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:

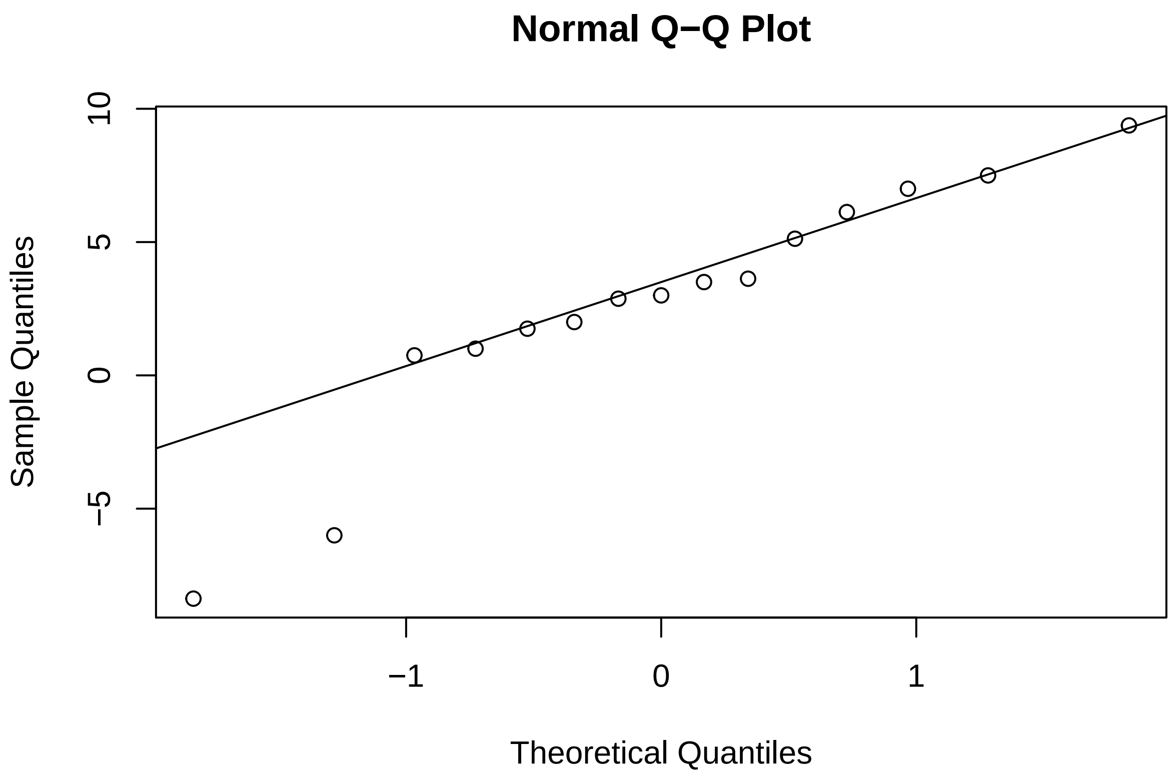
- `Batch` indicates which group of Darwin's plants the pair came from.
- `Cross` are the heights of the cross-fertilised plants.
- `Self` are the heights of the self-fertilised plants.

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

```
library(durham)
data(zeamays)
zeamays[,-1]
##     Cross   Self
## 1  23.500 17.375
## 2  12.000 20.375
## 3  21.000 20.000
## 4  22.000 20.000
## 5  19.125 18.375
## 6  21.500 18.625
## 7  22.125 18.625
## 8  20.375 15.250
## 9  18.250 16.500
## 10 21.625 18.000
## 11 23.250 16.250
## 12 21.000 18.000
## 13 22.125 12.750
## 14 23.000 15.500
## 15 12.000 18.000
```

- The paired t-test is the suitable test to analyse this data-set, according to the material covered in the course. This is because:
  - We deal with the scenario that the observations are in pairs, because the Cross and Self measurements are taken from the same plant.
  - The population variance is (obviously) unknown.
- We do not have enough evidence to support that any of the assumptions of the paired t-test is violated.
  - Each pair of the sample is independently drawn, according to the information given.
  - The QQ-plot, presented below, shows some evidence that the Normality assumption for the differences $D = Cross - Self$ may be violated. Particularly, we can see some asymmetry on the bottom-left part of the QQ-plot.

```
D = zeamays$Cross - zeamays$Self
qqnorm(D)
qqline(D)
```



