



# Stat Note 7

In the seventh of a series of articles about statistics for biologists, **Anthony Hilton** and **Richard Armstrong** discuss:

## **Chi-square contingency tables**

**P**REVIOUS Statnotes describing the application of statistical methods to microbiological problems have been applied to measurement data (Hilton & Armstrong, 2005 a,b).

Measurement data are expressed in units; they are continuous variables and, in many cases, fulfil the requirements of the normal distribution (Hilton & Armstrong, 2005a). In some studies, however, the data are not measurements but comprise counts or frequencies of particular events. Such data are often analysed using the chi-square ( $\chi^2$ ) distribution. An example of the use of this statistic to test whether an observed distribution of frequencies came from a normal distribution ('goodness of fit test') was described in Statnote 1 (Hilton & Armstrong, 2005c). The objective of this Statnote is to extend these methods to the analysis of two different variables.

### **The Scenario**

Methicillin-resistant *Staphylococcus aureus* (MRSA) is a significant cause of nosocomial and community morbidity and mortality that, over the past two decades, has become a worldwide problem exacerbated by the emergence of multidrug-resistant isolates. Such isolates demonstrate a reduced susceptibility to almost all clinically available antibiotics. It is generally accepted that sub-lethal exposure of bacteria to antibiotics can promote the rapid development of resistance and that this situation may be more likely to occur in a hospital setting than in the community. It might be hypothesised, therefore, that isolates of MRSA from a hospital (HA-MRSA) would demonstrate an enhanced resistance profile to antibiotics compared to MRSA isolated from the community (CA-MRSA).

To test this hypothesis, 197 isolates of MRSA consisting of 95 HA-MRSA and 102 CA-

MRSA were isolated from soft tissue infections and screened for their sensitivity to a panel of 10 antibiotics using the British Society for Antimicrobial Chemotherapy (BSAC) disc diffusion method. Isolates were designated as resistant (R) or sensitive (S). If the hospital is providing an environment which promotes the development of antibiotic resistance then it might be expected that HA-MRSA would demonstrate a greater than average spectrum of resistance (i.e.  $\geq 5$  antibiotics of the 10 screened) than those isolated from the community. The potential significance of the association between antibiotic sensitivity and location of an isolate can be investigated using the chi-square ( $\chi^2$ ) test.

### **How is the test carried out?**

First, the data are tabulated in the form of a 2 x 2 contingency table (Table 1). In Table 1, 44% of the HA-MRSA isolates were resistant to  $\geq 5$  antibiotics as against 4.9% of

the CA-MRSA isolates. Is this difference sufficient to conclude that there is an association between the antibiotic sensitivity profile of the isolate and its location? Second, the expected frequencies are calculated for each cell of the 2 x 2 table and subtracted from the observed frequencies. Chi-square is the sum of the squares of these deviations divided by the appropriate expected frequency. The value of  $\chi^2$  is taken to the  $\chi^2$  table for 1 degree of freedom (DF) to obtain the probability that the value of the statistic would occur by chance if there were no differences between the isolates.

### **Interpretation of the results**

The calculated value of  $\chi^2$  ( $\chi^2 = 41.84$ ) is considerably greater than the value tabulated at the 5% level of probability. This is a value that would occur rarely by chance, in fact less than 1 in a 1000, and hence, we conclude that there is an association

**Table 1.** Is there an association between Hospital-Acquired and Community-Acquired MRSA antibiotic sensitivities (N >20)?

2 x 2 contingency table			
MRSA Isolate	Resistant to $\geq$ 5 Antibiotics	Resistant to $\geq$ 5 Antibiotics	Total
HA-MRSA	42	53	95
CA-MRSA	5	97	102
Total	47	150	197 = Grand tot.

1. The expected frequency (EF) in each cell is calculated as (Row Total x Column Total)/ Grand Total
2. Hence, the expected frequency of HA-MRSA isolates resistant to  $\geq$  5 antibiotics is  $(95 \times 47)/197 = 22.66$ . This calculation is repeated for each of the four cells of the table.
3. Calculate  $\chi^2 = \sum (O_i - E_i)^2/E_i$ . In this case  $\chi^2 = 41.84$  (39.70 with Yate's correction) ( $P < 0.001$ ) with 1DF.

between the antibiotic sensitivity profile of an isolate and its location. Caution is necessary when interpreting the results of  $\chi^2$  tests in observational studies (Snedecor & Cochran, 1980). There may be many factors that vary between a hospital and community setting that could influence the antibiotic resistance profile of a MRSA strain, some of which may be wholly or partly responsible

deviation of the observed from the expected frequency. Another statistic that is sometimes given by statistical software is called 'phi-square' and is a measure of the degree of correlation between the two variables in a 2 x 2 table.

