TEACHING STUDENTS TO USE THE CHI-SQUARE TEST WHEN OBSERVATIONS ARE DEPENDENT

Austina S. S. Clark
University of Otago, New Zealand
aclark@maths.otago.ac.nz

The purpose of statistical analysis is to extract and assess information contained in data and draw inferences based on the analysis. Students, especially those doing experimental work and not majoring in statistics, need to apply the statistical methods they have learned to their collected data. These data often arise in diverse situations and standard methods cannot be applied directly. The statistical methodology must be developed to address problems associated with such data. Here we use the chi-square test with dependent data to illustrate methodology modification in order to analyse data correctly. The underlying mathematics is summarised, but students are only expected to analyse such data using software and to check modification procedure using simulations. The aim here is to show students that statistical procedures can be modified then applied even though the data does not fit assumptions for a standard procedure. The importance of simulation is evident as an additional benefit for learning.

INTRODUCTION

For the last five years I have been teaching statistics to two groups of postgraduate students, who major in either ecology or marine sciences. These students do a lot of experimental work as well as field work, but have limited knowledge in mathematics and statistics. Indeed most of them have only completed a first year course in statistics before joining the class. Their primary objective is to use statistical procedures to analyse data and draw inferences. They often have a simplistic view about some of the statistical procedures, which is fine if the underlying theoretical assumptions are met. Unfortunately some of the data collected by the students are diverse, and in this case using standard procedures could result in an incorrect analysis and invalid inference. In what follows, we use the chi-square test with dependent data as an example to illustrate how to modify and then apply a test correctly.

When the chi-square test is applied to test the association between two binomial distributions, each with p cells, we usually assume that cell observations are independent. If some of the cells are dependent we would like to investigate (i) how to implement the chi-square test and (ii) how to find the test statistics and the associated degrees of freedom. The test statistics and degrees of freedom are developed from results by Geisser and Greenhouse (1959), Huynh and Feldt (1976) and Satterthwaite (1941, 1946).

Students take more interest in data which is topical or comes from a context that relates to the areas they are working in. For example, since this course material was taught in winter when influenza was around, we chose the following example in class to illustrate the methods. The example considered two groups of patients suffering from influenza symptoms. One group suffered from H1N1 influenza 09 and the other from seasonal influenza, with data recorded by Chang, et al (2010). There were twelve symptoms collected for each patient and these symptoms were not totally independent, which is the feature that has led to the procedure described below.

METHODS

The medical records of sixty-four adult patients with a laboratory confirmed diagnosis of two types of influenza, seasonal influenza (F) and H1N1 influenza 09 (S), were used between 17 June and 31 July, 2009, in an Australian hospital. Twelve symptoms were extracted from each patient's records using 0 for no symptom and 1 for the symptom. These symptoms were as follows:

S1 (coryza), S2 (fever), S3 (cough), S4 (breathlessness), S5 (chest pain), S6 (sore throat), S7 (lethargy), S8 (myalgia), S9 (vomiting), S10 (diarrhoea), S11 (abdominal pain) and S12 (gastro-intestine upset).

Some of the symptoms are not independent: for example, abdominal pain and vomiting both could be associated with fever. Therefore it is not appropriate to use the chi-square test to check whether these symptoms are independent for these two types of influenza.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

We examined the correlation matrices for the two groups of patients, F (seasonal influenza) and S (H1N1 09). The correlation was significant so we calculated the two covariance matrices respectively and then pooled them together to form a pooled covariance matrix $\sum$. This procedure was taught in a computer laboratory, where both the lecturer and students had computer access and there was also a projector screen for the lecturer's computer. Thus students had the opportunity for hands-on experience and were able to ask questions if they had any doubt during the development of the process.

We calculated the mean proportion for each of the symptoms, say p, as follows:

$\overline{Y}_F = [\overline{y}_{F1,} \quad \overline{y}_{F2,} \quad \ldots \quad \overline{y}_{Fp}]'$ and $\overline{Y}_S = [\overline{y}_{S1,} \quad \overline{y}_{S2,} \quad \ldots \quad \overline{y}_{Sp}]'$.

