Further Statistical Analysis using R

Mark Dunning, Matt Eldridge and Sarah Vowler *

Last Document revision: September 23, 2015

Contents

*Acknowledgements: Sarah Dawson

1 Statistics Introduction

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.". R.A. Fisher, 1938

The goals of statistical methods could be summarised as follows:

- drawing conclusions about a population by analysing data on just a sample;
- evaluating the uncertainty in these conclusions; and,
- designing the sampling approach so that valid and accurate conclusions can be made from the data collected.

1.1 Statistical Tests

The statistical approach used is dependent on the data type. In this document we will describe **ANOVA** (analysis of variance), which can be used when we have two or more groups of continuous numerical data, and **Linear Regression**

1.2 Exploratory Analysis

Before conducting a formal analysis of our data, it is always a good idea to run some exploratory checks of the data:

- To check that the data has been read in or entered correctly;
- To identify any outlying values and if there is reason to question their validity, exclude them or investigate them further;
- To see the distribution of the observations and whether the planned analyses are appropriate.

It's always a good idea to calculate some summary statistics for your data, such as the mean and standard deviation, or the median and inter-quartile range if your data is skewed. You should also consider whether there may be outliers in your data (but do not remove them from the analysis without good reason) or whether there may be missing data. Summary statistics were covered in detail in Part 1 of the course.

1.3 Statistical Tests - basic setup

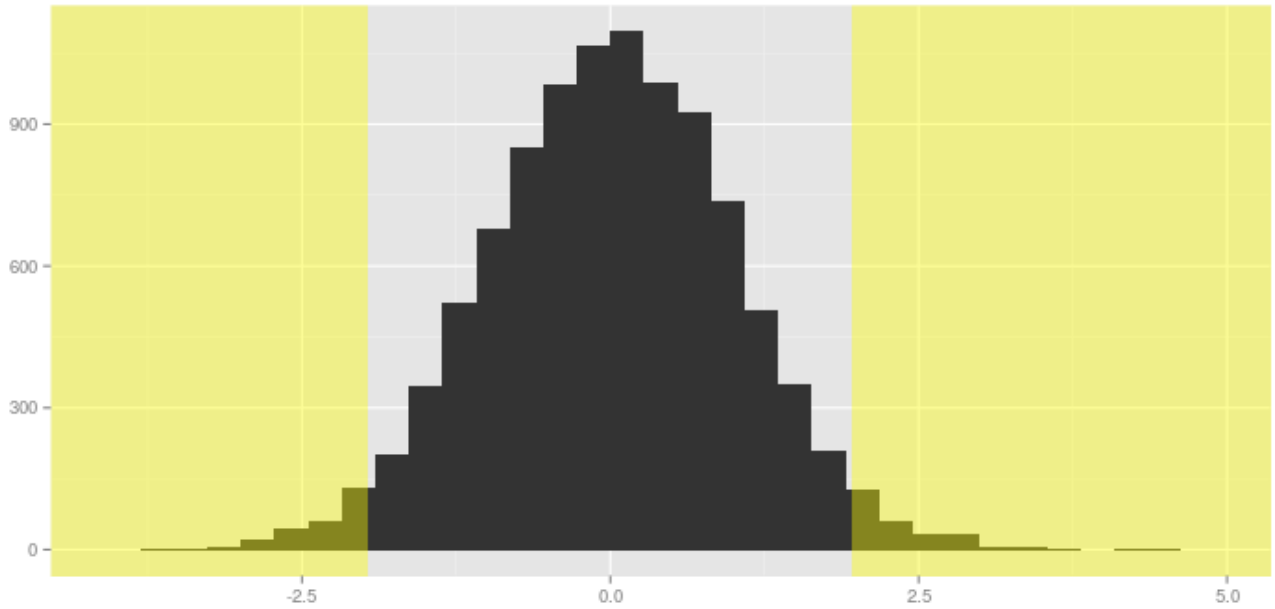
There are four key steps in every statistical test:

- 1 Formulate a **null hypothesis**, H_0. This is the working hypothesis that we wish to disprove.
- 2. Under the assumption that the **null hypothesis** is true, calculate a **test statistic** from the data.
- 3. Determine whether the **test statistic** is more extreme than we would expect under the **null hypothesis**, i.e. look at the **p-value**.
- 4. Reject or do not reject the **null hypothesis**.

As the name suggests, the null hypothesis typically corresponds to a **null** effect.

For example, there is **no difference** in the measurements in group 1 compared with group 2. A small p-value indicates that the probability of observing such a test statistic as small under the assumption that the null hypothesis is true. If the p-value is below a pre-specified **significance level**, then this is a **significant result** and, we would conclude, there is evidence to reject the null hypothesis.

