Chapter 18

One-Way Analysis of Variance

In this chapter we will discuss the following topic:

- How to use one-way analysis of variance (ANOVA), to compare the population means from two or more normal distributions. In R we will use the command **anova(lm())**.

In chapter 12, we compared the population means of two normal distributions. In this chapter, we will compare the means of several normal distributions. We will use a one-way analysis of variance (ANOVA), to test for differences in two or more means.

F-Distributions

The F-statistic is defined as

$$F = \frac{U_1/df_1}{U_2/df_2},$$

where U_1 and U_2 are independent chi-square random variables with df_1 and df_2 degrees of freedom, respectively. The distribution of F is called the Fisher's F distribution with df_1 and df_2 degrees of freedom. We denote the distribution by $F(df_1, df_2)$.

Problem. Let $df_1 = 2$ and $df = 7$. Find $P(F < 4.74)$.

Solution. We obtain

```
> pf(4.74,2,7)
[1] 0.9500549
```

Hence, $P(F < 4.74) = 0.95$.

We denote the critical value for which there is probability α to the right for it under the density curve for the F-distribution by $F_\alpha(df_1, df_2)$.

Problem. Find $F_{0.05}(2, 5)$

Solution. We obtain

```
> qf(0.05,2,5,lower.tail=F)
[1] 5.786135
```

Hence, $F_{0.05}(2, 5) = 5.786135$.

One-Way ANOVA

Suppose that we have m treatment means in an experiment that we would like to compare. We can also think about this as m different levels of a single factor.

For each $i = 1, 2, .., m$, let Y_{i1}, Y_{i2},..., Y_{in_i} be a simple random sample of size n_i from a Normal distribution $N(\mu_i, \sigma^2)$ with unknown variance σ^2 and unknown mean μ_i.

We assume that each of the m Normal distributions have the same variance. We are interested in the means model

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{for} \quad i = 1, 2, ..., m, \quad \text{and} \quad j = 1, 2, ..., n_i. \tag{0.1}$$

Here y_{ij} is the j^{th} observation for the i^{th} population, μ_i is the population mean of the i^{th} treatment or population, and ϵ_{ij} is a random error. The errors ϵ_{ij} are independent and normally distributed random variables with mean zero and unknown variance σ^2. The errors include all experimental variation such as measurements error and background noise. Notice that the expected value of Y_{ij} is μ_i. Equation (**??**) is called a one-way ANOVA since only one factor is under consideration. The experimental design used for such experiments is completely randomized design.

We wish to test the hypothesis that all the populations means are equal against the alternative hypothesis that at least one of the population means differ, i.e.,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m \quad \text{against} \quad H_a : \mu_i \neq \mu_j$$

for some pair (i, j).

We start with a simple example to get an idea on how to proceed.

Example. Suppose that we are performing an experiment for which we have three treatment means to compare. Suppose that we have two observations from group/population 1, three observations from group/population 2, and three observations from group/population 3. Here are the data:

Table 1: One-Factor Random Samples

				Means of i^{th} Group
Group1	1	3		$\overline{Y}_{1.} = 2$
Group2	4	5	6	$\overline{Y}_{2.} = 5$
Group3	6	8	7	$\overline{Y}_{3.} = 7$

We will first show how can we decompose the total variability in the data. Let n_i denote the sample size of the i^{th} group such that $n_1 = 2$, and $n_2 = n_3 = 3$. We let $\overline{Y}_{i.}$ denote the mean of the observations in the i^{th} treatment/group which is here given by

$$\overline{Y}_{i.} = \frac{1}{n_i}(Y_{i1} + Y_{i2} + \cdots + Y_{in_i}) = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

for $i = 1, 2, 3$, and where the dot indicates the index over which the average is taken. We have

$$\overline{Y}_{1.} = \frac{1+3}{2} = 2,$$

$$\overline{Y}_{2.} = \frac{4+5+6}{3} = 5,$$

and

$$\overline{Y}_{3.} = \frac{6+8+7}{3} = 7.$$

The total number of observations is then $n = n_1 + n_2 + n_3 = 2 + 3 + 3 = 8$. Let $\overline{Y}..$ denote the **grand mean** of all the n observations, which is given by

$$\overline{Y}.. = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{ij} = \frac{1+3+4+5+6+6+8+7}{8} = 5,$$

where the average is taken over both indices.

The total sum of squares, SS_T, is a measure of the total variability in the data and is given by

$$SS_T = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}..)^2 = (1-5)^2+(3-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(6-5)^2+(8-5)^2+(7-5)^2 = 36.$$

We notice here that we add all the square deviations of each observation from the grand mean.

