Nonparametric Comparison of Two Histograms

D. J. Best


Stable URL: 
http://links.jstor.org/sici?sici=0006-341X%28199406%2950%3A2%3C538%3ANCOTH%3E2.0.CO%3B2-D

*Biometrics* is currently published by International Biometric Society.
THE CONSULTANT’S FORUM

Nonparametric Comparison of Two Histograms

D. J. Best
Biometrics Unit, CSIRO Food Research Laboratory, P.O. Box 52, N. Ryde 2113, Australia

Summary

Scholz and Stephens (1987, Journal of the American Statistical Association 82, 918–924) proposed a nonparametric $k$-sample Anderson–Darling statistic for grouped data. This note demonstrates that a partition-of-$\chi^2$ method may give a more powerful nonparametric test, particularly when alternatives other than location shift are important. A taste-test example provides motivation.

1. Introduction

At the CSIRO Food Research Laboratory, the Japan Project was set up to look at, among other things, how Japanese and Australian consumers rated various sweet foods. Details concerning the taste-test methodology used are given in Laing et al. (1994). For Japanese chocolate, the data in Table 1 were obtained. A score of 7 indicates a consumer thought the sweetness to be very good, whereas a score of 1 indicates a consumer did not like the sweetness at all. For this $2 \times 7$ table the usual Pearson's chi-squared statistic for independence is 10.70, with approximate $(\chi^2)$ $P$-value of .10. A Student’s $t$-test for the difference between Australian and Japanese mean scores gives $P = .92$. A Mann–Whitney nonparametric test gives $P = .44$. Both “central tendency” tests indicate no difference in scores for the two consumer panels.

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese chocolate responses</td>
</tr>
<tr>
<td>Sweetness liking score</td>
</tr>
<tr>
<td>Consumers</td>
</tr>
<tr>
<td>Australian</td>
</tr>
<tr>
<td>Japanese</td>
</tr>
</tbody>
</table>

However, histograms based on the Table 1 data suggest—as did the chi-squared test—the possibility of a “spread” difference. Such a difference might indicate a difference in market segmentation. For example, the scores might be spread because young consumers liked the sweetness and older consumers did not. Although mean scores are similar, a significant number of Australian consumers do not like the sweetness of Japanese chocolate.

Comparing two histograms is the same as comparing two distributions of grouped data or looking at homogeneity of two samples of grouped data. This note compares two relevant homogeneity tests. Another approach would be to follow the methods in Chapter 5 of McCullagh and Nelder (1989) but attention here will be given only to nonparametric methods which make fewer model assumptions.

In textbooks, the most commonly described nonparametric statistic for comparing two distributions appears to be the Kolmogorov–Smirnov statistic. See, for example, a grouped data application given as Example 6.6b in Siegel and Castellan (1988). For the Japan Project data, the Kolmogorov–Smirnov statistic is $D_{33,31} = .20$ with $P = .29$ based on the chi-squared approximation given by equation (6.19) of Siegel and Castellan (1988).

Key words: Anderson–Darling statistic; Homogeneity of fit; Partition of chi-square; Rank tests.
Scholz and Stephens (1987, §5) suggest testing homogeneity of grouped data via the Anderson-Darling statistic. They give a statistic $A^2_{2N}$ in their equation (6) and suggest that critical values will be the same as for ungrouped data. Simulations below investigate this point. For the Japan Project data given above, define $k = \text{number of rows} = 2$ and $N = \text{total number of observations} = 64$. Thus, $A^2_{2N} = 2.27$ which, using Table 1 of Scholz and Stephens (1987), gives $0.10 > P > 0.05$. This $P$-value is smaller than any produced so far but still may not be regarded as significant.

Another technique for comparing two histograms is to use the location and scale (or spread) tests given in Section 5 of Nair (1986). These tests can be regarded as extensions of the one-sample goodness-of-fit test for grouped ordinal data given by Best and Rayner (1987, 1991) and Section 5.3 of Rayner and Best (1989), who used “natural number” integer scores rather than midranks. This test relied on partitioning the usual $x^2$ statistic into relevant components of location, spread, etc., and so Nair’s technique could be called a partition-of-$x^2$ method. Details are given in the Appendix. Pettitt (1976) discussed statistics identical to Nair’s for two-sample continuous data, whereas Boos (1986) gave similar statistics, also for continuous data, for the $k$-sample case.

