METRON (2014) 72:125–132
DOI 10.1007/s40300-014-0036-1

Two Bayesian/frequentist challenges for categorical data analyses

Alan Agresti

Received: 31 December 2013 / Accepted: 8 March 2014 / Published online: 22 March 2014

Abstract We discuss two challenging scenarios for frequentist and/or Bayesian inference for categorical data. First, for parameter space regions described by order restrictions, frequentist methods are less straightforward than Bayesian methods, especially for interval estimation. Second, for marginal modeling, frequentist inference is currently feasible only in relatively simplistic settings and Bayesian solutions seem to be essentially non-existent. Both areas have substantial scope for future research.

Keywords Marginal models · Order-restricted inference · Ordered categorical data · Stochastic orderings

1 Introduction

In this article we discuss two types of problem that seem to be insufficiently developed for frequentist and/or Bayesian inference. First, for parameter space regions described by order restrictions, frequentist methods are available for significance testing, but they are more complex than Bayesian methods and have rarely been extended to interval estimation. Second, marginal generalized linear modeling of multivariate responses is available with the method of generalized estimating equations (GEE), but likelihood-based frequentist methods are currently feasible only in simplistic settings and there seems to be almost no Bayesian literature. Our discussion uses the context of categorical response variables, for which much of the existing state-of-the-art literature originated in Italy, but the issues seem relevant for any type of data.

In Sect. 2 we discuss certain aspects of order-restricted inference, in the context of a categorical response variable. We illustrate for the comparison of two groups, such as comparisons of proportions for binary data and inference about stochastic ordering for ordinal data. In Sect. 3 we discuss implementation of marginal modeling with frequentist and Bayesian approaches. Each section presents challenges for future research.

A. Agresti (✉)
Department of Statistics, University of Florida, Gainesville, FL 32611, USA
e-mail: aa@stat.ufl.edu

2 Order-restricted inference

In 2000 and in 2010 the General Social Survey in the US asked subjects whether they had ever taken drugs (such as heroin and cocaine) by injecting themselves with a needle. Of subjects aged 20–39 in 2000, let π_1 denote the population proportion who would answer 'yes.' For this same cohort, let π_2 denote the population proportion who would answer 'yes' in 2010 (i.e., of those then having age 30–49). With a stable population, necessarily $\pi_1 \leq \pi_2$. How can we find a confidence interval for $\pi_2 - \pi_1$ that respects such a condition?

There is a substantial literature on order-restricted inference, as summarized by [27–29,33]. However, that literature is dominated by significance testing, with little attention paid to interval estimation. Asymptotic sampling distributions for order-restricted tests are non-standard, typically based on chi-bar-squared distributions (weighted averages of chi-squared distributions). This has both limited the speed with which they have been adopted in practice and made it awkward to implement standard methods of inverting results of tests to construct confidence intervals. For example, Robertson et al. [25] briefly describes the interval estimation problem before stating that the lack of results are "primarily due to the general intractability of these types of problems" and Silvapulle and Sen mention that "the confidence set estimation problem becomes more unmanageable in the constrained statistical inference cases" reviewed the few articles that address order-restricted interval estimation, some of which show how methods (such as the bootstrap) that are typically used when traditional methods fail are inadequate. To find a confidence interval for a parameter such as $\pi_2 - \pi_1$ that respects an order condition, most methodologists would probably merely form an ordinary interval and then truncate results.

By contrast, it is straighforward to use Bayesian methods for this purpose, combining prior distributions that impose the order restriction with the product binomial likelihood function for independent samples. For the survey mentioned above, 28 of 938 subjects of age 20–39 in 2,000 reported taking drugs by needle injection, and 25 of 667 subjects of age 30–49 in 2010 reported taking drugs by needle injection. Using a triangular uniform prior density over the region $0 \leq \pi_2 - \pi_1 \leq 1$ yields a 95 % highest posterior density interval for $\pi_2 - \pi_1$ of (0.0, 0.026).

Ordered categorical data provide a more general setting in which a variety of order-restricted questions naturally occur. We illustrate with Table 1, analyzed by [1]. This 2×5 contingency table results from a clinical trial that observed the outcome for patients who experienced trauma due to subarachnoid hemorrhage. A study objective was to determine whether a more favorable outcome tends to occur with the drug treatment group than with placebo. In the sample, the two rows are strictly stochastically ordered, with the drug treatment being higher than the placebo on the ordinal outcome measure. How can we summarize the evidence that the corresponding population conditional distributions have the same strict stochastic ordering? In a $2 \times c$ table, this is equivalent to the condition that the log of the cumulative odds ratio θ_j obtained by collapsing the table to a 2×2 table by combining the first j columns and combining the last $c - j$ columns is positive, $j = 1, \ldots, c - 1$. Alternatively, we could analyze this for a different ordinal odds ratio, such as the local odds ratios based on pairs of adjacent columns.