Since we have $n = 8$ observations, we have 7 variables that can vary freely so the degrees of freedom is $df = n - 1 = 8 - 1 = 7$.

We will now show that we can decompose the total variability in the data into two parts, in which one part is coming from the variation within the groups and the other part from variation between groups.

We will first consider the variation within the treatments or groups. We add the square deviations of each observation within a group from its sample mean and add these contributions from each group to obtain the **error sum of squares**, SS_E, which is given by

$$SS_E = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i.})^2 = (1-2)^2+(3-2)^2+(4-5)^2+(5-5)^2+(6-5)^2+(6-7)^2+(8-7)^2+(7-7)^2 = 6.$$

The degrees of freedom for each group is $n_i - 1$ so the total degrees of freedom is $df = (n_1 - 1) + (n_2 - 1) + (n_3 - 1) = n - m = 8 - 3 = 5$.

We next consider the variation between treatments or groups and look at the weighted square deviations of the sample means from the grand mean to obtain the **between**

treatment sum of squares, SS_{Trt}, where

$$\begin{aligned} SS_{Trt} &= \sum_{i=1}^{m} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \\ &= (2-5)^2 + (2-5)^2 + (5-5)^2 + (5-5)^2 + (5-5)^2 + (7-5)^2 + (7-5)^2 + (7-5)^2 \\ &= 2(2-5)^2 + 3(5-5)^2 + 3(7-5)^2 \\ &= 30. \end{aligned}$$

Since we have $m = 3$ groups, the degrees of freedom is $df = m - 1 = 3 - 1 = 2$.

Now notice that we have the following **ANOVA identity**:

$$SS_T = 36 = 6 + 30 = SS_E + SS_{Trt} \tag{0.2}$$

with degrees of freedom:

$$n - 1 = 7 = 5 + 2 = (n - m) + (m - 1).$$

The question we now wish to ask is how much of the total variation in the data is due to the variation between groups versus the variation within groups?

If we assume that the three treatments means are equal, i.e. that the null hypothesis is true, then we have the following F-statistics,

$$F = \frac{\frac{SS_{Trt}}{m-1}}{\frac{SS_E}{n-m}} = \frac{\frac{30}{2}}{\frac{6}{5}} = 12.5$$

which has 2 and 5 degrees of freedom. If the three treatment means have similar values, the variation between treatments will be small and hence the numerator of F will be small. Then more of the variation in the data will be due to SS_E so the denominator will be large. On the other hand, if the three treatments means differ, the numerator of F will be large and the denominator will be small such that F will be large. Thus, we reject the null hypothesis at the α-level if the observed value of F is larger than the critical value $F_\alpha(m-1, n-m)$.
Here the critical value is $F_{0.05}(2,5) = 5.7861$ which is smaller than $F = 12.5$ so we reject the null hypothesis. We can also calculate the p-value which is given by $P(F > 12.5) = 0.01134023 < 0.05$.

Summary of ANOVA
We have the ANOVA identity,

$$SS_T = SS_{Trt} + SS_E, \tag{0.3}$$

where

$$SS_T = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2,$$

$$SS_E = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i}.)^2 \quad \text{and} \quad SS_{Trt} = \sum_{i=1}^{m} n_i(\overline{Y}_{i}. - \overline{Y}..)^2.$$

Define the **mean squared for treatments**,

$$MS_{Trt} = \frac{SS_{Trt}}{m-1} = \frac{\sum_{i=1}^{m} n_i(\overline{Y}_{i}. - \overline{Y}..)^2}{m-1} = \frac{n_1(\overline{Y}_{1}. - \overline{Y}..)^2 + \cdots + n_m(\overline{Y}_{m}. - \overline{Y}..)^2}{m-1}$$

which is a measure of the variation among the m sample means. It is a weighted average of the square deviations of the sample means from the grand mean, $\overline{Y}..$, and can be shown to be an estimate of the variance σ^2 if the means are equal.
Define the **mean squared for errors**,

$$MS_E = \frac{SS_E}{n-m} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i}.)^2}{n-m} = \frac{(n_1-1)S_1^2 + \cdots + (n_m-1)S_m^2}{n-m},$$

where $n = n_1 + n_2 + \cdots + n_m$, and where $S_i^2 = \frac{\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i}.)^2}{(n_i-1)}$ is the i^{th} sample variance. MS_E is a measure of the variation of observations within a group given as a weighted average of the sample variances, and can be shown to be an estimate of σ^2.

