Introductory Statistics with R-Studio

Workshop 4 - tests

1 Learning Objectives

- To identify appropriate statistical tests for the question under investigation
- To perform appropriate statistical tests and power calculations.
- To understand how to examine data for normality.

2 Statistical testing

Statistical testing is the process of assessing the probability of a set of results arising by chance. The application of appropriate statistical tests can indicate whether a seemingly important result is actually likely to be that, or whether it is a fluke, occurring by chance.

This workshop builds on the previous three workshops and provides the tools to assess the reliability of the results obtained in the laboratory.

2.1 The caveat

No statistical test exists which will confirm that your inferred conclusions are correct. Statistics merely works with the numbers and calculates the probability of particular numbers occuring by chance. Any conclusions you draw from that information is entirely your own fault.

3 Inputs and outputs

The tests we apply will depend on the question we are asking and how the inputs and outputs are measured. We have already looked at two types of variable - ***continuous variables***, where the data may take any value within a distribution; and ***categorical variables*** or ***factors*** where the data may only take one of a specific (and not necesarily ordered) set of values. A third variable type is a ***rank order***, where the explicit variable measurement is not able to be used in calculations but different measurements can be placed in a rank order.

This workshop will concern itself with the following question types (we consider a *variable* to be continuous and a *category* to be a factor) :

- Is variable A covariant with variable B?
- Does a treatment (factor A) affect the response (variable B)?
- Do my two sets of variable measurements have different means?
- Are these two variable distributions different?
- What is the distribution of my data?

3.1 Is my data normal?

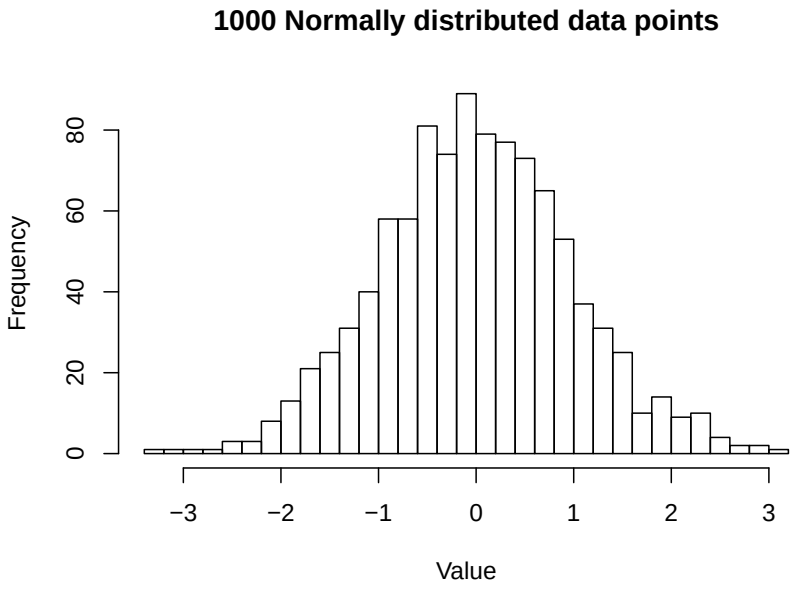
A sample which follows the normal distribution will have a characteristic distribution of values. Most (67%)[1] will be clustered within one standard deviation of the mean. A further 30% will be within 2 standard deviations. Less than 5% will lie outside this range.

Load up the dataset `distributions.txt` as the variable `dists`. Each column in this dataset is composed of 1000 points sampled from a different distribution. Column `Normal` is a normal distribution.

```
> dists <- read.table("distributions.txt", sep="\t", header=T)
```

The simplest first examination is to view the data as a histogram. The `breaks=` parameter specifies the number of bins in which to split the data.

```
> h<-hist(dists$Normal, breaks=25, xlab='Value', main="1000 Normally distributed data points")
> plot(h, xlab='Value', main="1000 Normally distributed data points")
> lines(h$mids, 200*dnorm(h$mids), col="red", lwd=2)
```