The results were summarized in a table with following layout:

	S1	S2	S3	. . . . .		Sp
F						
S						

In order to find the true proportion difference between the two groups, we need to find the difference between $\overline{Y}_F$ and $\overline{Y}_S$. Since there is correlation between the p variables, we cannot use the Penrose distance discussed by Manly (1994). However, we have instead two alternative ways of incorporating the correlation.

Firstly we can apply the Mahalanobis distance, D^2_{FS}, which takes into account the correlations between variables, where $D^2_{FS} = (\overline{Y}_F - \overline{Y}_S)' \Sigma^{-1} (\overline{Y}_F - \overline{Y}_S)$ with $\sum$ as the pooled covariance matrix for the two populations, and D^2_{FS} can be thought of as a multivariate difference for the two observations $\overline{Y}_F$ and $\overline{Y}_S$, taking account of all p variables. We assume that the populations which $\overline{Y}_F$ and $\overline{Y}_S$ come from are multivariate normally distributed - then the values of D^2_{FS} will follow a chi-square distribution with p degrees of freedom.

Alternatively we may apply the method suggested by Greenhouse & Geisser (1959) by transforming $\overline{Y}_F - \overline{Y}_S$ using the eigenvectors, A, and eigenvalues, D, of the covariance matrix $\sum$.

If $Z = \overline{Y}_F - \overline{Y}_S = [\overline{y}_{F1} - \overline{y}_{S1,} \quad \overline{y}_{F2} - \overline{y}_{S2,} \quad \ldots \quad \overline{y}_{Fp} - \overline{y}_{Sp}]' = [z_{1,} \ldots, z_p]'$ and

$W = A'Z = [w_{1,} \quad w_{2,} \ldots, w_p]$, then $W \sim MVN(0, D)$, where $w_i's$ are independent.

Next let $D^2_W = W'W = \| W \|^2$ and $\| W \|^2 = \| Z \|^2$; this indicates that the values of $\| W \|^2$ follow a chi-square distribution $m\chi^2_n$, where m is a multiplier and n can be approximated according to Satterthwaite (1941, 1946).

We used the eigenvalues $\lambda_i's$ of the covariance matrix $\sum$ to work out m and n as follows:

$$m = \sum_1^p \lambda_i^2 / \sum_1^p \lambda_i \text{ and } n = (\sum_1^p \lambda_i)^2 / \sum_1^p \lambda_i^2 .$$

Most students were content to accept the calculations of m and n. For the more inquisitive,
a detailed development of the approximation was handed out as reference.

As mentioned earlier, for this example, it was clear that the influenza symptoms were not totally independent, and so the methods detailed above were applied. Using MINITAB, the results were:

Method 1: $D^2_{FS} = (\bar{Y}_F - \bar{Y}_S)' \Sigma^{-1} (\bar{Y}_F - \bar{Y}_S)$ = 0.9384, which follows a χ^2_{12} distribution with a p-value of 0.9999.

Method 2: $D^2_W = W'W = \| W \|^2$ =0.1215, which follows a $m\chi^2_n$ distribution with m =0.2873, n =7.2596 and p-value = 0.9997.

Students were comfortable using either of the packages MINITAB or SPSS to work out the p-values.

Both methods showed that there is no significant difference in the proportions of the twelve symptoms between the two types of influenza. However, patients with H1N1 09 (S) were significantly younger than patients with seasonal influenza (F), $mean_S = 45$ versus $mean_F = 64$ with p-value < 0.01. The mean duration of symptoms prior to presentation was 4 days, with fever, cough and dyspnoea being the most common symptoms in both groups. Pneumonia occurred in 44% and 38% of H1N1 09 and seasonal influenza patients respectively. The students were able to carry out these calculations on their own.

