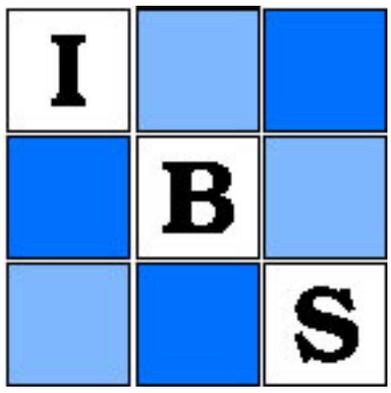

BIOMETRICS 39, 505–510
June 1983

THE CONSULTANT'S FORUM

Testing Marginal Homogeneity for Ordinal Categorical Variables

Alan Agresti
Department of Statistics, University of Florida, Gainesville, Florida 32611, U.S.A.

SUMMARY

The standard chi square tests of marginal homogeneity take no account of possible category ordering. We describe several strategies that make use of category order and which tend to yield more powerful tests for certain common alternatives.

1. Introduction

Consider a two-way contingency table with naturally ordered row categories that are identical to the column categories. Denote the unknown cell probabilities in the $r \times r$ table by $\{p_{ij}\}$ and let $\{p_{ij}\}$ be the observed cell proportions based on a sample of size n. We will consider methods of using the sample data to compare the marginal distributions $\{p_{i+}\}$ and $\{p_{+i}\}$.

Stuart (1955) proposed a test of marginal homogeneity, with $H_0: p_{i+} = p_{+i}$, $i = 1, \ldots, r$. He used the test statistic $Q = n\mathbf{d}'\hat{\mathbf{V}}^{-1}\mathbf{d}$, where $\mathbf{d}' = (d_1, \ldots, d_{r-1})$ with $d_i = p_{i+} - p_{+i}$ and where $\hat{\mathbf{V}}$ is the maximum likelihood estimate of the null covariance matrix of $n^{1/2}\mathbf{d}$. The elements of $\hat{\mathbf{V}}$ are $\hat{V}_{ij} = -(p_{ij} + p_{ji})$ for $i \neq j$ and $\hat{V}_{ii} = p_{i+} + p_{+i} - 2p_{ii}$. For random samples, the asymptotic joint normality of $\mathbf{d}$ induces a null asymptotic chi square distribution for Q, with $r - 1$ degrees of freedom (df). Very similar tests have been proposed by other authors. Bhapkar (1966, 1979) gave a Wald-type statistic which is asymptotically equivalent to Stuart's test statistic since it has the same form but uses the estimated non-null covariance matrix $\hat{\mathbf{W}}$ of $\mathbf{d}$. In Bhapkar's statistic, $\hat{w}_{ij} = -(p_{ij} + p_{ji}) - (p_{i+} - p_{+i})(p_{j+} - p_{+j})$ for $i \neq j$, and $w_{ii} = p_{i+} + p_{+i} - 2p_{ii} - (p_{i+} - p_{+i})^2$. See also Grizzle, Starmer and Koch (1969) and Koch *et al.* (1977) for more details of this approach. Ireland, Ku and Kullback (1969) used minimum discrimination information estimation to find an iterative solution for expected cell frequencies that satisfy marginal homogeneity and to obtain another χ^2_{r-1} statistic. Bishop, Fienberg and Holland (1975, pp. 294–5) suggested a χ^2_{r-1} statistic based on obtaining maximum likelihood estimates of expected frequencies that satisfy marginal homogeneity.

These chi square tests of marginal homogeneity are invariant to any like permutations of the variable categories. If the variable categories are ordered, these tests ignore that information. It is interesting to note, however, that many textbooks on contingency tables illustrate the tests by using tables with ordered categories. These tests are consistent against all alternatives, and it is not incorrect to apply them to tables with ordered categories. Nevertheless, the ordering of the categories is additional information which we can utilize to obtain more powerful tests, at least for certain important alternatives to the null hypothesis. In this paper we describe briefly several strategies that do utilize the ordering. Each method will be illustrated with the vision data in Table 1. The Stuart test yields $Q = 11.96$ based on df = 3 for these data, corresponding to $P = .01$.

Key words: Square contingency tables; Ordered categories; Matched pairs; Chi square statistics; Mann–Whitney test; Wilcoxon test; Ridits; Ordinal models.