The **significance level** is most commonly set at 5% and may also be thought of as the **false positive rate**. That is, there is a 5% chance that the null hypothesis is true for data-sets with test statistics corresponding to p-values of less than 0.05 i.e. we may wrongly reject the null hypothesis when the null hypothesis is true (false positive).



Equally, we may make **false negative** conclusions from statistical tests. In other words, we may not reject the null hypothesis when the null hypothesis is, in fact, not true. When referring to the false negative rate, statisticians usually refer to **power**, which is 1-false negative rate.

The **power** of a statistical test will depend on:

- The **significance level** - a 5% test of significance will have a greater chance of rejecting the null than a 1% test because the strength of evidence required for rejection is less.
- The **sample size** the larger the sample size, the more accurate our estimates (e.g. of the mean) which means we can differentiate between the null and alternative hypotheses more clearly.
- The **size of the difference or effect** we wish to detect bigger differences (i.e. alternative hypotheses) are easier to detect than smaller differences.
- The **variability**, or standard deviation, of the observations the more variable our observations, the less accurate our estimates which means it is more difficult to differentiate between the null and alternative hypotheses.

	Null hypothesis does not hold	**Null hypothesis holds**
Reject null hypothesis	Correct *True Positive*	Wrong *False positive*
Do no reject null hypothesis	Wrong *False negative*	Correct *True negative*

Table 1: Error definitions

2 R Introduction

To install R visit www.r-project.org. In the 'Getting Started' box half-way down the page follow the 'download R' link. Scroll down to the UK and select any one of the three links. On the next page choose the appropriate operating system for your computer from the three 'Download R for...' options.

This manual, and the accompanying practical will assume some familiarity with the R statistical language. In particular, you should be familiar with the following concepts:

- Using the RStudio program
- Setting your working directory
- Creating variables and basic object types; in particular vectors and data frames
- Using built-in R functions
- Using R to get help on functions
- Subset operations for vectors and data frames using the [] notation
- Reading tabular data into R
- Basic plots; scatter plots, boxplot and histogram

Several Online videos are available that cover this materials. For example

- http://shop.oreilly.com/product/0636920034834.do
- http://blog.revolutionanalytics.com/2012/12/coursera-videos.html
- http://bitesizebio.com/webinar/20600/beginners-introduction-to-r-statistical-software

2.1 Installation of R

After clicking on the 'Download R for Windows' link, select 'install R for the first time' on the following page. The version of R used to write this manual is 3.2.2, the version number you download may be

different as new versions are released every six months. Following this link will start the installation of R. If you get a security warning select 'Run'. Follow the directions in the install wizard to install R. We haven chosen to run `R` through the `RStudio` interface, which you will also need to install.

2.2 Installation of RStudio

To install RStudio visit http://www.rstudio.com/products/RStudio/ and follow the links to download `RStudio Desktop` for your operating system.

2.3 R packages Used

In order to run the examples in this manual, and the practical, you will need to execute the following command in R to install the required packages.

```
install.packages(c("tidyr","beeswarm","RColorBrewer"))
```

3 ANOVA

ANOVA stands for *an*alysis *o*f *v*ariance. There are three main types of ANOVA: one-way, two-way and repeated measures. In this course our main focus will be on the one-way ANOVA.

3.1 One-way ANOVA

The two-sample t-test is useful when we have just two groups of continuous data to compare, but when we want to compare more than two groups a one-way ANOVA can be used to simultaneously compare all groups rather than carrying out several individual two-sample t-tests. The main advantage of doing this is that it reduces the number of tests being carried out, meaning that the type I error rate (the probability of seeing a significant result just by chance) does not become inflated.

A one-way ANOVA compares group means by partitioning the variation in the data into **between group** variance and **within group variance**(see Table 2).

Source of Variation	**Degrees of Freedom**	**Sum of Squares**	**Mean Squares**	$F_{k-1,N-k}$
Between groups	$k-1$	$BSS = \sum_k^{i=1} n_i y_i^2 - \frac{T^2}{N}$	$S_B^2 = \frac{BSS}{k-1}$	
Within groups	$N-k$	$WSS = S - \sum_{i=1}^{k} n_i y_i^2$	$S_W^2 = \frac{WSS}{N-k}$	
Total	$N-1$	TSS = BSS + WSS		$\frac{S_B^2}{S_W^2}$

Table 2: One-way ANOVA table

The between group variance is divided by the within group variance to give an F statistic. This tells us the ratio of between group variation to within group variation. A large F-value implies that there are

significant differences between groups and conversely a small F-value implies there are not significant differences between groups. Giving this a little bit of thought, the idea becomes more intuitive:

- If the variance within groups is small, but between groups the variance is very large, we can infer that there are likely to be differences between groups. Following this theory, the F statistic is calculated by dividing the between group variation by the within group variation. So in this scenario, we divide a large number by a comparatively small number and this leaves us with a large(ish) number for our F statistic, corresponding to a small p-value.
- If the variance within groups is large, but between groups the variance is also large, it is more difficult to know whether the groups truly differ in their mean value. In this scenario, the F statistic is calculated by dividing a large number by another large number - depending on the relative size of these two large numbers we may be left with a large or small number for our F statistic. The same is true when the variance within and between groups are both small.
- Finally, if the variance within group is large, but between groups the variance is very small, we can infer that there are unlikely to be differences between groups. In this scenario, the F statistic is calculated by dividing a small number by a large number and this will result in a small(ish) number for our F statistic, corresponding to a large p-value.

Luckily, you wont need to do the calculations from Table 2 by hand as R will do these all for you, but its good to have an appreciation of what the test is actually doing in the background.

Obviously the outcome of the one-way ANOVA depends on the data available. If there is high variation in the data, a much larger sample will be needed to detect a difference between groups. Likewise, if we are interested in detecting very small differences between groups a larger sample size will be required.

3.2 ANOVA assumptions

There are several assumptions behind the one-way ANOVA:

- Normally distributed response (assessed for each group separately)
- Approximately equal variance across the groups
- Independent observations

The main assumption is that the distribution of the response variable should be normally distributed for each group being compared. This can be assessed prior to fitting the ANOVA by constructing a histogram of the response variable for each group being compared. This can also be assessed after fitting the ANOVA by constructing a normal probability plot of the residuals (sometimes called a Q-Q plot).

Another important assumption is that there is approximately equal variance across the groups being compared. This assumption is important because of the way the F-test in the ANOVA uses the pooled variance across groups. If one group has a much larger variance than another group, the results of the F-test may not be valid. The equal variance assumption can be assessed using either Bartlett's or Levene's test (REMEMBER! this adds to the multiple testing problem), or visually using a histogram plotted separately for each group.

A third assumption of the one-way ANOVA is the independence of observations. There is no easy way of assessing independence, so a lot of people overlook this assumption. However, a little thought about where the data comes from and how it was collected can give us a good indication of whether the observations are independent or not. Things like taking observations from related individuals or having multiple measurements per subject will cause the independence assumption to be invalid.

If the F-test provides a significant result, we may be interested in making comparisons between pairs of groups to identify where the difference lies and estimate the effect size. This can be done by using unpaired two-sample t-tests. If we wish to make multiple comparisons we must be careful to adjust for multiple testing and R has several options to do this. There are several different types of multiple-testing adjustment than can be made, each suiting different types of comparisons. These are discussed in more detail in the section 3.3.

3.3 Choosing the correct post-test

Tukey	Compare all pairs of columns
Bonferroni	Compare all pairs of columns OR compare selected pairs of columns
Dunnett	Compare all columns vs. control column
Trend test	Test for linear trend between mean and column number

Table 3: Multiple-testing adjustment methods

3.4 One-way ANOVA Example

Example: The protein expression level was measured in 5 cell types from a single cell line. We want to know whether there are any differences in the expression level between the five different cell types. The raw data are given in Table 4. These data come from the Babraham Bioinformatics course Statistical Analysis using GraphPad Prism

Our **null hypothesis** is that the mean value is the same in each of the five groups.

Our **alternative hypothesis** is that the mean value is different in one or more of the five groups.

These data can be read using the `read.csv` function in R, which will create a *data frame* representation.

```
proteinData <- read.csv("protein-expression.csv")
```

At this point, it is a good idea to inspect the data to make sure they have been imported correctly. Sometimes R will read data without complaint, but create an object that you can't actually use for analysis. If you are using RStudio, the command `View(proteinData)` will bring-up a display of the dataset. Otherwise the following commands will tell you about the dimensions of the data, first few lines and numerical summary of each column.