In conclusion, students were confident going through the various stages of calculation and from this concurred that, for those patients admitted to the hospital, the H1N1 09 influenza virus caused clinical disease in humans comparable to the seasonal influenza strains in this Australian city during the period 17 June to 31 July, 2009.

SIMULATION

Most students would stop the statistical procedure once the p-values were obtained. However they needed to be reminded that any modification of the procedure must be verified for adequacy, and so the simulation idea was introduced to emphasize this. We used MATLAB; I wrote the code with detailed comments so students could modify it when needed later. We simulated 200,000 times the proportions of the twelve symptoms for the two groups of influenza respectively, using both methods. The results are displayed in the graphs in Figure 1.

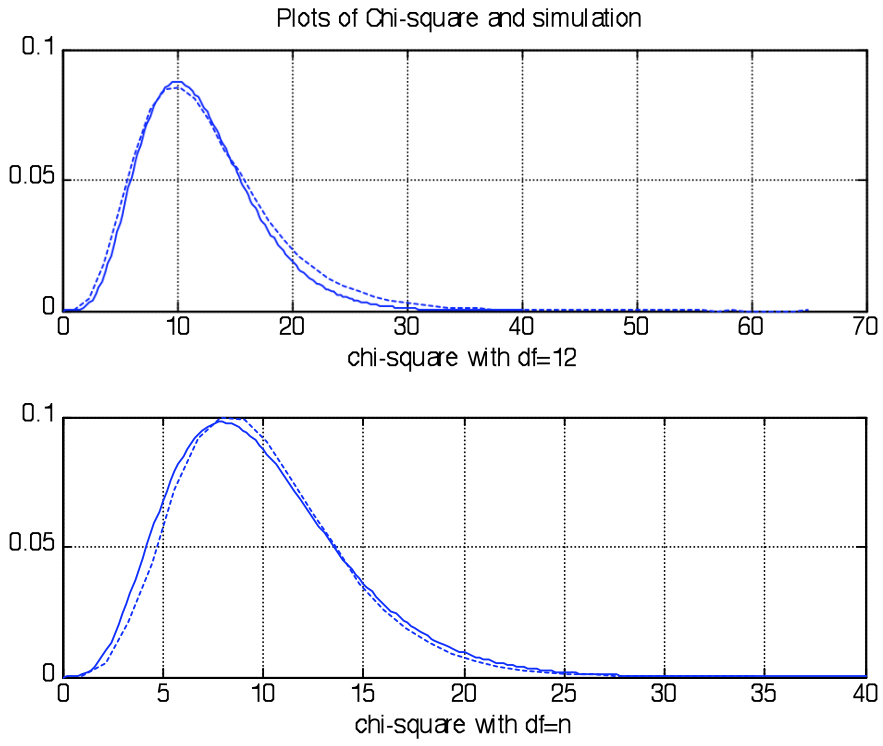

Figure 1. Results

From the simulation graphs students can observe that the simulated results, represented by broken lines, are very closed to the theoretical ones, represented by solid lines.

CONCLUSION

This study showed students that the chi-square test can still be applied even though some of the observations are dependent. The simulation results confirmed that the modified chi-square test was a reasonable tool to use. The study also drew attention to the students of (i) the need to test assumptions before using a statistical procedure and (ii) the importance of simulation. After detailed instruction in the computer laboratory using the MINITAB or SPSS packages, the students in the class were able to use the method confidently.

REFERENCES

Chang, Y., van Hal, S. J., Spencer, P. M., Gosbell, I. B., & Collett, P. W. (2010). Comparison of adult patients hospitalised with pandemic (H1N1) 2009 influenza and seasonal influenza during the "PROTECT' phase of the pandemic response. *The Medical Journal of Australia, 192*, 1-4.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degree of freedom from sample data in randomized block and split plot designs. *Journal of Educational Statistics, 1*(1), 69-82.

Manly, B. F. J. (1994). *Multivariate Statistical Methods* (2nd edition). London: Chapman & Hall.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6*, 309-316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110-114.