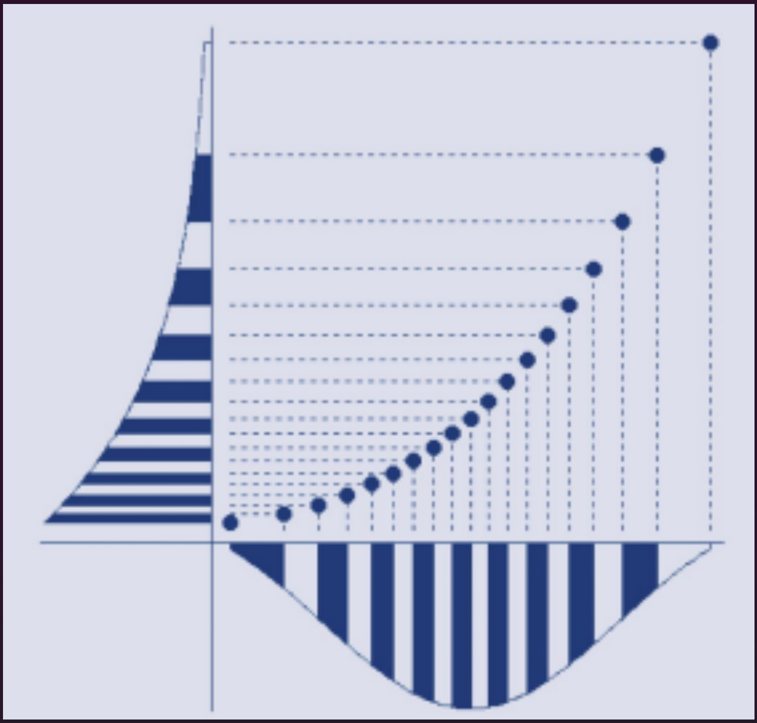
The R Series

Using R for Introductory Statistics

Second Edition



John Verzani

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

12

Analysis of variance

Analysis of variance, *ANOVA*, is a method of comparing means across samples based on variations from the mean. We begin by illustrating an ANOVA carried out in the traditional way, but we will see that the ANOVA model is just a special form of the linear model discussed in the previous chapter and R provides a common interface.

12.1 One-way ANOVA

A one-way analysis of variance is a generalization of the t-test for two independent samples, allowing us to compare population means for several independent samples. Suppose we have k populations of interest. From each we take a random sample. These samples are independent if the knowledge of one sample does not effect the distribution of another. Notationally, for the ith sample, let $x_{i1}, x_{i2}, \ldots, x_{in_i}$ designate the sample values.

The one-way analysis of variance applies to normally distributed populations. Suppose the mean of the ith population is μ_i and its standard deviation is σ_i. We use a σ if these are equivalent across the groups. A statistical model for the data with a common standard deviation is

$$x_{ij} = \mu_i + \epsilon_{ij},$$

where the error terms, ϵ_{ij}, are independent with $\mathsf{Normal}(0,\sigma)$ distribution.

• **Example 12.1: Number of calories consumed by month**
Consider 15 subjects split at random into three groups. Each group is assigned a month. For each group we record the number of calories consumed on a randomly chosen day. Figure 12.1 shows the data. We assume that the amounts consumed are normally distributed with common variance but perhaps different means. From the figure, we see that there appears to be more clustering around the means for each month than around the grand mean or mean for all the data. Perhaps more calories are consumed in the winter?

The goal of one-way analysis of variance is to decide whether the difference in the sample means is indicative of a difference in the population means or is attributable to sampling variation. ••