Table 1
Unaided distance vision for women; from Stuart (1955)

Grade of right eye	Grade of left eye			
	Highest	Second	Third	Lowest
Highest	1520	266	124	66
Second	234	1512	432	78
Third	117	362	1772	205
Lowest	36	82	179	492

2. Ordinal Strategies

2.1 *Mann–Whitney Test*

Probably in most comparisons of marginal distributions for ordered categorical variables, one is interested in whether one marginal distribution is stochastically larger than the other. We can adapt the Mann–Whitney test to focus on such alternatives.

Let X be selected at random from the marginal distribution $\{\rho_{i+}\}$, and let Y be selected independently at random from the marginal distribution $\{\rho_{+j}\}$. Consider the measure

$$\begin{aligned}\tau &= \text{pr}(Y > X) - \text{pr}(X > Y)\\ &= \sum_{j>i} \rho_{i+}\rho_{+j} - \sum_{i>j} \rho_{i+}\rho_{+j}\\ &= \sum \rho_{+i}\gamma_{i+} - \sum \rho_{i+}\gamma_{+i},\end{aligned}$$

where $\gamma_{i+} = \sum_{a\leqslant i}\rho_{a+}$, $\gamma_{+i} = \sum_{a\leqslant i}\rho_{+a}$ and $\gamma_{0+} = \gamma_{+0} = 0$. When Y is stochastically larger than X, $\tau > 0$, and when X is stochastically larger than Y, $\tau < 0$. Marginal homogeneity implies that $\tau = 0$.

The sample version, $\hat{\tau} = \sum_{j>i} p_{i+}p_{+j} - \sum_{i>j} p_{i+}p_{+j}$, is the difference between discrete analogs of the Mann–Whitney statistics. Several alternative expressions can be given for τ or $\hat{\tau}$, which utilize the equivalent ways of comparing distributions through Mann–Whitney statistics, Wilcoxon-type mean rank statistics, or mean ridit statistics. For example, let $\bar{R}_U(V)$ denote the mean ridit for the distribution of V when the distribution of U is the 'identified distribution' for calculating the ridits (see Bross, 1958). Then

$$\begin{aligned}\tau &= \bar{R}_X(Y) - \bar{R}_Y(X)\\ &= 2\{\bar{R}_Y(Y) - \bar{R}_Y(X)\}\\ &= 2\{\bar{R}_X(Y) - \bar{R}_X(X)\},\end{aligned}$$

where

$$\begin{aligned}\bar{R}_X(X) &= \tfrac{1}{2}\sum \rho_{i+}(\gamma_{i+} + \gamma_{i-1,+})\\ &= .5 = \bar{R}_Y(Y).\end{aligned}$$

Now τ might not seem as relevant as the corresponding measure for a pair, (X_i, Y_i), selected at random for the *joint* distribution. However, marginal homogeneity does not imply that $\text{pr}(Y_i > X_i) = \text{pr}(X_i > Y_i)$ for a matched pair. Also, the difference in probabilities, τ, (or, equivalently, the difference in mean ridits) is a meaningful summary of the difference between two stochastically ordered distributions, regardless of whether that difference is estimated with matched samples or with independent samples.

The delta method (see Goodman and Kruskal, 1972) can be applied to obtain a large-sample normal distribution for $\hat{\tau}$ when the samples that comprise the marginal distributions

are matched. Assume that the proportions $\{p_{ij}\}$ result from full multinomial sampling, and let $\phi_{ij} = \hat{\gamma}_{j+} + \hat{\gamma}_{j-1,+} - \hat{\gamma}_{+i} - \hat{\gamma}_{+,i-1}$, where $\{\hat{\gamma}_{i+}\}$ and $\{\hat{\gamma}_{+i}\}$ are the sample marginal distribution functions. It follows that

$$(\hat{\tau} - \tau)/\hat{\sigma}_{\hat{\tau}} \xrightarrow{d} \mathrm{N}(0, 1),$$

where

$$\hat{\sigma}^2_{\hat{\tau}} = \left\{ \sum_{i,j} \phi^2_{ij} p_{ij} - \left(\sum_{i,j} \phi_{ij} p_{ij} \right)^2 \right\} \Big/ n.$$

For large n, therefore, the null hypothesis of marginal homogeneity may be tested by means of the statistic $z_1 = \hat{\tau}/\sigma_{\hat{\tau}}$, which has an approximately standard normal null distribution. For the vision data (Table 1), we obtain $\hat{\tau} = .0169$, $\hat{\sigma}_{\hat{\tau}} = .0046$ and $z_1 = 3.65$.