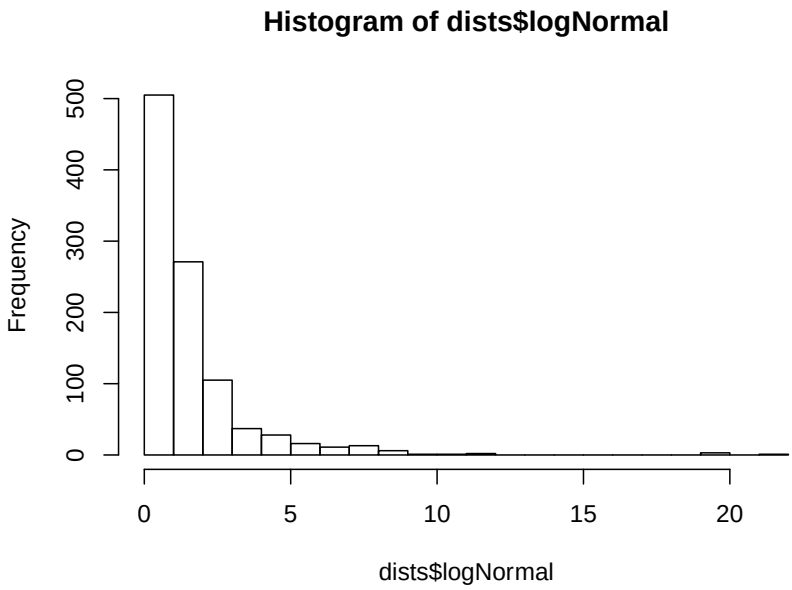


The graph has a suitably scaled curve corresponding to the normal distribution superimposed.

Not all the distributions give such nice data, though they may superficially appear normal.

```
> h2<-hist(dists$logNormal, breaks=25)
> plot(h2, xlab='Value', main="1000 data points from a logarithmic distribution")
> lines(h2$mids, 1000*dnorm(h2$mids, mean(dists$logNormal), sd(dists$logNormal)), col="red", lwd=2)
```

[1] these numbers are approximate



Given a sample of size n we can predict from the mean and standard deviation the values of n perfectly normally distributed points. We can then compare these points with the sample and plot it. This kind of plot where we split the data into n quantiles and compare to the distribution is a quantile-quantile plot, or qqplot.

R has a function for generating a qq plot from arbitrary data.

```
> qqnorm(dists$Normal)
> qqline(dists$Normal, col='red', lwd=2)
```

As you can see the points lie neatly (pretty much) along the line indicating a good fit to the normal distribution.

Some distributions look superficially normal but are subtly different. Students t distribution looks normal but is slightly platykurtic (a bit flatter on the peak and with thicker tails) The precise nature of the t distribution depends on the number of degrees of freedom for which it is calculated, the fewer the degrees of freedome the more extreme the kurtosis (flattening).

Challenge: *obtain the q-q plot and line for the t distribution with 7 degrees of freedom (column t7). What do you notice about the fit of the data to the ends of the line?*

Sometimes the data may fit a normal distribution once transformed. Obtain the q-q plot for the dataset `dists$logNormal`. Next try transforming it using `log()` as `log(dists$logNormal)`.

Sometimes the data may not fit a normal distribution at all. `binom0.2` is a binomial distribution and `pois15` is a Poisson distribution. The last two are discrete value distributions where the quantisation can be clearly seen due to the small number of trials or low number of counts respectively. At high trial numbers or high counts both these distributions tend to normal.

3.2 Are my two means different?

We presume that the data is normally distributed. If this is not the case then a sign/rank test such as Wilcoxon should be used.

Are the variances the same? If so then a Student's t test is in order. If not then use a Kolmogorov-Smirnov test. The t-test is more specific than the KS test so use it in preference *if* your data is appropriate.

The next question to be asked is whether the individual measurements are *independent* or whether they are *paired*. *Paired* measurements are those taken on the same sample, i.e. before and after treatment, or upstream/downstream on the same river. This can be known as a one sample test. *Independent* samples have no explicit relationship with any particular measurement in the other set. These can be known as two sample measurements. If your samples are paired then always do a paired test as the power is much greater.

Student's t-test calculates the t-statistic. This can be simply stated as:

$$t = \frac{\text{difference between the two means}}{\text{s.e. of the difference}} = \frac{\bar{y}_A - \bar{y}_B}{\text{s.e.}_{\text{diff}}}$$

Checking this value against the *Student's t distribution* for the appropriate number of *degrees of freedom* allows us to decide whether to accept the null hypothesis (the two means are not significantly different) or to reject it and accept that the meansare different.

