Analysis of Variance

Gaëlle Cordier

24/11/2014

ANOVA compares (two or more) means by comparing variances; one-Way ANOVA compares (two or more) means based on one factor by comparing variances.

Data example: crop yields per unit area measured from 10 randomly selected fields on each of 3 soil types (sand, clay, and loam)

```
yields<-data.frame(sand=c(6,10,8,6,14,17,9,11,7,11),
                   clay=c(17,15,3,11,14,12,12,8,10,13),
                   loam=c(13,16,9,12,15,16,17,13,18,14))
yields
```

```
##    sand clay loam
## 1     6   17   13
## 2    10   15   16
## 3     8    3    9
## 4     6   11   12
## 5    14   14   15
## 6    17   12   16
## 7     9   12   17
## 8    11    8   13
## 9     7   10   18
## 10   11   13   14
```

We may want to know whether soil types (categorical explanatory variable or factor) significantly affects crop yield (numerical respone variable):

$H_0 : \mu_s = \mu_c = \mu_l$

H_1 : at least one mean is different

If we take a look at the sample means:

```
yields_l<-melt(data = yields,
               value.name = "yield",
               measure.vars = c("sand","clay","loam"),
               variable.name = c("soil"))

head(yields_l); tail(yields_l)
```

```
##   soil yield
## 1 sand     6
## 2 sand    10
## 3 sand     8
## 4 sand     6
## 5 sand    14
## 6 sand    17
```

```
##    soil yield
## 25 loam    15
## 26 loam    16
## 27 loam    17
## 28 loam    13
## 29 loam    18
## 30 loam    14
```

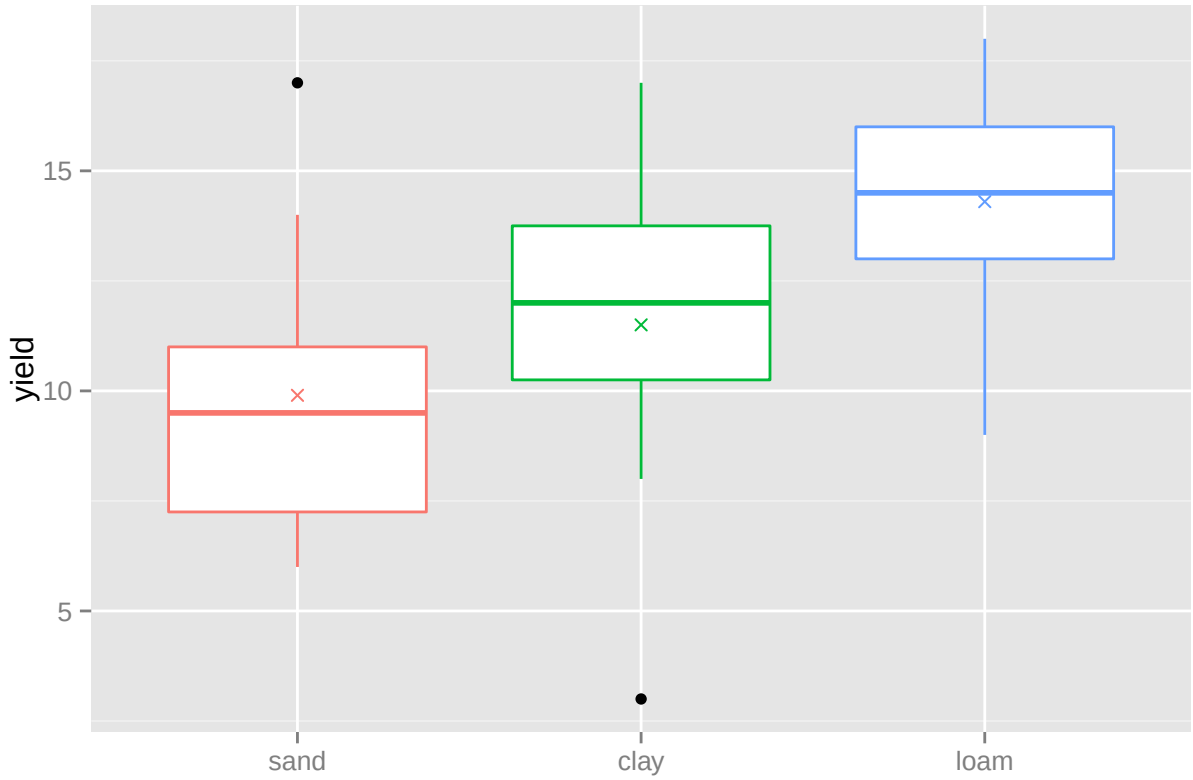
```r
means<-summaryBy(formula = yield~soil,data = yields_l,FUN = mean)
means
```

```
##   soil yield.mean
## 1 sand        9.9
## 2 clay       11.5
## 3 loam       14.3
```

```r
# or:
# aggregate(yield~soil, yields_l, mean)
```

and at the distribution of the yield values within the 3 soils:

```r
ggplot(data = yields_l,aes(x = soil,y = yield, color = soil))+
    geom_boxplot()+
    stat_summary(fun.y=mean, geom="point", shape=4)+
    labs(x="")+
    scale_color_discrete(guide=F)
```



we can see that yield may turn out to be significantly different between sand and loam soils (their boxes don't overlap), but it is not as clear whether clay yield will be significantly greater/lower than sand/loam yield.

