Further Statistical Analysis using R

Mark Dunning, Matt Eldridge and Sarah Vowler *

Last Document revision: September 29, 2015

Contents

1 Statistics Introduction **2**
- 1.1 Statistical Tests . . . 2
- 1.2 Exploratory Analysis . . . 3
- 1.3 Statistical Tests - basic setup . . . 3

2 R Introduction **4**
- 2.1 Installation of R . . . 5
- 2.2 Installation of RStudio . . . 5
- 2.3 R packages Used . . . 5

3 ANOVA **5**
- 3.1 One-way ANOVA . . . 6
  - 3.1.1 ANOVA assumptions . . . 7
  - 3.1.2 Choosing the correct post-test . . . 7
  - 3.1.3 One-way ANOVA Example . . . 8
  - 3.1.4 Checking the model assumptions . . . 9
  - 3.1.5 Fitting the model . . . 11
- 3.2 The Kruskal-Wallis test . . . 15
  - 3.2.1 Example . . . 16
- 3.3 Friedman's test . . . 18
  - 3.3.1 Examples from the literature . . . 18
  - 3.3.2 Null Hypothesis . . . 19
  - 3.3.3 Assumptions . . . 19
  - 3.3.4 Method . . . 19
  - 3.3.5 Example . . . 20
  - 3.3.6 Post-hoc testing . . . 21
  - 3.3.7 Presentation of the Results . . . 22
  - 3.3.8 Advantages and Limitations . . . 22

*Acknowledgements: Sarah Dawson

3.3.9 Summary . . . 22
3.4 Median Test . . . 23
3.4.1 Null Hypothesis . . . 23
3.4.2 Assumptions . . . 23
3.4.3 Method . . . 23
3.4.4 Example . . . 24
3.5 Jonckheere-Terpstra Test . . . 25
3.5.1 Null Hypothesis . . . 25
3.5.2 Assumptions . . . 25
3.5.3 Method . . . 25
3.6 Large Sample Size . . . 26
3.6.1 Example . . . 26
3.6.2 Presentation of results . . . 27
3.6.3 Analysis in R . . . 27
3.6.4 Advantages and limitations . . . 28
3.6.5 Summary . . . 28
3.6.6 Summary of Several Independent Samples . . . 28

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 our Introductory Statistics course.

1.3 Statistical Tests - basic setup

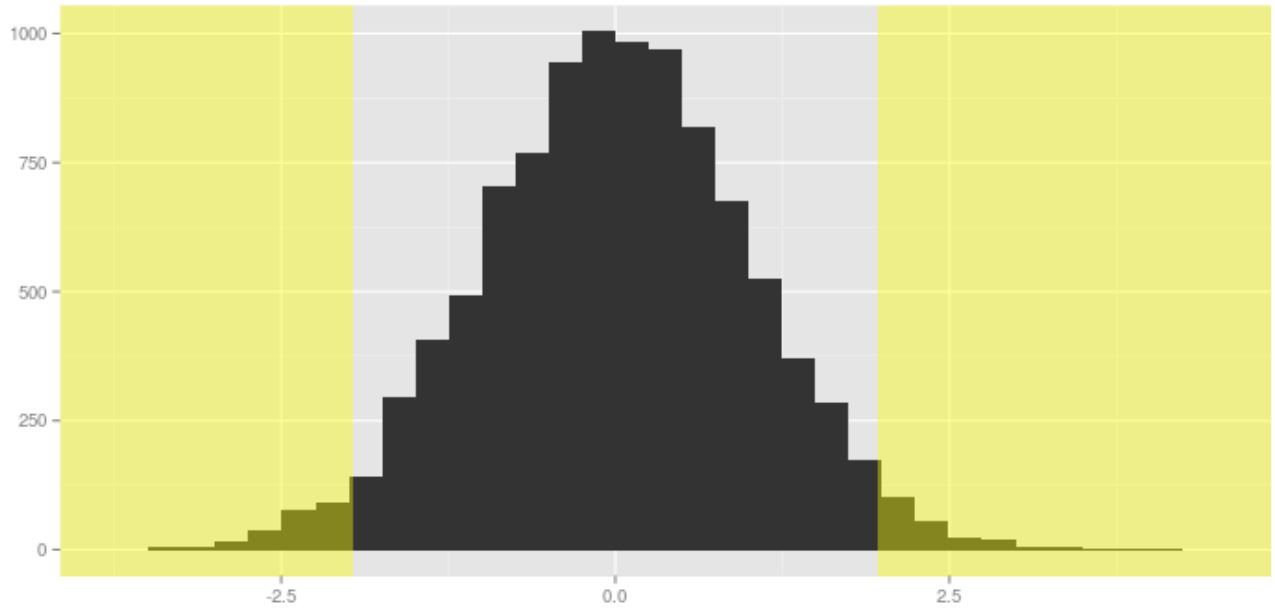
There are four key steps in every statistical test:

- 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.

3.1.1 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 3.1.2.

3.1.2 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

3.1.3 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

	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

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.

```
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
```

3.1.4 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 normality and constant variance which are two of the assumptions behind ANOVA (Section 3.1.1).