Calculating the standard error of the difference
The standard error of the differnece between two means is calculated as the sum of the standard errors of each of the means. This gives rise to the equation

$$\text{s.e.}_{\text{difference}} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

For a detailed derivation see *Crawley: An Introduction to Statistics using R, p76*

With the t statistic calculated we can determine the significance for our sample size (which determines the total number of degrees of freedom)

What is a degree of freedom? It is a value whose value is not constrained. In our case we have two samples. Each sample has a number of measurements, n. These measurements can vary but in order to maintain the parameter $\bar{y}$, our mean, the last measurement is constrained. Thus the number of degrees of freedom is the number of measurements, minus the number of estimated parameters. In the case of the t-test this is the number of measurements in both samples - 2 (for the two means).

Performing the t-test We could perform this longhand, calculating the mean of each dataset, calculating the variances, the standard errors and so on. R has all the necessary functions. This can take a while and be prone to error so there are simple functions that perform all the mechanics for us.

Start by importing the dataset `glucose1.txt` and attaching it.

```
> gluc <- read.table("glucose1.txt", sep="\t", header=T)
> attach(gluc)
```

This imports two columns `KO` and `WT`. We'll compare them as unpaired samples to see whether the KO is significantly different to the WT. First check that the data is normal (with `qqnorm()`) and the variances are about the same (by comparing `sd()`. We will then calculate a two-sided t-test to see whether the null hypothesis holds.

```
> t.test(WT, KO)


        Welch Two Sample t-test

data:  WT and KO
t = 0.9665, df = 26.099, p-value = 0.3427
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.808643 13.347214
sample estimates:
```

```
mean of x mean of y
 185.0239  180.7546
```

This is not significant - we would accept the null hypothesis that our experiment has not demonstrated the samples to be significantly different.

3.2.1 How many samples do I need?

R provides a mechanism for calculating how many samples we would need to establish a significant difference between our two test sets. We could do this by hand, adjusting the t-statistic in accordance to a theoretical number of samples and record each significance till we find p<0.05 (or whatever cutoff we are prepared to accept), or we can let R do the donkey work.

If we take our experiment and assume the delta (`mean(WT)-mean(KO)`) and variance to be a realistic estimate of the population, then we can use the power calculation to determine the number of samples we need.

```
> power.t.test(power=0.8, sd=sd(WT), delta=mean(WT)-mean(KO))


     Two-sample t test power calculation

              n = 161.006
          delta = 4.269286
             sd = 13.63172
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

That is a *minimum* number.

What is statistical power? Typically we look at the probability α that we have got a difference by chance (a type I error, or false positive). Statistical power looks at the probability of a type II, or false negative, error. In other words the probability β that we accept the null hypothesis when it does not hold. Typically power is calculated as $1-\beta$ and a reasonable cutoff is 0.8, or 80%. *'Given an experiment with that delta and those variances, how many samples do I need to give me at least an 80% chance of not falsely accepting the null hypothesis'*

3.2.2 The paired t-test

Fortunately in our experiment our samples are paired - we carry out a process between measuring WT and KO *on the same sample* and we are measuring the effect of that process. We can therefore use a paired test.

Paired or unpaired? With the unpaired test we are comparing every sample in group A to every sample in group B. If there is a substantial intrinsic variance in the groups then any difference may well be masked. To measure the paired differnece we can measure the standard error of mean of the differences, rather than the standard error of the difference of the means. No difference would give a mean of 0. A significant difference would give a mean away from 0 and a 95% CI that does not overlap 0.

The R code for a paired test is very similar to that we have seen before. We just add the parameter `paired=T` to the command.

```
> t.test(KO,WT, paired=T)
```

```
        Paired t-test

data:  KO and WT
t = -3.8009, df = 14, p-value = 0.001948
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.678405 -1.860166
sample estimates:
mean of the differences
              -4.269286
```

Aha! A significant result. How sure were we (what was the statistical power) that we would be able to detect a significant result with data like this?

If we set the `n=` parameter in the power calculation and leave the power unset, we can calculate this. Note that SD is now the standard deviation of the differences between paired measurements, not the SD of the measurements themselves.

```
> power.t.test(n=15, sd=sd(WT-KO), delta=mean(WT)-mean(KO), type="paired")


     Paired t test power calculation

              n = 15
          delta = 4.269286
             sd = 4.350306
      sig.level = 0.05
          power = 0.9417327
    alternative = two.sided

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

15 samples was plenty to guarantee that we would see a result of that magnitude.

