Handout 9.1. Analysis of variance

In this section, we consider comparisons among more than two groups parametrically, using analysis of variance.

1. One way analysis of variance

Let x_{ij} denote the *j*th observation in the *i*th group, $\bar{x}_{i.}$ is the mean of the ith group, and $\bar{x}_{..}$ is the mean of all observations. We can decompose the observations as

$$x_{ij} = \bar{x}_{..} + (\bar{x}_{i.} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{i.})$$

Informally corresponding to the model

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2)$$

in which the hypothesis is that all the group means are same. Now consider the sums of squares of the underbraced terms, known as variation within groups

$$\text{SSW} = \sum_{ij} (x_{ij} - \bar{x}_{i.})^2$$

and variation between groups

$$\text{SSB} = \sum_i n_i (x_{ij} - \bar{x}_{..})^2$$

Note that

$$\text{SSW + SSB = SST} = \sum_{ij} (x_{ij} - \bar{x}_{..})^2 \qquad .$$

Let MSW=SSW/(N-k), and MSB=SSB/(k-1). We calculate the F-statistics F=MSB/MSW.

Example 1:

```
y1 <- c(18.2, 20.1, 17.6, 16.8, 18.8, 19.3, 19.1)
y2 <- c(17.4, 18.7, 19.1, 16.4, 15.2, 18.4)
y3 <- c(15.2, 18.8, 17.7, 16.5, 15.9, 17.1, 16.3)
y<-c(y1, y2, y3)
n<-c(7, 6, 7)
group<-c(rep(1,7), rep(2,6), rep(1,7))

ydata<-data.frame(y=y, group=factor(group))
 fit<-lm(y~group, ydata)
 anova(fit)

Analysis of Variance Table

Response: y
```

```
           Df Sum Sq Mean Sq F value
 Pr(>F) group  1      0.080   0.0801
            0.0376 0.8484
 Residuals 18 38.322  2.1290
```

Example 2:

```
data(red.cell.folate)
attach(red.cell.folate)
summary(red.cell.folate)

anova(lm(folate~ventilation))
mean(red.cell.folate[1:8,1])
mean(red.cell.folate[9:17,1])
mean(red.cell.folate[18:22,1])
```

The specification of a one-way analysis of variance is analogous to a regression analysis. The only difference is that the **descriptive variable needs to be a factor** and not a numeric variable. We calculate a model object using lm and extract the analysis of variance table with anova.

Example 3:

```
data(juul)
attach(juul)
juul[1:20,]
anova(lm(igf1~tanner))      #this is wrong,
#This does not describe a grouping of data but a linear regression
on the group number. Notice the telltale 1 DF for the effect of
tanner.

juul$tanner<-
factor(juul$tanner,labels=c("one","two","three","four","five"))
detach(juul)
attach(juul)
summary(tanner)
anova(lm(igf1~tanner))
```

1.1 pairwise comparisons and multiple testing

```
summary(lm(folate~ventilation))

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)            316.62      16.16  19.588 4.65e-14 ***
ventilationN2O+O2,op   -60.18      22.22  -2.709   0.0139 *
ventilationO2,24h      -38.62      26.06  -1.482   0.1548
```

These coefficients do not have their usual meaning as the slope of a regression line but have a special interpretation:
The interpretation of the estimates is that the intercept is the mean in the first group(N20+02,24h), whereas the other two describe the difference between the relevant group and the first group.

```
mean(red.cell.folate[1:8,1])
[1] 316.625
 mean(red.cell.folate[9:17,1])
[1] 256.4444
 mean(red.cell.folate[18:22,1])
[1] 278
```

Among the t tests in the table, you can immediately find a test for the hypothesis that the first two groups have the same true mean(p=0.0139) and also whether the first and the third might be identical (p=0.1548). However, a comparison of the last two groups cannot be found.

A function called pairwise.t.test computes all possible two-group comparisons:

```
pairwise.t.test(folate,ventilation,p.adj="bonferroni")
pairwise.t.test(folate,ventilation)
```

1.2 relaxing the variance assumption

the traditional one way ANOVA requires an assumption of equal variances for all groups

```
sd(red.cell.folate[1:8,1])
sd(red.cell.folate[9:17,1])
sd(red.cell.folate[18:22,1])
```

```
oneway.test(folate~ventilation)
```

In this case the p-value increased to a nonsignificant value 0.09277.

```
pairwise.t.test(folate,ventilation,pool.sd=F)
```

1.3 graphical presentation

```
xbar<-tapply(folate,ventilation,mean)
s<-tapply(folate,ventilation,sd)
n<-tapply(folate,ventilation,length)
sem<-s/sqrt(n)
stripchart(folate~ventilation,vert=T,pch=16 ,method="jitter")
arrows(1:3,xbar+sem,1:3,xbar-sem,angle=90,code=3,length=0.1)
lines(1:3,xbar,pch=4,type="b",cex=2)
```

1.4 bartlett's test

```
bartlett.test(folate~ventilation)
Bartlett test of homogeneity of variances

data:  folate by ventilation
Bartlett's K-squared = 2.0951, df = 2, p-value = 0.3508
```

In this case, nothing id data contradicts the assumption of equal variances in the three groups.

Barlett's test, although like the F test for comparison of two variances, it is rather nonrobust against departures from the assumption of normal distribution