How does ANOVA allows us to make inferences about differences between means by looking at differences between variances?

Deviation (variability) and variance:

The distance from any point in a collection of data, to the mean of the data (sample mean), is the deviation. This can be written as:

$y_i - \overline{y}$

where y_i is the *ith* data point, and $\overline{y}$ is the estimate of the mean. If all such deviations are squared (so all differences are positives), and then summed, as in:

$\sum_{i=1}^{n} (y_i - \overline{y})^2$

this gives the "sum of squares" for these data.

The deviation is unscaled (the sum of squares will grow with the size of the data collection), so to compare samples of different sizes we need to scale it by dividing it by the degrees of freedom (the number of parameters of the system that may vary independently, or to simplify, the sample size minus 1). In fact, the sample variance of a discrete random variable is defined as:

$S^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}$

So, the deviation is an unscaled measure of dispersion (or variability), that when scaled for the number of degrees of freedom estimates the variance.

Partition of Sum of Squares:

The analysis of variance involves calculating the total variation or variability in the response variable (yield in this case) and partitioning it into two components: the intra-groupal or unexplained variability (variability in each group due to unknown factors) and the inter-groupal or explained variability (variability due to the explanatory variable, soil in this case). This is called the partition of sum of squares, and it allows us to quantify the relative importance of each one of said sources of variability: in our case, if the factor soil has an effect over crop yield, we would expect the total variability to be explained in greater measure by the explained variability than by the unexplained variability:

TOTAL VARIABILITY = EXPLAINED VARIABILITY + UNEXPLAINED VARIABILITY

Calculating the Sums of Squares

SSE

We can define the variability in the response variable within each group as:

$\text{SSE} = \sum_{i=1}^{k} \sum_{i=j}^{n} (y_{ij} - \overline{y})^2$

that is, the sum of squares of the differences between the observation j (n=*10* replicates) within each group and the mean of said group i (k=*3* factor levels). This would be the unexplained variation or residual variability (error sum of squares) since it is not explained by the differences between groups.

```
yields_l2<-data.frame(yields_l,mean=rep(means$yield.mean,each = 10))

yields_l2
```

```
##    soil yield mean
## 1  sand     6  9.9
## 2  sand    10  9.9
## 3  sand     8  9.9
## 4  sand     6  9.9
## 5  sand    14  9.9
## 6  sand    17  9.9
## 7  sand     9  9.9
## 8  sand    11  9.9
## 9  sand     7  9.9
## 10 sand    11  9.9
## 11 clay    17 11.5
## 12 clay    15 11.5
## 13 clay     3 11.5
## 14 clay    11 11.5
## 15 clay    14 11.5
## 16 clay    12 11.5
## 17 clay    12 11.5
## 18 clay     8 11.5
## 19 clay    10 11.5
## 20 clay    13 11.5
## 21 loam    13 14.3
## 22 loam    16 14.3
## 23 loam     9 14.3
## 24 loam    12 14.3
## 25 loam    15 14.3
## 26 loam    16 14.3
## 27 loam    17 14.3
## 28 loam    13 14.3
## 29 loam    18 14.3
## 30 loam    14 14.3
```

```
SSE<-with(data = yields_l2,expr = sum((yield-mean)^2))

SSE
```

```
## [1] 315.5
```

SSA

In a similar way, we can define the variability in the response variable between groups as:

$\text{SSA} = \sum_{i=1}^{k} \sum_{i=j}^{n} (\overline{y_i} - \overline{\overline{y}})^2 = n * \sum_{i=1}^{k} (\overline{y_i} - \overline{\overline{y}})^2$

that is, the sum of squares of the differences between the individual treatment means $n\ x\ i$ and the overall mean (the mean of all observations, or the mean of the group means). This would be the explained variability (the treatment sum of squares).

```
means
```

```
##   soil yield.mean
## 1 sand        9.9
## 2 clay       11.5
## 3 loam       14.3
```

```r
SSA<-10*sum((means$yield.mean-mean(means$yield.mean))^2)

SSA
```

```
## [1] 99.2
```

```r
# or
with(data = yields_l2,expr = sum((mean-mean(yields_l2$yield))^2))
```

```
## [1] 99.2
```

SST

The total variability (the total sum of squares) would be then:

$\text{SST} = \sum_{i=1}^{k} \sum_{i=j}^{n} (y_{ij} - \overline{\overline{y}})^2$

that is, the sum of squares of the differences between the observation ij and the overall mean.