Take home message Use a paired experimental design wherever possible. The significance of the results is much greater.

3.3 When the t-test is not suitable

The t-test is only useful under conditions where the variances are similar. For the rest of the time you will need to try an alternative method.

3.3.1 Continuous data

If we have continuous data that does not have the right properties for a t-test then we have to fall back to the Kolmogorov-Smirnov test. This test uses the cumulative distribution function to determine whether two distributions are significantly different.

The KS test can be run in a simple manner, just like the t-test, but in this case the relevant function is `ks.test`. This is purely a two sample test. If you have a paired sample you can (and should) use a paired t-test.

If the KS test indicates a positive result (ie a rejection of the null hypothesis) it can be interesting to explore why that is. One test is that of variance - are the variances significantly different? This can be answered with the `var.test()`.

Example: If we take the exam results for any particular year, is there a difference between English and Scottish students?[2]

Load the file `students.txt` as the variable `students`

```
> students<-read.table("students.txt", sep="\t", header=T)
> attach(students)
> tapply(grade, nation, mean)


 England Scotland
56.93333 60.36667


> tapply(grade, nation,sd)


  England  Scotland
 7.075618 16.447373


> ks.test(grade[nation=="England"], grade[nation=="Scotland"])


        Two-sample Kolmogorov-Smirnov test

data:  grade[nation == "England"] and grade[nation == "Scotland"]
D = 0.3667, p-value = 0.03543
alternative hypothesis: two-sided


> var.test(grade[nation=="England"], grade[nation=="Scotland"])


        F test to compare two variances

data:  grade[nation == "England"] and grade[nation == "Scotland"]
F = 0.1851, num df = 29, denom df = 29, p-value = 1.918e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.08808672 0.38883094
sample estimates:
ratio of variances
         0.1850698
```

We can see from this combination of tests that the means are not different but the variances are significantly greater for the scottish students.

3.4 Testing multiple treatments

Lets start off with a single treatment. Import the dataset `gardens.txt` to the variable `gardens` and attach it. There should be two columns, *garden* which is a factor describing which garden; and ozone, which is a measurement of ozone concentration.

[2] Yes these are invented data to illustrate the point..

```
> gardens<-read.table("gardens.txt", sep="\t", header=T)
> attach(gardens)
```

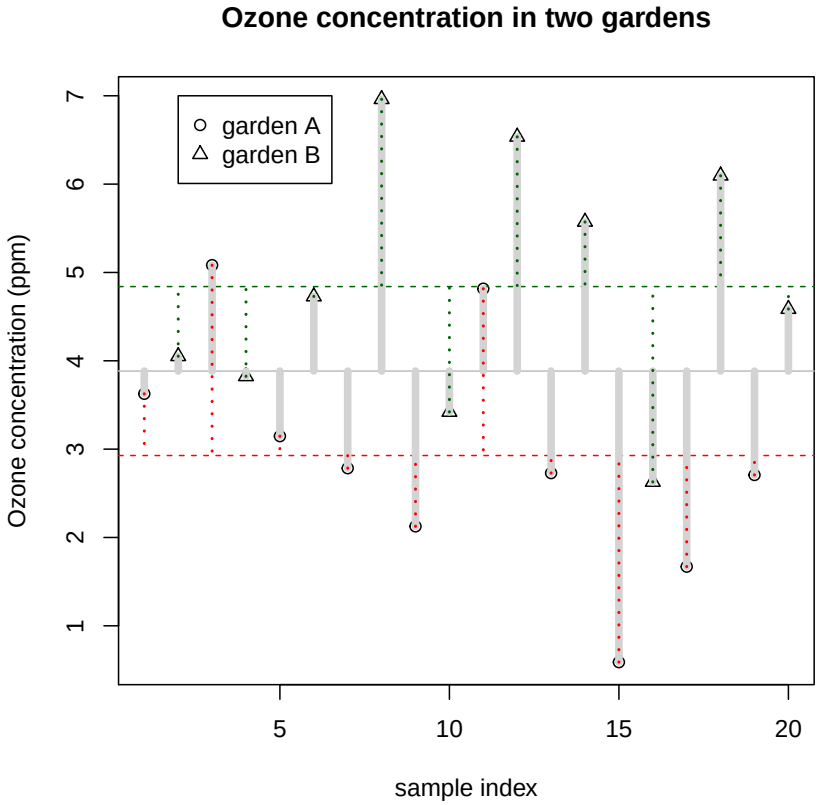
We can compare the two gardens with a t-test.

```
> t.test(ozone[garden=="A"], ozone[garden=="B"])


        Welch Two Sample t-test

data:  ozone[garden == "A"] and ozone[garden == "B"]
t = -3.0811, df = 17.963, p-value = 0.006452
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.2166919 -0.6082075
sample estimates:
mean of x mean of y
 2.927706  4.840156
```