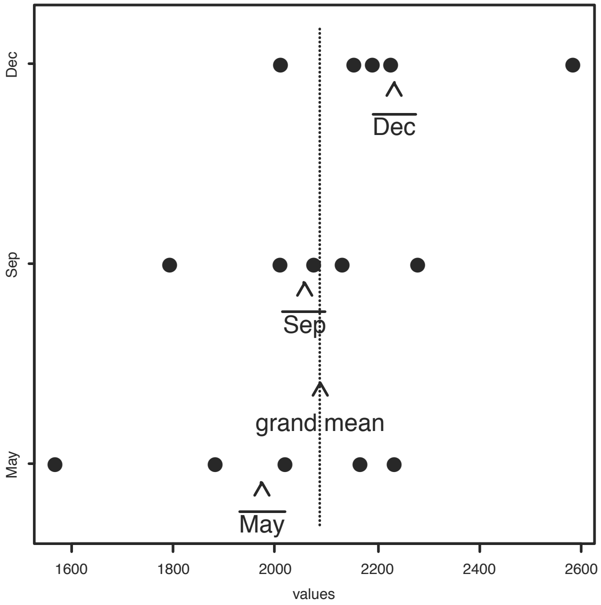


Figure 12.1: Amount of calories consumed by subjects for different months. Sample means are marked, as is the grand mean. Are the differences in the monthly means due to sampling variation or seasonal differences?

This problem is approached as a significance test. Let the hypotheses be

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, \quad H_A : \mu_i \neq \mu_j \text{ for at least one pair } i \text{ and } j.$$

A test statistic is formulated that compares the variations within a single group to those among the groups.

Let $\bar{x}$ be the grand mean, or mean of all the data, and $\bar{x}_i$ the mean for the ith sample. Then the total sum of squares is given by

$$\mathsf{SST} = \sum_i \sum_j (x_{ij} - \bar{x})^2.$$

This measures the amount of variation from the center of all the data.

An analysis of variance breaks SST up into two sums:

$$\mathsf{SST} = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + \sum_i n_i (\bar{x}_i - \bar{x})^2. \tag{12.1}$$

The first sum is called the *error sum of squares*, or SSE. The interior sum, $\sum_j (x_{ij} - \bar{x}_i)^2$, measures the variation within the ith group. The SSE is then a measure of the within-group variability. The second term in (12.1) is called the *treatment sum of squares* (SSTr). The word treatment comes from medical experiments where the population mean models the effect of some treatment. The SSTr compares the means for each group, $\bar{x}_i$, with the grand mean, $\bar{x}$. It measures the variability among the means of the samples. We can re-express Equation 12.1 as

$$\mathsf{SST} = \mathsf{SSE} + \mathsf{SSTr}.$$

From looking at the data in Figure 12.1 we expect that the SSE is smaller than the SST, as there appears to be more variation among groups than within groups. If the data came from a common mean, then we would expect SSE and SST to be roughly the same. If SSE and SST are much different, it would be evidence against the null hypothesis. How can we tell whether the differences are due to the null hypothesis being false or merely to sampling variation? As usual, we tell by finding a test statistic that can discriminate.

Based on the above observation, a natural test statistic to test whether $\mu_1 = \mu_2 = \cdots = \mu_k$ would be to consider the value SST − SSE = SSTr. "Large" values would be in the direction of the alternative. The F-statistic this comparison, but divides by the error sum of squares.

$$F = \frac{\mathsf{SSTr}/(k-1)}{\mathsf{SSE}/(n-k)}, \tag{12.2}$$

Large values are still consistent with a difference in the means. To get the proper scale, each term is divided by its respective degrees of freedom, yielding the mean sum of squares. The degrees of freedom for the total sum of squares are $n-1$. For the SSE the degrees of freedom are $n-k$, so the degrees of freedom for SSTr are $k-1$.

Under the assumption that the data is normally distributed with common mean and variance, this statistic will have a known distribution: the F-distribution with $k-1$ and $n-k$ degrees of freedom. This is a consequence of the partial F-test discussed in Chapter 11.[1]

The one-way analysis-of-variance significance test

Suppose we have k independent, *i.i.d.* samples from populations with $\mathsf{Normal}(\mu_i, \sigma)$ distributions, $i = 1, 2, \ldots, k$. A significance test of

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, $\quad H_A : \mu_i \neq \mu_j$ for at least one pair, i and j,

[1] This can be shown by identifying RSS(k) with the total sum of squares and RSS(p) with SSE in (11.11) and simplifying.

can be performed with test statistic

$$F = \frac{\overset{\mathsf{SSTr}}{\cancel{\mathsf{SST}}}/(k-1)}{\mathsf{SSE}/(n-k)}.$$

Under H_0, F has the F-distribution with $k-1$ and $n-k$ degrees of freedom. The p-value is calculated from $\mathsf{P}(F \geq \text{observed value} \mid H_0)$.