We would expect this test to be more powerful than the chi square tests when the marginal distributions are stochastically ordered. Unlike the chi square tests, however, this z test is not consistent for the class of all alternatives, but only for those with $\tau \neq 0$. For example, $\tau = 0$ for many tables whose marginal distributions differ in scale but not in location.

2.2 *Model Parameters*

For square contingency tables, McCullagh (1977, 1978) and Goodman (1979) have presented models which contain parameters describing the degree of marginal inhomogeneity. For both McCullagh's logistic model and his palindromic symmetry model, the parameter is given by

$$\Delta = \log\left(\sum_{i' \leq i} \sum_{j' > i} \rho_{i'j'} \Big/ \sum_{i' > i} \sum_{j' \leq i} \rho_{i'j'} \right),$$

which is assumed constant for $i = 1, 2, \ldots, r - 1$. For the palindromic symmetry model applied to the vision data, the maximum likelihood estimate of the parameter is .167 with standard error .046, leading to $z_2 = 3.63$ for testing marginal homogeneity, i.e. for testing that the parameter equals zero. In an unpublished report (Institute of Statistics, University of North Carolina, Mimeo Series No. 1323, 1980). K.A. Semenya and G. G. Koch (1980) gave weighted least-squares solutions for two logistic models applied to these data, and obtained results similar to those of McCullagh.

2.3 *Weighted Sum of Differences*

Consider the family of statistics of the form $\{M_1 = \sum w_i p_{i+},\ M_2 = \sum w_i p_{+i}\}$ for fixed scores $\{w_i\}$. A flexible approach to testing marginal homogeneity is to use $M_1 - M_2 = \sum w_i (p_{i+} - p_{+i})$, with scores $\{w_i\}$ chosen according to the alternative one wishes to detect. For four categories, we might choose, for example, the scores $\{3, 1, -1, -3\}$ to detect differences in location and $\{1, -1, -1, 1\}$ to detect differences in dispersion. The estimated variance of $M_1 - M_2$ is

$$\hat{\sigma}^2_{M_1 - M_2} = \frac{1}{n} \left\{ \sum_{i,j} (w_i - w_j)^2 p_{ij} - (M_1 - M_2)^2 \right\},$$

and for large n, $z_3 = (M_1 - M_2)/\hat{\sigma}_{M_1 - M_2}$ has approximately the standard normal distribution under H_0. For the vision data with the scores $\{3, 1, -1, -3\}$, we find $M_1 - M_2 = .0599$, $\hat{\sigma}_{M_1 - M_2} = .0173$ and $z_3 = 3.46$. Koch and Reinfurt (1971), Fleiss and Everitt (1971), and Bhapkar (1970) have also mentioned versions of this statistic which is consistent for those alternatives with $\sum w_i \rho_{i+} \neq \sum w_i \rho_{+i}$. Weighted sums of differences can also be constructed on an alternative scale, such as a logit scale, through

$$\sum_{i=1}^{r-1} w_i [\log\{\hat{\gamma}_{i+}/(1 - \hat{\gamma}_{i+})\} - \log\{\hat{\gamma}_{+i}/(1 - \hat{\gamma}_{+i})\}].$$

3. Power Comparisons

The efficiency of the approaches we have discussed, relative to Stuart's chi square test, depends on $\{\rho_{ij}\}$. To illustrate the higher power that an ordinal procedure can have, we made some power comparisons between a z test of the Mann–Whitney type and the version of Stuart's chi square test in which the non-null covariance matrix, $\mathbf{W}$, of $\mathbf{d}$ is used. We assumed random sampling from an underlying bivariate normal distribution having correlation ρ, for 16 different cases. These cases represent all combinations of $r = 3$ and $r = 6$, $\rho = .2$ and $\rho = .8$, $n = 200$ and $n = 400$, and two shifts Δ that denote where $r - 1$ category boundaries of Y are placed relative to those of X. For $r = 6$, the boundaries for the X categories are at μ_X, $\mu_X \pm 0.6\sigma_X$ and $\mu_X \pm 1.2\sigma_X$, and the boundaries for the Y categories are at $\mu_Y + 1.3\sigma_Y$, $\mu_Y + 0.7\sigma_Y$, $\mu_Y + 0.1\sigma_Y$, $\mu_Y - 0.5\sigma_Y$ and $\mu_Y - 1.1\sigma_Y$ for $\Delta = .1$, and at $\mu_Y + 1.4\sigma_Y$, $\mu_Y + 0.8\sigma_Y$, $\mu_Y + 0.2\sigma_Y$, $\mu_Y - 0.4\sigma_Y$ and $\mu_Y - 1.0\sigma_Y$ for $\Delta = .2$. These categorizations yield the marginal distribution of X, (.1151, .1592, .2257, .2257, .1592, .1151), and the marginal distribution of Y, (.0968, .1452, .2182, .2313, .1728, .1357) for $\Delta = .1$ and (.0808, .1311, .2088, .2347, .1859, .1587) for $\Delta = .2$. The tables with $r = 3$ are obtained by combining the first two, middle two, and last two categories of X and of Y from the case with $r = 6$.