	A	B	C	D	E
1	0.40	0.26	0.24	1.04	0.74
2	1.50	0.47	0.25	2.78	0.99
3	0.98	0.42	1.01	0.82	1.26
4	0.33	0.64	0.77	1.65	1.50
5	0.75	0.32	0.47	0.49	0.30
6	1.48	0.65	0.47	0.97	0.34
7	1.18	0.43	0.46	1.39	0.77
8	0.33	0.67	0.65	3.24	1.94
9	1.42	0.43	0.41	1.12	2.62
10	2.09	0.70	0.81	2.82	1.42
11	1.37	0.79	1.20	1.27	0.73
12	1.23	0.89	1.08	1.60	2.09
13			0.34	1.98	1.52
14			1.98	9.32	1.67
15			1.39	2.31	3.40
16			1.12	4.19	2.16
17			3.14	1.73	2.31
18			2.78	5.16	1.32

Table 4: Protein Expression data

```
head(proteinData)
```

```
##      A    B    C    D    E
## 1 0.40 0.26 0.24 1.04 0.74
## 2 1.50 0.47 0.25 2.78 0.99
## 3 0.98 0.42 1.01 0.82 1.26
## 4 0.33 0.64 0.77 1.65 1.50
## 5 0.75 0.32 0.47 0.49 0.30
## 6 1.48 0.65 0.47 0.97 0.34
```

```
dim(proteinData)
```

```
## [1] 18  5
```

```
summary(proteinData)
```

```
##        A                B                C                D               E
##  Min.   :0.3300   Min.   :0.2600   Min.   :0.2400   Min.   :0.490   Min.   :0.300
##  1st Qu.:0.6625   1st Qu.:0.4275   1st Qu.:0.4625   1st Qu.:1.157   1st Qu.:0.825
##  Median :1.2050   Median :0.5550   Median :0.7900   Median :1.690   Median :1.460
##  Mean   :1.0883   Mean   :0.5558   Mean   :1.0317   Mean   :2.438   Mean   :1.504
##  3rd Qu.:1.4350   3rd Qu.:0.6775   3rd Qu.:1.1800   3rd Qu.:2.810   3rd Qu.:2.053
##  Max.   :2.0900   Max.   :0.8900   Max.   :3.1400   Max.   :9.320   Max.   :3.400
##  NA's   :6        NA's   :6
```

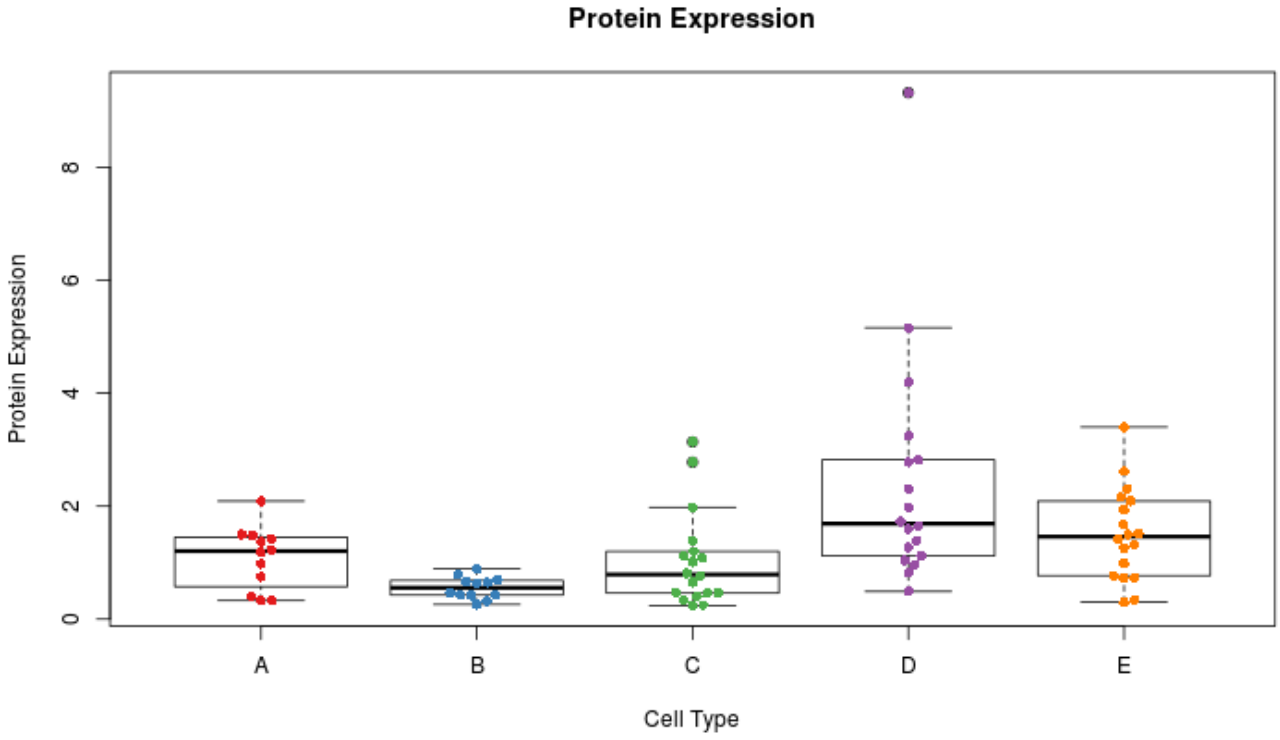


Figure 1: Boxplot of the protein expression levels for five cell types

3.5 Checking the model assumptions

A boxplot of these data can be created using the `boxplot` function, which allows us to compare the median and inter-quartile range (IQR) of each cell type. Optionally, we can use the *beeswarm* package to overlay individual points on the plot. See Figure 1

```
library(beeswarm)
library(RColorBrewer)
boxplot(proteinData,xlab="Cell Type",ylab="Protein Expression",main="Protein Expression")
beeswarm(proteinData, add=TRUE,pch=16,col=brewer.pal(5,"Set1"),method="swarm")
```