In a $2 \times c$ contingency table with a column response variable, it is common to assume that each row of the table $\mathbf{n}_i = (n_{i1}, n_{i2}, \ldots, n_{ic})$ has an independent multinomial distribution, with parameter $\boldsymbol{\pi}_i$ being the vector of conditional probabilities in row i. Altham [3] gave a simple Bayesian solution to the stochastic ordering question, using independent Dirichlet priors for $\boldsymbol{\pi}_2$ and $\boldsymbol{\pi}_2$ with hyperparameter vector $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ic})$, $i = 1, 2$. The posterior distribution of $(\boldsymbol{\pi}_i \mid \mathbf{n}_i)$ is then Dirichlet with parameters $(\mathbf{n}_i + \boldsymbol{\alpha}_i)$, $i = 1, 2$. Sym-

metric Dirichlet priors with $\{\alpha_{ij} = c\}$ are non-informative in terms of association directions, satisfying prior $\log(\theta_j) = 0$ for all j. Common special cases are the uniform distribution ($c = 1$) and the Jeffreys prior ($c = 0.5$). For them, the prior probability of a stochastic order in a particular direction is $1/c$ [3]. For the posterior Dirichlet, using the equivalence of a Dirichlet random variable with differences between certain order statistics from a uniform distribution over [0, 1], Altham derived the posterior probability of a stochastic order. Let $\{\mu_j = n_{1j} + \alpha_{1j}\}$ and $\{\nu_j = n_{2j} + \alpha_{2j}\}$. Let $\mu = (\mu_1 + \cdots + \mu_c)$ and $\nu = (\nu_1 + \cdots + \nu_c)$ be the effective posterior sample sizes. The posterior probability that row 2 is stochastically larger than row 1 equals

$$\sum_{s_1} \cdots \sum_{s_c} \frac{\binom{\mu_1+\nu_1-1}{s_1}\binom{\mu_2+\nu_2}{s_2} \cdots \binom{\mu_{c-1}+\nu_{c-1}}{s_{c-1}}\binom{\mu_c+\nu_c-1}{s_c}}{\binom{\mu+\nu-2}{\nu-1}},$$

where each s_j index varies between 0 and the upper limit in the corresponding binomial coefficient, but such that $(s_1 + \cdots + s_j) \leq (\mu_1 + \cdots + \mu_j - 1)$ for $1 \leq j \leq c - 1$. For Table 1, with uniform prior densities, the posterior probability that the response for drug is stochastically higher than for placebo is 0.705.

Altham's derivation does not extend to alternative types of ordering (such as positive local log odds ratios) for $2 \times c$ tables or to $r \times c$ or multiway contingency tables. However, with Dirichlet priors, the posterior probability of a condition such as a particular stochastic ordering of r rows on an ordinal response can be simulated simply and accurately. Kateri and Agresti [17] gave details and examples.

How might one answer the question about a stochastic ordering from a frequentist perspective? As in [1,5], we could test the null hypothesis of identical distibutions against an order-restricted alternative hypothesis such as stochastic ordering. However, because the union of H_0 and H_a is not the entire parameter space, such inference is not well suited for determining whether the ordinal structure truly holds. Agresti and Coull [1] showed that the P value from a likelihood-ratio test can be very small even when the sample badly violates the order restriction, merely because the ordinal condition is more consistent with the data than is the narrow null hypothesis.