Table 2 shows the approximate probabilities of rejecting the null hypothesis of marginal homogeneity for six α-levels. The values for the Mann–Whitney z test are given by $2 - \Phi(z_{\frac{1}{2}\alpha} - \tau/\sigma_{\hat{\tau}}) - \Phi(z_{\frac{1}{2}\alpha} + \tau/\sigma_{\hat{\tau}})$. For the chi square test the values are the probabilities that $\chi^2_{r-1,\lambda}$ random variables with noncentrality parameters $\lambda = n\boldsymbol{\delta}'\mathbf{W}^{-1}\boldsymbol{\delta}$ (where $\delta_i = \rho_{i+} - \rho_{+i}$) exceed the $100(1 - \alpha)$ percentage point of a χ^2_{r-1} random variable. From Table 2 we make the following observations:

(i) The power for the chi square test is uniformly smaller than that for the z test for all combinations of α, Δ, ρ, n and r which we have considered.

(ii) The power for both tests is substantially greater at $\rho = .8$ than at $\rho = .2$ for fixed α, Δ, n and r. Of course, the power for both tests is also greater for larger values of α, Δ and n.

(iii) The power for the z test is uniformly greater at $r = 6$ than at $r = 3$, for fixed α, Δ, ρ and n.

(iv) The ratio (1 − power for chi square test)/(1 − power for z test) is
  (a) uniformly greater at $\Delta = .2$ than at $\Delta = .1$, for fixed α, ρ, n, r,
  (b) uniformly greater at $\rho = .8$ than at $\rho = .2$, for fixed α, Δ, n, r,
  (c) uniformly greater at $n = 400$ than at $n = 200$, for fixed α, Δ, ρ, r,
  (d) uniformly greater at $r = 6$ than at $r = 3$, for fixed α, Δ, ρ, n.

Observation (iv) suggests that the use of the ordinal approach becomes relatively more advantageous as Δ, ρ, n and r increase. These tendencies make intuitive sense, since as Δ, ρ and n increase, the sample marginal distributions are more likely to be stochastically ordered, thus reflecting the type of deviation from H_0 that the Mann–Whitney test naturally detects. We would also expect an ordinal approach to tend to be relatively more efficient the larger the number of categories. When $r = 2$ there is no advantage to be gained from knowing the category order. As r increases, the data become 'more continuous' and a test which ignores quantitative information becomes more disadvantageous. As a further illustration of these tendencies, it is interesting to note that the probability that the z test yields a smaller P-value than does the chi square test is about .69 when $\Delta = .1$, $\rho = .2$, $n = 200$ and $r = 3$, and about .94 when $\Delta = .2$, $\rho = .8$, $n = 400$ and $r = 6$.

The sample sizes reported in Table 2 were chosen to be large so that the asymptotic chi square and normal distributions used would closely approximate the true sampling distributions. Hence, the nominal α-levels used in Table 2 should also be close to the actual levels. We also performed Monte Carlo simulations of 10 000 tables for each of the 16 cases as a

Table 2

Powers of z test and χ^2 test for $r \times r$ table representing sample of size n from underlying normal distribution with correlation ρ and marginal shift Δ