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

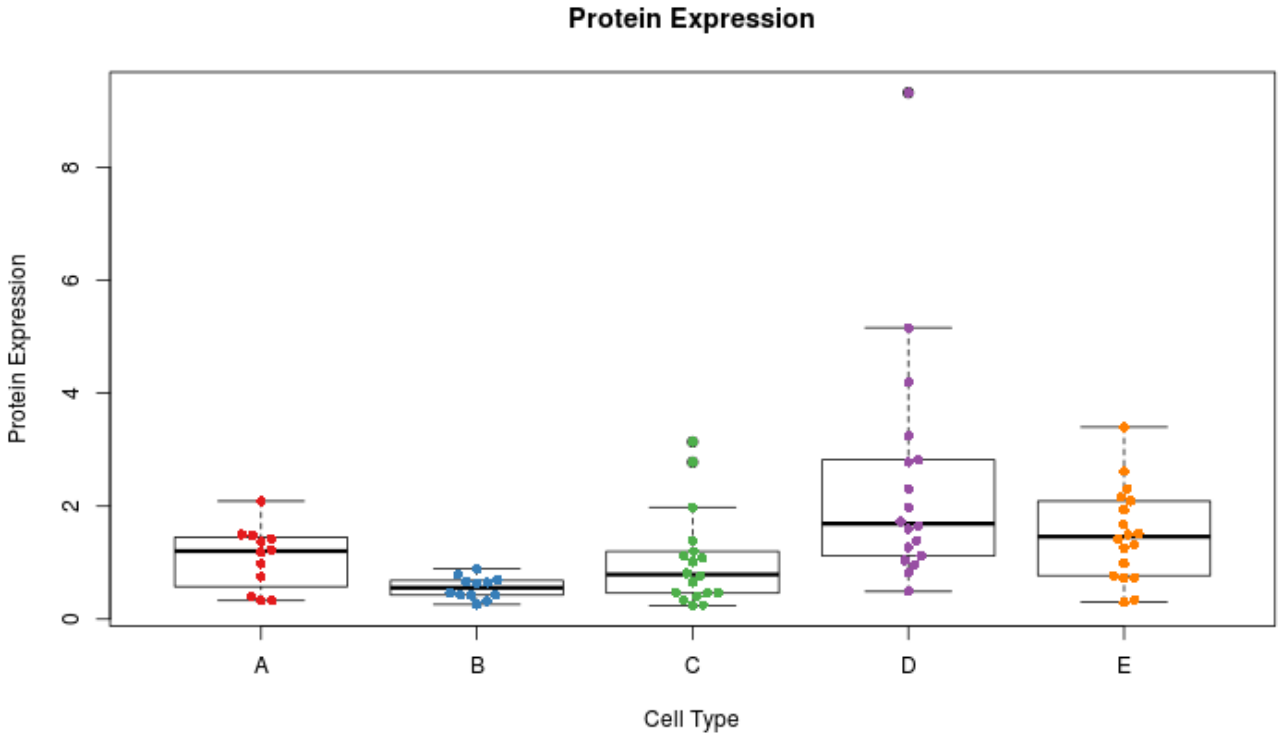


Figure 1: Boxplot of the protein expression levels for five cell types

```
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
## 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
```

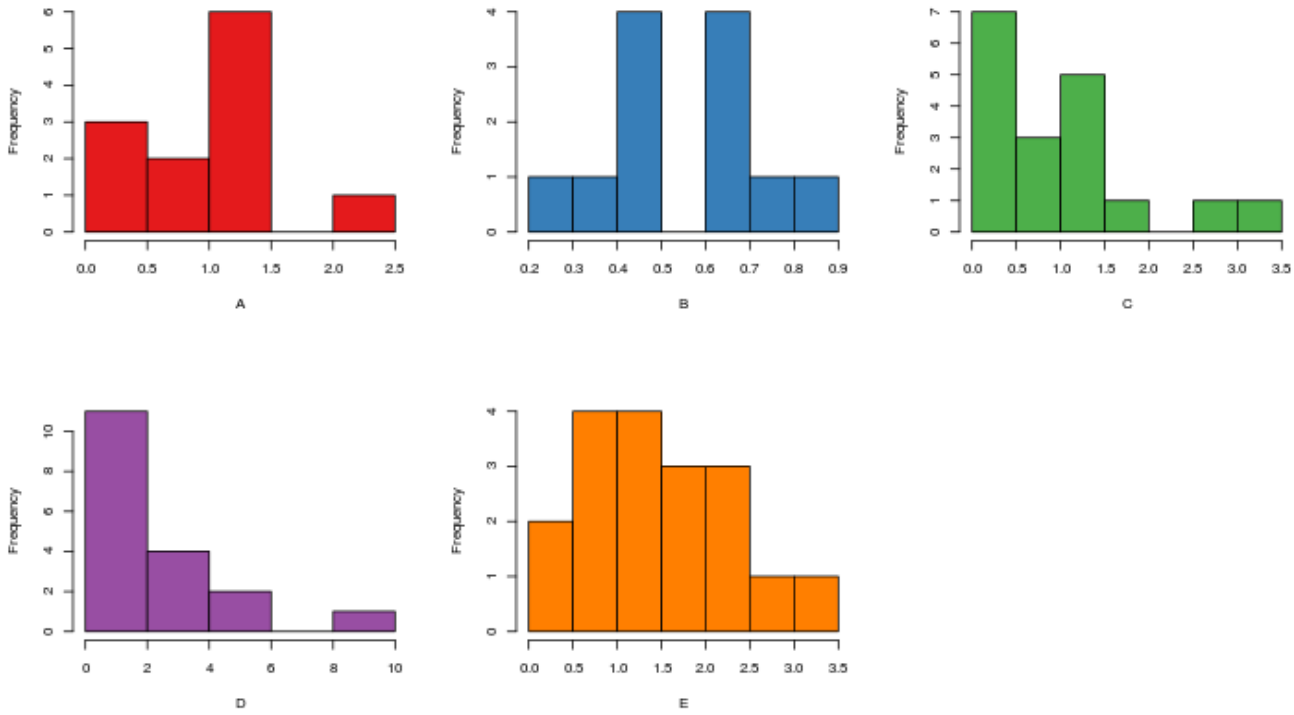


Figure 2: Histogram of the protein expression levels for five cell types

```
## 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="")
```

