BIOMETRICS 54, 620–629
June 1998

Testing and Viewing Symmetry in Contingency Tables, with Application to Readers of Fish Ages

Geoffrey T. Evans* **and John M. Hoenig**[1]

Department of Fisheries and Oceans, Science Branch,
P.O. Box 5667, St. John's, Newfoundland A1C 5X1, Canada

SUMMARY

If there are several methods for assigning an integer score to something and the true value is unknown (or even nonexistent), then one can compare the methods only with each other. We develop a new approach to detecting nonrandom differences among methods based on subtracting the smallest reading from all the readings on a specimen before combining counts into a contingency table. When there are three methods, the approach projects the cube of possible triples of scores into a regular hexagon. It conceals some information, but patterns that survive projection can become easier to detect both formally and visually. Summarizing data by projection may be necessary for achieving sufficient power to detect that methods are not equivalent. We illustrate with data on age determination of scallops from shell markings.

1. Introduction

We wish to compare methods (tests, devices, experimenters) for assigning an integer value to something. For example, the age of many animals can be estimated by counting periodic growth marks in skeletal hard parts (otoliths of fish, shells of mollusks). However, observers may interpret the evidence differently if spawning activity or environmental perturbations produce marks that can be confused with true age marks. Different methods for preparing the specimens may also produce different results. The aging example prompted this work and will influence its terminology; similar problems might arise in comparing intelligence tests or determining if different analysts provide comparable assessments of anxiety or depression. If differences among readers cannot be eliminated, it is desirable that the methods at least be symmetric or interchangeable. Otherwise, if one reader is replaced part way through a time series by another, nonequivalent reader, readings in the second part of the series will not mean the same as readings in the first part. This paper is about tests for symmetry, especially testing among three readers. The null hypothesis is that the result 'A reads a and B reads b' is as likely as 'A reads b and B reads a.'

In comparisons of age readers, the number of otoliths or shells read is typically not large, and this often leads one to pool cells in the resulting contingency table. In Section 2, we consider how to compare two readers. We describe two well-known tests in a way that naturally suggests a new test, intermediate between them, based on pooling cells whose signed differences in age readings are equal. The power to detect age-independent differences among methods is enhanced by this pooling. In Section 3, we consider how to compare three readers and develop generalizations of the test statistics of Section 2. Here the pooled statistics have visual as well as potential analytic advantages. Age determinations of Iceland scallops are considered as an example in Section 4. Section 5 contains a small step toward comparing four readers. Section 6 explores the possibility of determining whether errors tend to occur in a particular direction.

---

[1] *Present address:* Virginia Institute of Marine Science, P.O. Box 1346, Gloucester Point, Virginia 23062, U.S.A.

* *Corresponding author's email address:* evans@athena.nwafc.nf.ca

Key words: Bowker's test; Interchangeability of subscripts; McNemar's test; Test of symmetry; Three-way contingency table.

2. Two Readers

If two readers read the same collection of individually identified otoliths, we can summarize the results in a contingency table whose (a, b) cell contains the number n_{ab} of otoliths assigned age a by reader A and age b by reader B. We condition the analysis on $n_{ab} + n_{ba}$, the total in the cells to be compared (Bowker, 1948; Hettmansperger and McKean, 1973). We wish to test the null hypothesis H_0 that n_{ab} and n_{ba} have the same expected value, i.e., that the table is symmetric (in the sense of matrix symmetry).

Bowker (1948) tested for symmetry with the statistic

$$\sum_{b=1}^{N-1} \sum_{a=b+1}^{N} \frac{(n_{ba} - n_{ab})^2}{n_{ba} + n_{ab}}, \tag{1}$$

where possible readings are scaled to range from 1 to N. If all sums $n_{ab} + n_{ba}$ are large, (1) has an asymptotic χ^2 distribution under H_0 with $N(N-1)/2$ degrees of freedom. When $n_{ab} + n_{ba}$ is small, the χ^2 approximation breaks down. Any pair for which $n_{ab} + n_{ba} = 0$ contributes neither to the statistic nor to the degrees of freedom (Hoenig, Morgan, and Brown, 1995). Small or zero $n_{ab} + n_{ba}$ could occur because of the small sample size (relative to the number of ages) typical in fish aging studies, but they are also likely to occur in large samples—we would hardly expect one reader to assign age 1 and another age N to the same specimen.