The ANOVA identity in (**??**) leads to two estimates of σ^2, one due to variation within treatments or groups, and the other due to variation between treatments or groups. If the treatment means are equal, these two estimates should be similar and unbiased estimator of σ^2. If the treatment means differ, it can be shown that the estimate provided by MS_{Trt} is greater than σ^2 and its estimate will differ from the estimate of MS_E.

We define

$$F = \frac{MS_{Trt}}{MS_E} = \frac{\frac{SS_{Trt}}{m-1}}{\frac{SS_E}{n-m}}$$

which has the F distribution with $m-1$ and $n-m$ degrees of freedom if H_0 is true. We reject the null hypothesis that the means are equal if $F > F_\alpha(m-1, n-m)$, where $F_\alpha(m-1, n-m)$ is the critical value with $m-1$ and $n-m$ degrees of freedom and significance level α.
The conditions for inference for ANOVA are:

- The samples from each population are independent random samples that provide a measure of the same variable.
- The populations under questions are Normally distributed with *unknown* population means. Because of the robustness of the ANOVA F-test, it is enough to check that the samples are approximately Normal. Also with large sample sizes the normality assumption can be relaxed unless the distribution is severely skewed with outliers. For large sample sizes the sample means will be approximately normally distributed due to the Central limit theorem.

- The populations have the same *unknown* standard deviations. As a rule of thumb, the largest sample standard deviation should be no more than twice as large as the smallest sample standard deviation.

The Regression Approach to the ANOVA Method
We will now return to the means model for three treatments:

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{for} \quad i = 1, 2, 3, \quad \text{and} \quad j = 1, 2, ..., n_i. \tag{0.4}$$

We will here consider the three different treatments as three *levels* of a *factor* or a qualitative variable. We use indicator variables x_{1j}, x_{2j} to represent the levels such that

$$x_{1j} = \begin{cases} 1 & \text{if the } j^{th} \text{ observed value is from treatment 2;} \\ 0 & \text{otherwise.} \end{cases}$$

and

$$x_{2j} = \begin{cases} 1 & \text{if the } j^{th} \text{ observed value is from treatment 3;} \\ 0 & \text{otherwise.} \end{cases} .$$

Since we have three treatment levels, we use two indicator variables. The means model in (**??**) can be written in terms of a regression model,

$$Y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \epsilon_{ij} \quad \text{for} \quad i = 1, 2, 3, \quad j = 1, 2, ..., n_i.$$

(This model is a type of a multiple regression model that we will return to in the next chapter). The relationship between β_0, β_1, β_2, and μ_1, μ_2, and μ_3 will now be explained. Consider first the observations from treatment 1 for which both $x_1 = x_2 = 0$. Then

$$Y_{1j} = \beta_0 + \epsilon_{1j} \quad \text{with} \quad \mu_1 = \beta_0.$$

Now for observations from treatment 2, we have $x_1 = 1$ and $x_2 = 0$ such that

$$Y_{2j} = \beta_0 + \beta_1 + \epsilon_{2j} \quad \text{with} \quad \mu_2 = \beta_0 + \beta_1.$$

Finally, for observations from treatment 3, we have $x_1 = 0$ and $x_2 = 1$ such that

$$Y_{3j} = \beta_0 + \beta_2 + \epsilon_{3j} \quad \text{with} \quad \mu_3 = \beta_0 + \beta_2.$$

We use least squares to estimate the parameters β_0, β_1, and β_3. It can be shown that the ANOVA method is equivalent to the regression approach.

Next we will illustrate how we can solve the previous problem with R. We will test the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \text{ against } H_1 : \mu_i \neq \mu_j \text{ for at least one pair } (i, j) \text{ for } i, j = 1, 2, 3.$$

Solution. We obtain:

```
> Group1=c(1,3)
> Group2=c(4,5,6)
> Group3=c(6,8,7)
```

```
> treatment=c(rep(1,length(Group1)),rep(2,length(Group2)),rep(3,length(Group3)))
> treatmentfactor=factor(treatment,labels=c(1,2,3))
> y=c(Group1,Group2,Group3)
> g=lm(y~treatmentfactor)
> anova(g)
Analysis of Variance Table

Response: y
                Df Sum Sq Mean Sq F value  Pr(>F)
treatmentfactor  2     30    15.0    12.5 0.01134 *
Residuals        5      6     1.2
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
```

We obtain that the F-value is 12.5 with a p-value of $0.01134 < 0.05$ so we reject the null hypothesis at the 5% significance level. Thus the treatments means differ from each other. Notice that we are here not testing which of the individual treatment means that differ. We are only testing the fact that there is at least one pair of treatment means that differ.