3.1.5 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

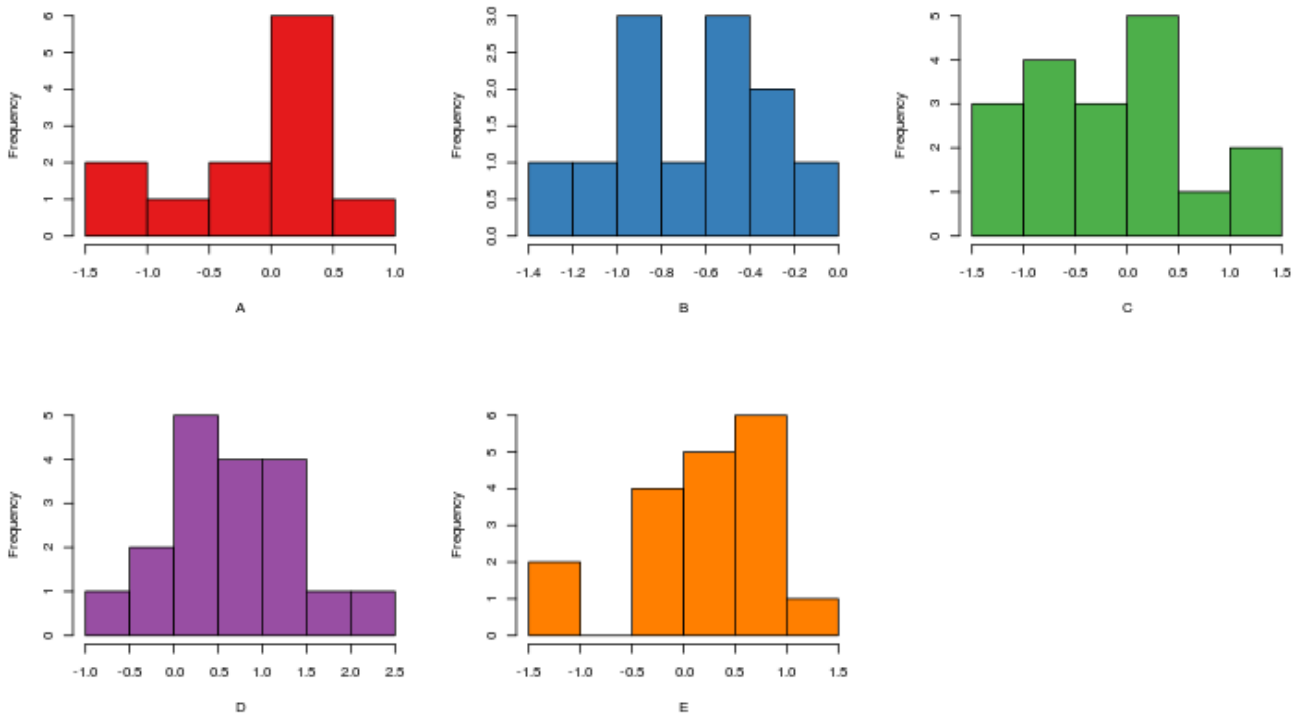


Figure 3: Histogram of the natural-log-transformed protein expression levels for five cell types

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, which can be done using the `gather` function from the package *tidyr*. For comprehensive descriptions of data manipulation and tidy data, see Hadley Wickham's paper [1] or video [2] on the subject.

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
```

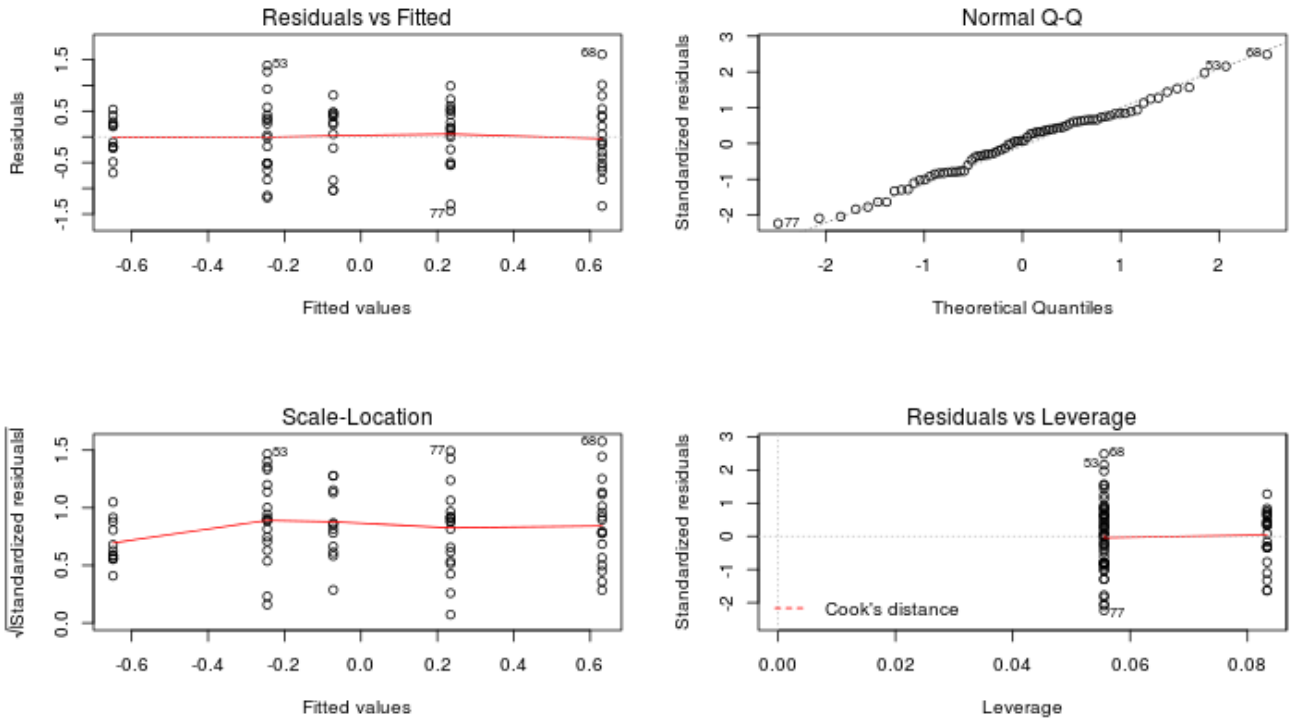


Figure 4: Visualising the results of the ANOVA model

```
## 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)
```

The summary function allows us to assess the significance of the model:-

```
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 $e^{1.28} = 0.562$.