A (maximally) pooled test compares the sum of all elements on one side of the main diagonal with the sum of the elements on the other side. The pooled test statistic

$$\frac{\left(\sum_{b=1}^{N-1} \sum_{a=b+1}^{N} (n_{ba} - n_{ab}) \right)^2}{\sum_{b=1}^{N-1} \sum_{a=b+1}^{N} (n_{ba} + n_{ab})} \tag{2}$$

has an asymptotic χ^2 distribution under H_0 with 1 degree of freedom (assuming there are any off-diagonal readings at all) (Hettmansperger and McKean, 1973; Bishop, Feinberg, and Holland, 1975; generalized from McNemar, 1947). If (2) is significant, then we can reject the null hypothesis.

For what follows, it is useful to introduce a new variable $p = a - b$ representing the difference between readings. We can then rewrite the Bowker and maximally pooled test statistics as

$$\sum_{p=1}^{N-1} \sum_{a=1}^{N-p} \frac{(n_{p+a,a} - n_{a,p+a})^2}{n_{p+a,a} + n_{a,p+a}}, \qquad \frac{\left(\sum_{p=1}^{N-1} \sum_{a=1}^{N-p} (n_{p+a,a} - n_{a,p+a}) \right)^2}{\sum_{p=1}^{N-1} \sum_{a=1}^{N-p} (n_{p+a,a} + n_{a,p+a})}, \tag{3}$$

respectively. The inner sums are now performed along a direction parallel to the main diagonal of the table instead of, as before, parallel to one of its sides.

Suppose we believe that the size of errors, when they occur, is nearly independent of the true state of the specimen. If there are few enough observations that pooling is indicated, we might then choose to pool cells with the same value of p. We can express this step both algebraically and geometrically. In terms of formal algebra, two sorts of operations are involved in forming the test statistics: comparing (forming a weighted squared deviation between observed and expected values) and summing. In the Bowker statistic, comparing is done first; in the pooled statistic, it is done last. It is then natural to perform the comparison after the first sum but before the second to obtain the diagonally projected test statistic

$$\sum_{p=1}^{N-1} \frac{\left(\sum_{a=1}^{N-p} (n_{p+a,a} - n_{a,p+a}) \right)^2}{\sum_{a=1}^{N-p} (n_{p+a,a} + n_{a,p+a})}. \tag{4}$$

Viewed geometrically, the inner sums parallel to the diagonal project the matrix into a line segment perpendicular to the diagonal (Figure 1). Observations on this line segment above the diagonal are compared with corresponding observations below. The statistic has an asymptotic χ^2 distribution

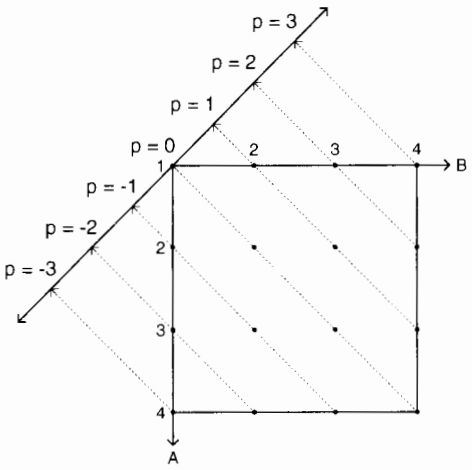

Figure 1. A new test for symmetry is based on projecting the data in a square table onto a line segment perpendicular to the main diagonal. The index $p = b - a$ is the difference between the readings. The test consists of comparing sums along the lines $p = i$ and $p = -i$ for $i > 0$.

under H_0 with degrees of freedom equal to the number of terms in the sum of squares for which the denominator is positive. There are potentially $N - 1$ degrees of freedom. If large errors are rare so that few pairs of readings differ by more than one unit, then there will be few nonzero cells more than one unit from the main diagonal, and the diagonally projected test will be almost the same as the maximally pooled test.

