Chi-Square Test for Goodness of Fit

By: Erin Norton & Maheen Kibria

About the Test

At times, it is necessary to compare an experimental data set to that of expected values or a predicted ratio under the circumstances given by the experiment or trial. The chi-square test for goodness of fit is useful in this circumstance, because it incorporates observed and expected values for data and the R function for this test provides an p-value that we can use to either reject or retain the null hypothesis. In other words, by using this test we are able to conclude whether the data is statistically consistent with the expected value or ratio, or if it has a statistically significant difference from the expected value or ratio.

Organizing the Data

The best way to organize the experimental (observed) values and the expected values is in the following way, using a program like Excel:

Class	Observed	Expected
A		
B		
C		
D		
E		

The names of the classes can be manipulated to better fit the data, and the chart will be filled in with the values from the experimental data set. The last cells of each column may also be utilized to tabulate the total number of observations in the observed and expected groups (which would be the same number). However, if using a statistical program such as R, it is unnecessary because R will do these calculations for you.

Calculating Expected Values According to an Expected Ratio

A key example of a chi square test using an expected ratio is in Mendelian genetics. The expected ratio of genotypes for a cross between two heterozygous individuals, is 9:3:3:1. Now we must extrapolate this ratio to a larger number of subjects in order to arrive at expected values for the conditions given.

Example

If you had a data pool of 400 subjects and you wanted to determine how many individuals would fall into each group of the ratio, you would first determine the ratio that each group represents (number of individuals of the group divided by the total number of individuals), as in the following calculations:

$$9/(9+3+3+1) = 9/16$$
$$3/(9+3+3+1) = 3/16$$
$$3/(9+3+3+1) = 3/16$$
$$1/(9+3+3+1) = 1/16$$

You would then multiply each group's ratio by the number of subjects in the set to which you are applying the ratio. In this example, this number is 400. See the following calculations:

$$400 \times (9/16) = 225$$
$$400 \times (3/16) = 75$$
$$400 \times (3/16) = 75$$
$$400 \times (1/16) = 25$$

Notice that these four expected values (225 + 75 + 75 + 25) add up to 400. All of our possible groups are accounted for, and the subjects of the data pool are partitioned into expected values based on the ratio given. These expected values would then be placed in the cells of the table corresponding to the expected values.

Transferring the data to R

Copy and paste the Excel data into a text editing program such as Notepad and save it as a text (.txt) file, and then use the following command in R:

```
> read.table(file="filename.txt")
```

This commands R to read the table, and then we are able to attach it in order to perform the chi-square for goodness of fit test. We can assign this dataset a name, which makes it easier to view in R without typing in a long command every time. To do so you would type something like the following for a dataset titled "reservoir:"

```
> res=read.table(file="reservoir.txt")
```

Test Logistics

Assumptions
The test assumes that observed values are random

Null Hypothesis
H_0: P_J probabilities are distributed according to the model
This is to say that the expected and observed values are the same, because they follow the expected pattern in the model or set of expected values.

Alternative Hypothesis

H_1: P_J probabilities differ from the model
This is to say that the expected and observed values are statistically different, as they do not exhibit contingency with the expected pattern or set of expected values.

The Chi-Square Test Statistic
The chi-square value itself is the following calculation:

$$\chi sq = \sum \frac{(observed - expected)^2}{expected}$$

This calculation is interpreted as the difference between observed and expected value, squared, and then divided by the expected value; each of these calculations (one for each individual data entity, or pair of observed/expected values) summed together. The resulting chi-square value is the same as the one R will produce for the data set, and knowing the origin of the value can help us to be sure the right chi-square value is produced.

How to run the test in R

After loading the dataset onto R, make sure that the headings of the tables are "observed" and "expected." If they aren't, you can manipulate this by entering the following in the table-reading command:

```
> read.table(file="filename.txt", header=TRUE)
```

Simply enter this command in order to tell the program that the titles "observed" and "expected" need to move above set and out of the first row of data, where they tend to end up. After attaching the data, you can then use the following command to run a chi-square test:

```
> chisq.test(observed,p=expected,rescale.p=TRUE)
```

The program will make the calculation and display it…which leads us to the next section.

How do I interpret these results?

R displays an X-squared value, the degrees of freedom, and a p-value. The p-value is of most value to us in terms of making a decision. Assuming you have set an alpha value before completing the test, you can then compare the p-value to the alpha value. The p-value represents the probability that the apparent difference between the actual data and the expected values is due to chance alone. Therefore, if this value is lower than a certain threshold—either 0.05 or 0.01—we can reject the null hypothesis because this probability is so low.

Rejecting the null hypothesis

In rejecting the null hypothesis (p-value is less than alpha), we are accepting the alternative hypothesis, and that is to say that the expected and observed values have a statistically significant difference. The data does not follow the pattern of the expected values or the predicted ratio.

