Lab week 6

Testing ANOVA assumptions in R

Example. Five analysts each made 10 determinations of the paracetamol content of the same batch of tables. The results are shown below.

A	84.55	84.61	84.26	84.36	84.66	84.31	84.65	84.41	84.52	84.44
B	84.12	84.04	83.95	84.51	84.08	84.07	84.35	83.99	84.25	84.14
C	84.44	84.48	84.14	84.17	84.31	84.60	84.44	84.24	84.64	84.47
D	84.05	84.14	84.53	84.07	84.45	83.95	84.10	84.29	84.13	83.98
E	84.09	84.53	84.60	84.48	84.42	84.57	84.35	84.30	84.37	84.63

We are interested in finding out whether there is a significant difference between the paracetamol contents obtained by the five analysts. Before testing the null hypothesis that the mean paracetamol contents are the same for the five experimenters

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

H_1: not all the means are equal

we must test whether the assumptions behind ANOVA are sustained by the data.

There are five groups, all of size n_j=10 which we store in R into five vectors: a,b,c,d and e.

$> \ a \ <- \ c(84.55, 84.61, 84.26, 84.36, 84.66, 84.31, 84.65, 84.41, 84.52, 84.44)$

$> \ b \ <- \ c(84.12, 84.04, 83.95, 84.51, 84.08, 84.07, 84.35, 83.99, 84.25, 84.14)$

$> \ c \ <- \ c(84.44, 84.48, 84.14, 84.17, 84.31, 84.60, 84.44, 84.24, 84.64, 84.47)$

$> \ d \ <- \ c(84.05, 84.14, 84.53, 84.07, 84.45, 83.95, 84.10, 84.29, 84.13, 83.98)$

$> \ e \ <- \ c(84.09, 84.53, 84.60, 84.48, 84.42, 84.57, 84.35, 84.30, 84.37, 84.63)$

Next we combine them into one long vector, $\mathbf{y}$:

$> \ y \ <- \ c(a, b, c, d, e)$

We create a second long vector, called **group** to identify the group membership. One way to create the group vector is using the c command and type 10 values of 1 for group a, 10 values of 2 for group b, 10 values of 3 for group c, 10 values of 4 for group d and 10 values of 5 for group e.

A faster way to create this vector is by using the **rep** function

$> \ rep(1, 10)$

[1] 1 1 1 1 1 1 1 1 1 1

creates 10 replicates of the value 1.

So the vector **group** containing the 50 indexes can be created with the command

$> \ group <- c(rep(1, 10), \ rep(2, 10), \ rep(3, 10), \ rep(4, 10), \ rep(5, 10))$

Or even simpler with

$> \ group <- rep(1 : 5, each = 10)$

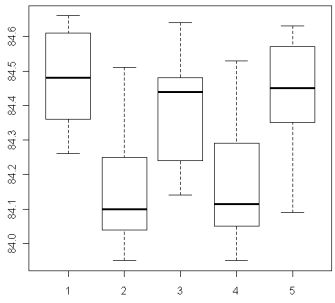
Redefine the *group* variable as a factor:

$> \ group \ <- factor(group)$

Before running the analysis of variance, you should graph the means and standard errors of each group of data. Illustrate these results with parallel boxplots (one for each treatment).

$> \ plot(group, y)$

The side-by-side boxplots indicate that the data in the five groups might have different means but



similar variances. We can highlight different degrees of dispersion when the spread of the boxplot

for one group is twice as large as the spread for another group. In our example the spread of the five boxplots is similar. Groups b and d appear to have means lower than those of groups a, c and e. The graphical analysis is subjective and further investigations is necessary.

Place the two long vectors y and $group$ together in a unifying dataframe called **paracetamol**

$> \ paracetamol = data.frame(y, group)$

and use the analysis of variance function in R to obtain the list of residuals and predicted values.

$> \ model <- aov(y \sim group, det)$

$> \ resid <- residuals(model)$

$> \ resid$

Resid contains the list of residuals (observed values - fitted values).

$> \ pred <- predict(model)$

$> \ pred$

Pred contains the list of fitted values which in ANOVA are the group means.

Assumptions

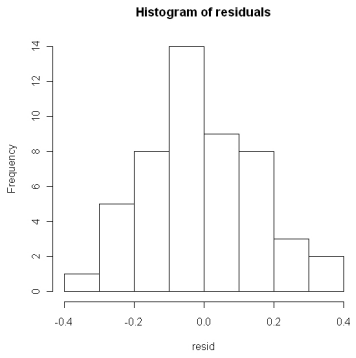
- Each population from which a sample is taken is assumed to be normal.
- Each sample is randomly selected and independent.
- The populations have approximately equal variances(standard deviations).