For the Japan Project data the partition-of-$x^2$ method, using midrank scores, gives the analysis shown in Table 2. This suggests there is indeed a spread effect. The $P$-values shown are based on 10,000 Monte Carlo simulations. If the residual had been significant then perhaps higher-order partitions would have been interesting. Boos (1986) looked at skewness and kurtosis partitions in the continuous data case. If the independence hypothesis is rejected, then density estimates or fitted values similar to those defined in Best and Rayner (1991, p. 591) are available for the partition-of-$x^2$ method. The Rank Stat (1994) package does relevant calculations.

Which statistic should be used for comparing two histograms? The following section uses the algorithm of Patefield (1981) to compare sizes and powers for the $A^2_{2N}$ and partition-of-$x^2$ methods for selected tables with fixed marginal totals. Such conditional inference is conventional for rank tests such as the Wilcoxon and Mood tests when data are tied and, as Pettitt (1976) and Nair (1986) respectively note, these well-known two-sample rank tests are intimately related to both the Anderson–Darling (A–D) and partition-of-$x^2$ tests. Powers were also calculated for the two-sample Kolmogorov–Smirnov tests but these were not competitive and are not shown.

### 2. Size and Power Comparisons

This section compares the A–D and partition-of-$x^2$ tests via powers. To make sure these power comparisons are “fair” it was first checked that the critical values used give fairly similar sizes. For the A–D test we used critical values from Table 1 of Scholz and Stephens (1987) and for the partition-of-$x^2$ test we used $x^2$ distribution critical values. In practice, we suggest that use of exact or almost exact Monte Carlo $P$-values is preferable to use of tables of critical values. Section 5.7 of the StatXact (1991) manual describes how to do this for the location component of the partition-of-$x^2$ method, whereas the Rank Stats (1994) package gives such a $P$-value for both location and spread components. Sizes based on 10,000 simulations were calculated and as these were quite similar and close to their nominal values, it is sensible to calculate powers based on the same critical values as used to obtain sizes. The random $2 \times c$ tables ($c = 5, 7, 9$) were generated using the algorithm of Patefield (1981).

To produce tables with fixed margins, alternatives were produced by moving columns in the random $2 \times c$ tables ($c = 5, 7, 9$) so that (i) in row 1 the column entries are in descending order, (ii) in row 1 the column entries are biggest in the centre and smallest in columns 1 and $c$. Table 3 gives powers based on 10,000 simulations for these alternatives for the case of equal column totals. Comparisons, not using fixed margins, may be relevant but these are not considered here.

In Table 3 we see that the location component, $V^1_1$, say, of $x^2$ is good at detecting alternative (i) and the spread component, $V^2_1$ say, is good at detecting alternative (ii), whereas $V^1_1 + V^2_1$ has reasonable power for both sorts of alternative. The A–D statistic is good for the linear alternative (i)
Table 3

Powers for A–D and partition-of-$\chi^2$ statistics

(i) Linear alternative, equal column totals, and $\alpha = .05$

<table>
<thead>
<tr>
<th># categories</th>
<th>N</th>
<th>$X^2$</th>
<th>Location ($V_1^2$)</th>
<th>Spread ($V_2^2$)</th>
<th>$V_1^2 + V_2^2$</th>
<th>A–D</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>50</td>
<td>.06</td>
<td>.46</td>
<td>.00</td>
<td>.18</td>
<td>.31</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>.05</td>
<td>.62</td>
<td>.00</td>
<td>.42</td>
<td>.50</td>
</tr>
<tr>
<td>9</td>
<td>90</td>
<td>.04</td>
<td>.82</td>
<td>.00</td>
<td>.64</td>
<td>.73</td>
</tr>
</tbody>
</table>

(ii) Quadratic alternative, equal column totals, and $\alpha = .05$

<table>
<thead>
<tr>
<th># categories</th>
<th>N</th>
<th>$X^2$</th>
<th>Location ($V_1^2$)</th>
<th>Spread ($V_2^2$)</th>
<th>$V_1^2 + V_2^2$</th>
<th>A–D</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>50</td>
<td>.06</td>
<td>.00</td>
<td>.25</td>
<td>.15</td>
<td>.01</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>.05</td>
<td>.00</td>
<td>.40</td>
<td>.20</td>
<td>.01</td>
</tr>
<tr>
<td>9</td>
<td>90</td>
<td>.04</td>
<td>.00</td>
<td>.45</td>
<td>.23</td>
<td>.02</td>
</tr>
</tbody>
</table>

but not very powerful for the quadratic alternative (ii). We would expect that $V_1^2 + V_2^2$ would perform very similarly to $S$ in Pettitt (1976). See his Figure 1 and associated discussion. The performance of $V_1^2$, $V_2^2$, $V_1^2 + V_2^2$ is similar to the performance of these statistics in the multinomial or $1 \times c$ case discussed in Best and Rayner (1987, 1991) and Section 5.3 of Rayner and Best (1989).