Tests with minimal pooling are able to detect more differences in principle when sample sizes are large. On the other hand, pooled tests are more able to detect differences in practice when sample sizes are small, if the differences reinforce each other upon pooling. Consider the tables

$$X = \begin{pmatrix} 0 & 0 & 0 \\ 4 & 0 & 4 \\ 0 & 0 & 0 \end{pmatrix}, \qquad Y = \begin{pmatrix} 0 & 2 & 1 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix}, \qquad Z = \begin{pmatrix} 0 & 3 & 0 \\ 0 & 0 & 3 \\ 1 & 0 & 0 \end{pmatrix}.$$

Only Bowker's test will detect the asymmetry of X (Table 1) because the individual differences are large but cancel each other with either method of pooling. Only the maximally pooled test will detect the asymmetry of Y because the individual differences are small but reinforce each other upon pooling. For Z, the differences reinforce each other under diagonal projection but interfere with each other under maximal pooling: only the diagonally projected test will detect asymmetry.

Table 1

Tests for the contingency tables described in the text. p represents the probability that would be calculated assuming a χ^2 distribution, which is suspect with such small numbers. Exact probabilities (computed as described in the text) that are <0.05 are in boldface type.

Table		Bowker	Projected	Pooled
X	χ^2	8	0	0
	d.f.	2	2	1
	p	0.0183	1	1
	Exact p	**0.0156**	1	1
Y	χ^2	6	6	6
	d.f.	3	2	1
	p	0.1116	0.0498	0.0143
	Exact p	0.125	0.0625	**0.03125**
Z	χ^2	7	7	3.57
	d.f.	3	2	1
	p	0.0719	0.0302	0.0588
	Exact p	0.125	**0.03125**	0.125

Because the cells have small expected values, it is prudent to compute the exact probability of an outcome whose test statistic is at least as high as that of the example and not just the approximate χ^2 probability. For example, the diagonally projected test of Y has two comparisons, 5 vs. 0 and 1 vs. 0, making the test statistic (4) equal to 6. The only other ways to make the statistic ≥ 6 entail replacing $\{5,0\}$ by $\{0,5\}$ or $\{1,0\}$ by $\{0,1\}$, and so the probability of a test statistic at least that large is $(.5^5 + .5^5) \times (.5 + .5) = .0625$. Thus, the test does not suggest asymmetry (at the 5% level), although the χ^2 approximation did.

3. Three Readers

When there are three readers, two problems arise. First, it is harder to visualize which cells in the contingency table are to be compared. Secondly, the extra space in three dimensions means that there will be many more empty or nearly empty cells that vitiate Bowker's test and make pooling more necessary. The diagonal projection idea extends naturally to three readers and helps with both of these problems. It projects the cube of the three-way contingency table into a regular hexagon in which corresponding cells are easily identified (Figure 2).

In generalizing the Bowker statistic to three readers, Haberman (1978) took the null hypothesis of interchangeability of readers to mean that the result 'A reads a, B reads b, C reads c' is equally probable for any permutation of the readings a, b, and c. Writing test statistics calls for some preliminary investment in notation. Let $\pi(a,b,c)$ denote the set of permutations of $\{a,b,c\}$ and $\Pi(a,b,c)$ the number of such permutations, six if all indices are distinct and three if two indices are the same. Let m_{abc} be the mean of n_{abc} and its $\Pi - 1$ permuted kin and $d_{abc} = n_{abc} - m_{abc}$

