9

Analysis of Variance

Analysis of variance is the technique we use when all the explanatory variables are categorical. The explanatory variables are called **factors**, and each factor has two or more **levels**. When there is a single factor with three or more levels we use one-way Anova. If we had a single factor with just two levels, we would use Student's test (see p. 76), and this would give us exactly the same answer that we would have obtained by Anova (remember the rule that $F = t^2$). Where there are two or more factors, then we use two-way or three-way Anova, depending on the number of explanatory variables. When there is replication at each level in a multi-way Anova, the experiment is called a **factorial** design, and this allows us to study **interactions** between variables, in which we test whether the response to one factor depends on the level of another factor.

One-way Anova

There is a real paradox about analysis of variance, which often stands in the way of a clear understanding of exactly what is going on. The idea of analysis of variance is to compare two or more means, but it does this by comparing variances. How can that work?

The best way to see what is happening is to work through a graphical example. To keep things as simple as possible, we shall use a factor with just two levels at this stage, but the argument extends to any number of levels. Suppose that we have atmospheric ozone concentrations measured in parts per hundred million (pphm) in two commercial lettuce-growing gardens (we shall call the gardens A and B for simplicity).

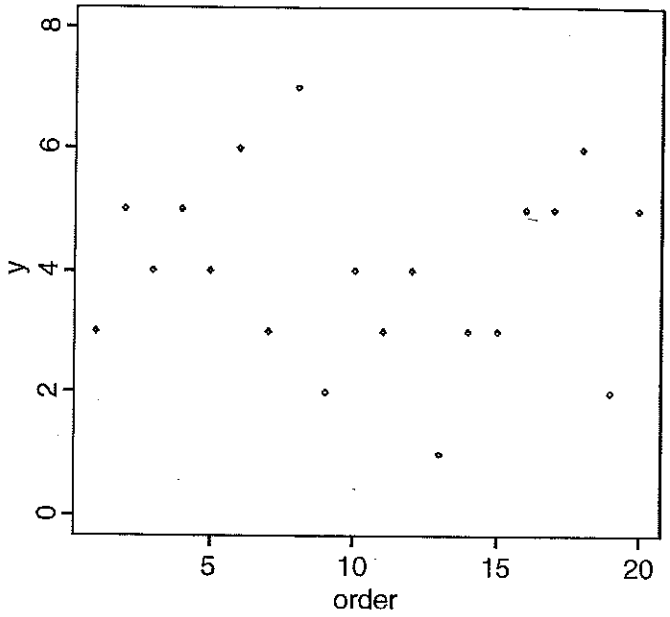
```
oneway<-read.table("c:\\temp\\oneway.txt",header=T)
attach(oneway)
names(oneway)
[ 1] "ozone"  "garden"
```

As usual, we begin by plotting the data, but here we plot the y values (ozone concentrations) against the order in which they were measured:

```
plot(1:20,ozone,ylim=c(0,8),ylab="y",xlab="order")
```

Statistics: An Introduction using R M. J. Crawley
 ISBNs: 0-470-02298-1 (PBK); 0-470-02297-3 (PPC)

lots of scatter = lots of variance in y



There is lots of scatter, indicating that the variance in y is large. To get a feel for the overall variance, we can plot the mean value of y and indicate each of the residuals by a line from mean(y) to the value of y:

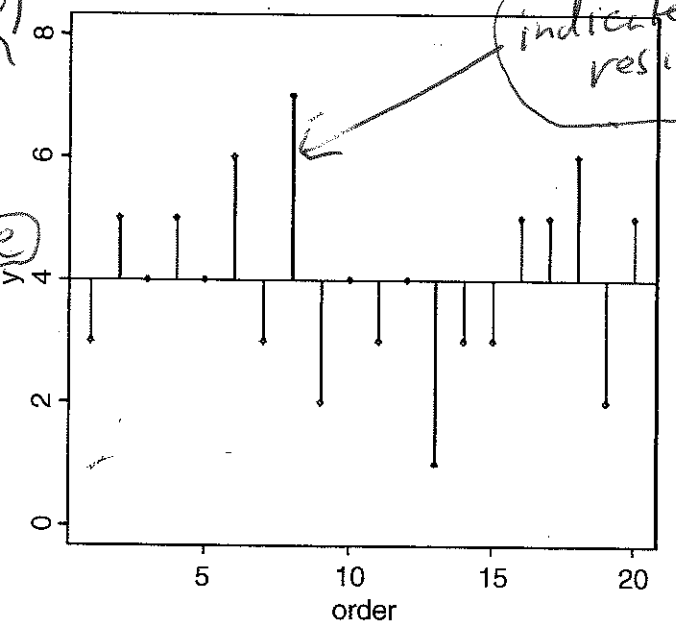
what does abline do?

```
abline(mean(ozone),0)
for(i in 1:20) lines(c(i,i),c(mean(ozone),ozone[i]))
```

abline(mean(ozone),0)

slope is 0

indicate each residual by a line



We refer to this overall variation as the **total sum of squares**, *SSY*, which more formally is given by:

SSY

$$SSY = \sum (y - \bar{y})^2$$

which should look familiar, because it is the formula used in defining the variance of y (s^2 = sum of squares/degrees of freedom; see p. 37).

This next step is the key to understanding how analysis of variance works. Instead of fitting the overall mean value of y through the data, and looking at the departures of all the data points from the overall mean, let's fit the individual treatment means (the mean for garden A and the mean for garden B in this case), and look at the departures of data points from the appropriate treatment mean. It will be useful if we have different plotting symbols for the different gardens; say open circles (pch = 1) for garden A and solid circles (pch = 16) for garden B. Note the type of type = "n" to suppress plotting when we first draw the axes:

```
plot(1:20,ozone,ylim=c(0,8),type="n",ylab="y",xlab="order")
```

Now add the points for garden A:

```
points(seq(1,19,2),ozone[garden==="A"],pch=1)
```

To space out the points, we put data from the two gardens in alternating positions on the graph, using seq(1,19,2) for garden A and seq(2,20,2) for garden B:

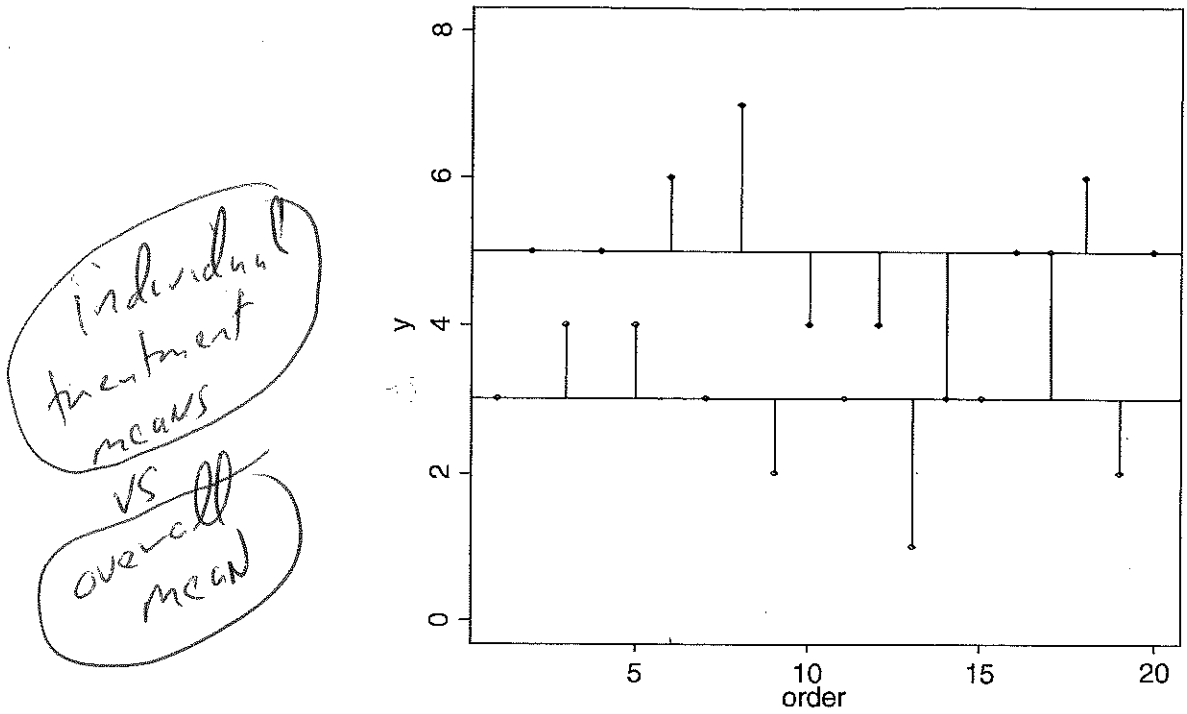
```
points(seq(2,20,2),ozone[garden==="B"],pch=16)
```