comment: The RColorBrewer package is also used to define a colour pallette for the dataset.

Figure 1 shows us that the median value varies between the five groups. We can also see that the data is skewed for cell types A and D, as the bar in the middle, which shows the median, is not equally between the two outer bars, which show the lower and upper quartiles. In addition, there are extreme values present for cell types C and D. We can also see that protein expression levels in some groups are much more variable than in others; the protein expression levels in group B are very consistent, but for groups C, D and E they are much more varied.

We can make a similar assessment of the data using a histogram (plotted separately for each group in our dataset). See Figure 2. In particular we can use these histograms to make an assessment of

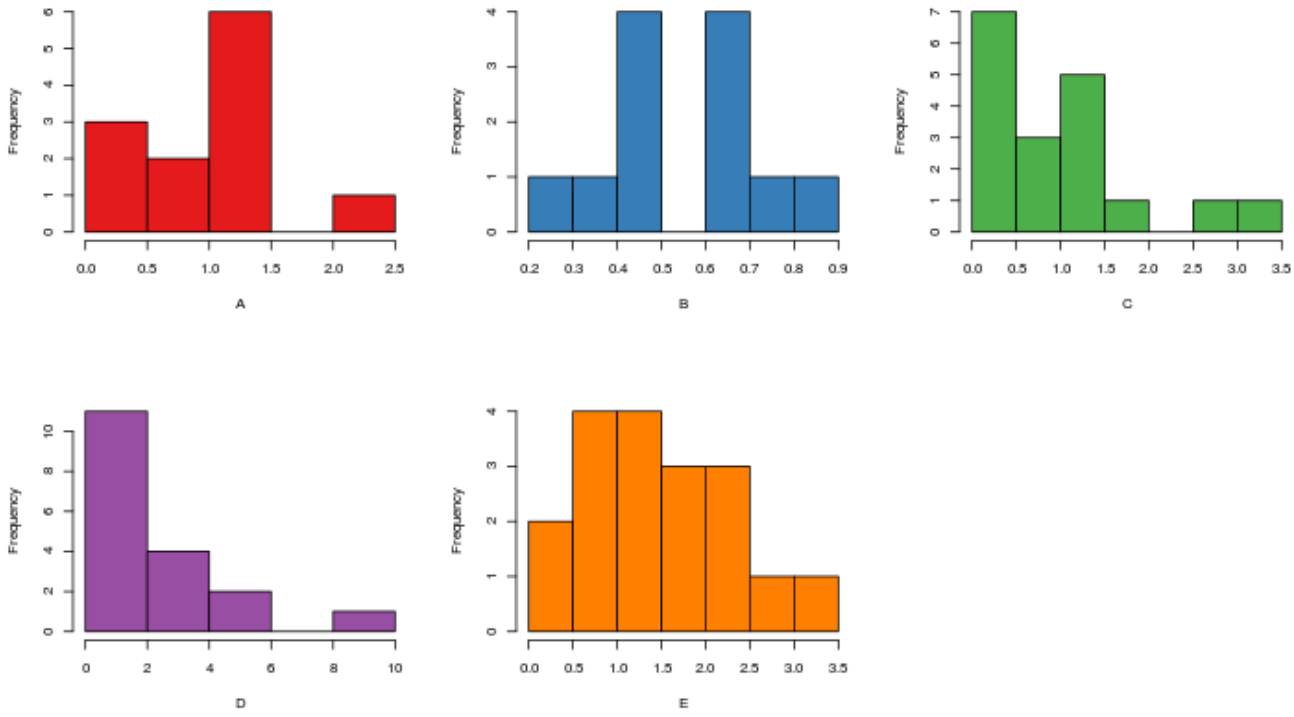


Figure 2: Histogram of the protein expression levels for five cell types

normality and constant variance which are two of the assumptions behind ANOVA (Section 3.2).

```
par(mfrow=c(2,3))
cols <- brewer.pal(5,"Set1")
hist(proteinData$A,xlab="A",col=cols[1],main="")
hist(proteinData$B,xlab="B",col=cols[2],main="")
hist(proteinData$C,xlab="C",col=cols[3],main="")
hist(proteinData$D,xlab="D",col=cols[4],main="")
hist(proteinData$E,xlab="E",col=cols[5],main="")
```

comment: If you are more-familiar with programming, you could write this chunk of code with a for loop (or similar)

