National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

12-18

A Note on Bowker's Symmetry Statistic, its Components and Bootstrap P-Values

D.J. Best and J.C.W. Rayner

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

A NOTE ON BOWKER'S SYMMETRY STATISTIC, ITS COMPONENTS AND BOOTSTRAP P-VALUES

D.J. BEST[1] AND J.C.W. RAYNER[2,1]

University of Newcastle and University of Wollongong

Abstract

Two statistics for matched pairs categorical, possibly nominal, data with more than two categories, are the Stuart (1955) statistic of marginal homogeneity and the Bowker (1948) statistic for bivariate symmetry. The Stuart statistic is possibly more widely known but the Bowker statistic has easily understood components with approximate χ_1^2 p-values and deserves to be better known and used. Like many similar statistics its approximate χ^2 p-values can be inaccurate for small counts. Here we suggest bootstrap p-values in such cases. We also consider a Wald statistic for symmetry and give examples where use of the Bowker statistic seems to improve on the Stuart statistic. Both these statistics are nonparametric and should be more robust than the parametric alternatives.

Keywords: Categorical data; cross-classified data; Stuart's test.

1. Introduction

A statistic of Bowker (1948) helps examine symmetry in square $r \times r$ contingency tables where row and column variables have the same categories. In this expository note categorical data are considered where two dependent samples are obtained and each observation in one sample matches that in the other. Agresti (2013) and others call this matched pairs data. A square table of matched pairs data is symmetric if $p_{ij} = p_{ji}$ for all $i \neq j$, $i, j = 1, 2, \ldots, r$ where p_{ij} is the probability of an observation belonging to the i, jth cell of the table. Let n_{ij} be the table count for the i, jth cell. Bowker's statistic is defined as

$$X^2 = \sum_{j>i} (n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji}) .$$

[1] School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia

[2] National Institute for Applied Statistics Research Australia, University of Wollongong, NSW 2522, Australia

Rayner and Thas (2005) show that X^2 is a score statistic, with $r(r-1)/2$ components $V_{ij}^2 = (n_{ij} - n_{ji})^2 / (n_{ij} + n_{ji})$, $X^2 = \sum_{j>i} V_{ij}^2$. Bowker's statistic has an asymptotic χ^2 distribution with $r(r-1)/2$ degrees of freedom, $\chi^2_{r(r-1)/2}$. The V_{ij}^2 are themselves score statistics and are asymptotically mutually independent each with asymptotic χ_1^2 distribution. A V_{ij}^2 can be significant even if X^2 is not.

The counts in a square contingency table are often described as cross-classified data. If $n = \sum_i \sum_j n_{ij}$ then the maximum likelihood estimates of the p_{ij} under a null symmetry hypothesis are $\hat{p}_{ii} = n_{ii}/n$ and $\hat{p}_{ij} = (n_{ij} + n_{ji})/(2n)$ for $i \neq j$. As above we emphasise that each count in the square tables we consider results from two correlated responses from the same subject and so it is not appropriate to use the classic chi-squared test for independence for a contingency table. The Bowker statistic is a categorical data analogue of the continuous data paired t-test statistic. In the following we look at sensory evaluation, marketing, traffic noise annoyance and health examples.

2. A taste-test example

We base this example on Gacula et al. (2009, Exercise 9.16, p.454). Suppose in a central location consumer taste-test each of 131 consumers evaluated two products, A and B, for saltiness using a five point just-about-right (JAR) intensity scale from 1 = not at all salty enough, through 3 = just about right to, ultimately, 5 = much too salty. Note that the 1, 3, 5 here are codes, not numerical values. The cross-classified counts are in Table 1.

Table 1
JAR data for the saltiness of two products, A and B

	B1	B2	B3	B4	B5
A1	1	1	1	1	2
A2	1	4	2	13	2
A3	1	2	6	6	2
A4	0	1	2	31	11
A5	0	0	2	8	31

For these data $r = 5$. We find $X^2 = 17.759$ with p-value 0.059 using the χ^2_{10} approximation and $V_{24}^2 = 72/7 = 10.3$ with p-value 0.00014 based on the χ_1^2 approximation. We could list all 10 components but clearly V_{24}^2 is the largest. Component tests may be more sensitive than the X^2 test. The X^2 value is close to significance at the 0.05 level while the V_{24}^2 p-value suggests that when product A is not salty enough then product B is too salty. However some of the cross-classified counts are small and the χ^2 approximations may not be accurate. Computer routines for X^2 often only give the approximate χ^2 p-value which, as noted, may not be accurate. The R routine 'mcnemar_test', for example, only gives the approximate chi-squared probability for X^2.