Now it is clear that the mean ozone concentration in garden B is substantially higher. The aim of analysis of variance is to determine whether it is significantly higher, or whether this kind of difference could come about by chance alone, when the mean ozone concentrations in the two gardens was really the same.

Now we draw the residuals–the differences between the measured ozone concentrations, and the means for the gardens involved:

```
abline(mean(ozone[garden==="A"]),0)
abline(mean(ozone[garden==="B"]),0)
k<- -1
for (i in 1:10){
k<-k+2
lines(c(k,k),c(mean(ozone[garden==="A"]),ozone[garden==="A"] [i]))
lines(c(k+1,k+1),c(mean(ozone[garden=="B"]),ozone[garden=="B"] [i]))
}
```

This raises some questions. If the means in the two gardens are not significantly different, what should be the difference in the lengths of the residual lines in this figure and the figure before? After a bit of thought, you should see that if the means were the same, then the two horizontal lines in this figure would be in the same place, and hence the lengths of the residual lines would be the same as in the previous figure. We're half way there. Now, suppose that mean ozone concentration is different in the two gardens. Would the residual lines be bigger or smaller when we compute them from the individual treatment means (as above), or from the overall mean (as in the previous figure)? They would be **smaller** when computed from the individual treatment means **if the individual treatment means were different.**



So there it is. That is how analysis of variance works. **When the means are significantly different, then the sum of squares computed from the individual treatment means will be smaller than the sum of squares computed from the overall mean.** We judge the significance of the difference between the two sums of squares using analysis of variance.

The analysis is formalized by defining this new sum of squares: it is the sum of the squares of the differences between the individual y values and the relevant treatment mean. We shall call this *SSE*, the **error sum of squares** (there has been no error in the sense of a mistake; 'error' is used here as a synonym of 'residual'):

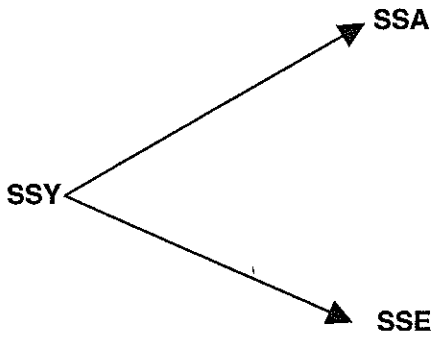
$$SSE = \sum_{j=1}^{k} \sum (y - \bar{y}_j)^2.$$

We compute the mean for the jth level of the factor in advance, and then add up the squares of the differences. Given that we worked it out this way, can you see how many degrees of freedom should be associated with *SSE*? Suppose that there were n replicates in each treatment ($n = 10$ in our example). And suppose that there are k levels of the factor ($k = 2$ in our example). If you estimate k parameters from the data before you can work out *SSE*, then you must have lost k degrees of freedom in the process. Since each of the k levels of the factor has n replicates, there must be $k \times n$ numbers in the whole experiment ($2 \times 10 = 20$ in our example). So the degrees of freedom associated with *SSE* is $k.n - k = k(n - 1)$. Another way of seeing this is to say that there are n replicates in each treatment, and hence $n - 1$ degrees of freedom for error in each treatment (because 1 d.f. is lost in estimating each treatment mean). There are k treatments (i.e. k levels of the factor) and hence there are $k \times (n - 1)$ d.f. for error in the experiment as a whole.