A more appropriate frequentist approach for the posed question about a stochastic ordering uses an intersection–union test [9]. Consider a particular type of ordinal condition in a $2 \times c$ table expressed in terms of $c - 1$ ordinal odds ratios, such as uniformly positive cumulative log odds ratios for all columns j. One regards the null hypothesis H_0 that the order restriction does not strictly hold as the union of $c - 1$ events, where event k says that the kth of the log odds ratios is >0. The alternative hypothesis states that the order restriction strictly holds, that is, it is the intersection event that all $c - 1$ of the log odds ratios are positive. To achieve overall size α, for each individual log odds ratio k we conduct an α-level one-sided test that it is ≤ 0 versus the alternative that it is >0. For example, we could use the signed square root of the Pearson statistic for the relevant 2×2 table, which is the one-sided two-sample z statistic for comparing two proportions using the pooled standard error. We then reject the overall H_0 if each of the $c - 1$ individual tests is significant. From [9], this test has large-sample size of α. Moreover, Berger showed that this test is also the likelihood-ratio test for this decomposition of the parameter space. One can regard the P value for the test as equaling the maximum of the P values for the individual tests. For Table 1, using the two-sample one-sided z test statistic, the minimum individual z score occurs for the test of $\log(\theta_4) = 0$, with $z = 0.66$ generating a P value = 0.25. Here, we have conducted the individual tests under the null condition that $\log(\theta_j) = 0$ and acted as if this gives the same P-value as testing under $\log(\theta_j) \leq 0$, which seems intuitively reasonable but requires further justification. In

Table 1 Cross classification of treatments and extent of trauma due to subarachnoid hemorrhage, with conditional distributions in parentheses

Treatment group	Outcome				
	Death	Vegetative state	Major disability	Minor disability	Good recovery
Placebo	59 (28 %)	25 (12 %)	46 (22 %)	48 (23 %)	32 (15 %)
Drug	135 (23 %)	39 (7 %)	147 (25 %)	169 (29 %)	102 (17 %)

a somewhat different order-restricted setting, Dardanoni and Forcina [13] used this type of approach to evaluate Lorenz curve orderings.

Compared to the Bayesian approach, an unsatisfying aspect of this test is its severe discreteness. For example, if all except one of the sample log odds ratios take large positive values, the P value is governed essentially completely by that one exception; that is, we could change the counts in the table in any way whatever that keeps that one corresponding 2×2 table (for the exception) the same but for which all the other individual P values remain smaller than the P value for the exception, and the overall P value of the intersection–union test will not change. Also, it is awkward to handle more complex structures, such as when H_a is an unspecified stochastic ordering among several unordered groups (i.e., a union of intersections of events). In future research, it is of interest to extend frequentist methods such as the intersection-union test to more complex settings. These include a union of intersections of events for mixtures of nominal and ordinal variables in two-way contingency tables, and further generalizations for multiway contingency tables and for marginal distributions of a multivariate ordinal response. Also, ultimately, estimating effect size has greater importance than significance testing. Assuming a particular ordinal structure, how can we construct good frequentist confidence intervals for measures such as assumed positive cumulative log odds ratios? As already mentioned, for a Bayesian approach it is straightforward to construct posterior intervals over constrained parameter spaces.

3 Marginal models

We next consider a problem for which neither frequentist nor Bayesian methodology is well developed, and the Bayesian approach may be the more difficult of the two to implement. We introduce basic issues using Table 2, which shows data from a recent General Social Survey: Each subject was asked about the success of the US government in $y_1 =$ providing health care and $y_2 =$ protecting the environment. To analyze whether responses tend to be more positive on one issue than the other, we could apply a standard ordinal model with a location shift to the two marginal distributions. One such possibility is the cumulative logit model

$$\text{logit}[P(y_1 \le j)] = \alpha_j, \quad \text{logit}[P(y_2 \le j)] = \alpha_j + \beta, \quad j = 1, 2.$$

The simple special case $\beta = 0$ is marginal homogeneity. Fitting the model of marginal homogeneity and testing whether it is adequate is a standard method of square contingency table analysis. Yet, there seems to be little (if any) literature on Bayesian approaches for such models and for their extensions to multiway contingency tables.

In practice, the context is usually much more complex. We may have a cluster of observations $(y_{i1}, y_{i2}, \ldots, y_{iT})$ for subject i (e.g., repeated measures or longitudinal data), the cluster sizes T_i may vary, and it is relevant to explore effects of explanatory variables on each y_{it} marginally. Moreover, in longitudinal studies there are usually some missing observations,

Table 2 Cross classification of General Social Survey responses about success of US government in providing health care and in protecting the environment

Health care	Environment			
	Successful	Mixed	Unsuccessful	Total
Successful	199	81	83	363
Mixed	129	167	112	408
Unsuccessful	164	169	363	696
Total	492	417	558	1,467

because some subjects may drop out during the study, thus contributing only partial sets of observations.

