NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

12-13

BOOTSTRAP P-VALUES FOR COCHRAN'S Q, STUART and BOWKER TESTS

D. J. BEST and J.C.W RAYNER

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

BOOTSTRAP P-VALUES FOR COCHRAN'S *Q*, STUART and BOWKER TESTS

D.J. BEST[1] and J.C.W. RAYNER*[1,2]
University of Newcastle and University of Wollongong

Summary

Cochran's *Q* assesses treatment differences in randomized block designs with binary data. We suggest using bootstrap p-values rather than p-values based on the chi-squared distribution for tests based on *Q*. These chi-squared p-values for *Q* are the only ones usually given in statistical software and can be inaccurate. The same approach allows improved p-values to be given for sparse two-way cross-classification data.

Key Words: Binary data, cross-classification data, marginal homogeneity, nonparametric, symmetry in two-way tables.

1. Introduction

For binary data in randomized blocks we obtain improved p-values for Cochran's *Q* test. Cochran's *Q* is available in many statistical packages or else can be calculated by a Friedman's rank test routine which adjusts for ties. However the distribution of Cochran's *Q* can be poorly approximated by the chi-squared distribution. See, for example, Bhapkar and Somes (1977) and section 4 below. For two treatments and binary data Suissa and Shusta (1991) and Berger and Sidik (2003) have given exact unconditional p-values but we are unaware that such has been given for more than two treatments. For t treatments and binary data there are 2^t possible responses for each block. Following Bennett (1967) we can use maximum likelihood and a multinomial model to estimate the probabilities of each of these 2^t responses. These can be used in a parametric bootstrap simulation which, we suggest, improves upon the chi-squared p-values. Sometimes the iterations needed for the maximum likelihood estimates (MLEs) do not converge. In such cases the approximation of Bhapkar and Somes (1977) can be used. We do not advocate conditional p-values as these assume that in a repetition of an experiment or trial that the number of 'successes' will be fixed within each block. This seems unreasonable.

For completeness we now give an explicit formula for Cochran's *Q*. Suppose we have a randomised block design in which t treatments are applied to r blocks. Let X_{ij} be the outcome for treatment i on block j, and suppose that $X_{ij} = 1$ if the outcome is a

* Author to whom correspondence should be addressed.
Telephone: 61 2 49215737; Fax: 02 4921 6898; e-mail: John.Rayner@newcastle.edu.au
[1] School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia
[2] Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia

'success', and $X_{ij} = 0$ otherwise. Using standard 'dot' notation, $\sum_{i=1}^{t} X_{ij} = X_{\bullet j}$, $\sum_{j=1}^{r} X_{ij} = X_{i\bullet}$ and $\sum_{i=1}^{t}\sum_{j=1}^{r} X_{ij} = X_{\bullet\bullet}$. Cochran's Q is given by

$$Q = t(t-1)\frac{\sum_{i=1}^{t}(X_{i\bullet} - X_{\bullet\bullet}/t)^2}{\sum_{j=1}^{r} X_{\bullet j}(t - X_{\bullet j})}.$$

For $t = 2$ we can also define Q in terms of the frequencies n_1, n_2, n_3 and n_4 for the responses (1, 1), (1, 0), (0, 1) and (0, 0). This gives $Q = (n_2 - n_3)^2/(n_2 + n_3)$. Similarly for $t = 3$ we can also define Q in terms of the frequencies n_1, ... , n_8 of the responses (1, 1, 1), (1, 1, 0), (1, 0, 1), (1, 0, 0), (0, 1, 1), (0, 1, 0), (0, 0, 1) and (0, 0, 0). This gives

$$3(n_2 + n_3 + n_4 + n_5 + n_6 + n_7)\,Q =$$
$$(n_2 + n_3 + 2n_4 - 2n_5 - n_6 - n_7)^2 + (n_2 + n_5 + 2n_6 - 2n_3 - n_4 - n_7)^2 +$$
$$(n_3 + n_5 + 2n_7 - 2n_2 - n_4 - n_6)^2.$$

Sections 2 and 3 look at the cases of two and three treatments. Davis (2002, p.169) says the underlying multinomial model implies Cochran's Q is a nonparametric statistic and hence the basis of a robust procedure. Section 4 looks at the Bhapkar and Somes (1977) approximation while section 5 considers cross-classified data.