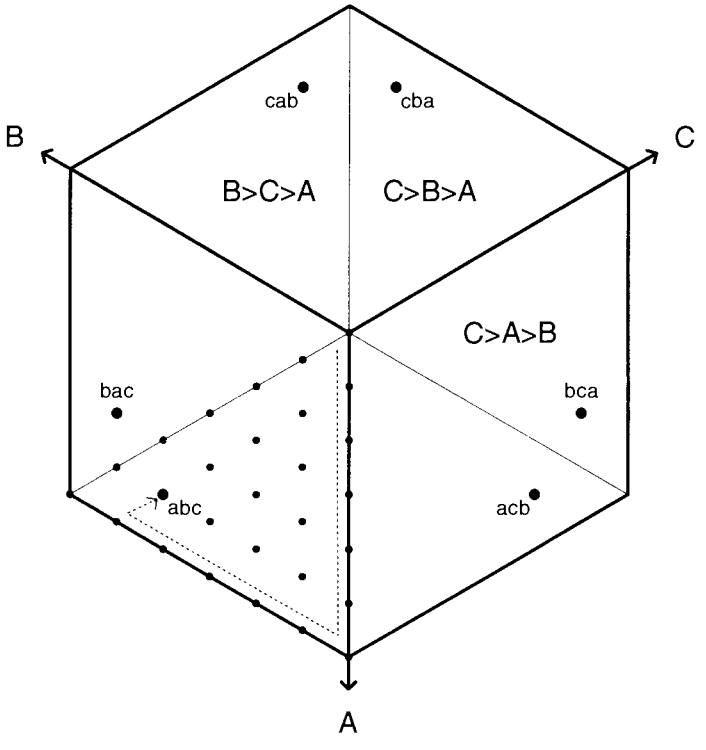

Figure 2. Representations of triples of age readings for three readers, A, B, and C. This figure can be seen as both a cube and its projection onto a hexagon. Plotting a point in a cube means following the directions of the three axes of the cube the appropriate distances. The same is true of plotting a point in the two-dimensional projection of the cube, following the projected axes. (Readers should therefore see the plotting path in Figure 2 both as along three mutually orthogonal directions in three-space and as along three directions mutually at 120° in two-space.) The dotted lines indicate how to plot the point ($a = 6$, $b = 5$, $c = 1$). Plotting the permuted kin means taking the same three distances but permuting the axes along which they are plotted. Notice that plotting a point whose three components are all equal just follows a triangle back to the starting point. It is therefore easier in practice to subtract the smallest reading of the triple from all three readings, leaving just two positive displacements to plot. Any point with $a - c = 5$ and $b - c = 4$ would be plotted at the same place.

the deviation from this mean. The Bowker test statistic is

$$\sum_{c=b}^{N}\sum_{b=a}^{N}\sum_{a=1}^{N-1}\frac{\sum_{\pi}\left(d_{c,b,a}\right)^2}{m_{c,b,a}}, \tag{5}$$

where the sum is performed only for cells such that $c > a$. Making substitutions $p = c - a$ and $q = b - a$ to prepare for diagonal projections, the Bowker statistic is also

$$\sum_{p=1}^{N-1}\sum_{q=0}^{p}\sum_{a=1}^{N-p}\frac{\sum_{\pi}\left(d_{p+a,q+a,a}\right)^2}{m_{p+a,q+a,a}}. \tag{6}$$

The summation over a is along the direction parallel to the main diagonal of the cube of possible observations, and p is the difference between the largest and smallest readings. The index q indicates a direction along which the difference between maximum and minimum readings is constant. As before, we sum only when the denominator is positive, and each sum over π contributes $\Pi - 1$ degrees of freedom. There are $N^3 - (N+2)(N+1)N/6$ degrees of freedom when all cells are included (Haberman, 1978).

We can equally well regard π as operating on the equivalence class consisting of all points $(a + x, b + x, c + x)$ for arbitrary x. The diagonally projected test statistic analogous to (4) is then

$$\sum_{p=1}^{N-1}\sum_{q=0}^{p}\frac{\sum_{\pi}\left(\sum_{a=1}^{N-p} d_{p+a,q+a,a}\right)^2}{\sum_{a=1}^{N-p} m_{p+a,q+a,a}}. \tag{7}$$

There are $(N-1)(N-2)/2 + 4(N-1)$ degrees of freedom when all cells are included.

Constructing a maximally pooled test analogous to (2) consists of moving the comparison operation all the way to the left. We treat separately the cases where all three readers differ $(0 < q < p)$, where the two lowest agree $(q = 0)$, and where the two highest agree $(q = p)$, obtaining the pooled statistic

$$\frac{\sum_{\pi}^{6}\left(\sum_{p=1}^{N-1}\sum_{q=1}^{p-1}\sum_{a=1}^{N-p} d_{p+a,q+a,a}\right)^2}{\sum_{p=1}^{N-1}\sum_{q=1}^{p-1}\sum_{a=1}^{N-p} m_{p+a,q+a,a}} + \frac{\sum_{\pi}^{3}\left(\sum_{p=1}^{N-1}\sum_{a=1}^{N-p} d_{p+a,a,a}\right)^2}{\sum_{p=1}^{N-1}\sum_{a=1}^{N-p} m_{p+a,a,a}} + \frac{\sum_{\pi}^{3}\left(\sum_{p=1}^{N-1}\sum_{a=1}^{N-p} d_{p+a,p+a,a}\right)^2}{\sum_{p=1}^{N-1}\sum_{a=1}^{N-p} m_{p+a,p+a,a}} \tag{8}$$

with $5 + 2 + 2 = 9$ degrees of freedom if all denominators are positive. Geometrically, the first term compares the sum of the numbers of observations inside a triangle (Figure 2) with the mean of the sums of the six triangle interiors; the second term compares the sums of the observations on the projections of the three positive axes; the third term compares sums on the three negative axes.