### Yate's correction

Statistical software usually includes the option of calculating  $\chi^2$  with Yate's correction. This correction

**Table 2.** Is there an association between Hospital-Acquired and Community-Acquired MRSA antibiotic sensitivities (N < 20)?

2 x 2 contingency table			
MRSA Isolate	Resistant to $\geq$ 5 Antibiotics	Resistant to $\geq$ 5 Antibiotics	Total
HA-MRSA	A	B	(A + B)
CA-MRSA	C	D	(C + D)
Total	(A + C)	(B + D)	Total (T) = N

1. If T < 12, calculate the probability (P) of this particular outcome among all possible outcomes with the same row and column totals: i.e.,  $P = A! \times B! \times C! \times D! / CA! \times CB! \times DA! \times DB! \times T!$
2. If T is larger than 12, then a calculation based on logarithms can be used (see Fisher and Yates Table XXX, Dawkins 1975).

for an observed significant difference.

To understand why a 2 x 2 table has only 1 DF, examine the deviations of the observed from the expected frequencies for each cell of the table. Examination of these deviations will show that they are all the same apart from their sign, i.e., in a 2 x 2 table there is only a single *independent* estimate of the

improves the estimate of  $\chi^2$  in a 2 x 2 table when expected frequencies are small (e.g., when expected frequencies < 10). The absolute value of the difference between the observed and expected frequencies is reduced by 0.5 before squaring. The effect of this is to make the estimate of  $\chi^2$  slightly more conservative when the table contains small frequencies. Yate's correction

applied to the above example gives a value of  $\chi^2 = 39.70$ .

### R x C contingency tables

It is possible to analyse two variables with a greater number of categories per variable and this is termed a rows (R) x columns (C) contingency table. For example, antibiotic resistance may have been tested for several different strains simultaneously. To make the test, the expected frequency is calculated for each cell of the table as in Table 1. The value of  $\chi^2$  is then calculated using the usual formula and the value of  $\chi^2$  compared with the  $\chi^2$  distribution, entering the table for (Number of rows - 1)(Number of columns - 1) DF (Snedecor & Cochran, 1980). If a significant  $\chi^2$  is obtained, the R x C table may need to be broken down into smaller tables to compare some of the isolates in more detail.

### Fisher's 2 x 2 'Exact Test'

The  $\chi^2$  test described above is only an approximate test when applied to a 2 x 2 table and the approximation becomes poorer as sample size decreases. Hence, the test is inaccurate when the expected frequencies are low and it is usually considered inappropriate if the expected values fall below 5. One remedy is to apply Yate's correction as described above. An alternative to  $\chi^2$ , called Fisher's 2 x 2 exact test, can be used, however, and is illustrated in Table 2. This test should be applied if the total sample size is less than 20 or if N lies between 20 and 40 and the smallest expected frequency is less than 5. When the total of the observations is small, say less than 12, the probability of a particular distribution of values in a 2 x 2 table being obtained, given the particular row and column totals, can be calculated directly from the data. If the

total is larger than 12, then a more complex calculation can be made using logarithms (Dawkins, 1975).

### Conclusions

When the data are counts or the frequencies of particular events and can be expressed as a contingency table, then they can be analysed using the  $\chi^2$  distribution. When applied to a 2 x 2 table, the test is approximate and care needs to be taken in analysing tables when the expected frequencies are small either by applying Yate's correction or by using Fisher's exact test. Larger contingency tables can also be analysed using this method. Note that it is a serious statistical error to use any of these tests on measurement data!

### References

- Dawkins, HC (1975) Staffrooms: 'Pro-formas' for the Guidance of Statistical Calculations. Edward Arnold, London.
- Fisher RA & Yates F (1963) Statistical Tables for Biological, Agricultural and Medical Research. Sixth Ed. Longman, Edinburgh.
- Hilton A & Armstrong RA (2005a) Statnote 2: Describing the normal. *Microbiologist* Vol 6: No.3, pps 30 - 33.
- Hilton A & Armstrong RA (2005b) Statnote 3: Testing the difference between two groups. *Microbiologist* Vol 6: No.4, pps 30 - 32.
- Hilton A & Armstrong RA (2005c) Statnote 1: Is the data normal? *Microbiologist* Vol 6: No.2, pps 34-35.
- Snedecor GW & Cochran WG (1980) Statistical Methods, 7th Ed. Iowa State University Press, Ames Iowa.

### Dr Anthony\* Hilton and Dr Richard Armstrong\*\*

\*Pharmaceutical Sciences and  
\*\*Vision Sciences, Aston University, Birmingham, UK