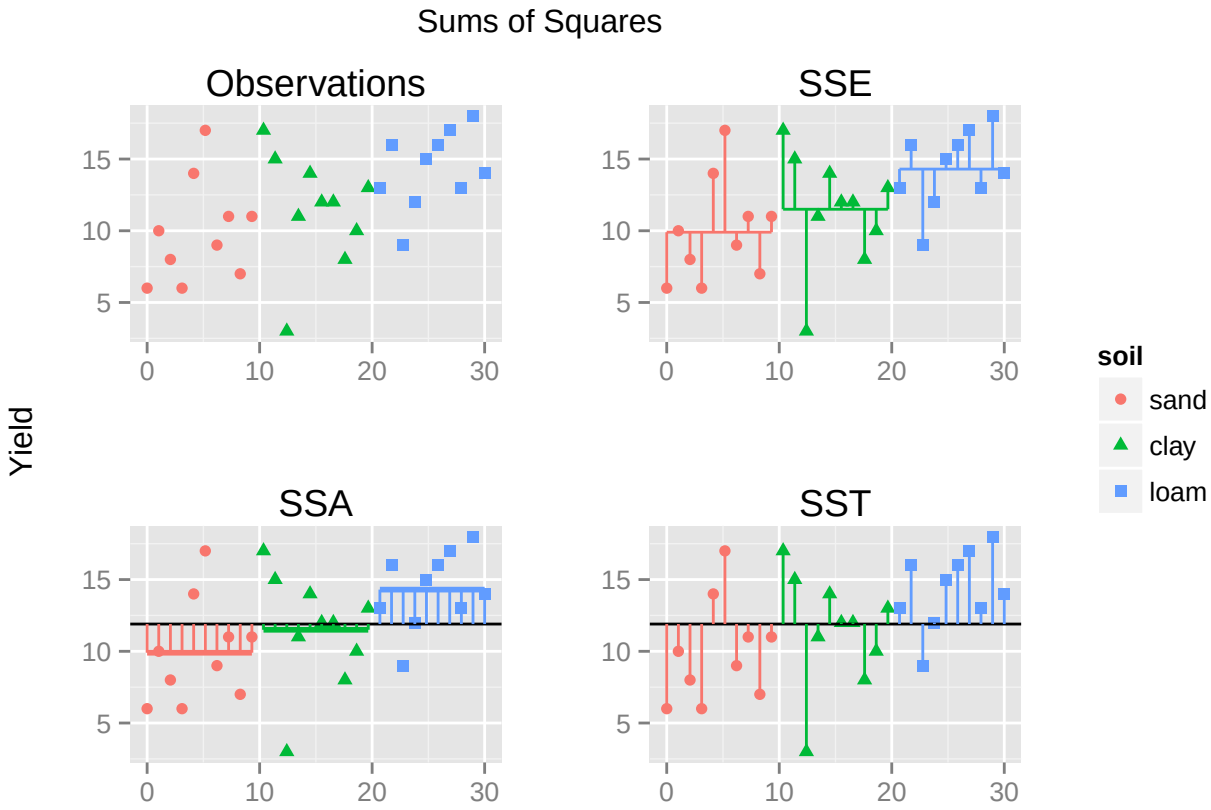
```r
SST<-with(data = yields_l2,expr = sum((yield-mean(yield))^2))

SST
```

```
## [1] 414.7
```

Plotting the variability

```r
plotdata<-data.frame(yields_l2,ov.mean=rep(mean(means$yield.mean)),
                     x=seq(from = 0,to = 30,length.out = nrow(yields_l2)))

p1<-ggplot(data = plotdata,aes(x=x,y=yield,shape=soil,color=soil))+
    geom_point(size=2)+
    labs(x="",y="",title="Observations")

p2<-ggplot(data = plotdata,aes(x=x,y=yield,shape=soil,color=soil))+
    geom_point(size=2)+
    geom_line(aes(y=mean))+
    geom_linerange(aes(ymin=yield,ymax=mean))+
    labs(x="",y="",title="SSE")

p3<-ggplot(data = plotdata,aes(x=x,y=yield,shape=soil,color=soil))+
    geom_point(size=2)+
    geom_line(aes(y=mean),size=1)+
    geom_abline(intercept=11.9,slope=0)+
    geom_linerange(aes(ymin=mean,ymax=ov.mean))+
    labs(x="",y="",title="SSA")

p4<-ggplot(data = plotdata,aes(x=x,y=yield,shape=soil,color=soil))+
    geom_point(size=2)+
    geom_abline(intercept=11.9,slope=0)+
    geom_linerange(aes(ymin=yield,ymax=ov.mean))+
    labs(x="",y="",title="SST")
```

```
legend<-gtable_filter(ggplot_gtable(ggplot_build(p1)), "guide-box")

grid.arrange(arrangeGrob(p1 + theme(legend.position="none"),
                         p2 + theme(legend.position="none"),
                         p3 + theme(legend.position="none"),
                         p4 + theme(legend.position="none"),
                         nrow = 2,
                         main = textGrob("Sums of Squares", vjust = 1),
                         left = textGrob("Yield", rot = 90, vjust = 1)),
             legend,
             widths=unit.c(unit(1, "npc") - legend$width, legend$width),
             nrow=1)
```