2. Two treatments

Following Simonoff (2003, p.291), Table 1 classifies 261 boys under five years of age in Ngamiland, Botswana on the basis of the W (WHO) standard and the E (Ehrenberg) standard. In Table 1 A represents malnourished (1) and B represents normal (0).

The information in Table 1, a 2 × 2 table, can be displayed in Table 2, a 2 × 4 table.

Table 1. Frequencies for W and E

	E		
W	A	B	
A	20 (n_1)	0 (n_2)	20 ($n\pi_S$)
B	4 (n_3)	237 (n_4)	241
	24 ($n\pi_T$)	237	261

Table 2. Frequencies for the classifications W and E for candidates A and B

W	A	A	B	B
E	A	B	A	B
	20 (n_1)	0 (n_2)	4 (n_3)	237 (n_4)

In Table 2 classification A is shown as A and classification B is shown as B. The research question is, "Is the proportion (π_W) who are classified A by the W standard different to the proportion (π_E) who are classified A by the E standard?"

For i = 1, 2, 3 and 4 let $\hat{p}_i$ be the maximum likelihood estimator (MLE) of p_i, the unknown probability of cell i in Tables 1 and 2. We seek to test the marginal homogeneity hypothesis H_0: $\pi_W - \pi_E = 0$ against K: $\pi_W - \pi_E \neq 0$. Bennett (1967) assumes a multinomial model with n_i, i = 1, 2, 3 and 4 counts in Tables 1 and 2 and shows that $\hat{p}_i = n_i/(\lambda_1 + \lambda_2 \mathbf{a}_i)$, where λ_1 and λ_2 are Lagrange multipliers and $\mathbf{a}_i$ is the ith column of matrix $\mathbf{A} = (0, 1, -1, 0)$.

If $\mathbf{p} = (p_1, p_2, p_3, p_4)^T$ then marginal homogeneity implies $\mathbf{Ap} = 0$. To see this observe that $\pi_W = p_1 + p_2$, $\pi_E = p_1 + p_3$, and so H_0 requires $\pi_W - \pi_E = p_2 - p_3 = 0$. Matrix $\mathbf{A}$ gives the coefficients of p_i in H_0 and so here the coefficient of p_1 is 0, while the coefficients for p_2, p_3 and p_4 are 1, –1 and 0 respectively. Here $\mathbf{A}$ has only one row as the null hypothesis only involves one linear constraint. In maximizing the likelihood by choice of $\mathbf{p}$ we need to involve the marginal homogeneity constraint $\mathbf{Ap} = 0$ and this implies the need to involve the Lagrange multipliers λ_1 and λ_2. As $p_1 + p_2 + p_3 + p_4 = 1$ it follows as in Bennett (1967) that $\lambda_1 = n$. Alternatively we could put $p_4 = 1 - \sum_{i=1}^{3} p_i$ in the multinomial.

Under marginal homogeneity $\hat{p}_2 - \hat{p}_3 = 0$, whence $n_2/(n + \lambda_2) - n_3/(n - \lambda_2) = 0$ giving $\lambda_2 = n(n_2 - n_3)/(n_2 + n_3)$ and ultimately $\hat{p}_1 = n_1/n$, $\hat{p}_2 = (n_2 + n_3)/(2n) = \hat{p}_3$ and $\hat{p}_4 = n_4/n$.

Cases i = 1, 2, 3 and 4 in Tables 1 and 2 are independent because different boys are classified. We can test H_0 using the Pearson statistic $X^2 = \sum_{i=1}^{4} (n_i - n\hat{p}_i)^2 / n\hat{p}_i = (n_2 - n_3)^2/(n_2 + n_3) = Q = 4.000$ for these data. The test statistic is called McNemar's statistic after McNemar (1947). McNemar's statistic has an approximate χ_1^2 distribution. Thus an approximate p-value for these data would be $P(X^2 > 4.000) = 0.046$. There is some evidence that classification differs from W to E. Notice that this X^2 statistic reflects neither Pearson's test for independence nor Pearson's test for homogeneity as the same boys are classified by standards W and E.