3.2 The Kruskal-Wallis test

This tests if k independent samples are drawn from the same population. As the Kruskal-Wallis test ranks the values, it is more powerful than the Median test. The Kruskal-Wallis test is derived from the one-way ANOVA, but uses ranks rather than actual observations. It is also the extension of the Mann-Whitney U test.

The assumptions of the Kruskal-Wallis test are:

- 1. The data have been collected from a randomly selected set of observations.
- 2. The dependent variable is at least at the ordinal level of measurement.
- 3. There are more than two independent groups.
- 4. There is independence of observations within each group and between the groups. There are no repeated measures or multiple response categories.
- 5. The shapes of the distributions of the groups are similar.

If the last assumption holds then the hypotheses are:

H_0: The medians in the k groups are equal. H_A: There is a difference in medians between the k groups.

If the last assumption does not hold:

H_0: The k groups have the same shape and location H_A: The k groups have a different shape and location.

The alternative hypothesis can be directional or non-directional. If a significant result is obtained then post hoc testing can be used to see where any differences lie. If the assumptions are met the test can be used in the following way:

- 1. Determine the null and alternative hypothesis and α the level of significance for the test.
- 2. Rank the whole sample from lowest to highest.
- 3. Calculate the sum of the ranks for each group.
- 4. Calculate the average rank in each group, R_i, and the average rank for the whole sample,R .
- 5. Calculate the test statistic H,

$$H = 12\frac{\sum n_i(R_i - R)^2}{N(N+1)} \tag{1}$$

  where n_i = the number of observations in group i N = the total sample size
- 6. Compare this value with the χ^2 distribution with $k1$ degrees of freedom. If the statistic is bigger than the critical value in the chi-square table, the result is significant. If the result is significant, then pairwise post hoc test can be carried out.

3.2.1 Example

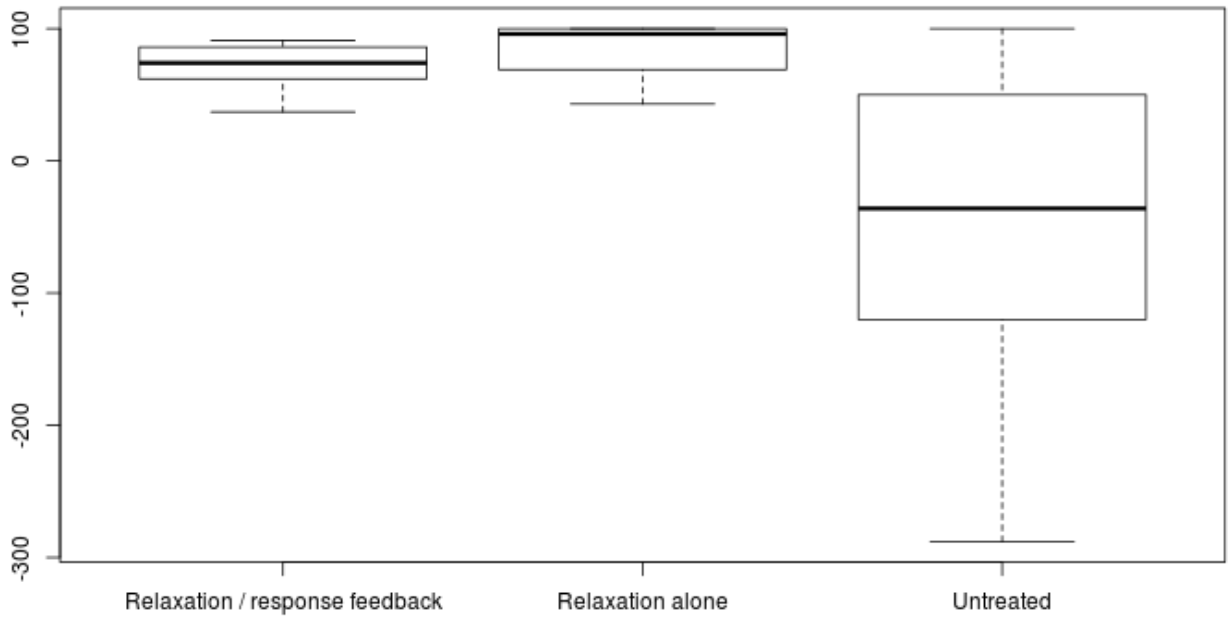
The data are the reduction in weekly headache activity for three treatment groups, expressed as a percentage of the baseline data (example from Altman).

Relaxation / response feedback	Relaxation alone	Untreated
62.00	69.00	50.00
74.00	43.00	-120.00
86.00	100.00	100.00
74.00	94.00	-288.00
91.00	100.00	4.00
37.00	98.00	-76.00

A quick boxplot of the raw data reveal that assumptions for a One-Way ANOVA are not satisfied and we need to take a non-parametric approach.

```
headache <- matrix(c(62,74,86,74,91,37,
                     69,43,100,94,100,98,
                     50,-120,100,-288,4,-76), ncol=3)

colnames(headache) <- c("Relaxation / response feedback","Relaxation alone","Untreated")
boxplot(headache)
```



The null and alternative hypothesis for our test are as follows:-

H_0: The three samples come from populations with the same median. H_A: At least one sample comes from a population with a different median.