This gives a non significant result, but is that all that can be said? What proportion of the observed variation is due to the garden, and what is otherwise unexplained?

Analysis of Variance (ANOVA) is the method of choice for this analysis. If the dataset for each garden is fitted to the mean for that garden, then the residual sum of squares can be calculated as the sum of the squared errors. Subtracting this from the residuals calculated when fitted to the whole dataset gives the sum of squares due to the garden.

We can see this represented in the following graph.



We can see the mean of the whole dataset (grey line) and the respective residuals. The mean for each garden is shown along with the residuals for that garden in the appropriate (dashed) colour. The grey outline is the total residual for the dataset (from the dataset mean).

This process is simplified in the `aov()` command.

```
> aov(ozone~garden)


Call:
   aov(formula = ozone ~ garden)

Terms:
                  garden Residuals
Sum of Squares  18.28732  34.67460
Deg. of Freedom        1        18

Residual standard error: 1.387936
Estimated effects may be unbalanced
```

That gives us the aw data. We can see that the total variance is just under 53 so the gardens explain about 1/3 of the variance in our data. Is this significant?

The ratio of the garden mean sum of squares to the residual mean sum of squares gives the F-statistic. This can then be assesed to see if the garden parameter has a signficant contribution to the measured ozone value.

Once again, R has this packaged with the `summary()` function

```
> summary(aov(ozone~garden))


            Df Sum Sq Mean Sq F value  Pr(>F)
garden       1  18.29  18.287   9.493 0.00644 **
Residuals   18  34.67   1.926
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

The garden selection makes a significant contribution to the measurement of ozone. If we treat the analysis as a linear model (as we did for the regression fit in workshop 3) then we can calculate an effect size through the R^2 value. Again, R provides a function for this.

```
> summary.lm(aov(ozone~garden))


Call:
aov(formula = ozone ~ garden)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3390 -0.8555 -0.1713  0.8630  2.1550

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9277     0.4389   6.670 2.94e-06 ***
gardenB       1.9124     0.6207   3.081  0.00644 **
---
```

```
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 1.388 on 18 degrees of freedom
Multiple R-squared:  0.3453,        Adjusted R-squared:  0.3089
F-statistic: 9.493 on 1 and 18 DF,  p-value: 0.00644
```

The R^2 value can be clearly seen and shows that the garden selection contributes about 30% to the variance.

The ANOVA summary produced with `summary.lm()` may be a little confusing. What is this *Intercept*? Why is the estimate for garden B nothing like `mean(ozone[garden=="B"])`?
The Intercept is the estimate for the first category (garden A in this case). The value shown for garden B is the difference between garden A and garden B, ie 4.8402 - 2.9277 If there were more categories then these would also be listed with the difference between them and the Intercept.

`plot(aov(ozone garden))` gives several informative plots.

Remember from workshop 3 that ***significance*** refers to the reliability of the result. A result can be highly significant but of very low magnitude. In this case the *explanatory power* is the ratio of the explained variance (from the gardens) to the total variance in the ozone measurement

3.4.1 More than one condition

Often we have more than two categories for a variable, or more than one variable (with multiple categories in each). These are known as factorial experiments. Applying the same sort of analysis of variance works well in these cases.

Load the file `growth2.txt` as the variable `weights`[3]