We note that H_0 is often examined assuming $n\hat{p}_2 + n\hat{p}_3$ to be fixed. This is the basis of the so-called exact conditional test of H_0 providing an 'exact' conditional p-value as opposed to a bootstrap p-value to be described shortly. For the Table 1 data the exact two-sided conditional p-value 0.125, found using binomial probabilities with n = 4 and p = 0.5.

We consider a better p-value would be based on a parametric bootstrap approach. Suppose we generate many Table 1 type data sets using random multinomials with n = 20 and $(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4) = (0.0766, 0.0076, 0.0076, 0.9080)$ and calculate new X^2 values for each such table. The proportion of these new tables with $X^2 > 4.000$ is then a bootstrap estimate of the p-value. When 100,000 such tables were generated we found a bootstrap p-value of 0.043. Berger and Sidik (2003) discuss unconditional p-values which should

be similar to those of Suissa and Shusta (1991). Our experience is that our bootstrap p-values are very close to these unconditional p-values.

3. Three treatments

Table 3 gives data from Davis (2002, p.186) concerning three drugs A, B, and C from 46 patients who respond with either an F (favourable) or U (unfavourable). There are $2^3 = 8$ possible responses for each of the 46 patients and we assume that for each patient a response is independent of the response of every other patient.

Table 3. Response triples for drug therapy

Drug	Response triples								
A	F	F	F	F	U	U	U	U	
B	F	F	U	U	F	F	U	U	
C	F	U	F	U	F	U	F	U	
Frequency	6 (n_1)	16 (n_2)	2 (n_3)	4 (n_4)	2 (n_5)	4 (n_6)	6 (n_7)	6 (n_8)	46 (n)

The marginal homogeneity null hypothesis is H_0: $\pi_A = \pi_B = \pi_C = \pi$, say, where the π's refer to F's and not U's.

Under marginal homogeneity H_0: $\pi_A - \pi_C = 0$ and $\pi_B - \pi_C = 0$. We have

$$\pi_A = p_1 + p_2 + p_3 + p_4,\ \pi_B = p_1 + p_2 + p_5 + p_6, \text{ and } \pi_C = p_1 + p_3 + p_5 + p_7,$$

where p_i is the unknown probability associated with n_i in Table 3. Thus under H_0

$$\pi_A - \pi_C = p_2 + p_4 - p_5 - p_7 = 0 \text{ and } \pi_B - \pi_C = p_2 + p_6 - p_3 - p_7 = 0,$$

and so we can define the constraint matrix

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 1 & -1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 \end{pmatrix}$$

corresponding to the coefficients of the p_i in $\pi_A - \pi_C$ and $\pi_B - \pi_C$.

To find the MLEs of the p_i we need to maximize the likelihood subject to $\mathbf{Ap} = 0$. As in the previous section, following Bennett (1967) but now using three Lagrange multipliers λ_1, λ_2 and λ_3, we find $\lambda_1 = n$ and now if $\boldsymbol{\lambda} = (\lambda_2, \lambda_3)^T$ and $\mathbf{a}_i$ is the ith column of $\mathbf{A}$, then we consider

$$\hat{p}_2 + \hat{p}_4 - \hat{p}_5 - \hat{p}_7 = \frac{n_2}{n + \boldsymbol{\lambda}^T a_2} + \frac{n_4}{n + \boldsymbol{\lambda}^T a_4} - \frac{n_5}{n + \boldsymbol{\lambda}^T a_5} - \frac{n_7}{n + \boldsymbol{\lambda}^T a_7} = 0 \text{ and}$$

$$\hat{p}_2 + \hat{p}_6 - \hat{p}_3 - \hat{p}_7 = \frac{n_2}{n + \boldsymbol{\lambda}^T a_2} + \frac{n_6}{n + \boldsymbol{\lambda}^T a_6} - \frac{n_3}{n + \boldsymbol{\lambda}^T a_3} - \frac{n_7}{n + \boldsymbol{\lambda}^T a_7} = 0.$$

These two non-linear equations can be solved simultaneously for λ_2 and λ_3 to obtain the MLE of **p**. Other approaches for obtaining the MLE are available. See, for example, Bergsma et al. (2009). However we do not discuss such approaches here.