Δ	ρ	n	r	Test	α-level					
					.20	.10	.05	.02	.01	.001
.1	.2	200	3	z	.394	.258	.166	.090	.056	.011
				χ^2	.347	.213	.130	.066	.039	.007
			6	z	.422	.283	.185	.103	.065	.013
				χ^2	.315	.184	.107	.051	.029	.004
		400	3	z	.547	.400	.285	.175	.118	.029
				χ^2	.475	.326	.219	.125	.081	.017
			6	z	.588	.443	.323	.204	.141	.037
				χ^2	.425	.276	.176	.095	.059	.011
	.8	200	3	z	.621	.476	.354	.229	.161	.044
				χ^2	.544	.392	.275	.166	.112	.027
			6	z	.745	.616	.492	.349	.262	.088
				χ^2	.564	.407	.286	.174	.118	.029
		400	3	z	.831	.724	.610	.466	.368	.147
				χ^2	.761	.630	.506	.363	.275	.097
			6	z	.928	.864	.783	.661	.566	.292
				χ^2	.800	.677	.555	.410	.318	.121
.2	.2	200	3	z	.754	.627	.503	.360	.272	.093
				χ^2	.675	.530	.403	.270	.195	.059
			6	z	.801	.685	.566	.421	.326	.122
				χ^2	.616	.461	.335	.213	.148	.040
		400	3	z	.934	.873	.795	.677	.582	.306
				χ^2	.888	.801	.701	.566	.468	.218
			6	z	.958	.913	.852	.752	.667	.388
				χ^2	.853	.748	.636	.494	.397	.170
	.8	200	3	z	.969	.934	.883	.795	.717	.414
				χ^2	.943	.887	.815	.703	.614	.341
			6	z	.995	.986	.971	.936	.898	.711
				χ^2	.969	.933	.882	.796	.721	.460
		400	3	z	.999	.998	.994	.983	.970	.878
				χ^2	.998	.993	.985	.965	.942	.809
			6	z	1.000	1.000	1.000	.999	.998	.984
				χ^2	1.000	.999	.997	.991	.984	.928

check on how similar the powers reported in Table 2 would be when, as in practice, estimated standard errors are used in the z statistic and when estimated covariance matrices are used in the chi square statistic. Nearly all the observed powers obtained from the Monte-Carlo simulations fell within .01 of the theoretical ones given in Table 2, and all fell within .03.

Two added advantages of the ordinal approach should be mentioned. First, each statistic can be modified in an obvious manner to yield a useful descriptive measure of the degree of marginal inhomogeneity. Secondly, the ordinal approach may (unlike Stuart's Q) be used to obtain one-sided P-values.

ACKNOWLEDGEMENTS

The author would like to thank the referees for their valuable criticisms, and Dr David Faulkenberry for helpful discussions.

Résumé

Les tests du χ^2 de l'homogénéité marginale standards ne prennent pas en compte l'ordre sur les classes. Nous décrivons plusieurs stratégies qui tiennent compte de cette contrainte d'ordre et qui permettent de produire des tests plus puissants pour certaines alternatives souvent rencontrées.

References

Bhapkar, V. P. (1966). A note on the equivalence of two criteria for hypotheses in categorical data. *Journal of the American Statistical Association* **61,** 228–235.

Bhapkar, V. P. (1970). Categorical data analogs of some multivariate tests. In *Essays in Probability and Statistics*, R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao and K. J. C. Smith (eds), 95–110. Chapel Hill: University of North Carolina Press.

Bhapkar, V. P. (1979). On tests of marginal symmetry and quasi-symmetry, in two- and three-dimensional contingency tables. *Biometrics* **35,** 417–426.

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

Bross, I. D. J. (1958). How to use ridit analysis. *Biometrics* **14,** 18–38.

Fleiss, J. R. and Everitt, B. S. (1971). Comparing the marginal totals of square contingency tables. *British Journal of Mathematical and Statistical Psychology* **24,** 117–123.

Goodman, L. A. (1979). Multiplicative models for square contingency tables with ordered categories. *Biometrika* **66,** 413–418.

Goodman, L. A. and Kruskal, W. (1972). Measures of association for crossclassifications IV: Simplification of asymptotic variances. *Journal of the American Statistical Association* **67,** 415–421.

Grizzle, J. E., Starmer, C. F. and Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* **25,** 137–156.

Ireland, C. T., Ku, H. H. and Kullback, S. (1969). Symmetry and marginal homogeneity of an $r \times r$ contingency table. *Journal of the American Statistical Association* **64,** 1323–1341.

Koch, G. G. and Reinfurt, D. W. (1971). The analysis of categorical data from mixed models. *Biometrics* **27,** 157–173.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H. and Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* **33,** 133–158.

McCullagh, P. (1977). A logistic model for paired comparisons with ordered categorical data. *Biometrika* **64,** 449–453.

McCullagh, P. (1978). A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika* **65,** 413–418.

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

Received September 1980; *revised May* 1981 *and March and April* 1982