There are kinds of asymmetry that can be detected with a maximally pooled test with three readers that could not be with two. Recall table X, which might be obtained if each reader was symmetric about the true value but one was more variable than the other. The diagonally projected and maximally pooled tests could detect no difference. If there are three readers, one of whom is variable whereas the other two are not, then the projected cube looks like

$$\begin{array}{ccccc} & & 0 & 0 & 0 \\ & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 4 & 0 \\ & & 0 & 0 & 0 \end{array}$$

and even the maximally pooled test will detect asymmetry.

4. Example of Scallop Ages

Naidu (1993) compared the results from eight readers of Iceland scallop (*Chlamys islandica*) ages. From the large number of possible comparisons, we chose one set of three readers who were not obviously different at first glance (Table 2). Even when only two readers are compared, few cells have more than one observation, and there are many pairs of the form n_{ab}=1, n_{ba}=0. As we saw in the example at the end of Section 2, a $\{1,0\}$ pair by itself is uninformative about symmetry. Its contribution to χ^2 will always be 1; but to preserve significance (for $\alpha < 1/2$) with the extra degree of freedom, a number >1 must be added to the χ^2 statistic. So when we use the χ^2 approximation, such pairs produce a spurious loss of power. Table 3 shows all pairwise comparisons of readers by diagonal projection. There is an indication that reader A tends to have the highest readings. However, even the projected data have $\{0,1\}$ pairs. Therefore, we analyzed the data three ways: by (1) ignoring $\{0,1\}$ pairs, (2) pooling the data at the ends of the projected vector corresponding to the $\{0,1\}$ pairs, and (3) using the maximally pooled test. No significant differences were detected by any of the tests.

We now consider all three readers in Table 2 together. No triple of ages occurred more than once, so the three-dimensional version of Bowker's test cannot show anything. Figure 3 plots the right-hand, projected half of Table 2. Visually, some patterns emerge from the projection: there are 13 observations for which A and C agree (along the projected B axis, excluding the 5 observations at the origin for which all readers agree), 3 for which B and C agree, and 3 for which A and B

Table 2

Original age determinations and their projections obtained by subtracting the lowest reading for three readers of ages of Iceland scallops (readers 5, 6, and 7 in Table 8 of Naidu (1993)). The six sets of readings shown in boldface type lie at the same place in the hexagon in Figure 3.

Original			Projected		
A	B	C	A	B	C
3	3	3	0	0	0
6	7	6	**0**	**1**	**0**
8	12	8	0	4	0
6	5	7	1	0	2
11	10	11	1	0	1
12	12	12	0	0	0
7	5	7	2	0	2
16	14	17	2	0	3
10	10	10	0	0	0
6	6	7	0	0	1
5	5	5	0	0	0
7	8	7	**0**	**1**	**0**
16	19	16	0	3	0
30	19	22	11	0	3
9	10	9	**0**	**1**	**0**
12	13	13	0	1	1
20	14	17	6	0	3
10	11	10	**0**	**1**	**0**
8	9	8	**0**	**1**	**0**
9	12	9	0	3	0
15	12	14	3	0	2
17	13	16	4	0	3
11	12	11	**0**	**1**	**0**
15	14	14	1	0	0
12	15	12	0	3	0
19	14	14	5	0	0
14	13	14	1	0	1
14	14	13	1	1	0
7	7	8	0	0	1
7	7	7	0	0	0

Table 3

Comparisons of scallop age readings by two readers after projection into a one-dimensional array and after further pooling. The projected data give rise to some $\{1,0\}$ comparisons, which individually provide no information about asymmetry. These cells can be eliminated (truncated case) to increase power. Alternatively, it seems reasonable to pool all $\{1,0\}$ pairs at the extremes of the projection (further pooled case), although in general, and especially in higher dimensions, post hoc decisions about pooling can be problematic.