We find $\lambda_2 = \lambda_3 = 11.040$ and so $\hat{\mathbf{p}}^{\mathrm{T}}$ = (0.130, 0.236, 0.057, 0.070, 0.057, 0.070, 0.249, 0.130). For the data here Cochran's Q = 8.470 with chi-squared p-value 0.015. Using the $\hat{\mathbf{p}}$ just given we find, after 100,000 simulations, that the bootstrap p-value is 0.021. As above tests based on Q may be considered nonparametric, and so have the advantage of making few assumptions. Observe that unlike the $t = 2$ case discussed above $Q \neq X^2$ when $t = 3$. We will not compare Q and X^2 here.

4. The Bhapkar and Somes approximation

Depending on the counts n_i, when $t > 2$ the maximum likelihood estimates of the cell probabilities may fail to converge. In such cases approximate p-values due to Bhapkar and Somes (1977) can be used. In their Table 2 they show that a $\theta\chi^2_\phi$ distribution gives better sizes for Q than a χ^2_3 distribution for $t = 4$. As they only used 1,000 simulations some of their test sizes are slightly in error. Further, they only presented values for $n = 100$. Our Table 4 below looks at n = 10, 30, 50, 100 and 1,000 with nominal test size 5% and uses 100,000 Monte Carlo simulations. The asymptotic distribution of Q does not appear to be χ^2_3. Clearly the $\theta\chi^2_\phi$ approximation does better than χ^2_3 for n = 30, 50, 100 and 1,000. For $n = 10$ the discrete nature of the distribution of Q can defeat both approximations. See, for example, the results for P3 and P4. The $\theta\chi^2_\phi$ approximation is reasonable for $n \geq 30$ and excellent for $n \geq 50$.

Table 4. Test sizes for nominal 5% sizes for χ^2_3 and $\theta\chi^2_\phi$ approximations

Simulation	$n = 10$		$n = 30$		$n = 50$		$n = 100$		$n = 1000$	
Approximation	χ^2_3	$\theta\chi^2_\phi$	χ^2_3	$\theta\chi^2_\phi$	χ^2_3	$\theta\chi^2_\phi$	χ^2_3	$\theta\chi^2_\phi$	χ^2_3	$\theta\chi^2_\phi$
P1	0.070	0.043	0.070	0.047	0.072	0.048	0.070	0.048	0.072	0.049
P2	0.074	0.043	0.075	0.047	0.076	0.048	0.076	0.050	0.076	0.050
P3	0.057	0.038	0.061	0.045	0.061	0.047	0.061	0.048	0.062	0.050
P4	0.061	0.040	0.065	0.047	0.065	0.047	0.065	0.048	0.065	0.050
P5	0.085	0.043	0.083	0.048	0.084	0.049	0.084	0.048	0.083	0.049
P6	0.085	0.044	0.086	0.049	0.086	0.049	0.087	0.050	0.088	0.050
P7	0.046	0.034	0.050	0.045	0.052	0.048	0.052	0.049	0.052	0.050
P8	0.046	0.034	0.052	0.046	0.053	0.048	0.053	0.049	0.053	0.050

The parameters θ and ϕ in $\theta\chi^2_\phi$ are estimated by

$$\hat{\theta} = (t-1)S_2 / S_1^2 \text{ and } \hat{\phi} = S_1^2 / S_2$$

where

$$S_1 = \text{tr}(\mathbf{BV}^*) \text{ and } S_2 = \text{tr}(\mathbf{BV}^*)^2$$

in which, putting T_{jk} = total number of simultaneous 'successes' ('1' or 'F' above) in the jth and kth treatments, for j and $k = 1, \ldots, t$, $p_{jk} = T_{jk}/n$, $\bar{p} = \sum_{i=1}^{t} p_{ii}/t$, $\mathbf{V}^* = (p_{jk} - \bar{p}^2)$ and $\mathbf{B} = \mathbf{I}_t - \mathbf{1}_t\mathbf{1}_t^{\mathrm{T}}/t$.