The R function `oneway.test` will perform this significance test.

• Example 12.2: Number of calories consumed by month, continued
The one-way test can be applied to the example on caloric intake. The two sums can be calculated directly as follows:

```
may <- c(2166, 1568, 2233, 1882, 2019)
sep <- c(2279, 2075, 2131, 2009, 1793)
dec <- c(2226, 2154, 2583, 2010, 2190)
#
xbar <- mean(c(may, sep, dec))
SST <- (15-1) * var(c(may, sep, dec))   # (n-1) * var(.) is SST
SSE <- (5-1) * var(may) + (5-1) * var(sep) + (5-1) * var(dec)
SSTr <- 5 * ((mean(may) - xbar)^2 + (mean(sep) - xbar)^2 +
            (mean(dec) - xbar)^2)
#
c(SST=SST, SSTr=SSTr, SSE=SSE)

##    SST   SSTr    SSE
## 761384 174664 586720

#
n <- 15; k <- 3
F.obs = (SSTr/(k-1)) / (SSE/(n-k))
F.obs

## [1] 1.786

pf(F.obs, df1=k-1, df2=n-k, lower.tail=FALSE)

## [1] 0.2094
```

We get a p-value that is not significant. Despite the graphical evidence, the differences can reasonably be explained by sampling variation. ••

Using R's model formulas to specify ANOVA models

The calculations to perform an analysis of variance need not be so complicated, as R has functions to compute the values desired. These functions use model formulas. If `x` stores all the data and `f` is a *factor* indicating which group the data value belongs to, then

```
x ~ f
```

represents the statistical model

$$x_{ij} = \mu_i + \epsilon_{ij}.$$

The default behavior for `plot` of the model formula `x ~ f` was to make a boxplot. This is because this graphic easily allows for comparison of centers for multiple samples. The dot plot in Figure 12.1 is good for a small data set, but the boxplot is preferred for larger data sets.

Using `oneway.test` to perform ANOVA

The function `oneway.test` is used as

```
oneway.test(x ~ f, data=..., var.equal=FALSE)
```

As with the `t.test` function, the argument `var.equal` is set to `TRUE` if appropriate. By default it is `FALSE`.

Before using `oneway.test` with our example of caloric intake, we put the data into the appropriate form: a data vector containing the values and a factor indicating the sample the corresponding value is from. This can be achieved with `stack`.

```
d <- stack(list(may=may, sep=sep, dec=dec)) # need names for list
names(d)                                    # two variables

## [1] "values" "ind"

oneway.test(values ~ ind, data=d, var.equal=TRUE)

##
##  One-way analysis of means
##
## data:  values and ind
## F = 1.786, num df = 2, denom df = 12, p-value = 0.2094
```

We get the same *p*-value as in our previous calculation, but with much less effort.

Using `aov` for ANOVA

The alternative `aov` function will also perform an analysis of variance. It returns a model object similar to `lm` but has different-looking outputs for the `print` and `summary` extractor functions. These are analysis-of-variance tables that are typical of other computer software and statistics books.

Again, it is called with a model formula.

```
res <- aov(values ~ ind, data = d)
res

## Call:
##    aov(formula = values ~ ind, data = d)
##
## Terms:
##                    ind Residuals
## Sum of Squares  174664    586720
## Deg. of Freedom      2        12
##
## Residual standard error: 221.1
## Estimated effects may be unbalanced
```

The function returns the two sums of squares calculated in Example 12.1 with their degrees of freedom. The `Residual standard error`, $\widehat{\sigma}$, is found by the square root of $\mathsf{RSS}/(n-k)$, which in this example is

```
sqrt(586720/12)

## [1] 221.1
```

The result of `aov` has more information than shown, just as the result of `lm` does. For example, the `summary` function returns

```
summary(res)

##             Df Sum Sq Mean Sq F value Pr(>F)
## ind          2 174664   87332    1.79   0.21
## Residuals   12 586720   48893
```

These are the values needed to perform the one-way test. This tabular layout is typical of an analysis of variance.