```
> weights <-read.table("growth2.txt", sep="\t", header=T)
> attach(weights)
```

This dataset describes the response (weight gain) of farm animals to three grammene crop diets with four additional supplements.

Let's look at the aggregate data by diet and supplement. We can use the `tapply()` function to apply a function to a set of data grouped according to a factor or group of factors.

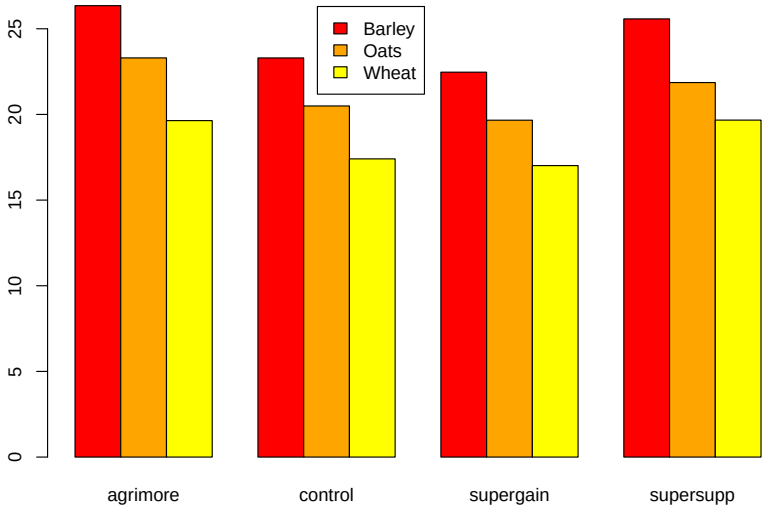
```
> tapply(gain, list(diet,supplement),mean)
```

```
       agrimore  control supergain supersupp
barley 26.34848 23.29665  22.46612  25.57530
oats   23.29838 20.49366  19.66300  21.86023
wheat  19.63907 17.40552  17.01243  19.66834
```

Plotting this table gives:

```
> barplot(tapply(gain, list(diet,supplement),mean), beside=T, col=c("red", "orange", "yellow"))
> legend(6.3,26.3,c("Barley","Oats","Wheat"), fill=c("red", "orange", "yellow"))
```

[3]This example is taken from the Crawley book. The raw data is available at http://www.bio.ic.ac.uk/research/crawley/statistics/data.htm



We can carry out the ANOVA in much the same way as before, using `aov()` and `summary.lm()`. However we have to be careful how we specify the model.

```
> summary.lm(aov(gain~diet+supplement))


Call:
aov(formula = gain ~ diet + supplement)

Residuals:
     Min       1Q   Median       3Q      Max
-2.30792 -0.85929 -0.07713  0.92052  2.90615

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         26.1230     0.4408  59.258  < 2e-16 ***
dietoats            -3.0928     0.4408  -7.016 1.38e-08 ***
dietwheat           -5.9903     0.4408 -13.589  < 2e-16 ***
supplementcontrol   -2.6967     0.5090  -5.298 4.03e-06 ***
supplementsupergain -3.3815     0.5090  -6.643 4.72e-08 ***
supplementsupersupp -0.7274     0.5090  -1.429     0.16
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 1.247 on 42 degrees of freedom
Multiple R-squared:  0.8531,        Adjusted R-squared:  0.8356
F-statistic: 48.76 on 5 and 42 DF,  p-value: < 2.2e-16
```

This gives us a detailed breakdown of the *independent* effects of diet and supplement. Each parameter is estimated and cahracterised independently. But what if there is some interaction between the two sets of factors?

If we change the model from `gain diet+supplement` to `gain diet*supplement` we can evaluate the inter-dependence of the factors.

```
> summary.lm(aov(gain~diet*supplement))
```

```
Call:
aov(formula = gain ~ diet * supplement)

Residuals:
     Min       1Q   Median       3Q      Max
-2.48756 -1.00368 -0.07452  1.03496  2.68069