In Table 4 we use a subset of the **p** given in Bhapkar and Somes (1977, Table 1). There are a few obvious typographical errors in this table. We look at $\mathbf{p} = (p_1, \ldots, p_{16})^{\mathrm{T}}$ defined by P1, …, P8 where

P1 = (.03, .03, .03, .03, .03, .03, .03, .29, .29, .03, .03, .03, .03, .03, .03, .03),
P2 = (.03, .03, .03, .29, .03, .03, .03, .03, .03, .03, .03, .03, .29, .03, .03, .03),
P3 = (.04, .04, .04, .04, .04, .04, .04, .22, .22, .04, .04, .04, .04, .04, .04, .04),
P4 = (.04, .04, .04, .22, .04, .04, .04, 04, .04, .04, 04, .04, .22, .04, .04, .04),
P5 = (.02, .02, .02, .02, .02, .02, .02, .36, .36, .02, .02, .02, .02, .02, .02, .02),
P6 = (.02, .02, .02, .36, .02, .02, .02, 02, .02, .02, .02, .02, .36, .02, .02, .02),
P7 = (.07, .07, .07, .07, .07, .07, .07, .01, .01, .07, .07, .07, .07, .07, .07, .07) and
P8 = (.07, .07, .07, .01, .07, .07, .07, .01, .01, .07, .07, .07, .01, .07, .07, .07).

We estimated θ and ϕ from the simulated data sets for each of the 100,000 Monte Carlo trials. It isn't apparent that Bhapkar and Somes (1977) did this. The results in Table 4 are typical of other scenarios we have investigated. Even for n = 10,000 the χ^2_3 approximation can be poor. Notice that, as required by H_0, $\pi_A = \pi_B = \pi_C = \pi_D$ for P1, …, P8. Wallenstein and Berger (1981, Table 2) also give test sizes for P2 and P6. Their results are similar to ours but again they only use 1,000 simulations.

In the two examples of sections 2 and 3 the chi-squared and bootstrap p-values for Q are similar; see Table 5. Although it is not a dramatic effect, we see that the p-value from $\theta\chi^2_\phi$ is equal to or closer to the bootstrap p-value than χ^2_{t-1} is. We would expect sometimes that χ^2_{t-1}, $\theta\chi^2_\phi$ and bootstrap would not agree on significance at a given level. In such cases we suggest using the bootstrap p-value.

Table 5. P-values for Q for the examples of sections 2 and 3

Example	χ^2_{t-1}	$\theta\chi^2_\phi$	Bootstrap
1, with $t = 2$	0.035	0.035	0.041
2, with $t = 3$	0.015	0.021	0.021

It is apparent from Table 4 that the Bhapkar and Somes (1977) approximation can be safely used when $n \geq 30$ and the MLEs have convergence problems. For $n < 30$ and MLEs having convergence problems it also appears from Table 4 that the Bhapkar and Somes (1977) approximation, while not always having the correct test size, is a better approximation than the χ^2_{t-1} approximation for obtaining p-values.

5. Two-way cross-classifications

Suppose 33 consumers were asked to rate two coffees X and Y for their flavour on a three point 'just right' category scale: too weak, just right and too strong. Table 6 gives a cross classification of their responses. Further discussion of 'just right' data is given, for example, in Lawless and Heymann (2010, 334-339).

Table 6. Cross-classification of coffee responses

Coffee Y\Coffee X	Too weak	Just right	Too strong
Too weak	4 (n_1)	10 (n_2)	4 (n_3)
Just right	3 (n_4)	3 (n_5)	4 (n_6)
Too strong	1 (n_7)	2 (n_8)	2 (n_9)

Table 7. Alternative display of coffee responses

i	1	2	3	4	5	6	7	8	9
Coffee X	1	1	1	2	2	2	3	3	3
Coffee Y	1	2	3	1	2	3	1	2	3
n_i	4	10	4	3	3	4	1	2	2

