.1 Dot-Subscript notation

Throughout this document, dot subscript notation is used to symbolize some data components. In this notation, the position of the dot represents the total summation of the subscript it replaces. For example, $\bar{y}_{i.}$, represents the mean of observations belonging to the ith level and $\bar{y}_{..}$, represents the grand total mean (pooled across all levels). For a more formal view:

$$y_{i.} = \sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{i.} = \frac{\bar{y}_{.}}{n} \quad i = 1, 2, ..., a$$
$$y_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{..} = \frac{y_{..}}{N}$$

5.2 Residual

A residual is defined as follows:

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

where $\hat{y}_{ij}$ is an estimate for the observation y_{ij} and $y_{ij} = \bar{y}_{i}$.

Through residual analysis, the normality assumption of the model can be verified with a normal probability plot of residuals. The pattern in this plot should be approximately linear if the residuals are indeed normally distributed. The equal variance assumption can be verified with residuals vs fitted values plot. Other formal statistical tests such as the Levenes test also exist to test the validity of the constant variance assumption.

6 References

Borkowski, J. (2015). The Random effects model. Montana State University. http://www.math.montana.edu/jobo/st541/sec2d.pdf

Findlay, S and Morin, A and Rundle, H. (2016). Applied Biostatistics. Department of Biology, University of Ottawa.

Montgomery, Douglas C. (2012). Experiments with a Single Factor: The Analysis of Variance. Design and Analysis of Experiments. 8th ed. N.p.: John Wiley & Sons. 65-130. Print.

Motulsky, H. (2014). Intuitive Biostatistics: A Nonmathematical guide to statistical thinking (3rd ed.). Oxford University Press. 381-389.Print

PennState Eberly College of Science. (2017). Design of Experiments. https://onlinecourses.science.psu.edu/stat503/node/12

Tangren, J. (2002). A Field Guide to Experimental Designs. Washington State University. http://www.tfrec.wsu.edu/ANOVA/index.html

Xu, Jing.(2014). Birkbeck, University of London. Statistical Modelling. http://www.bbk.ac.uk/ems/fac downloads/SML4.pdf