Computing the ranks then gives the following table:

	Relaxation / response feedback	Rank	Relaxation alone	Rank	Untreated	Rank
	62.00	8.00	69.00	9.00	50.00	7.00
	74.00	10.50	43.00	6.00	-120.00	2.00
	86.00	12.00	100.00	17.00	100.00	17.00
	74.00	10.50	94.00	14.00	-288.00	1.00
	91.00	13.00	100.00	17.00	4.00	4.00
	37.00	5.00	98.00	15.00	-76.00	3.00
Rank sum		59.00		78.00		34.00
(mean)		9.83		13.00		5.67

We then have all the varaibles that we need in order to compute the test statistic

$R = \frac{N+1}{2} = \frac{18+1}{2} = 9.5$

$R_1 = 9.833;\ R_2 = 13;\ R_3 = 5.667$

$H = \frac{12(6(9.833-9.5)^2+6(13-9.5)^2+6(5.667-9.5)^2)}{18(18+1)} = 5.7254495$

To see how to compute the statistic in R, we first have to re-format into a more convenient format using the `gather` function from the package *tidyr* package. We can then use the `kruskal.test` function,

which applies the kruskal-wallis test.

```
library(tidyr)
headache <- data.frame(headache)
headache <- gather(headache)
kt <- kruskal.test(value~key,data=headache)
kt
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  value by key
## Kruskal-Wallis chi-squared = 5.7254, df = 2, p-value = 0.05711
```

```
names(kt)
```

```
## [1] "statistic" "parameter" "p.value"   "method"    "data.name"
```

```
kt$statistic
```

```
## Kruskal-Wallis chi-squared
##                    5.72545
```

```
kt$p.value
```

```
## [1] 0.05711293
```

3.3 Friedman's test

The Friedman test extends the previously mentioned Wilcoxon signed ranks test to more than two repeated values at more than two time points. Alternatively, to more than two matched groups where the individuals of each group are randomly assigned to a group.

The test examines the ranks at the different time points or matched pairs and tests whether the continuous underlying distribution of the variables is the same. It is the non-parametric equivalent of the repeated measures ANOVA.

3.3.1 Examples from the literature

> The Friedman test is used by Kraemer et al in their paper: Time-of-day variations of indicators of attention: performance, physiologic parameters, and self-assessment of sleepiness (Biol Psychiatry 2000 Dec 1; 48(11):1069-80). The objective of this study was to analyse time-of-day variations of different indicators of attention and their interrelations. Time-of-day variations were tested non-parametrically with Friedman's test for repeated measurements.
>
> Gustafson et al also use the Friedman test in their paper: Effects of 4 hand-drying methods for removing bacteria from washed hands: a randomised trial (Mayo Clin Proc 2000 Jul;

75(7): 705-8). The objective of this study was to evaluate the effects of 4 different drying methods to remove bacteria from washed hands. The Friedman test was used to show that there was no significant difference in the efficiency of the four methods at removing bacteria.

3.3.2 Null Hypothesis

H_0: There is no difference in median between the groups being tested. H_A: There is at least one difference in median between the groups.

This is a non-directional alternative hypothesis. If the alternative hypothesis is to be directional a more powerful way of analysing the data would be to carry out planned comparisons, with the appropriate correction for multiple testing. Alternatively, overall tests of significance followed by post hoc tests can be used.

3.3.3 Assumptions

- 1. The data to be analysed are continuous and at least at the ordinal level of measurement.
- 2. The data from a randomly selected sample are either multiple observations from a single sample across more than two time periods or conditions. Otherwise, the data are blocks of matched subjects in which the subjects from a given block are each randomly assigned to one of the three or more conditions.
- 3. The subjects or blocks of subjects are independent; that is, the results within one block do not have an influence on the results within the other blocks

3.3.4 Method

- 1. Construct the null and alternative hypotheses.
- 2. Construct a two-way table with N (the number of subjects or matched sets of subjects) rows and k (the number of conditions or data collection periods) columns.
- 3. Rank each row from lowest to highest and sum ranks in each column.
- 4. If the null hypothesis is not true then the sum of the columns will vary from column to column. The Friedman test examines the extent to which these column sums vary from what is expected using the following formula:

$$F_r = \frac{12}{Nk(k+1)} \sum_j R_j^2 - 3N(k+1) \tag{2}$$

  where: R_j = the sum of the ranks for column j N = the number of subjects k = the number of periods or conditions.
- 5. Look F_r up in tables of Friedmans distribution.
- 6. Reject the null hypothesis in favour of the alternative hypothesis if the Fr value is greater than (or equal to) the value in the tables.

Note: If N and k are sufficiently large, then F_r can be compared to a χ^2 distribution on $k-1$ degrees of freedom.