Frequentist likelihood-based inference, such as maximum likelihood model fitting, is feasible for simple data such as in Table 2, but not for the more general context. Such inference is awkward for marginal models, even without explanatory variables, mainly because the multinomial likelihood refers to the joint distribution rather than the marginal distributions. For example, for a 3×3 table such as Table 2, the multinomial likelihood for cell probabilities $\{\pi_{ij}\}$ and cell counts $\{n_{ij}\}$ is

$$L(\boldsymbol{\pi}) \propto \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \cdots \pi_{33}^{n_{33}}$$

but the model parameters refer to marginal probabilities such as $P(y_1 \leq 1) = \pi_{11}+\pi_{12}+\pi_{13}$. More generally, with T observations on a multicategory response with explanatory variables, the likelihood function uses a product of multinomial distributions at the predictor settings, each having c^T probabilities, but we cannot substitute the marginal model formula in this likelihood to utilize the standard fitting methods such as Fisher scoring.

Frequentist approaches for marginal modeling of multivariate categorical responses can handle relatively simple cases, such as a common but not very large T for all subjects and relatively few categorical explanatory variables. Williamson and Kim [32] proposed a probit model in terms of a latent multivariate normal model, Masarotto and Varin [22] used Gaussian copula marginal regression and implemented it in the R package gcmr, Lang and Agresti [20] and several follow-up papers by Lang used methods for maximizing a likelihood subject to constraints, that is, maximizing the joint multinomial likelihood function while regarding the model formula as a constraint equation. For other approaches, such as using a correspondence with an equivalent saturated loglinear model or with the equivalent set of (univariate, bivariate, trivariate, ...) distributions, see [10,14–16,23]. There is also now a considerable literature on order-restricted inference for marginal models, such as [4,6–8].

Lang's [18,19] research incorporates methods for solving Lagrangian likelihood equations with a refinement of the Newton-Raphson algorithm, and is useful for a wide variety of models of form $\mathbf{L}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ for probabilities or expected frequencies $\boldsymbol{\mu}$ in a contingency table, with link function $\mathbf{L}$. An important special case is the generalized loglinear model $\mathbf{C}\log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$,which includes marginal logit models. Lang's R function mph.fit (mph = multinomial Poisson homogeneous) can fit many such models. Colombi and Cazzaro [12] extended his approach to include inequality constraints, and those authors and S. Giordano extended Lang's R function in their impressive hierarchical multinomial marginal models (hmmm) R package.

These approaches for maximum likelihood model fitting are not feasible for implementation with large T or with data sets that also contain continuous predictors and different cluster sizes because of issues such as missing data. So in practice, for frequentist inference, methodologists usually use the generalized estimating equations (GEE) approach, which is a multivariate generalization of quasi-likelihood methods. For a multivariate response $\boldsymbol{y}_i$ for

subject i with $E(\boldsymbol{y}_i) = \boldsymbol{\mu}_i$ and $\text{var}(\boldsymbol{y}_i) = \boldsymbol{V}_i$ and with models expressed simultaneously for the marginal means, the estimates are solutions of generalized estimating equations

$$\sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0},$$

with $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$. The GEE generalize likelihood equations for a univariate response, without fully specifying the joint distribution. That is, the GEE method provides estimates of regression parameters without assuming a full multivariate distribution. In $\boldsymbol{V}_i$, it assumes a "working" correlation structure (e.g., exchangeable, autoregressive) for $\boldsymbol{y}_i$. The estimator of $\boldsymbol{\beta}$ is consistent even if the correlation structure is misspecified, when the marginal model is correct. The method uses empirical robust estimates of standard errors that are valid even if the correlation structure is misspecified, based on a sandwich estimated covariance matrix. Originally specified by [21] for univariate responses [30] extended GEE methods to handle multinomial (ordinal or nominal) responses, potentially increasing efficiency by basing working correlations on pairwise ordinal odds ratios.

From a frequentist perspective, more research would be useful here for ordinary likelihood methods, because GEE methods have limitations from not being likelihood-based. Related to this, GEE methods rely on Wald methods of inference, which are not optimal. It is desirable to expand the scope to handle marginal models that are currently not feasible to fit. It may be easier to develop further compromises between maximum likelihood and GEE, such as by using composite likelihood methods, as surveyed by [31]. Future research with marginal modeling could also possibly generalize [2], who proposed a Pearson-type "pseudo-score" statistic that does not require the full likelihood and may generalize to multivariate settings.

As mentioned, in the general marginal modeling format, Bayesian approaches also must deal with the awkward form of the likelihood function relative to the model formula. The relatively sparse Bayesian literature includes using a latent multivariate normal model [11], using a multivariate logistic with correlation parameters and marginal logistic distributions [24], and marginal models with inequality constraints [8]. See [26] for the Bayesian approach to composite likelihood analyses.