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   26.3485     0.6556  40.191  < 2e-16 ***
dietoats                      -3.0501     0.9271  -3.290 0.002248 **
dietwheat                     -6.7094     0.9271  -7.237 1.61e-08 ***
supplementcontrol             -3.0518     0.9271  -3.292 0.002237 **
supplementsupergain           -3.8824     0.9271  -4.187 0.000174 ***
supplementsupersupp           -0.7732     0.9271  -0.834 0.409816
dietoats:supplementcontrol     0.2471     1.3112   0.188 0.851571
dietwheat:supplementcontrol    0.8183     1.3112   0.624 0.536512
dietoats:supplementsupergain   0.2470     1.3112   0.188 0.851652
dietwheat:supplementsupergain  1.2557     1.3112   0.958 0.344601
dietoats:supplementsupersupp  -0.6650     1.3112  -0.507 0.615135
dietwheat:supplementsupersupp  0.8024     1.3112   0.612 0.544381
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 1.311 on 36 degrees of freedom
Multiple R-squared:  0.8607,        Adjusted R-squared:  0.8182
F-statistic: 20.22 on 11 and 36 DF,  p-value: 3.295e-12
```

We see from the table that only five parameters are significantly different from the intercept. The interaction terms are not significant, indicating no interaction between diet and supplement.

In addition, some factors are not significantly different to one another (less than 2 s.e. apart) so could be grouped in a final model. The factor in question is the supplement, and both agrimore and the control can be grouped, as can supersupp and supergain.

What is an interaction? An interaction occurs when the effect seen for a particular combination of factors is significantly different to the sum of the effects seen for the two factors independently.

3.5 Categorical outcomes

A categorical outcome is one where the response is classified into discrete categories.

3.5.1 Non-calculable values

A non-calculable value is one where performing arithmetic (e.g. the calculating of means) makes no sense, either because the number itelf is a convenient label for an ordering and has no quantitative merit beyond that, or the model distribution is not suitable for such calculations. In these cases a rank test is in order.

What is an incalculable number? '*Please indicate your satisfaction with this explaination on a scale of 1-10 where 1 is extremely dissatisfied and 10 is completely satisfied*' Is a value of 3 really half that of a value of 6 for this question? The finishing order of athletes in a race is another uncalculable number. Is finishing 3rd really 1.5 times finishing second? Does average finishing position actually mean anything? These are numbers thatwe shoudl not do calculations on but which the relative ordering canbe used to infer results.

Every year the Hawkhill Harriers, a local running club, take on the Dundee Wheelers, a cycling club, in a cross country race. There is always some debate over whether the course that year favours runners or cyclists

and any attempt to determine this from the top placings is subject to influence by one or two exceptional athletes. A signed rank test is the statisticians way to resolve this age old dispute.

The finish results are listed in `race.txt`. We need to split them into two groups and then perform the Wilcoxon test on them.

```
> race <-read.table("race.txt", sep="\t", header=T)
> cyclists<-race$finish[race$Club=="Wheelers"]
> runners<-race$finish[race$Club=="Harriers"]
> wilcox.test(cyclists, runners)


        Wilcoxon rank sum test

data:  cyclists and runners
W = 49, p-value = 0.007544
alternative hypothesis: true location shift is not equal to 0
```

The slight significance in the data lends itself to meany interpretations. One might be that cyclists are generally better than runners. An alternative hypothesis may be that the course was slightly more favourable this year for the cyclists.

3.5.2 Statistics without numbers

In some cases we have no numerical data, just a category choice. This can be modelled with a binomial distribution. Is the proportion of A to B something we would expect by chance?

A coin is flipped 6 times. What is the probability of at least 5 heads? we use the `binom.test()` method to test this. This takes two parameters, the number of 'successes' and the number of tests. An optional parameter specifies the likelihood of any one test failing (by default `p=0.5`)

In our case we have 1 failure (tails) out of 6 attempts. This gives 5 successes.

```
> binom.test(5, 6)


        Exact binomial test

data:  5 and 6
number of successes = 5, number of trials = 6, p-value = 0.2188
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3587654 0.9957893
sample estimates:
probability of success
             0.8333333
```

As can be seen, the tested hypothesis is that the probability of heads is not equal to 50%. In this case the occurence of 5 out of 6 heads is not significant. ***Challenge:*** *Would five heads in a row be significantly suspicious?*[4]

Instead of coins, let's turn to dice. A dice is rolled 6 times and 5 of those times a 6 is rolled. Is this suspicious?

In this case we need to set the probability of success. For a standard 6-sided die[5] this is 1 in 6, or 0.1667.

[4] For the socially accepted significance level of 0.05
[5] You did check the die has the normal numbers, didn't you?