3.3.5 Example

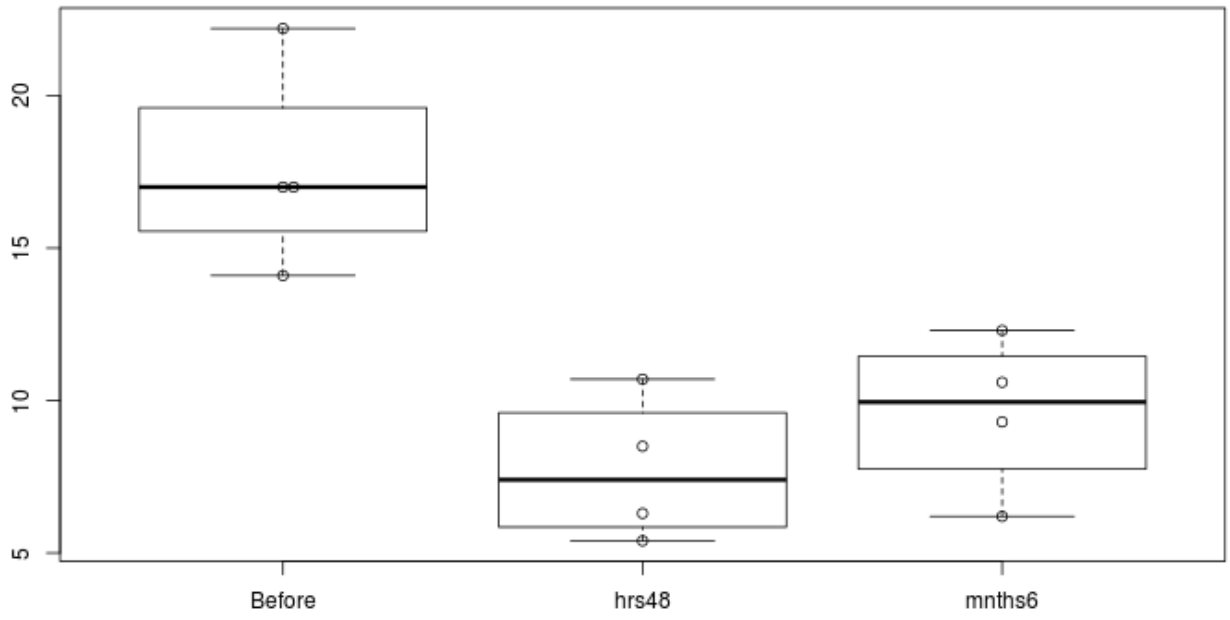
The data for this example is taken from Rubin and Peters paper. The Friedman test will be used to study whether or not hydralazine would relieve high blood pressure in the lungs.

Person	Before	After48Hours	After6Months
1	22.20	5.40	10.60
2	17.00	6.30	6.20
3	14.10	8.50	9.30
4	17.00	10.70	12.30

Table 5: Total pulmonary resistance before and after hydralazine

We create a data frame representation in R, and produce the boxplot:-

```
rubinPeters <- data.frame(Person = 1:4,
                          Before=c(22.2,17,14.1,17),
                          hrs48 = c(5.4,6.3,8.5,10.7),
                          mnths6 = c(10.6,6.2,9.3,12.3))
boxplot(rubinPeters[,2:4])
beeswarm(rubinPeters[,2:4],add=TRUE)
```



$F_r = \frac{12}{4\times3(3+1)}[12^2 + [5^2 + 7^2] - [3 \times 4(3+1)] = 6.5$

As 6.5 is the same as the value in the table there is sufficient evidence to reject the null hypothesis and conclude that at least one group is different from the others.

Person	Before	Rank	After48Hours	Rank	After6Months	Rank
1.00	22.20	3.00	5.40	1.00	10.60	2.00
2.00	17.00	3.00	6.30	2.00	6.20	1.00
3.00	14.10	3.00	8.50	1.00	9.30	2.00
4.00	17.00	3.00	10.70	1.00	12.30	2.00
		12.00		5.00		7.00

```
fr <-  friedman.test(as.matrix(rubinPeters[,-1]))
fr

##
##  Friedman rank sum test
##
## data:  as.matrix(rubinPeters[, -1])
## Friedman chi-squared = 6.5, df = 2, p-value = 0.03877
```

3.3.6 Post-hoc testing

If the Friedman test shows that there is a difference in medians in the groups it is possible to carry out post hoc testing to see which groups there is actually a difference between. This is done by comparing average ranks in all the pairs or comparing to baseline. The null hypothesis that there is no difference in mean ranks between the pairs will be rejected if the absolute value of these differences is greater than a specified critical value. If the following condition holds, the null hypothesis will be rejected:

$$|\bar{R}_i - \bar{R}_j| \geq Z_\alpha/[k(k-1)]\sqrt{\frac{k(k+1)}{6N}} \tag{3}$$

where; $R_i =$ the mean rank in period or condition i $R_j =$ the mean rank in period or condtion j $Z_\alpha =$ the critical z value for α' $\alpha' = \alpha/[k(k-1)]$ $k =$ the number of periods or conditions $N =$ the number of subjects

In the example above, the average ranks for the three time points are 3, 1.25 and 1.75. Since k= 3 and $\alpha = 0.05$ critical value of the z-statistic is a z for which $\alpha' = 0.05$ / 3(2) $= 0.0083$. Looking this value up in the normal tables gives z $= 2.39$. The critical value

$$2.39\sqrt{\frac{3(4)}{6(4)}} = 1.68$$

The absolute values of the three comparisons are:

$$|\bar{R}_1 - \bar{R}_2| = |3 - 1.25| = 1.75 > 1.68$$

$$|\bar{R}_1 - \bar{R}_3| = |3 - 1.75| = 1.25 < 1.68$$

$$|\bar{R}_2 - \bar{R}_3| = |1.25 - 1.75| = 0.5 < 1.68$$

The comparison between before treatment and 48 hours after treatment is the only one that is greater than the critical value of 1.68. Therefore, we can conclude that according to the post hoc approach, hydralazine only relieves high blood pressure in the lungs 48 hours after the treatment. This effect was not maintained 6 months after treatment.

It is also possible to use the Wilcoxon ranked sign tests for post hoc testing. The procedure is carried out in the same way as described before. However, the Bonferroni correction must be applied to allow for multiple testing. That is the critical value of α becomes $\alpha' = \alpha/k$ where k is the number of tests to be carried out and α is the original significance level. The value of α' is the one looked at in the table or that the output p-value is compared against.

3.3.7 Presentation of the Results

The results of the Friedman test could be reported in the following way:

> The results of the Friedman test indicate that there is a significant difference in median total pulmonary resistance across the three time periods. Therefore, we can conclude that hydralazine alters total pulmonary resistance (p=0.042).
>
> Post hoc analyses with the adjustment of the two-tailed level to 0.0083 indicated that there were decreases in total pulmonary resistance from before treatment (Md=17.0) to 48 hours after treatment (Md= 7.4). No other significant differences were found.