In its current format, the data are unsuitable for analysis using one-way ANOVA as both the normality assumption and the constant variance assumption are violated. We can sometimes overcome this issue by transforming the data, and in this case we can use a log-transformation to normalise the data (Note: you do not need to do this step if the assumptions of normality and constant variance hold). This can be carried out within R using the `log` function.

```
proteinData.ln <- log(proteinData)
head(proteinData.ln)

##              A          B          C          D          E
```

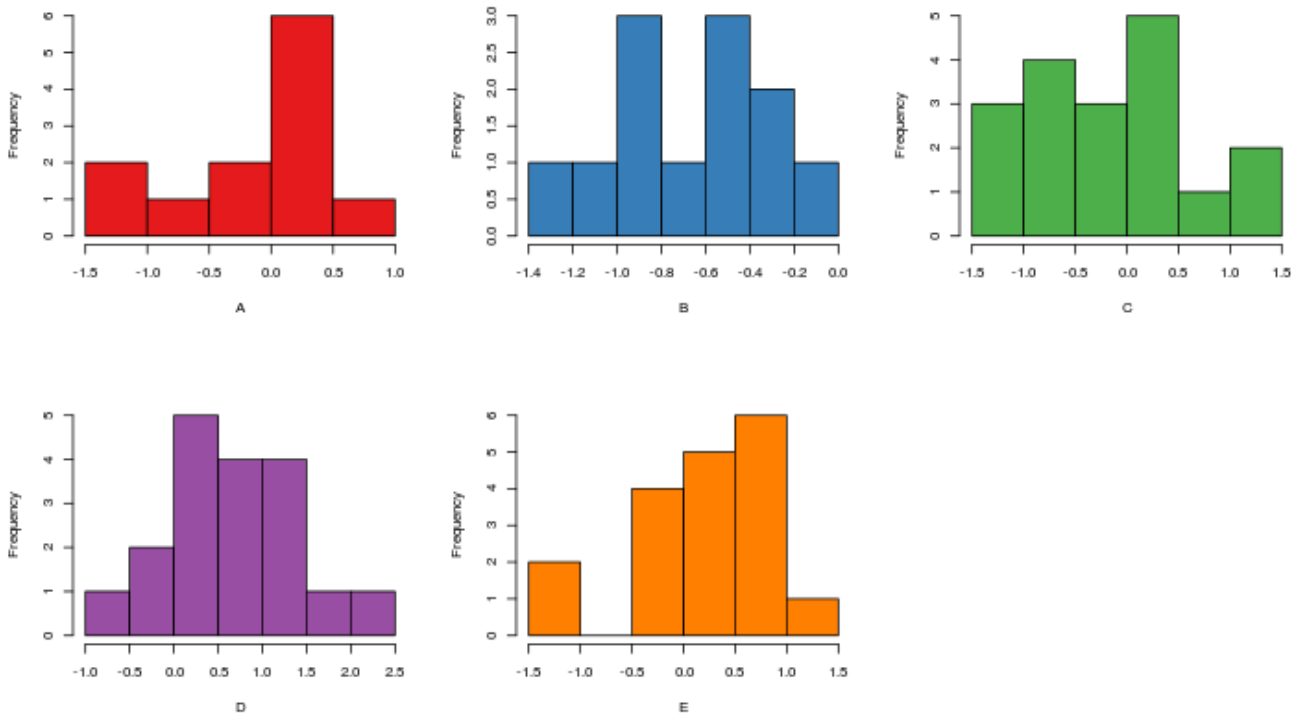


Figure 3: Histogram of the natural-log-transformed protein expression levels for five cell types

```
## 1 -0.91629073 -1.3470736 -1.427116356  0.03922071 -0.30110509
## 2  0.40546511 -0.7550226 -1.386294361  1.02245093 -0.01005034
## 3 -0.02020271 -0.8675006  0.009950331 -0.19845094  0.23111172
## 4 -1.10866262 -0.4462871 -0.261364764  0.50077529  0.40546511
## 5 -0.28768207 -1.1394343 -0.755022584 -0.71334989 -1.20397280
## 6  0.39204209 -0.4307829 -0.755022584 -0.03045921 -1.07880966
```

comment: If you are not sure what the `log` is doing, don't forget that you can bring-up the help page: `?log`