- Below we pretend that the Normality assumption is not violated, and hence we perform paired t-test. (More safely, we could use the corresponding non-parametric test).

```
alpha = 0.05
t_test = t.test(x=zeamays$Cross,
                y=zeamays$Self,
                paired=TRUE,
                conf.level = 1-alpha)
t_test
```

```
##
##  Paired t-test
##
## data:  zeamays$Cross and zeamays$Self
## t = 2.148, df = 14, p-value = 0.0497
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.003899165 5.229434169
## sample estimates:
## mean of the differences
##                2.616667
```

- The paired t-test shows that there is a significant difference between the Cross height (in inches) and Self height (in inches) at sign. level $\alpha = 0.05$.
- This is because the p-value of the test is p-value= 0.0497029, the the value of the t-statistic is 2.1479875, the degrees of freedom of the statistics are 14, and the critical value at sig. level 0.05 is 2.1447867.

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?

```
library(datasets)
data(sleep)
sleep
```

```
##    extra group ID
## 1    0.7     1  1
## 2   -1.6     1  2
## 3   -0.2     1  3
## 4   -1.2     1  4
## 5   -0.1     1  5
## 6    3.4     1  6
## 7    3.7     1  7
## 8    0.8     1  8
## 9    0.0     1  9
## 10   2.0     1 10
## 11   1.9     2  1
## 12   0.8     2  2
## 13   1.1     2  3
## 14   0.1     2  4
## 15  -0.1     2  5
## 16   4.4     2  6
## 17   5.5     2  7
## 18   1.6     2  8
## 19   4.6     2  9
## 20   3.4     2 10
```

- The t-test for two independent sample is the suitable test to analyse this data-set, according to the material covered in the course, and based on the description of the data given in this document. Note that the distribution of the data-set in R package *datasets* is slightly different implying that the observations are in pairs – we ignore that description here.

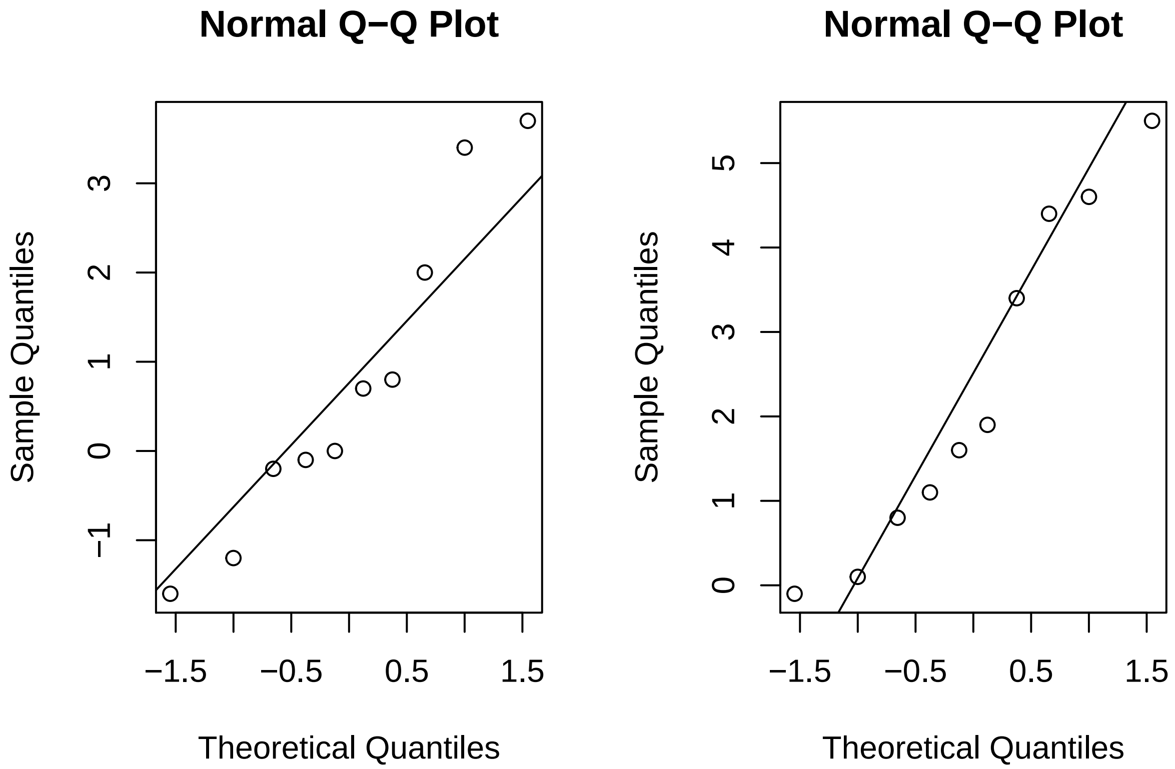
- Assumptions:

  - The population variance is (obviously) unknown. So I use the T statistic.

