# Computer practical 12

*Statistics 1, Term 2*

Aim

In this computer practical, we look at using R to carry out some common analyses. We consider three such analyses:

1. the matched pairs t test, used to compare two dependent samples
2. two-sample t tests, used to compare two independent samples
3. analysis of variance, used to compare three or more independent samples.

In the case of the latter, you may not yet have seen this in the lectures and should treat that component of the practical as a foretaste.

## Statistical techniques

Remember that you should always explore your data graphically first before calculating confidence intervals or performing hypothesis tests as it may suggest making a transformation or a change of approach.

### Two independent samples

Appropriate plots:

- Side by side box-plots of the two samples.
- Quantile plot of each sample.

The R function `t.test` can be used to compute the confidence interval for the difference between the population means and the corresponding hypothesis test. If x and y are the two samples, one types

```
> t.test(x, y)
```

The result is Welch's confidence interval and test which is like the conservative procedure described in lectures except that a more complex calculation is used for the degrees of freedom instead of basing it on the smaller of the two sample sizes.

If one is willing to assume that the population variances are equal, one can compute the pooled procedure from lectures by

```
> t.test(x, y, var.equal=TRUE)
```

### Matched pairs

Appropriate plots:

- A histogram and/or quantile plot of the differences of the pairs y-x.
- Plot y versus x and add the line y=x to the plot.

The R function `t.test` can be used to compute the confidence interval for the mean difference of pairs and the corresponding hypothesis test. If x and y are the two samples (so that first element of x is paired with the first element of y and so on), one types

```
> t.test(x, y, paired=TRUE)
```

### Multiple samples (Analysis of Variance)

Appropriate plots:

- Side by side box-plots of the samples.
- Quantile plot of each sample.
- Effects and residuals plot (first term material, see below).
- Location-scale plot (first term material, see below).
- Quantile plot of the residuals (first term material, see below).

Suppose that the data are in a data frame df with a variable y containing the values of the response and another variable g defining the groups (population from which each measurement comes). Then one can compute a one-way analysis of variance by

```
> aov(y~g, data=df)
```

However, the output will be a bit cryptic and we would be better off saving the result of aov by giving it a name, say fit, and using some other R functions to examine fit:

```
> fit = aov(y~g, data=df)
> anova(fit)
> plot(fit)
```

The anova command prints out the conventional analysis of variance table, part of which was described in the first term and part of which is taught in this term. The plot command produces the effects and residuals plot from the first term (provided you have loaded the durham package).

To obtain a location scale plot, you need to first compute the means and standard deviations:

```
> means = tapply(df$y, df$g, mean)
> sds = tapply(df$y, df$g, sd)
> plot(log(means), log(sds))
```

To obtain a quantile plot of the residuals from fit, type

```
> qqnorm(resid(fit))
```

## Applications

### Darwin's Zea Mays study

Dataset: zeamays {durham}

The data set was produced by Charles Darwin. Darwin was exploring the effects of in-breeding as opposed to cross-breeding on the strength of species. One of his experiments was performed on Zea Mays plants. Several pairs of seeds were taken from Zea Mays plants grown under controlled conditions. The two seeds in each pair were taken from the same plant; however one of the pair was taken from a self-fertilised flower (one fertilised by the same plant) and the other from a cross-fertilised flower (one fertilised by a different plant). The two seeds were subsequently grown together under identical conditions and their heights (in inches) were measured.

There are three variables in the zeamays data frame:

- `Batch` indicates which group of Darwin's plants the pair came from.
- `Cross` are the heights of the cross-fertilised plants.
- `Self` are the heights of the self-fertilised plants.

Ignoring the batch variable, analyse the data. What do you do and what do you find?

```
# INSERT YOUR CODE HERE ...
```

- COMMENTS…

### Student's sleep data

Dataset: sleep {datasets}

Student was the nom de plume of William Gosset who discovered the t-distribution and some of its early uses. The sleep data record the effect of two soporific drugs (group variable) in terms of the extra number of hours of sleep compared to normal for 10 subjects in each group.

Analyse the data. What do you do and what do you find?

```
# INSERT YOUR CODE HERE ...
```

- COMMENTS…

### Chlorpheniramine data

Dataset: chlorph {durham}

In the first term, you analysed the chlorpheniramine maleate (chlorph) data graphically. The data record measurements by seven laboratories of the amount chlorpheniramine maleate in a composite of a number of tablets. Each lab was sent a single piece of the composite and made 10 measurements. The whole experiment was repeated twice, for tablets from two manufacturers: A and B. Here you will analyse just the data for manufacturer A.

Create a data frame from chlorph which only contains the data for manufacturer A. Use the Multiple samples technique described above to analyse the data. What do you find and how?

```
# INSERT YOUR CODE HERE ...
```

- COMMENTS…

## Save me

Generate the document as a Notebook, PDF, Word, or HTML by choosing the relevant option (from the pop-up menu next to the Preview button). Then save your Markdown code by choosing the relevant option (from the task bar menu).

Save the *.Rmd script, so that you can edit it later.