Hopefully the transformation will result in normally distributed data. To check this, create another histogram in the same way you did on the untransformed data (See Figure 3)

```
par(mfrow=c(2,3))
hist(proteinData.ln$A,xlab="A",col=cols[1],main="")
hist(proteinData.ln$B,xlab="B",col=cols[2],main="")
hist(proteinData.ln$C,xlab="C",col=cols[3],main="")
hist(proteinData.ln$D,xlab="D",col=cols[4],main="")
hist(proteinData.ln$E,xlab="E",col=cols[5],main="")
```

3.6 Fitting the model

Figure 3 now shows that most of the cell types are approximately normally distributed, though cell type A is still a little questionable. We'll proceed for now, as the one-way ANOVA is quite robust to small deviations from normality, but we must bear this in mind when interpreting out results. We can also see from Figure 3 that the variance in each of the five groups is now much more similar. They do NOT need to be perfectly the same, and in this case they are similar enough for the assumption of equal variance to be reasonable. Now that we are happy that the assumptions are reasonable, we can perform the one-way ANOVA.

When perfoming linear regression or ANOVA in R, it is more convenient to transform our data into **long** format. The package *tidyr* can do this for us [1].

Once we have transformed our data, the `aov` function fits the analysis of variance model to our data. Diagnostic plots of the model fit can be visualised using the `plot` function (Figure 4). Of the most interest is the QQ-plot, which allows us to assess whether the distribution of the response variable should is normally distributed for each group being compared.

```
library(tidyr)
anovaData <- gather(proteinData.ln)
head(anovaData)
```

```
##   key       value
## 1   A -0.91629073
## 2   A  0.40546511
## 3   A -0.02020271
## 4   A -1.10866262
## 5   A -0.28768207
## 6   A  0.39204209
```

```
mod <- aov(value~ key, data=anovaData)
mod
```

```
## Call:
##    aov(formula = value ~ key, data = anovaData)
##
## Terms:
##                       key Residuals
## Sum of Squares  14.26860  32.05574
## Deg. of Freedom        4        73
##
## Residual standard error: 0.6626611
## Estimated effects may be unbalanced
## 12 observations deleted due to missingness
```

```
par(mfrow=c(2,2))
plot(mod)
```

[1] For a comprehensive description of tidy data, see Hadley Wickham's paper [1] or video [2] on the subject