  - We cannot find substantial evidence against the Normality assumption; i.e. that the two samples are drawn from Normally populated distributions. See the QQ-plots below do not show any serious evidence against the Normality assumption for the two samples.

```
x = sleep$extra[sleep$group==1]
y = sleep$extra[sleep$group==2]
```

```
par(mfrow=c(1,2))
qqnorm(x)
qqline(x)
qqnorm(y)
qqline(y)
```



- We do not find statistically significant evidence that the two population variances are not equal. Precisely, we perform the hypothesis F-test for the equality of variance in order to test if the two populations have equal variances.

```
alpha = 0.05
F_test <- var.test(x, y, ratio = 1,
         alternative = "two.sided",
         conf.level = 1-alpha)
F_test
```

```
##
##  F test to compare two variances
##
## data:  x and y
## F = 0.79834, num df = 9, denom df = 9, p-value = 0.7427
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.198297 3.214123
## sample estimates:
## ratio of variances
##          0.7983426
```

- We cannot reject the null hypothesis that the two population variances are equal at sig. level $\alpha = 0.05$.
- This is because the p-value of the test is p-value= 0.7427199, the value of the F-statistic is 0.7983426, the degrees of freedom of the statistic are (9, 9), and the critical values at sig. level 0.05 are 4.0259942 and 0.2483859.
- Also the 95% confidence interval for the ratio of the variances is (0.198297, 3.2141227) that includes 1.
- We perform the two independent samples t-test.

```r
alpha = 0.05
t_test= t.test(x, y,
       alternative = "two.sided",
       mu = 0,
       paired = FALSE,
       var.equal = TRUE,
       conf.level = 1-alpha)
t_test
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = -1.8608, df = 18, p-value = 0.07919
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.363874  0.203874
## sample estimates:
## mean of x mean of y
##      0.75      2.33
```

- We cannot reject the null hypothesis that the two population means are equal at sig. level $\alpha = 0.05$.
- This is because the p-value of the test is p-value= 0.0791867, the value of the T-statistic is -1.8608135, the degrees of freedom of the statistic are 18, and the critical values at sig. level 0.05 are 2.100922.
- Also the 95% confidence interval for the difference of the means is (-3.363874, 0.203874) that does include 0.
- Therefore, the mean effect (as defined by the experiment) of the two drugs under comparison are not significantly different at sig. level 0.05.

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.