Failing to reject the null hypothesis

If our p-value for the chi-square analysis is greater than the specified alpha value, we cannot reject our null hypothesis. In this case, we retain our null hypothesis, which states that the observed data set *does* coincide statistically with the expected values. This is to say that there is no statistical difference between the observed and expected values, or that the observed data set does fit the given expected ratio for the circumstances surrounding the data.

For 2x2 observed/expected data tables—the Yates Correction

A special adjustment in the chi-square calculation can be made for data sets in which there are only four values—2 observed and 2 expected. Essentially, the difference for each observed-expected value is reduced by 0.5 to control for the degrees of freedom being only 1. To run this in R, use the following commands after acquiring, labeling, and attaching the data set in the same way as described above:

```
> x=read.table(file="filename.txt")
> O=observed
> E=expected
> df=1
> YatesCHISQ=sum(((abs(O-E)-0.5)^2)/E)
> YatesCHISQ
> YatesProb=1-pchisq(YatesCHISQ,df)
> YatesProb
```

The resulting Yates probability is most likely going to be different than the p-value produced by R for a 2x2 table without this adjustment. The Yates Correction is just a more powerful way to produce a p-value for a data set of expected and observed values of such a small size.

An Example Annotated Script

We must first transfer data into R from our .txt file. We do this the same way we have in the past by defining our working directory and then specifying out specific data set of interest.

```
> setwd("c:/Rdata")
> reservoir=read.table("reservoir.txt",header=TRUE)
```

Next we need to attach the dataset we just defined using the `attach` function and then view the dataset.

```
> attach(reservoir)
> reservoir
```

Then we use the `chisq.test` function to conduct the chi-squared test for goodness of fit. Our formal indicates to R that the numbers being used are probabilities, which then converts expected values into probabilities that sum to 1.

```
> chisq.test(observed,p=expected,rescale.p=TRUE)
```

We use the p-value to make our final decision and compare the p-value from the output with our designated α-level.

Compiled Protoype in R

The dataset "reservoir" gives the observed levels (in feet) of a Tennessee reservoir at hourly intervals, as well as the corresponding expected levels for each of the hourly intervals for the same reservoir, under the same conditions. Perhaps we want to know if there is statistical difference between the observed and expected values. To determine this, we would run a chi-square test for goodness of fit, because in essence, we are testing the extent to which the observed data "fits" the expected data. For our purposes, we will apply an alpha value of 0.05. The following commands in R would be used:

```
> setwd("c:/Rdata")
> res=read.table("reservoir.txt",header=TRUE)
> attach(res)
> res
```

```
   observed expected
1    795.31   657.68
2    795.28   657.65
3    795.24   657.63
4    795.21   657.60
5    795.18   657.59
6    795.14   657.55
7    795.12   657.54
8    795.09   657.51
9    795.06   657.52
10   795.04   657.52
11   795.02   657.47
12   795.01   657.45
13   795.01   657.46
14   794.99   657.43
15   794.98   657.42
16   794.98   657.43
17   794.97   657.43
18   794.97   657.44
19   794.97   657.43
20   794.96   657.43
21   794.96   657.43
22   794.95   657.44
23   794.96   657.43
24   794.94   657.41
25   794.93   657.41
26   794.92   657.40
27   794.91   657.41
28   794.89   657.41
29   794.88   657.38
30   794.88   657.36
31   794.86   657.37
32   794.84   657.35
```

```
33   794.83   657.33
34   794.80   657.32
35   794.78   657.31
36   794.75   657.27
37   794.74   657.25
38   794.71   657.24
39   794.68   657.21
40   794.66   657.18
41   794.62   657.14
42   794.60   657.14
43   794.56   657.14
44   794.53   657.12
45   794.50   657.12
46   794.47   657.09
47   794.43   657.07
48   794.40   657.05
49   794.37   657.04
50   794.34   657.01
51   794.30   657.01
52   794.27   656.97
53   794.24   656.94
54   794.20   656.92
55   794.18   656.90
56   794.14   656.86
57   794.11   656.82
58   794.08   656.77
59   794.05   656.75
60   794.00   656.68
```

```
> chisq.test(observed,p=expected,rescale.p=TRUE)

        Chi-squared test for given probabilities

data:  observed
X-squared = 2e-04, df = 59, p-value = 1
```

To interpret these results, we will compare the p-value produced by R (p-value=1) to the value we set for alpha (alpha=0.05). Our p-value is greater than alpha, which disenables us to reject our null hypothesis. We retain our null hypothesis, which states that the experimental data (observed values of the reservoir levels) fits the expected ratio presented by the expected reservoir levels. We can conclude that there is no difference between the patterns of expected and observed data.