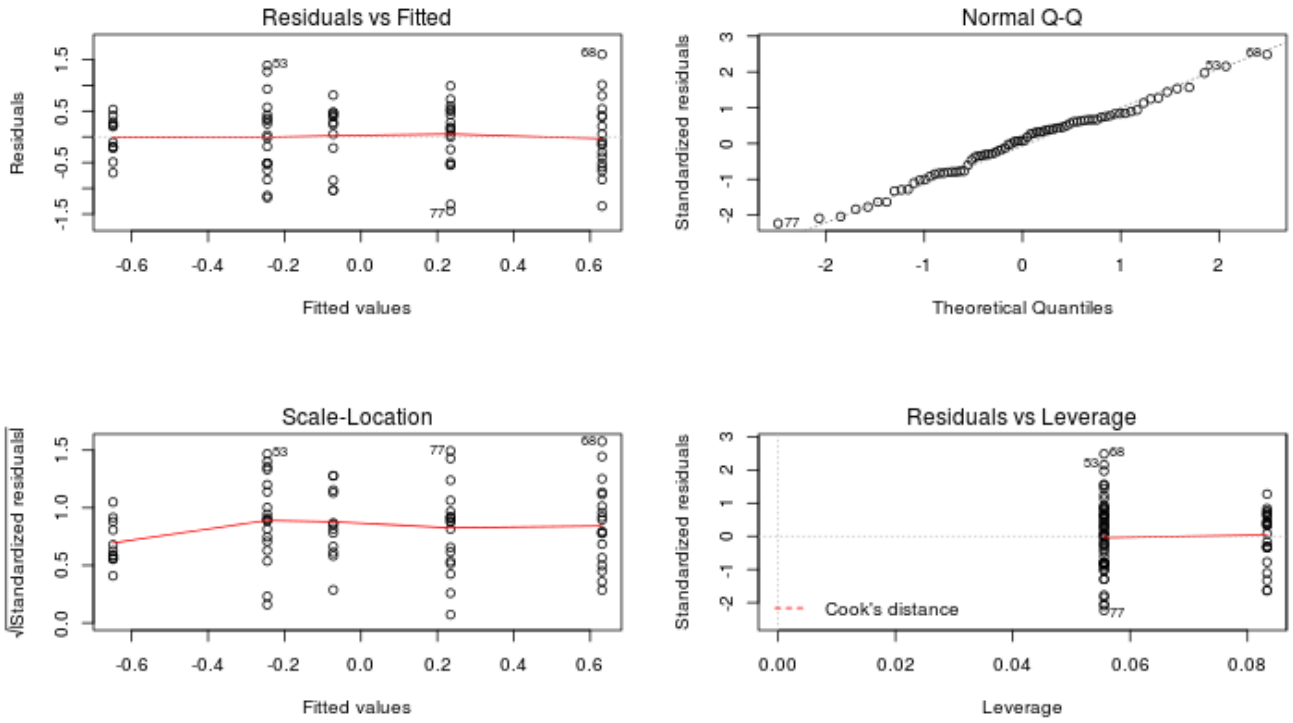


Figure 4: Visualising the results of the ANOVA model

The `summary` function allows us to assess

```
summary(aov(mod))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## key          4  14.27   3.567   8.123 1.78e-05 ***
## Residuals   73  32.06   0.439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 12 observations deleted due to missingness
```

One of the assumptions of the one-way ANOVA, which we have already explored with our scatter plot, is that the variances in each of the five groups are approximately equal. We can formally test this assumption when carrying out a one-way ANOVA, using a Bartlett's test. The results provide a p-value indicating whether equal variance can be assumed, with a value less than 0.05 suggesting that equal variance should not be assumed.

```
bt <- bartlett.test(value~key,data=anovaData)
bt
```

```
##
```

```
##  Bartlett test of homogeneity of variances
##
## data:  value by key
## Bartlett's K-squared = 5.8261, df = 4, p-value = 0.2125
```

In this example, the p-value was 0.213, so there is no evidence to suggest that the variances in each of the five groups aren't approximately equal to each other. If the Bartlett's test gives a significant p-value, we cannot assume equal variances across the groups and a non-parametric test, such as the Kruskal-Wallis test, should be used instead of a one-way ANOVA.

The equal variance assumption is reasonable in this example, so we can go ahead and use the one-way ANOVA to analyse the data. The results of the one-way ANOVA provide a p-value of < 0.0001 which is statistically significant. This suggests that there is evidence of a difference in the mean log-transformed protein expression levels between two or more of the five cell types. As the result of the one-way ANOVA was significant, we may be interested in making further comparisons between pairs of groups, and we can do this with the post- test results. Note: if the one-way ANOVA result had not been significant we would usually stop here and not look at the post-test results.

```
post.tests<- TukeyHSD(mod)
post.tests
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = value ~ key, data = anovaData)
##
## $key
##           diff        lwr       upr     p adj
## B-A -0.5762181 -1.3329262 0.1804901 0.2187264
## C-A -0.1726875 -0.8634644 0.5180894 0.9560187
## D-A  0.7034259  0.0126490 1.3942027 0.0438762
## E-A  0.3068774 -0.3838995 0.9976543 0.7265567
## C-B  0.4035305 -0.2872463 1.0943074 0.4809387
## D-B  1.2796439  0.5888671 1.9704208 0.0000183
## E-B  0.8830955  0.1923186 1.5738723 0.0054767
## D-C  0.8761134  0.2582638 1.4939630 0.0015431
## E-C  0.4795649 -0.1382847 1.0974145 0.2023355
## E-D -0.3965485 -1.0143981 0.2213012 0.3841989
```

The post-tests are actually just unpaired t-tests, but the results reported are adjusted for multiple testing. In this example, we want to compare all pairs of groups, but sometimes there may be specific groups that you wish to compare. You should plan your comparisons before starting your analysis and it is better, at least from a statistical angle, to perform the least number of comparisons that will sufficiently answer your question(s). As we want to perform more than one t-test on this data (in fact we are performing 10 pairwise comparisons!), we must be careful to adjust for multiple testing. Here we see that there is a significant difference between cell types A v D (p value: 0.0438762), a very

significant difference between cell types B v E (p value: 0.0054767) and C v D (p value: 0.0015431), and an extremely significant difference between cell types B v D (p value: 1.8302674×10^{-5}).

Just because a result is statistically significant does not mean that it is biologically or clinically important. You can refer to the mean difference column to judge whether a difference of the seen in the data is likely to be of biological or clinical importance, but remember that these values are now on the natural log scale! You can transform back to the original scale by taking the exponential of the mean difference: in this example, the mean difference between B and D is 1.28 on the natural log scale. On the original scale this translates to $e^{1.28} = 0.562$.

4 References

References

[1] Hadley Wickham. Tidy data. *Journal of Statistical Software*, VV. URL: http://vita.had.co.nz/papers/tidy-data.pdf.

[2] Hadley Wickham. Tidy data and tidy tools. URL: https://vimeo.com/33727555.