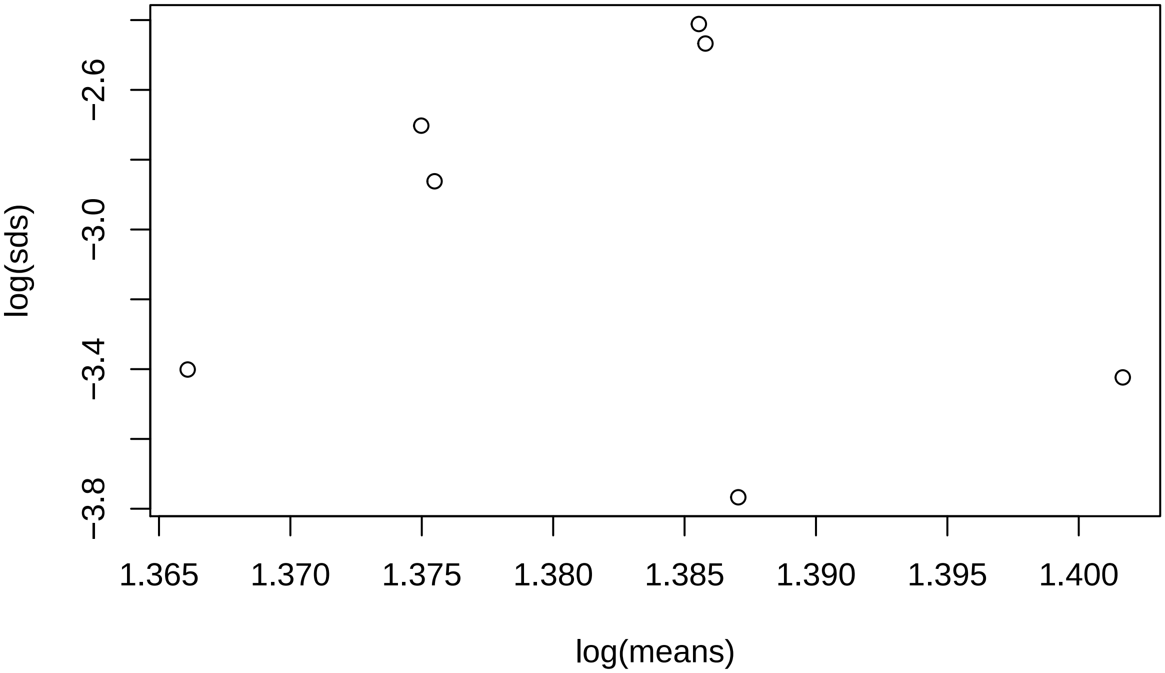
Create a data frame from chlorph which only contains the data for manufacturer A. Use the Multiple samples technique described above to analyse the data. What do you find and how?

```r
library(durham)
data(chlorph)
chlorphA <- chlorph[chlorph$manufacturer=='A',-2]
summary(chlorphA)
```

```
##  chlorpheniramine laboratory
##  Min.   :3.810    1:10
##  1st Qu.:3.935    2:10
##  Median :4.000    3:10
##  Mean   :3.985    4:10
##  3rd Qu.:4.037    5:10
##  Max.   :4.130    6:10
##                    7:10
```

- Location-scale plot is presented below:

```
means = tapply(chlorphA$chlorpheniramine, chlorphA$laboratory, mean)
sds = tapply(chlorphA$chlorpheniramine, chlorphA$laboratory, sd)
plot(log(means), log(sds))
```

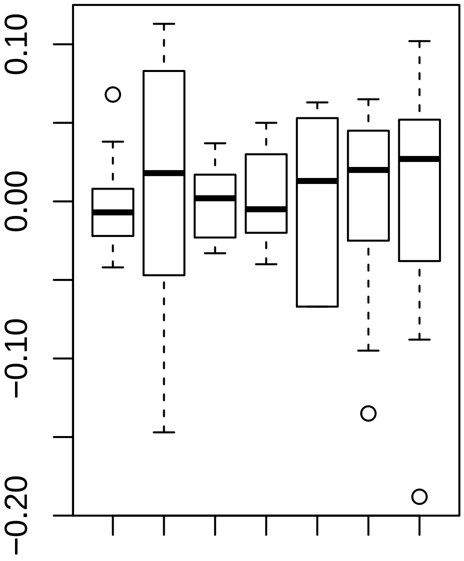


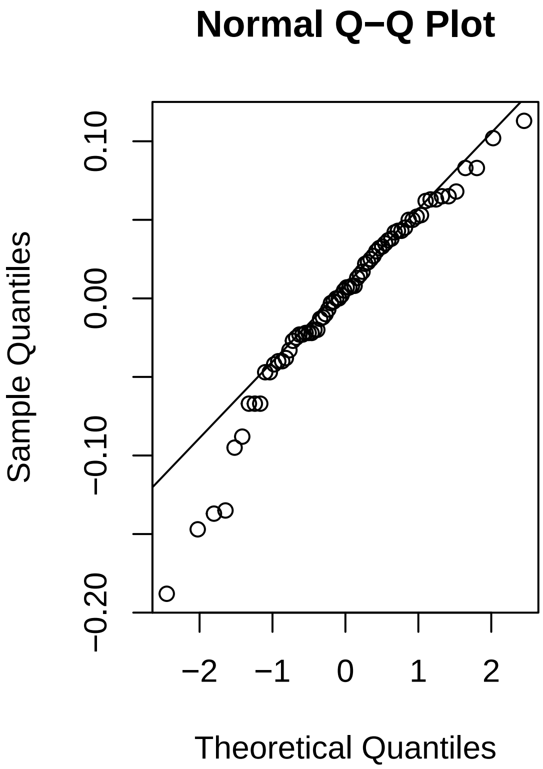
- It does not suggest that there is need for a transformation (of course this is arguable).
- We fit the model

```
fit = aov(chlorpheniramine~laboratory, data=chlorphA)
```

- Then we check if the assumptions are satisfied
  - The residuals need to have constant mean and variance with respect to the levels of the factors.