3. Bootstrap p-values

When there are small counts a better p-value can be obtained using a bootstrap procedure. Suppose we generate many sets, n_s say, of random counts from a multinomial with parameters n and p_{ij}, $i, j = 1, 2, ..., r$. Here we used $n_s = 100{,}000$. For each of the n_s sets generated calculate X^2 and then the bootstrap p-value for X^2 is the proportion of the n_s sets greater than or equal to the X^2 value for the cross- classified data. The bootstrap p-values for the components are defined similarly.

For the example data above the proportion or bootstrap p-value is 0.016 for X^2 and 0.0007 for V_{24}^2. The bootstrap p-value indicates X^2 is significant at the 0.05 level and we can conclude that A is less salty than B with V_{24}^2 the most important component. The generation of the n_s random sets can be quite computer intensive. Here there are 25 probabilities in the multinomial and $n = 131$. In calculating any of the V_{24}^2 for the observed data or the n_s data sets used to get the bootstrap p-value it may happen that $V_{24}^2 = 0/0$: V_{24}^2 is indeterminate. Depending on one's software this may cause a problem. One solution is to take V_{ij} to be zero in such cases. Note that the bootstrap procedure here gives unconditional bootstrap p-values unlike conditional bootstrap p-values which might be based on binomials with probability of success 0.5 and number of trials fixed at $n_{ij} + n_{ji}$ for each V_{ij}^2.

As an extreme example of how the χ^2 approximation can fail when counts are small we consider the Sock Test data for musculoskeletal pain given in Simonoff (2003, p.288). Two therapists categorized 21 subjects with musculoskeletal pain while they stretched to put on their socks. The subjects were from Bergen, Norway, and the results are shown in Table 2 where the categories ranged from 0 = perform task easily, to 3 = can hardly perform task. We find $X^2 = 5.000$ with p-value 0.546 using the χ_6^2 approximation. Simonoff (2003, p.288) also obtained a clearly non-significant p-value using a different statistic and said "there is very strong inter-tester reliability for the Sock Test". However we find a bootstrap p-value of 0.046 casting doubt on the inter-tester reliability.

Table 2

Therapists Categorization Counts

	Therapist 2 categories			
Therapist 1 categories	0	1	2	3
0	5	0	0	0
1	3	5	0	0
2	0	2	3	0
3	0	0	0	3

4. A Special Case

If $r = 2$ the Bowker statistic becomes the McNemar (1947) statistic. There are four counts for $r = 2$, namely n_{11}, n_{12}, n_{21}, n_{22} and only two probabilities, p_{12} and p_{21}, to compare. The McNemar statistic is $X^2 = (n_{12} - n_{21})^2 / (n_{12} + n_{21}) = V_{12}^2$. Again we use an

example from Gacula et al. (2009). Suppose 100 consumer panelists were presented with a food product with brand name and the same product without brand name, and then asked to give their purchase preference. The results are given in Table 3.

Table 3
Purchase preferences of 100 consumer panellists

	Buy (NB)	Not Buy (NB)
Buy (B)	26	23
Not Buy (B)	11	40

In Table 3 NB = no brand shown and B = brand shown. We find $X^2 = 4.235$ with χ_1^2 p-value 0.0396. As above, for small counts the χ^2 approximation is not always accurate.

A more accurate bootstrap p-value can be calculated as above using

$$\hat{p}_{11} = n_{11}/n,\ \hat{p}_{12} = (n_{12}+n_{21})/(2n) = \hat{p}_{21} \text{ and } \hat{p}_{22} = n_{22}/n.$$

A collection of n_s random multinomial counts are generated using these four $\hat{p}_{ij}$ and n as parameters. The proportion of X^2 values from the n_s sets with values greater than or equal to that for the observed data is again the bootstrap p-value. For the brand effect data the X^2 bootstrap p-value is 0.0401 in good agreement with the χ_1^2 p-value. It appears a brand effect exists for the food product.

5. A Wald Statistic

Krampe and Kuhnt (2007) compare a Wald statistic to the Bowker statistic for a number of data sets. The statistic values for their data sets were always close. The Wald statistic is defined as

$$\sum_{j>i} n(n_{ij}-n_{ji})^2 / \{n(n_{ij}+n_{ji}) - (n_{ij}-n_{ji})^2\}.$$

Aside. It is routine to show that the difference between the Wald and Bowker statistics is

$$\sum_{j>i} \frac{(n_{ij}-n_{ji})^4}{n(n_{ij}+n_{ji})^2} + \mathrm{O}(n^{-2})$$

so the numerical proximity is not surprising.

Krampe and Kuhnt (2007) apply this Wald statistic to a sparse data set concerning annoyance caused by traffic noise. The same subjects categorized their annoyance on two occasions, A1 and A2 say, with the cross-classified results in Table 4.