Explanation. The code can be explained as follows:

- The command **treatment=c(rep(1,length(Group1)),rep(2,length(Group2)),rep(3,length** creates a vector for which 1 is repeated 2 times (length of **Group1**), 2 is repeated 3 times (length of **Group2**), and 3 is repeated 3 times (length of **Group3**) for which 1 corresponds to treatment 1, 2 corresponds to treatment 2, and 3 corresponds to treatment 3.
- The command **factor(treatment,labels=c(1,2,3))** converts the numerical values of **treatment** to a factor with three levels labeled 1,2,3.
- The function

  ```
  g=lm(y~treatmentfactor)
  ```

  performs a linear regression using indicator variables.
- The function **anova(g)** performs an analysis of variance.

Explanation. The output can be explained as follows:

- The **treatmentfactor** is the *between treatments*, where **Sum Sq** and **Mean Sq** are the *between treatment sum of squares,* SS_{Trt}, and the *mean squared for treatments,* MS_{Trt}, respectively.
- For the **Residuals**, the **Sum Sq** and **Mean Sq** are the *error sum of squares,* SS_E, and the *mean squared for errors,* MS_E, respectively.

We next find estimates for the population mean values:

```
> summary(g)

Call:
lm(formula = y ~ treatmentfactor)

Residuals:
         1          2          3          4          5          6          7
-1.000e+00  1.000e+00 -1.000e+00  2.557e-16  1.000e+00 -1.000e+00  1.000e+00
         8
 1.447e-16

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.0000     0.7746   2.582   0.0493 *
treatmentfactor2   3.0000     1.0000   3.000   0.0301 *
treatmentfactor3   5.0000     1.0000   5.000   0.0041 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.095 on 5 degrees of freedom
Multiple R-squared: 0.8333,     Adjusted R-squared: 0.7667
F-statistic:  12.5 on 2 and 5 DF,  p-value: 0.01134
```

We obtain the regression line:

$$\hat{Y}_{ij} = 2 + 3x_{1j} + 5x_{2j},$$

where

$$\overline{Y}_{1.} = \beta_0 = 2,$$

$$\overline{Y}_{2.} = \beta_0 + \beta_1 = 2 + 3 = 5,$$

and

$$\overline{Y}_{2.} = \beta_0 + \beta_2 = 2 + 5 = 7$$

which corresponds to the values in table 1.
We will now use R to analyze the following problem:

Problem. Cuckoos lay eggs in the nest of other host bird species who adopt and hatch their eggs. The survival of the eggs are due to the ability to lay eggs in nests that are properly adopted by species who serve as foster-parents. The birds mate and lay their eggs in the same territory year after year such that geographical sub species develops adopted by particular host species. The following data [1] shows the length of Cuckoos's egg in millimeters versus the host-parent species for the eggs:

```
> data
     MDW  TREE  HDGE ROBIN  PIED  WREN
1  19.65 21.05 20.85 21.05 21.05 19.85
2  20.05 21.85 21.65 21.85 21.85 20.05
3  20.65 22.05 22.05 22.05 21.85 20.25
4  20.85 22.45 22.85 22.05 21.85 20.85
5  21.65 22.65 23.05 22.05 22.05 20.85
.. ..    ..    ..     ..   ..    ..
```

where MDW=Meadow Pipit , TREE=Tree Pipit , HGDE=Hedge Sparrow, ROBIN= Robin , PIED= Pied Wagtail, and WREN=Wren.
Is there a significant difference between the mean lengths of eggs from the six different host species?

Solution. We wish to test the null hypothesis that there is no significant difference between the mean lengths of eggs from the six different host species against the alternative hypothesis that at least one pair of the population mean lengths differ, i.e.,

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$$

against

$$H_1 : \mu_i \neq \mu_j \text{ for at least one pair } (i, j) \text{ for } i, j = 1, 2, 3, 4, 5, 6.$$