There are several ways to test the assumption of **normality**. One is to look at the distribution of the residuals and analyze the shape of the histogram and the boxplot

$> \ hist(resid)$

$> \ boxplot(resid)$

The histogram is symmetric and the boxplot does not indicate the presence of outliers. Both the

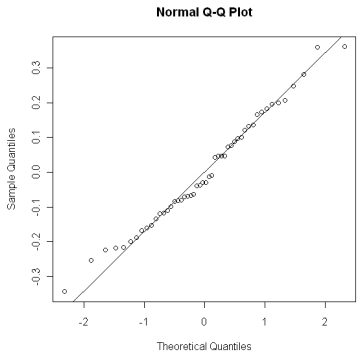


histogram and the boxplot indicate normality.

Normality can be visually assessed also from the Normal Quantile plot obtained with the **qqnorm()** function

> $qqnorm(resid)$

> $qqline(resid)$ The Normal quantile plot shows the residuals against the expected normal quan-



tiles. The expected quantile is the number of SDs from the mean where such an observation would be expected to lie in normal distribution with the sample mean and standard deviation. When the sample is normally distributed the points will form a straight-line. Deviation from the line indicates non-normality. The assumption of normality can be formally tested using the Anderson-Darling method. The goodness-of-fit test can be used to check the null hypothesis H_0 : data is normally distributed, against H_1 : data is not normally distributed.

$> \ ad.test(resid)$

The output from the Anderson-Darling test accepts the normality assumption since the p-value 0.5311 is greater than 0.05.

Anderson-Darling normality test

data: resid

A = 0.3158, p-value = 0.5311

Another test that can be used to identify significant departures from normality is the Shapiro-Wilk test. $> \ shapiro.test(resid)$

Shapiro-Wilk normality test

data: resid

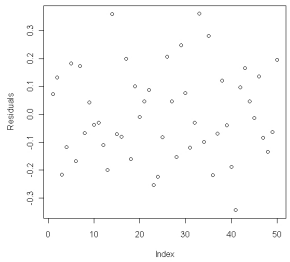
W = 0.9819, p-value = 0.6348

The output from the Shapiro-Wilk test accepts the normality assumption since the p-value 0.6348 is greater than 0.05.

Both the visual inspection and the results from the goodness-of-fit tests indicate the fact that the normality assumption holds.

The next assumption made by ANOVA is that of **independence**. This is the assumption that the errors associated with each observation are independent. We can check this by visually checking the plot of the residuals against the order of the data. $> \ plot(resid)$

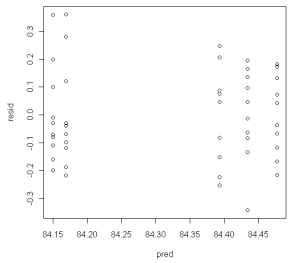
The plot does not exhibit any particular features(no pattern or trend), hence we can assume the



data are independent. This assumption depends on correct experimental design, and in particular on the correct identification of the experimental unit and appropriate randomization.

The final assumption made by ANOVA is that the variation is the same in each group (homoscedasticity). The plot of Residuals Versus the Fitted Values (the fitted values are the group means) shows whether the variation is the same in each group. In this case the residuals clearly vary similarly for all the fitted values. $> \ plot(pred, resid)$

Formal testing for the homogeneity of variances can be done with several tests. Bartlett's test



assesses equality of the variances of more the five samples assuming a normal distribution for each sample.

$> \ bartlett.test(y \sim \ group, paracetamol)$

Bartlett test of homogeneity of variances

data: y by index

Bartlett's K-squared = 0.8327, df = 4, p-value = 0.934

Bartlett's test is not reliable with departures from normality. In such cases use Levene's test as an alternative test. Levene's test is implemented in the **Rcmdr** package. Install it and then load it in *R* with $> \ library(Rcmdr)$.

$> \ leveneTest(y \sim \ index, paracetamol)$

Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	4	0.0425	0.9964

Bartlett's test (which assumes Normality within each factor level) is greater than 0.05. Levene's test does not assume Normality and also fails to reject the null hypothesis of equal variances.
So the this example we can consider ALL the ANOVA assumptions are correct.

The tests of significance should be thought of as supplemental information. Significance or non-significance of tests of assumptions should not be the primary information used to decide if the analysis of variance is appropriate.

We can continue with the analysis of variance that tests whether the mean paracetamol content is the same for the five analysts.
summary(model) or anova(model) contain the summary of the ANOVA performed by *R*.

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
group	4	0.94569	0.236423	8.1817	4.781e-05	⋆⋆⋆
Residuals	45	1.30035	0.028897			

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

The p-value of ANOVA is 0.00004781 which is very small, hence we reject the null hypothesis that the means are equal for all analysts.