Now we come to the 'analysis' part of the analysis of variance. The total sum of squares in y, *SSY*, is broken up (analysed) into components. The unexplained part of the variation

is called the error sum of squares, *SSE*. The component of the variation that is explained by differences between the treatment means is called the treatment sum of squares, and is traditionally denoted by *SSA*. This is because in two-way analysis of variance, with two different categorical explanatory variables, we shall use *SSB* to denote the sum of squares attributable to differences between the means of the second factor, *SSC* to denote the sum of squares attributable to differences between the means of the third factor, and so on.

Analysis of variance, therefore, is based on the notion that we break down the total sum of squares, *SSY*, into useful and informative components:



Typically, we compute all but one of the components, then find the value of the last component by subtraction of the others from *SSY*. We already have a formula for *SSE*, so we could obtain *SSA* by difference: $SSA = SSY - SSE$. Starting with *SSY* we calculate the sum of the squares of the differences between the y values and the overall mean:

```
SSY <-sum((ozone-mean(ozone))^2)
SSY
```

```
[1] 44
```

The question now is 'how much of this 44 is attributable to differences between the means of gardens A and B (*SSA* = explained variation) and how much is sampling error (*SSE* = unexplained variation)?'. We have a formula defining *SSE*; it is the sum of the squares of the residuals calculated separately for each garden, using the appropriate mean value. For garden A we get

```
sum((ozone[garden=="A"]-mean(ozone[garden=="A"]))^2)
```

```
[1] 12
```

and for garden B

```
sum((ozone[garden=="B"]-mean(ozone[garden=="B"]))^2)
```

```
[1] 12
```

so the error sum of squares is the total of these components $SSE = 12 + 12 = 24$. Finally, we can obtain the treatment sum of squares, *SSA*, by difference: $SSA = 44 - 24 = 20$.

At this point, we can fill in the Anova table (see p. 136):

Source	Sum of squares	Degrees of freedom	Mean square	F-ratio
Garden	20.0	1	20.0	15.0
Error	24.0	18	$s^2 = 1.3333$	
Total	44.0	19		

We need to test whether an F-ratio of 15.0 is large or small. To do this we compare it with the critical value of F from quantiles of the F-distribution, qf. We have one degree of freedom in the numerator, and 18 degrees of freedom in the denominator, and we want to work at 95% certainty ($\alpha = 0.05$):

```
qf(0.95,1,18)
```

```
[1] 4.413873
```

The calculated value of 15.0 is much greater than the critical value of $F = 4.41$, so we can reject the null hypothesis (equality of the means) and accept the alternative hypothesis (the two means are significantly different). We used a one-tailed F-test (0.95 rather than 0.975 in the qf function) because we are only interested in the case where the treatment variance is large relative to the error variance. This approach is rather old-fashioned; the modern view is to calculate the **effect size** (the difference between the means is 2.0 pphm ozone) and to state the probability that such a difference would arise by chance alone when the difference between the means was actually 0. For this we use cumulative probabilities of the F distribution, rather than quantiles, like this:

```
1-pf(15.0,1,18)
```

```
[1] 0.001114539
```

So the probability of obtaining data as extreme as ours (or more extreme) if the two means really were the same is roughly one tenth of 1%.

That was quite a lot of work. Here is the whole analysis in R in a single line:

```
summary(aov(ozone~garden))
```

```
            Df  Sum Sq Mean Sq F value   Pr(>F)
garden       1 20.0000 20.0000      15 0.001115 **
Residuals   18 24.0000  1.3333
```

The first column shows the sources of variation (*SSA* and *SSE* respectively); note that R leaves off the row that we had included for total variation, *SSY*. The next column shows degrees of freedom: there are two levels of garden (A and B) so there is $2 - 1 = 1$ d.f. for garden, and there are 10 replicates per garden, so $10 - 1 = 9$ d.f. per garden and two