Date like that in Table 6 are discussed, for example, in Lawless and Heymann (2010, pp334-339). The marginal frequencies for coffee X are (8, 15, 10), and for coffee Y are (18, 10, 5). Barplots of these frequencies would give a visual comparison of these two sets of marginal frequencies, indicating the flavour of coffee Y is too weak. To more formally check marginal homogeneity of the responses we can calculate the Stuart (1955) statistic, which, for two treatments, and three categories may be given as

$$S = \frac{c_1 d_3^2 + c_2 d_2^2 + c_3 d_1^2}{2(c_1 c_2 + c_2 c_3 + c_3 c_1)} = 6.069 \text{ here,}$$

where $c_1 = (n_2 + n_4)/2$, $c_2 = (n_3 + n_7)/2$, $c_3 = (n_6 + n_8)/2$, $d_1 = n_2 + n_3 - n_4 - n_7$, $d_2 = n_4 + n_6 - n_2 - n_8$ and $d_3 = n_7 + n_8 - n_3 - n_6$. Using the χ^2_2 approximation S has a p-value of 0.048. A bootstrap p-value can be calculated as above. First, rewrite Table 6 as in Table 7.

In Table 7 we have coded too weak as '1', just right as '2' and too strong as '3'. The marginal homogeneity null hypothesis is H_0: $\pi_{Y_1} - \pi_{X_1} = 0$ and $\pi_{Y_2} - \pi_{X_2} = 0$. In H_0, Y_j and X_j, $j = 1, 2$ refer to each coffee being scaled or scored as 1 or 2. Thus, for example, π_{X_2} is the probability coffee X is scaled as 2: just right.

To find the MLEs of the p_i one approach, as above, involves simultaneously solving the two non-linear equations

$$\hat{p}_2 + \hat{p}_3 - \hat{p}_4 - \hat{p}_7 = 0 \text{ and } \hat{p}_4 + \hat{p}_6 - \hat{p}_2 - \hat{p}_8 = 0$$

for λ_2 and λ_3 where $\hat{p}_i = n_i/(n + \lambda^{\mathrm{T}}\mathbf{a}_i)$, in which $\mathbf{a}_i$, as above, is the ith column of the constraint matrix $\mathbf{A}$ given by

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 1 & 0 & -1 & 0 \end{pmatrix}.$$

We find $\lambda_2 = 22.722$ and $\lambda_3 = 6.570$ from which

$$\hat{\mathbf{p}}^{\mathrm{T}} = (0.121, 0.203, 0.072, 0.178, 0.091, 0.101, 0.097, 0.076, 0.061).$$

The bootstrap p-value is 0.042, which is not too different from the chi-squared p-value 0.048 given above. There is some evidence the marginal distributions of the two coffees differ, in agreement with the visual inspection of the marginal frequencies.

Had the MLE iteration not converged a bootstrap p-value would no longer be available. The Bhapkar and Somes approximation is not available because it has only been developed and validated for binary data. The χ^2_{t-1} approximation is still available even if we may prefer something more robust.

Our example here is for two products scaled on three categories. Extensions to more products and/or more categories are possible. See, for example, Rayner and Best (2001, section 6.8)

Suppose we now consider testing for symmetry. We will use the coffee data example. The maximum likelihood estimator $\hat{\mathbf{p}}$ under the null symmetry hypothesis has elements given by

$$\hat{p}_1 = n_1/n,\ \hat{p}_2 = (n_2 + n_4)/(2n),\ \hat{p}_3 = (n_3 + n_7)/(2n),\ \hat{p}_4 = \hat{p}_2,$$
$$\hat{p}_5 = n_5/n,\ \hat{p}_6 = (n_6 + n_8)/(2n),\ \hat{p}_7 = \hat{p}_3,\ \hat{p}_8 = \hat{p}_6,\ \hat{p}_9 = n_9/n.$$

Clearly, when testing for symmetry no iteration is needed. We find

$$\hat{\mathbf{p}}^{\mathrm{T}} = (4/33, 13/66, 5/66, 13/66, 1/11, 1/11, 5/66, 1/11, 2/33).$$