In Table 4 -1 is very low, -2 is low, -3 moderate, -4 high and -5 very high. We find Bowker's statistic is 15.161 and the Wald statistic is 15.245. As in a number of other

data sets examined in Krampe and Kuhnt (2007), the Bowker and Wald statistics give fairly similar values. The χ^2_{10} approximate p-values here are close, being 0.126 and 0.123 respectively. Using the unconditional bootstrap p-values approach of section 2 above we find a more exact p-value of 0.040 for both statistics. These are close to the simulated p-values of Krampe and Kuhnt (2007). Again, as in the taste-test example above, the approximate chi-squared p-values are not significant at the 0.05 level but the more exact bootstrap p-values are.

Table 4
Annoyance cause by traffic noise on two occasions

	A2-1	A2-2	A2-3	A2-4	A2-5
A1-1	51	28	3	0	0
A1-2	15	68	40	5	1
A1-3	0	29	77	21	1
A1-4	0	4	19	80	14
A1-5	0	1	5	26	88

A look at the components V_{ij}^2 can assist in seeing why the Bowker and Wald tests indicate a lack of symmetry. The large V_{45}^2 suggests a bigger drift from very high (26) annoyance to high annoyance rather than the reverse (14). Thus this component suggests a bigger drift to more moderate annoyance (category 3) going from first occasion to second occasion of noise categorization. The large V_{12}^2 suggests a bigger drift from very low (28) to low rather than the reverse (15): drift from low to very low. Again there is a bigger drift to more moderate annoyance (category 3) going from first occasion to second.

6. Better than the Stuart statistic?

Another statistic which is useful for analysis of $r \times r$ square tables is the Stuart (1955) marginal homogeneity statistic which compares marginal row and column totals. Consider the following data from Agresti (2013, Table 11.16). A survey was made of the brand choice for a sample of purchases of instant coffee. At a later time the brand choice of the same brands using the same purchasers was surveyed. The counts are shown in Table 5 where we see there are large frequencies on the diagonal indicating many did not change their purchase brand.

The bracketed (1) and (2) indicate first and second survey. The marginal totals indicate a drift from High Point (171 down to 135) to Sanka (204 up to 231) with the Stuart (1955) statistic having a significant value of 12.256 and χ^2_4 p-value of 0.015. However the components of the X^2 Bowker statistic allow a more objective assessment, rather than just an indication, of the difference between the responses at the two survey times. We find X^2 = 20.412 with χ^2_{10} p-value of 0.026 and bootstrap p-value 0.019. However we also see that V_{13}^2 = 11.951 accounts for over 50% of X^2 and using χ^2_1 has p-

value of less than 0.001, which is highly significant. This is significant evidence that the drift from High Point to Sanka (44) was larger than the drift from Sanka to High Point (17). If we omit V_{13}^2 from X^2 we have $D = X^2 - V_{13}^2 = 8.461$ and using the χ_9^2 approximation D is not significant.

Table 5
Brand choice of five coffees on two occasions

	High Point (2)	Taster's Choice (2)	Sanka (2)	Nescafe (2)	Brim (2)
High Point (1)	93	17	44	7	10
Taster's Choice (1)	9	46	11	0	9
Sanka (1)	17	11	155	9	12
Nescafe (1)	6	4	9	15	2
Brim (1)	10	4	12	2	27

7. Concluding Remarks

Rayner and Thas (2005) showed that X^2 is a score test statistic, gave some powers, a medical example and discussed the use of the components V_{ij}^2 in detail. Our examples here are from the sensory evaluation, marketing and traffic noise areas but, of course, the methods used apply to many other scientific areas. A Wald statistic was defined but gave a similar value to the Bowker statistic. We leave a more comprehensive comparison of the Bowker and Wald statistics to future work. Also a comparison of Krampe and Kuhnt's simulated p-values and our unconditional bootstrap p-values is left for future work. In the annoyance data these two simulated p-values were the same. For smaller data sets we suggest finding p-values using the bootstrap method. The Bowker components V_{ij}^2 can be used to more closely examine the data than a Stuart (1955) statistic for marginal homogeneity.

References

Agresti, A. (2013*). Categorical Data Analysis* (3rd edition). Wiley, New York.

Bowker, A. (1948). A test for symmetry in contingency tables. *J.Amer.Statist.Assoc.*, **43**, 572-574.

Gacula, M., Singh, J., Bi, J. and Altan, S. (2009). *Statistical Methods in Food and Consumer Research* (2nd edition). Academic, New York.

Krampe, A. and Kuhnt, S. (2007). Bowker's test for symmetry and modifications within the algebraic framework. *Computational Statistics & Data Analysis,* **51**, 4134-4142.

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

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

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

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