• Example 12.3: Effect of grip on cross-country skiing

Researchers at Montana State University performed a study on how various ski-pole grips affect cross-country skiing performance. There are three basic grip types: classic, modern, and integrated. Suppose 9 skiers are assigned at random to the three grip-types and for each the skier has upper-body power

Grip type	classic	integrated	modern
	168.2	166.7	160.1
	161.4	173.0	161.2
	163.2	173.3	166.8

Table 12.1: Upper-body power output (watts) by ski-pole grip type.

output measured. The data is summarized in Table 12.1. Does there appear to be a difference in power output due to grip type?

We can investigate the null hypothesis that the three grips will produce equal means with an analysis of variance. We assume that the errors are all independent and that the data is sampled from normally distributed populations with common variance but perhaps different means.

First we enter in the data. Instead of using `stack`, we enter in all the data at once and create a factor using `gl` to indicate grip type.[2]

```
UBP <- c(168.2, 161.4, 163.2, 166.7, 173.0, 173.3,
         160.1, 161.2, 166.8)
grip.type <- gl(3, 3, 9, labels=c(”classic”, ”integrated”, ”modern”))
boxplot(UBP ~ grip.type, ylab=”Power (watts)”,
        main=”Effect of cross country grip”)
```

The boxplot in Figure 12.2 indicates that the integrated grip has a significant advantage. But is this due to sampling error? We use `aov` to carry out the analysis of variance.

```
res <- aov(UBP ~ grip.type)
summary(res)

##             Df Sum Sq Mean Sq F value Pr(>F)
## grip.type    2  116.7    58.3    4.46  0.065 .
## Residuals    6   78.4    13.1
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

We see that there is a small p-value that is significant at the 10% level. ●●

[2] We could also use `rep` with vectorized arguments to create the factor, but `gl` is designed for just this task.

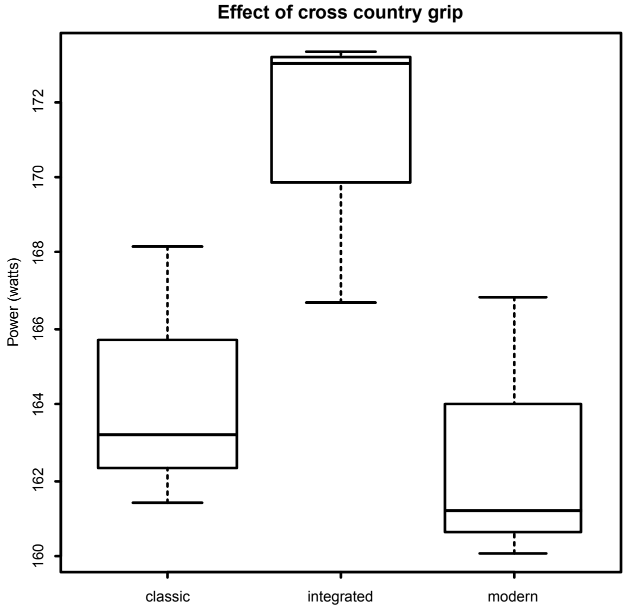


Figure 12.2: Effects of cross-country ski-pole grip on measured power output.

The nonparametric Kruskal–Wallis test

The Wilcoxon rank-sum test was discussed as a nonparametric alternative to the two-sample t-test for independent samples. Although the populations had no parametric assumption, they were assumed to have densities with a common shape but perhaps different centers.

The Kruskal–Wallis test, a nonparametric test, is analogous to the rank-sum test for comparing the population means of k independent samples.

In particular, if $f(x)$ is a density of a continuous random variable with mean 0, the assumption on the data is that x_{ij} is drawn independently of the others from a population with density $f(x - \mu_i)$. The hypotheses tested are

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, \qquad H_A : \mu_i \neq \mu_j \text{ for at least one pair } i \text{ and } j.$$

The test statistic involves the ranks of all the data. Let r_{ij} be the respective rank of a data point when all the data is ranked from smallest to largest, $\bar{r}_i$ be the mean of the ranks for each group, and $\bar{r}$ the grand mean. The test statistic is:

$$T = \frac{12}{n(n+1)} \sum_i n_i (\bar{r}_i - \bar{r})^2. \tag{12.3}$$

Statistical inference is based on the fact that T has an asymptotic χ^2-distribution with $k - 1$ degrees of freedom.