We can now proceed as before by generating many sets of random multinomial counts using a multinomial distribution with $n = \sum_{i=1}^{9} n_i$ and $\hat{\mathbf{p}}$ as parameters. For each such set of counts calculate the Bowker (1948) statistic

$$B = 2\{(n_2 - n\hat{p}_2)^2/(n\hat{p}_2) + (n_3 - n\hat{p}_3)^2/(n\hat{p}_3) + \{(n_6 - n\hat{p}_6)^2/(n\hat{p}_6)\} = \sum_{i=1}^{9} (n_i - n\hat{p}_i)^2/(n\hat{p}_i)$$

and find a bootstrap p-value as the proportion of the statistics greater than or equal to 6.236, the value of B for the Table 6 data. We find the bootstrap p-value to be 0.091 compared with the usual chi-squared p-value of 0.101. Following Rayner and Thas (2005) we also note that $2(n_2 - n\hat{p}_2)^2/(n\hat{p}_2) = 3.769$ and so this is the major component of the B value for the Table 6 data. The chi-squared p-value for this component is 0.052 with corresponding bootstrap p-value 0.044. As above, as the chi-squared p-value is an approximation and we suggest routinely validating it with the bootstrap p-value,

particularly for sparse tables. Simonoff (2003, p.288) gives a sparse table where use of the chi-squared approximation for the *B* test gives a p-value of 0.545, while our bootstrap p-value is 0.036.

We previously noted that for the Table 6 data coffee Y has a too weak flavour. Although we do not give details here, the test for conditional symmetry as defined, for example, in Simonoff (2003, p.301), is not significant for the Table 6 with a p-value of 0.854.

The reader with an $r \times r$ table with $r > 3$ should be able to adapt the above discussion.

6. Conclusion

We have explained how to obtain a bootstrap p-value for Cochran's *Q* and illustrated its use in examples involving two and three treatments. For some data sets the chi-squared p-values for *Q* commonly given in software can be inaccurate and so our bootstrap p-value can be used to validate the accuracy of the chi-squared p-values. Sometimes an approximation due to Bhapkar and Somes (1977) is needed. We hope that our explanation would guide the reader who had four or more treatments to compare. Brief illustrations of how to employ the same bootstrap technique to cross-classified data were also given. Our bootstrap p-values are not permutation p-values or conditional p-values and rely on few assumptions. Our approach could be labelled as nonparametric. Finally we note that some bootstrap procedures are becoming more accessible and, for example, recent versions of the SPSS statistical software package have some bootstrap procedures available.

References

Bennett, B.M. (1967). Tests of hypotheses concerning matched samples. *Journal of the Royal Statistical Society*, Series B, 29, 468-474.

Berger, R.L. and Sidik, K. (2003). Exact unconditional tests for a 2 × 2 matched pairs design. *Statistical Methods in Medical Research*, 12, 91-108.

Bergsma, W., Croon, M., Hagenaars, J., 2009. *Marginal Models for Dependent, Clustered, and Longitudinal Categorical Data*. New York: Springer.

Bhapkar, V.P. and Somes, G.W. (1977). Distribution of *Q* when testing equality of matched proportions. *Journal of the American Statistical Association*, 72, 658-661.

Bowker, A.H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, 43, 572-574.

Davis, C.S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer.

Lawless, H.T. and Heymann, H. (2010). *Sensory Evaluation of Food*. New York: Springer.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions of percentages. *Psychometrika*, 12, 153-157.

Rayner, J.C.W. and Best, D.J. (2001). *A Contingency Table Approach to Nonparametric Testing*. Boca Raton: Chapman & Hall/CRC.

Rayner, J.C.W. and Thas, O. (2005). More informative testing for bivariate symmetry. *Australian and NZ Journal of Statistics*, 47 (2), 211-217.

Simonoff, J.S. (2003). *Analyzing Categorical Data*. New York: Springer.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42, 412-416.

Suissa, S. and Shusta, J.J. (1991). The 2 × 2 matched pairs trial: exact unconditional design and analysis. *Biometrics*, 47, 361-372.

Wallenstein, S. and Berger, A. (1981). On the asymptotic power of tests for comparing *K* correlated proportions. *Journal of the American Statistical Association*, 76, 114-118.