Projected data (frequency distribution of disagreement)

Reader		Number of pairs for which $i-j$ equals															
i	j	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4
B	A	1					1	1	1	1	2	4	8	7		3	1
B	C									4	3	4	8	7		3	1
C	A				1			1		1		4	18	5			

Projected test statistics

	Truncated			Further pooled		
Comparison	χ^2	d.f.	p	χ^2	d.f.	p
B vs. A	3.82	4	0.43	6.82	5	0.23
B vs. C	3.96	3	0.27	4.96	4	0.29
C vs. A	0.11	1	0.74	3.11	2	0.21

Maximally pooled test statistics

	No. pairs with				
i vs. j	$i>j$	$j>i$	χ^2	d.f.	p
B vs. A	11	11	0	1	1.0
B vs. C	11	11	0	1	1.0
C vs. A	5	7	0.33	1	0.57

agree. Thus, B is the reader most likely to differ from the other two. Also, reader A's discrepancies are the largest.

The projected data have many sets of the form $\{1,0,0,0,0,0\}$ or $\{1,0,0\}$, which are not individually informative about symmetry. When we delete these, we are left with four informative sets: $\{6,2,1\}$, $\{2,1,1\}$, $\{3,0,0\}$, and $\{1,1,0,0,0,0\}$. The diagonally projected statistic (7) has $\chi^2 = 4.667 + 0.5 + 6 + 4 = 15.167$; d.f. $= 11$; $p = 0.175$. This is based on 18 scallop shells. The maximally pooled test allows us to use all 25 shells for which there was some difference among the 3 readings. The three sets of the maximally pooled test are $\{10,2,2\}$, $\{1,1,3\}$, and $\{0,0,4,2,0,0\}$; the pooled statistic (8) has $\chi^2 = 9.143 + 1.6 + 14 = 24.743$; d.f. $= 9$; $p = 0.0033$. Thus, the pooled test provides strong evidence that the three readers considered are not identical, although separate two-way tests did not. The p values obtained here are so different from 0.05 that we are not worried about the approximate nature of the χ^2 probabilities.

5. Four Readers

Projecting the hypercube of four readers along its main diagonal produces a rhombic dodecahedron in three dimensions. (More careful thought indicates that it should be considered as a figure with 24 faces: each rhombus divided into two triangles.) Including possibilities of equalities between readings, even for the maximally pooled test, there are 77 comparisons to be considered. This suggests that only the most aggregated test will be powerful enough to detect anything with the number of otolith readings we are ever likely to have. The test will be more useful in cases where large numbers of subjects can be examined or tested.

6. Bias and Asymmetric Errors

6.1 *Two Readers*

Although equal bias in both readers does not produce an asymmetric table, it may sometimes be detectable. Suppose most readings are correct and errors tend to be in one direction and are inde-

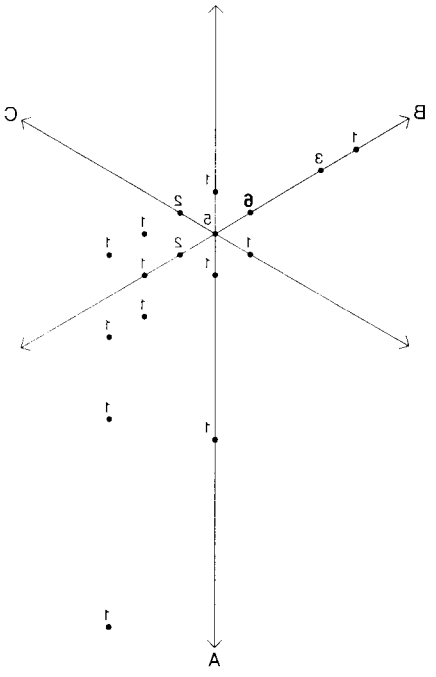

Figure 3. Data on the ages of 30 Iceland scallops determined by three readers (see Table 2) projected along the main diagonal. The six scallops plotted 1 year from the origin along the B axis (shown in bold type) are identified in Table 2. Shading indicates regions where the most different reading is higher than the other two. It has been added to show construction of the pooled bias statistic (expression (9)): the total number of scallops inside the shaded region (15) is compared to the number outside the shaded region (8). Observations on the boundaries between the two regions are not used.