(extraction of legend and grid.arrange with global Y-axis and common legend as seen here)

In conclusion

The total sum of squares is the sum of the treatment sum of squares and the error sum of squares:

$SST = SSA + SSE$

```
all.equal(SST,SSA+SSE)
```

```
## [1] TRUE
```

So the difference between SST and SSE is the treatment sum of squares, SSA:

$SSA = SST - SSE$

This is the amount of the variation in yield that is explained by the differences between the treatment means.

Drawing the ANOVA table

Source of variation

- Explained variability: type of soil
- Unexplained variability: error
- Total variability

```
Source<-c("Soil type","Error","Total")
```

Sum of squares

- SSA
- SSE
- SST

```
SS<-c(SSA,SSE,SST)
```

Degrees of freedom

The degrees of freedom are the number of parameters that may vary independently (the number of observations minus 1).

- Treatment degrees of freedom: we are comparing 1 mean per soil type with the overall mean (3 parameters), so we have: $3 - 1 = 2$ degrees of freedom.
- Error degrees of freedom: we are comparing 10 observations per soil type with the soil type mean (10 parameters by soil type), so we have: $(10 - 1) * 3 = 9 * 3 = 27$ degrees of freedom.
- Total degrees of freedom: we are comparing 30 observations with the overall mean (30 parameters), so we have: $30 - 1 = 29$ degrees of freedom.

```
df<-c(2,27,29)
```

Mean square

The mean square is obtained by divinding the sum of squares by the degrees of freedom, and is a measure of the treatment variance and the error variance:

```
MS<-SS/df
```

The treatment variance is the mean square between groups (MS_B):

```
MS_B<-MS[1]
MS_B
```

```
## [1] 49.6
```

The error variance is the mean square within groups (MS_W), and since there is equal replication in each soil type, it is equal to the mean of the variances of the soil types:

```
MS_W<-MS[2]
MS_W
```

```
## [1] 11.68519
```

```
vars<-aggregate(yield~soil, yields_l, var)
vars
```

```
##   soil     yield
## 1 sand 12.544444
## 2 clay 15.388889
## 3 loam  7.122222
```

```
mean(vars$yield)
```

```
## [1] 11.68519
```

The total variance is the total mean square, and it is equal to the variance of all the observations:

```
MS_T<-MS[3]
MS_T
```

```
## [1] 14.3
```

```
sum((yields_l2$yield-11.9)^2)/29
```

```
## [1] 14.3
```

```
# or
var(yields_l2$yield)
```

```
## [1] 14.3
```

F ratio and p-value

$F = \frac{MS_B}{MS_W}$

The F ratio tests the null hypothesis that the treatment means are all the same:

$H_0 : \mu_s = \mu_c = \mu_l$

H_1 : at least one mean is significantly different from the others

If the null hypothesis isn't true, we would expect the MS_B to be greater than the MS_W, so we expect the F-ratio to be > 1. On the contrary, if the null hypothesis is true, we expect the F-ratio to have a value close to 1.

```
FR<-MS_B/MS_W
FR
```

```
## [1] 4.244691
```

Is this value (> 1) significant? To answer this question we must compare the test statistic F=4.24 with the critical value of F, that is, the quantile of the probability distribution given the degrees of freedom of the numerator (df=2) and the denominator (df=27), for a given probability (0.95 if $\alpha = 0.05$):

```
qf(p = 0.95,df1 = 2,df2 = 27)
```

```
## [1] 3.354131
```

As F $>$ critical value, we would reject the null hypothesis. In order to be able to work independently of the confidence interval, we use the cumulative distribution, that is, we look for the probability of being greater than our quantile (that would be the inverse of the cumulative distribution $P(x <= X)$):

```
pf(q = FR,df1 = 2,df2 = 27)
```

```
## [1] 0.9750493
```