We start by checking the conditions for inference:

```
> data=read.delim("eggs.txt")
> attach(data)
> MDW=MDW
> TREE=TREE[1:15]
> HDGE=HDGE[1:14]
> ROBIN=ROBIN[1:16]
> PIED=PIED[1:15]
> WREN=WREN[1:15]
> sd(MDW)
[1] 0.9206278
> sd(TREE)
[1] 0.9014274
> sd(HDGE)
[1] 1.068737
> sd(ROBIN)
[1] 0.6845923
> sd(PIED)
[1] 1.067619
> sd(WREN)
[1] 0.7437357
> stem(MDW)
```

```
  The decimal point is at the |

  19 | 6
  20 | 1
  20 | 69
  21 |
  21 | 666999
  22 | 111111111133333333444
  22 | 669999
  23 | 1334
  23 | 69
  24 | 34

> stem(TREE)

  The decimal point is at the |

  21 | 19
  22 | 146
  23 | 3334469
  24 | 111

> stem(HDGE)

  The decimal point is at the |

  20 | 9
  21 | 6
  22 | 19
  23 | 11114999
  24 | 1
  25 | 1

> stem(ROBIN)

  The decimal point is at the |

  21 | 19
  22 | 1113446
  23 | 1111139

> stem(PIED)

  The decimal point is at the |
```

```
  21 | 1999
  22 | 146
  23 | 1134
  24 | 1119

> stem(WREN)

  The decimal point is at the |

  19 | 9
  20 | 13
  20 | 999
  21 | 11134
  21 |
  22 | 1113
```

The largest standard deviation is less than twice as large as the smallest standard deviation, so the sample standard deviations are similar. We notice from the stem plot that the samples are approximately Normal but with some irregularities. There are outliers in sample 1 and 3 and some samples show some slight skewness. We have relatively large sample sizes so we will proceed with the ANOVA F-test:

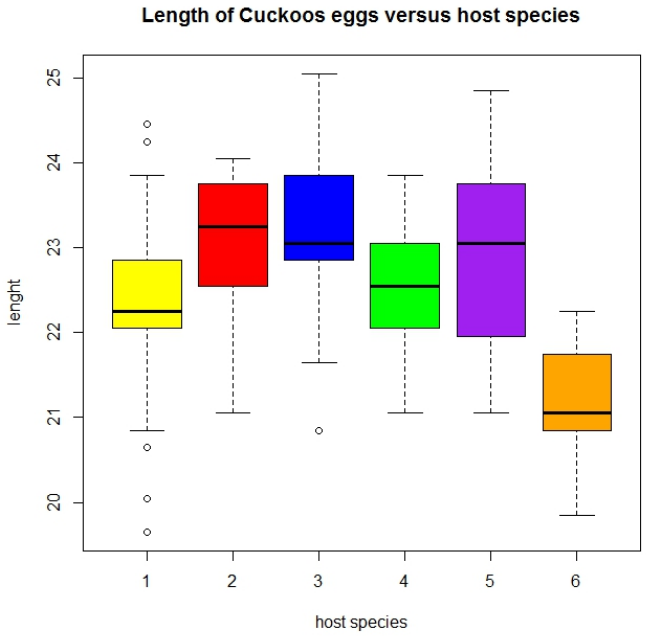
```
> treatment=c(rep(1,length(MDW)),rep(2,length(TREE)),
+ rep(3,length(HDGE)),rep(4,length(ROBIN)),rep(5,length(PIED)),rep(6,length(WREN)))
> treatmentfactor=factor(treatment,labels=c(1,2,3,4,5,6))
> y=c(MDW,TREE,HDGE,ROBIN,PIED,WREN)
> g=lm(y~treatmentfactor)
> anova(g)
Analysis of Variance Table

Response: y
                 Df Sum Sq Mean Sq F value    Pr(>F)
treatmentfactor   5 42.940  8.5879  10.388 3.152e-08 ***
Residuals       114 94.248  0.8267
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The p-value is much smaller than 0.01 so we reject the null hypothesis. Thus, there is a significant difference between the mean length of the eggs from different population of host species. We include a boxplot of the distribution of length of eggs versus host species:

```
< table=data.frame(y,treatment)
< boxplot(y~treatment,xlab="host species", ylab="lenght",
```

```
+ main= "Length of Cuckoos eggs versus host species", data=table,
+ col=c("yellow","red","blue","green","purple","orange"))
```



Explanation. The code can be explained as follows:

- The command TREE=TREE[1:15] selects the first 15 elements of the vector **TREE**. The remaining entries in the vector **TREE** are NA (missing values). There are other ways to select the values of the entries of the TREE variable by avoiding the NA but will not cover that technique in this tutorial.

References

[1] L.H.C. Tippett. *"The Methods of Statistics"*. 4th Edition, John Wiley and Sons, Inc., p. 176, 1952. The data and the story are also listed on DASL at http://lib.stat.cmu.edu/DASL/

00