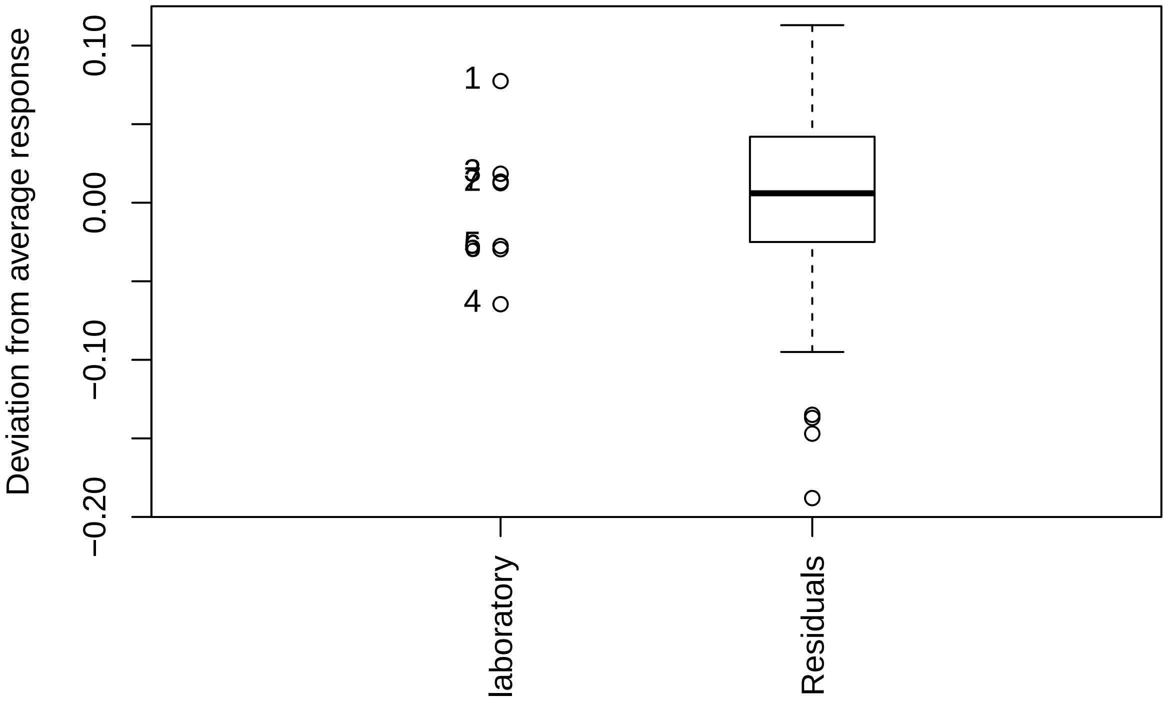
```
par(mfrow=c(1,2))
plot(chlorphA$laboratory, fit$residuals)
qqnorm(fit$residuals)
qqline(fit$residuals)
```





- The boxplot above shows that the spread of the residuals at some Laboratories are different than than others. Hence, this indicates that the homogeneity assumption may violated.
- The QQ-plot of the residuals indicates strong evidence that the distribution of the residuals is left skewed, and hence not Normal.
- Then a Non-parametric approach could be more appropriate for this problem. However, here we pretend that these assumptions are not violated, and proceed as if none of the assumptions was violated.
- The Effects and residuals plot is presented below.

```r
plot(fit)
```



- The ANOVA table is presented below

```
alpha = 0.05
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: chlorpheniramine
##            Df  Sum Sq  Mean Sq F value    Pr(>F)
## laboratory  6 0.12474 0.020789  5.6601 9.453e-05 ***
## Residuals  63 0.23140 0.003673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We reject the null hypothesis that the mean chlorpheniramine maleate amount is the same for all the labs at sig. level 0.05. Therefore, at least the mean of chlorpheniramine maleate amounts between two labs are different at sig. level 0.05.

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.