pendent so that coincident errors are rarer still. (Independence is quite a strong assumption; we might normally expect the occasional misleading otolith that fools all readers the same.) Let us say that the age associated with the cell n_{ab} is $(a + b)/2$. Then the mean age of cells for which readings differ, off the main diagonal, will differ from the mean of cells for which the readings agree. Suppose, for example, that any reader errs one third of the time and all errors overestimate by 1. For a set of 36 otoliths all of age 2, we would expect the table

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 16 & 8 \\ 0 & 8 & 4 \end{pmatrix}.$$

The mean age of diagonal entries is $(16 \times 2 + 4 \times 3)/(16 + 4) = 2.2$; the mean age of off-diagonal entries is 2.5. If, on the other hand, errors are equally likely to be 1 year too high or 1 year too low, then for the 36 otoliths, we would expect the table

$$\begin{pmatrix} 1 & 4 & 1 \\ 4 & 16 & 4 \\ 1 & 4 & 1 \end{pmatrix},$$

for which the mean age of diagonal and off-diagonal entries are both 2.

A measurement scale that is closed at one end (e.g., fish that cannot be less than age zero (first year of life)) is one possible source of asymmetric errors. However, the test is not very specific. The table of expected values

$$\begin{pmatrix} 16 & 0 & 0 & 8 \\ 0 & \underline{0} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & 4 \end{pmatrix},$$

where the underscore denotes the true age, is generated by an observational process without bias (for each reader the mean age assigned is 2), but the mean ages of diagonal (8/5) and off-diagonal (5/2) entries still differ. Therefore, we have a way to detect asymmetry of errors; bias is just one common way to produce such an asymmetry.

This test is not very powerful. When many ages are present in the sample, that fact alone makes the sets of diagonal and off-diagonal elements more variable and will make it harder to detect a difference in their means. It is interesting that it is possible in theory to detect asymmetric errors with just two readers.

6.2 *Three Readers*

The concept of the most different reading (the one furthest from the average) makes sense for three readers but not for two, and there is a corresponding symmetry in a hexagon that does not exist in a line. This raises the possibility of detecting bias from an asymmetry under reflection in the lines that bisect opposite sides (Figure 3).

We can test the hypothesis that the most different reading (when there is one) is as likely to be high as low. In the shaded region of Figure 3, the most different reading is the highest. The number of observations in this region is

$$S = \sum_{p=1}^{N-1} \left(\sum_{\pi}^{6} \sum_{q=1}^{\lfloor p/2 \rfloor} \sum_{r=1}^{N-p} n_{p+r,q+r,r} + \sum_{\pi}^{3} \sum_{r=1}^{N-p} n_{p+r,r,r} \right),$$

where $\lfloor p/2 \rfloor$ denotes the greatest integer strictly less than $p/2$. The number of observations in the unshaded region, where the most different reading is the lowest, is

$$U = \sum_{p=1}^{N-1} \left(\sum_{\pi}^{6} \sum_{q=\lceil p/2 \rceil}^{p-1} \sum_{r=1}^{N-p} n_{p+r,q+r,r} + \sum_{\pi}^{3} \sum_{r=1}^{N-p} n_{p+r,p+r,r} \right),$$

where $\lceil p/2 \rceil$ is the least integer $> p/2$. We omit the boundaries along which there is no most different reading. The χ^2 statistic

$$\frac{(U - S)^2}{U + S} \tag{9}$$

has one degree of freedom.

For the scallop example, $S = (1 + 1 + 1) + (6 + 3 + 1) + (2) = 15$ and $U = (1) + (2 + 1 + 1 + 1 + 1) + (1) = 8$ (Figure 3) and $\chi^2 = 2.13$; d.f. $= 1$; $p = 0.14$. This does not provide strong evidence that the most different reading, when there is one, tends to be high.

If errors are rare and independent (which they weren't for the scallop example), then when two of three readers agree, the third is probably wrong. In that case, we can infer the direction of bias; otherwise, we can only determine that bias exists.

7. Discussion

Bowker-type tests can detect any kind of departure from interchangeability of subscripts in principle. In practice, sparseness of the data can result in low power of the test and poor approximation to the asymptotic χ^2 null distribution. One way around these problems is to pool cells. For two-dimensional tables, there is the McNemar-like procedure, which is a maximal pooling. This eliminates the possibility of detecting an important class of alternatives to the null hypothesis: for a difference or pattern to disappear under maximal pooling, the cell probabilities need to satisfy only one condition, i.e., the sum of probabilities above the diagonal is equal to the sum of probabilities below. Thus, for example, no difference can be detected between two readers who have unequal variability but have symmetric errors about the true value, like table X of Section 2.