Although the power comparisons of Table 3 are very limited, they do illustrate that use of the A–D statistic cannot always be recommended. Similar comments are likely to pertain for the comparison of $k$ histograms and so the partition-of-$\chi^2$ method is preferred.

Acknowledgements

The author would like to thank an editor and the referee for helpful comments.

Résumé


References

Rank Stats (1994). Nonparametric statistics for sensory evaluation. CSIRO Biometrics Unit, P.O. Box 52, North Ryde, Australia.
Suppose we have a one-way table of counts with associated multinomial probabilities \( \{p_j\} \). Define \( \{g_j(x_j)\} \) to be the set of polynomials orthogonal on \( \{p_j\} \), where \( x_j \) is a “score” associated with the \( i \)th count. If \( c \) is the number of categories in the one-way table, we take

\[ g_0(x_j) = 1 \]

and

\[ g_1(x_j) = A(x_j - S_1), \quad g_2(x_j) = C(x_j^2 - A^2 Y x_j + Z), \]

in which

\[ S_i = \sum_j s_j^i p_j, \quad A = (S_2 - S_1^2)^{-\frac{1}{2}}, \quad Y = S_3 - S_1 S_2, \quad Z = A^2 Y S_1 - S_2, \]

\[ C = (S_4 + A^4 Y^2 S_2 + Z^2 - 2 A^2 Y S_3 + 2 Z S_2 - 2 A^2 Y Z S_1)^{-\frac{1}{2}}. \]

Suppose that \( \{n_j\} \) are the counts and that \( n = \sum_{j=1}^c n_j \). Then the usual Pearson \( X^2 \) statistic is

\[ X^2 = \sum_{j=1}^c (n_j - np_j)^2/(np_j). \]

It is easy to show as, for example, in Lancaster (1953), that

\[ X^2 = \sum_{\ell=1}^c U_{\ell}^2, \]

in which

\[ U_{\ell} = \sum_{j=1}^c n_j g_{\ell}(x_j)/\sqrt{n} \quad \text{for } \ell = 1, \ldots, c - 1. \]

Suppose now that we have a \( k \times c \) two-way table of counts, \( n_{ij} \), say, where the column categories are ordered and have associated scores \( \{x_j\} \). Putting \( e_{ij} = (S_{i-1}^k n_{ij})/(\sum_{j=1}^c n_{i-1} n_{ij})/N \), where \( N = \sum_{i=1}^k \sum_{j=1}^c n_{ij} \), the usual Pearson \( X^2 \) statistic is

\[ X^2 = \sum_{i=1}^k \sum_{j=1}^c (n_{ij} - e_{ij})^2/e_{ij}. \]

This \( X^2 \) statistic can be partitioned by calculating \( U_{ij} \), that is, by calculating \( U_{\ell} \) as above for each row, using \( p_{ij} = (S_{i-1}^k n_{ij})/n \) for \( j = 1, \ldots, c \). The statistic \( V_{1}^2 \) is equal to \( \sum_{j=1}^c U_{1j}^2 \) and \( V_{2}^2 \) is equal to \( \sum_{j=1}^c U_{2j}^2 \) when the \( \{x_j\} \) are midranks. As each row of the two-way table can be partitioned using the \( U_{ij} \), it follows that \( X^2 = \sum_{i=1}^k \sum_{j=1}^c U_{ij}^2 \). Notice in passing that \( (N - 1)V_{1}^2/N \) is the usual Kruskal–Wallis statistic adjusted for ties. Nair (1986) has given an alternative technique for calculating \( V_1^2 \) and \( V_2^2 \) but the above formulae clearly show the partition-of-\( X^2 \) link. For \( k > 2 \) a scatter plot of \( (U_{1i}, U_{2i}) \), \( i = 1, \ldots, k \) can sometimes provide a useful data summary.