Note: That post hoc testing was carried out here on a very small sample size as an illustration, in reality post hoc testing would not be carried out on such a small sample size.

3.3.8 Advantages and Limitations

The Friedman test is very versatile and can be used with randomised block designs and multiple observations of a single sample. It is useful when the dependent variable is skewed.

There are some drawbacks however, it is possible for the medians not to change and there still to be significant differences between groups. Although it is often referred to as the Friedman two-way ANOVA by ranks, it is restricted to within group comparisons. It is not possible to test between group comparisons. This is a major disadvantage in clinical research as it is not possible to make experimental-control group comparisons. Each group can be analysed separately and compare their results. However, it is not possible to test a group and time interaction with independent groups.

3.3.9 Summary

Friedmans test, tests the null hypothesis that k related variables come from the same population. For each case, the k variables are ranked from 1 to k; the test statistic is based on these ranks.

After establishing a difference between one of the variables, post hoc testing can be carried out to decide which of the variables are actually different. An appropriate method for allowing for multiple testing must be carried out.

3.4 Median Test

Tests whether two or more independent samples are drawn from populations with the same median using the χ^2 statistic. It can be used when the assumptions of similarity of distributions for the Mann-Whitney U and Kruskal-Wallis tests are not met.

3.4.1 Null Hypothesis

H_0: There is no difference in medians amongst the groups being studied.

H_A: There is at least one difference in medians amongst the groups being studied.

3.4.2 Assumptions

The assumptions of the Median test are:

- 1. The dependent variable is at least at the ordinal level of measurement. The data are from 2 or more groups
- 2. The groups are independent and a subject can only be in one of the groups.
- 3. The assumptions of the χ^2 test apply to the second half of the test.

3.4.3 Method

If these assumptions are met then the test can be carried out in the following way:

- 1. Construct the null and alternative hypotheses and decide on α the level of significance for the test.
- 2. Treat the data as a single sample and calculate the overall median.
- 3. Separate the data into the various groups and classify the observations in each group as either above, below or equal to the overall median. Calculate the number above and below or equal to the median, in each group.
- 4. Arrange these values into a $2 \times c$ contingency table, where the two rows are: ¿ or to the overall median. The c columns are the groups.
- 5. Calculate the χ^2 statistic for the table, if the assumptions hold.
- 6. Compare the value of the chi-square statistic with the value in the tables on $(c-1)$ degrees of freedom (where c is the number of groups) at the pre-specified level of α
- 7. Reject the null hypothesis of equal medians if X^2 exceeds the critical value of the χ^2 distributions.
- 8. If the null hypothesis is rejected, it is then possible to do post-hoc testing on the individual groups to see which ones are significantly different. This again will be using the median test, but applied to pairs of groups.

3.4.4 Example

There are three groups with different types of dementia (data from Sanjana Nyatsanza, Fulbourn hospital). Below are the patients scores on a mini mental state examination (MMSE). The median test will be used to see if there is a significant difference between the groups.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
Group1	19.00	7.00	17.00	28.00	21.00	6.00	21.00	19.00	27.00	8.00	25.00
Group2	16.00	22.00	30.00	24.00	22.00	23.00	22.00	28.00	29.00	29.00	0.00
Group3	4.00	9.00	30.00	29.00	25.00	22.00	25.00	26.00	27.00	18.00	10.00

1. Stating the null hypothesis and significant level

H_0: There is no difference in medians between the groups.

H_A: There is a difference in medians between the groups.

$\alpha = 0.05$

2. The overall median (i.e. the one in the middle when ranked in order) is 22.

3. Classify the values in each group as above or below 22.

19	7	17	28	21	6	21	19	27	8	25
<=22	<=22	<=22	>22	<=22	<=22	<=22	<=22	>22	<=22	>22
16	22	30	24	22	23	22	28	29	29	0
<=22	<=22	>22	>22	<=22	>22	<=22	>22	>22	>22	<=22
4	9	30	29	25	22	25	26	27	18	10
<=22	<=22	>22	>22	>22	<=22	>22	>22	>22	<=22	<=22

	Group1	Group2	Group3	Total
<=22	8.00	5.00	5.00	18.00
>22	3.00	6.00	6.00	15.00
Total	11.00	11.00	11.00	33.00

5. Find the expected values for each of the cells:

6.00	6.00	6.00
5.00	5.00	5.00

and calculate the χ^2 statistic

$\chi^2 = \sum_{ij} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

$= \frac{(8-6)^2}{6} + \frac{(5-6)^2}{6} + \frac{(5-6)^2}{6} + \frac{(3-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(6-5)^2}{5}$

$= 0.667 + 0.167 + 0.167 + 0.8 + 0.2 + 0.2 = 2.20$

6. The χ^2 statistic for $\alpha = 0.05$ is 5.99

7. $\chi^2 = 2.20 < 5.99$, there is insufficient evidence to reject the null hypothesis that the medians are the same for all three groups, $p = 0.33$

8. As this result is not significant, post-hoc testing could not be carried out as there are no significant differences between the groups.

The Median test has is very straightforward and easy to apply and is particularly useful when the exact values of the scores (especially those at the extremes) are unknown. The test only considers two states for the scores, above and below (or equal to) the median and does not take the size of the differences into account. Therefore, the Median test is less powerful than the Mann-Whitney U and Kruskal-Wallis tests.

3.5 Jonckheere-Terpstra Test

Also known as the test for ordered alternatives or a nonparametric test for trend. The Jonckheere-Terpstra test is used when the assumption that the independent variable is nominal in the Kruskal-Wallis test is violated i.e. the groups have an explicit order. Since it allows the independent variable (the groups) to have an order it is more powerful than the Kruskal-Wallis test when the groups are ordered.