It is interesting that the maximally pooled test can detect symmetric but unequal variability in three-dimensional tables. For a difference or pattern to disappear under maximal pooling with three readers, the cell probabilities must satisfy nine equations corresponding to the nine degrees of freedom of (8). This can happen, but only rarely. Thus, pooling does not narrow the alternative space as much in three- and four-dimensional tables as in two-dimensional tables.

The new projection statistics are algebraically and geometrically intermediate between the Bowker-type and maximally pooled statistics. From a biological point of view, they are appealing because they are more powerful than Bowker's test and more general than the maximally pooled test for detecting patterns that do not depend on the true value. In the current context, the projected test is ideal for age-independent errors.

Another way to increase power is to use a more efficient experimental design. One could obtain the specimens by simple random sampling, which gives rise to a multinomial distribution. However, the analysis is conditioned on the pair totals (or the total of all the permuted kin) so that the test can be conducted for any collection of specimens selected prior to determining ages. Thus, if age discrepancies are most likely to occur for old animals, which are generally large animals, one could select large animals to be examined by the readers.

A variety of models have been developed to describe various types of symmetry. Often, these involve imposing restrictions on parameters in a hierarchical loglinear model. However, this is not the only approach of interest. It does not appear suitable for modelling expectations of sums of cells (e.g., marginal totals or totals obtained by diagonal projection). Nor does it appear suitable for addressing the question of whether errors or discrepant observations tend to be in a particular direction. The projection technique reduces the dimensionality of the data and makes it easier to assess age-independent discrepancies visually.

At a purely formal algebraic level, swapping the order of summing and comparing suggests other possible test statistics. For example, one might combine the summation variables of (1) with the (sum, compare, sum) order of (4) to form a projected statistic in which the parallel projection lines of Figure 1 are replaced by a sort of herringbone pattern. Or with three readers, one might consider the order (sum, sum, compare, sum) intermediate between (7) and (8). However, in this paper, we have concentrated on those statistics for which formal algebraic possibility was reinforced by biological motivation.

Acknowledgements

We are grateful to Sam Naidu, Michel Giguère, and Marc Lanteigne for allowing us to take an example from their not-yet-published work on comparing methods for aging scallops. Choudary Hanumara, William Warren, Nicholas Barrowman, Colin Greene, and three anonymous reviewers provided helpful comments.

Résumé

Lorsqu'il existe plusieurs façon d'attribuer un score entier a une observation, et que la meilleure est inconnue (voire inexistante), il est possible de comparer les méthodes entre elles. Nous développons une nouvelle approche pour détecter des différences non aléatoires entre des méthodes, basée sur la soustraction de la plus petite observation de toutes les observations d'un spécimen, effectuée avant de combiner les fréquences dans une table de contingence. Pour trois méthodes, cette approche projette le cube des triplets obtenus sur un hexagone régulier. Cette méthode perd une partie de l'information, mais celle qui subsiste est facilement détectable, tant visuellement que formellement. Résumer les observations par projection s'avère nécessaire pour obtenir la puissance nécessaire à détecter que différentes méthodes ne sont pas équivalentes. Nous l'illustrons par des données relatives a la détermination de l'âge par marquage de la coquille chez des mollusques bivalves.

References

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Massachusetts: MIT Press.

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

Haberman, S. J. (1978). *Analysis of Qualitative Data*, Volume I. New York: Academic Press.

Hettmansperger, T. P. and McKean, J. W. (1973). On testing for significant change in C × C tables. *Communications in Statistics* **2,** 551–560.

Hoenig, J. M., Morgan, M. J., and Brown, C. A. (1995). Analyzing differences between two age determination methods by tests of symmetry. *Canadian Journal of Fisheries and Aquatic Sciences* **52,** 364–368.

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

Naidu, K. S. (1993). *Report of the Workshop on Age Determination of Iceland scallops* (Chlamys islandica), Institut Maurice Lamontagne, Mont Joli, Quebec, January 14–15, 1993. Available from K. S. Naidu, Department of Fisheries and Oceans, Science Branch, St. John's, Newfoundland, Canada.

Received May 1996; *revised January and July* 1997; *accepted August* 1997.