Stats with Sparrows - 17

Julia Schroeder and David Orme

17

This practical will look at some non-parametric statistics.

Housekeeping!

```
rm(list=ls())
```

Chi-square test

The χ^2 test tests for differences in a dataset of nominal data. For instance, you can test whether hair color and eye color are correlated with each other.

The steps to calculate the χ^2 value are: • Work out the column, row and table totals using the rowSums(), colSums() and sum() functions. • From these work out the expected counts for each class under the null hypothesis that there is no association between categories. We can use the outer() function to multiply each element in a vector with each element in another vector, rather than the usual multiplication of pairs of values. • For each cell, calculate, using the expected values (if there was no correlation) (O − E)2/E. • Sum the cells to get the total χ^2 for the test. Nicely enough, R provides us with a dataset for this:

```
hairEyes <- matrix(c(34, 59, 3, 10, 42, 47), ncol = 2, dimnames = list(Hair =
c("Black",
    "Brown", "Blond"), Eyes = c("Brown", "Blue")))
hairEyes
```

```
##        Eyes
## Hair    Brown Blue
##   Black    34   10
##   Brown    59   42
##   Blond     3   47
```

```
rowTot <- rowSums(hairEyes)
colTot <- colSums(hairEyes)
tabTot <- sum(hairEyes)
Expected <- outer(rowTot, colTot)/tabTot
Expected
```

```
##          Brown     Blue
## Black 21.66154 22.33846
## Brown 49.72308 51.27692
## Blond 24.61538 25.38462
```

Then you can calculate χ^2:

```
cellChi <- (hairEyes - Expected)^2/Expected
tabChi <- sum(cellChi)
tabChi

## [1] 54.63907
```

Now we have our test statistic but need to calculate the probability of getting that value (or one more extreme) under the null hypothesis. As with the pt() function, we need to provide our test statistic and the degrees of freedom. This is (number of rows-1)x(numberofcolumns-1)=(3-1)x(2-1)=2x1=2. Unlike the t test,wedon't have to worry about tails: the null hypothesis is that the two factors are independent; we don't have to think about a direction of independence. The p value we get from our χ2 value is a very small number — clearly, hair colour and eye colour are not independent:

```
1- pchisq(tabChi, df = 2)

## [1] 1.365463e-12
```

Once again, you won't be surprise to hear, all this is built-in to a function for us. This time we are going to save the results and look at the structure. The test calculates all sorts of things that it doesn't print out, including all the matrices we just calculated. One of the items is a matrix showing residuals — an indication of the extent of the departure from the expected values for each cell. Clearly, blonds are very unlikely to have brown eyes and people with black hair are fairly unlikely to have blue eyes; it doesn't seem to matter as much for brunettes.

```
hairChi <- chisq.test(hairEyes)
print(hairChi)

##
##  Pearson's Chi-squared test
##
## data:  hairEyes
## X-squared = 54.639, df = 2, p-value = 1.365e-12
```

χ^2: some warnings The χ^2 test is hugely useful for looking to see if categorical data are independent but it has some problems which you should be aware of if you are planning to use it.

• The test is less reliable if any of the combinations of factors are infrequent (the rule of thumb is 5 or fewer occurrences). Make sure that you collect enough data to avoid this and, if some combinations are genuinely rare, you may find that you need to lump categories together. • The calculation for 2x2 tables has also traditionally been corrected using Yates' correction in the cell calculations (|O − E| − 0.5)2/E. This correction tends to make the test less powerful. The chisq.test() function detects these conditions and corrects and warns about them. If you have a 2x2 table and counts smaller than 5, you should look up Fisher's exact test (?fisher.test).