3.5.1 Null Hypothesis

H_0: There is no difference in median values between the groups. H_A: The median values of the groups increase in a specific predetermined sequence.

3.5.2 Assumptions

- 1. The data have been collected from a randomly selected set of observations.
- 2. The data to be analysed are continuous and at least at the ordinal level of measurement.
- 3. The k groups must be ordinal with a predetermined order.
- 4. Under the null hypothesis it is assumed that each sample is from the same population.

3.5.3 Method

- 1. Construct the null and alternative hypotheses and determine the level of significance, α.
- 2. Specify the order of the groups, which need not be equal sized.
- 3. Cast the data into a two-way table with the groups in the pre-specified order, arranged from smallest to largest.
- 4. Within each group order the data from smallest to largest.
- 5. Count the total number of times each value in the first group precedes (is lower than) a value in the subsequent groups this is the precedent count for the group.
- 6. Add to each precedent count when a tie occurs between groups.

- 7. Find the precedent count for the remaining groups and sum over the groups to give J, the test statistic.
- 8. Compare this value to that in the tables for J. If the statistic is greater than or equal to the critical value in the Jonckheere-Terpstra test tables, the result is significant. If the result is significant, the null hypothesis is rejected in favour of the alternative hypothesis.

3.6 Large Sample Size

1. When the sample size is large the distribution of J tends to a normal distribution, with mean:-

$$\mu_j = \frac{N^2 - \sum_{j=1}^{k} n_j^2}{4} \tag{4}$$

and standard deviation

$$\sigma_j = \sqrt{\frac{N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)}{72}} \tag{5}$$

where,

N = total sample size n_j = sample size of group j k = number of groups $\sum$ = sum across groups

A z statistic can then be calculated as follows:

$$z = \frac{J - \mu_j}{\sigma_j} \tag{6}$$

This z statistic can then be compared to the normal tables

3.6.1 Example

Mcm-2 values were collected in a breast cancer study. The median Mcm-2 value was expected to increase with histological grade (data below). This hypothesis was tested using the Jonckheere-Terpstra test.

Grade1	Grade2	Grade3
1.99	4.40	6.94
3.01	9.82	8.04
4.17	10.23	9.82
7.13	11.99	15.75
9.82	11.99	18.30
9.91	13.17	25.01
	13.20	26.40
		28.17

Stating assumptions and significance level

1. H_0: The median Mcm-2 value is the same across histological grades.

H_A: There is an increase is median Mcm-2 value as histological grade increases. $\alpha = 0.05$

2. The order of the groups is that of increasing histological grade.

5. The precedent counts for each pair of groups is in the table below.

6. has been added to each precedent count with a tie between groups.

Grade1And2	Grade1And3	Grade2And3
7	8.00	8.00
7	8.00	5.50
7	8.00	5.00
6	7.00	5.00
5.5	5.50	5.00
5	5.00	5.00
		5.00
Total:37.5	41.50	38.50

7. $J = 37.5 + 41.5 + 38.5 = 117.5$

8. From the tables ($n_1 = 6, n_2 = 7, n_3 = 8, \alpha = 0.05$) the critical value is 99. As $117.5 > 99$, there is sufficient evidence to reject the null hypothesis and conclude that there is a significant increase in median Mcm-2 value as histological grade increases.

3.6.2 Presentation of results

The results of the Jonckheere-Terpstra test could be reported in the following way:

> The results of the Jonckheere-Terpstra test show that there is a trend for an increase in median Mcm-2 value as histological grade increases (J=117.5, p=0.003).

3.6.3 Analysis in R

```
library(clinfun)
grade <- data.frame(Grade1 = c(1.99,3.01,4.17,7.13,9.82,9.91,NA,NA),
                    Grade2 = c(4.40,9.82,10.23,11.99,11.99,13.17,13.20,NA),
                    Grade3 =c(6.94,8.04,9.82,15.75,18.30,25.01,26.40,28.17))
grade <- gather(grade)
grade$key <- as.numeric(gsub("Grade","",grade$key))
jonckheere.test(grade$value,grade$key)

##
##  Jonckheere-Terpstra test
```

```
##
## data:
## JT = 117.5, p-value = 0.004047
## alternative hypothesis: two.sided
```

3.6.4 Advantages and limitations

The main advantage of the Jonckheere-Terpstra test is that unlike the Kruskal-Wallis test it allows the groups to have an order therefore it is more powerful than the Kruskal-Wallis test if the groups have a pre-specified order. Since the Jonckheere-Terpstra test is a test for trend there is no need for post hoc tests to see where differences lie after a significant result.

The main limitation of the test is that the groups must have a pre-specified or explicit order. It is not possible to look for an order and then test for a trend. If there is no explicit order then a Kruskal-Wallis test should be used instead.

3.6.5 Summary

The Jonckheere-Terpstra test is a more powerful alternative to the Kruskal-Wallis test when there is an explicit order to the groups. It is a test for trend of increasing medians between the groups. It is a much under used nonparametric test often the Kruskal-Wallis test is used where it would have been more appropriate to use the Jonckheere-Terpstra test.

3.6.6 Summary of Several Independent Samples

In this section, five tests for use when the independent variable has more than two groups have been covered. The Chi-Square test and the Mantel-Haenszel test are to be used when the dependent variable is categorical. The Median and Kruskal-Wallis tests are to be used when the dependent variable is continuous but there is no order to the groups. The Jonckheere-Terpstra test is more powerful if there is an order to the groups.

If the assumptions of the Kruskal-Wallis test are satisfied, it is more powerful to use this test than the Median Test. That is the Kruskal-Wallis test is more likely to correctly reject the null hypothesis.

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.