```
> binom.test(5,6, p=1/6)


        Exact binomial test

data:  5 and 6
number of successes = 5, number of trials = 6, p-value = 0.0006644
alternative hypothesis: true probability of success is not equal to 0.1666667
95 percent confidence interval:
 0.3587654 0.9957893
sample estimates:
probability of success
             0.8333333
```

I'd definitely want a close look at that die.

3.5.3 Testing proportional data

Similarly, if we have two proportions, we can test to see whether they are significantly different. The `prop.test()` will take two distributions as two lists (`c(trials, successes)`) and assess whether the chance of success is significantly different.

Let us assume that we have a round of promotions and that 4 ladies and 196 men have been appointed. This might seem like a gender imbalance, but there were 3720 men applying for promotion and only 40 ladies. Is this a statistically different success rate between ladies (10%) and men(6%)?

`prop.test()` takes two arguments. The first is a list of successes, the second a list of trials.

```
> prop.test(c(4,196), c(40,3720))


        2-sample test for equality of proportions with continuity correction

data:  c(4, 196) out of c(40, 3720)
X-squared = 0.9449, df = 1, p-value = 0.331
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.05856862  0.15319228
sample estimates:
    prop 1     prop 2
0.10000000 0.05268817
```

Under the hood this is performing a Chi-squared test which is a categorical test.

3.5.4 The Chi-squared test (for when you have large counts)

The Chi-squared test examines distortion in contingency tables. By constructing the table we can establish the expected proportions in each cell from the row and column sums, and then determine how far our results deviate from that. The likelihood of this happening by chance can then be established to give the statistical significance of the result.

What is a contingency table? Where we have two (or potentially more) factors with two or more categories then we can construct a table where each axis represents the categories from one factor. Each box in the table is then a count of the number of data points with that combination of factor values.

The χ^2 value is the sum of the squared errors divided by the expected values. We can then test this value to see if it is significantly larger than expected.

	Blue eyes	Brown Eyes	Row total
Fair Hair	38	11	49
Dark Hair	14	51	65
Column totals	52	62	114

The row totals give the expected proportion for fair and dark hair (49/114 and 65/114 respectively) and the column totals the expected proportion for blue and brown eyes respectively. From this we can determine that the expected count using simple probabilities for fair hair and blue eyes is 49/114 * 52/114 *114, or 22.35. Completing the table can now be done by simple arithmetic so that the column and row totals are correct.

We can now calculate the χ^2 statistic

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

which evaluates to 35.33, and this value can be tested using the `pchisq()` function. We first need to establish the number of degrees of freedom. This is calculated as the product of the degrees of freedom in each factor. The degrees of freedom for a factor is the number of levels - 1.

For a 2×2 table the number of degrees of freedom is $(2-1) \times (2-1)$ which is 1.

```
> pchisq(35.33, 1)
```

```
[1] 1
```

This is a highly significant result. Hair colour and eye colour are strongly dependent.

R, as ever, has one function that will do all of this in one go. First we will construct our contingency table as a matix, then we will call the `chisq.test()` function with the matrix.

To construct the matrix we give R a list of values and the number of rows in the matrix. R fills the matrix up by column from top to bottom

```
> count <- matrix(c(28,14,11,51), nrow=2)
> chisq.test(count)
```

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  count
X-squared = 23.5265, df = 1, p-value = 1.232e-06
```

The chi-squared test is a reasonable model but not a perfect description of the distribution. There are some corrections made, in this case the Yates correction, that adjust the calculation from that described above. This is especially important for small counts (<30 in a cell)

3.5.5 Fisher's exact test (For when you have small counts)

What is a small count? The Chi-squared test breaks down completely if any of the table cells has an expected count of less than 5. The Fisher's exact test is far better in these cases.

Why don't we just use the Fisher's test? The exact test relies on the calculation of factorial numbers which can be very time consuming, especially as the numbers get bigger. We therefore restrict use of the Fisher's test to smaller numbers

R has a function `fisher.test()` that can take a matrix of counts as prepared previously. However it also has the ability to take two vectors (lists) of factors (which should refer to the same datapoints in the same order) and calculate the contingency tables for you.

In any year 10 essay titles are set from which students have a free choice. This means that some essay titles are selected in some years and not in others. Is there a significant difference betweent eh two years in the table below?

Year	2010	2011
titles chosen	6	2
not chosen	4	8

```
> x<-matrix(c(6,4,2,8), nrow=2)
> fisher.test(x)


        Fisher's Exact Test for Count Data

data:  x
p-value = 0.1698
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.6026805 79.8309210
sample estimates:
odds ratio
  5.430473
```