4 Conclusion

We have highlighted two challenging scenarios for frequentist and/or Bayesian inference for categorical data. The challenge of developing interval estimation methods for order-restricted inference seems relevant for any type of data, and the challenge of extending frequentist and Bayesian methods for marginal modeling seems relevant for all but multivariate normal modeling.

Of course, other challenges exist that are broadly recognized these days by the Statistics community. Perhaps foremost is further development of statistical analyses for "big data." Although considerable new methodology has been developed on regularization methods such as the lasso, much remains to be done. For categorical data, especially challenging is the issue of Bayesian inference, in particular, identifying appropriate noninformative priors when the number of parameters exceeds the number of observations (such as in much genetics research). For instance, at the March 2013 University of Padova workshop Recent Advances in Statistical Inference: Theory and Case Studies, the talk by Jim Berger on "Overall Objective Priors" which is online at http://homes.stat.unipd.it/sites/homes.stat.unipd.it.lauraventura/

files/berger showed that different highly disperse Dirichlet priors can lead to quite different posterior inferences about multinomial parameters.

Many of the issues highlighted in the Padova workshop are quite different from ones that might have been discussed in such a workshop a quarter-century ago. In another quarter century, we can only vaguely imagine what new challenges will have arisen, as our field continues its rapid growth to deal with new types of data and enormous magnitudes of it. An increasingly relevant question, for both frequentists and Bayesians, seems to be this: What is the core unifying theory that underlies methods for Statistical Science?

Acknowledgments Thanks to Sabrina Giordano and Maria Kateri for helpful comments.

References

1. Agresti, A., Coull, B.: The analysis of contingency tables under inequality constraints. J. Stat. Plann. Infer. **107**, 45–73 (2002)
2. Agresti, A., Ryu, E.: Pseudo-score confidence intervals for parameters in discrete statistical models. Biometrika **97**, 215–222 (2010)
3. Altham, P.M.E.: Exact Bayesian analysis of a 2 × 2 contingency table and Fisher's "exact" significance test. J. Roy. Statist. Soc. Ser B **31**, 261–269 (1969)
4. Bartolucci, F., Forcina, A.: Extended RC association models allowing for order restrictions and marginal modeling. J. Am. Statist. Assoc. **97**, 1192–1199 (2002)
5. Bartolucci, F., Scaccia, L.: Testing for positive association in contingency tables with fixed margins. Comput. Stat. Data Anal. **47**, 195–210 (2004)
6. Bartolucci, F., Colombi, R., Forcina, A.: An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. Stat. Sinica **17**, 691–711 (2007)
7. Bartolucci, F., Forcina, A., Dardanoni, V.: Positive quadrant dependence and marginal modeling in two-way tables with ordered margins. J. Am. Stat. Assoc. **96**, 1497–1505 (2001)
8. Bartolucci, F., Scaccia, L., Farcomeni, A.: Bayesian inference through encompassing priors and importance sampling for a class of marginal models for categorical data. Comput. Stat. Data Anal. **56**, 4067–4080 (2012)
9. Berger, R.: Likelihood-ratio tests and intersection-union tests. In: Panchapakesan, S., Balakrishnan, N. (eds.) Advances in Statistical Decision Theory and Applications, pp. 225–237. Birkhauser, Boston (1997)
10. Bergsma, W.P., Rudas, T.: Marginal models for categorical data. Ann. Stat. **30**, 140–159 (2002)
11. Chib, S., Greenberg, E.: Analysis of multivariate probit models. Biometrika **85**, 347–361 (1998)
12. Colombi, R., Cazzaro, M.: Multinomial-Poisson model subject to inequality constraints. Stat. Model. **9**, 215–233 (2009)
13. Dardanoni, V., Forcina, A.: Inference for Lorenz curve orderings. Economet. Journal **2**, 49–75 (1999)
14. Fitzmaurice, G.M., Laird, N.M.: A likelihood-based method for analysing longitudinal binary responses. Biometrika **80**, 141–151 (1993)
15. Glonek, G.: A class of regression models for multivariate categorical responses. Biometrika **83**, 15–28 (1996)
16. Glonek, G.F.V., McCullagh, P.: Multivariate logistic models. J. Roy. Stat. Soc. B **57**, 533–546 (1995)
17. Kateri, M., Agresti, A.: Bayesian inference about odds ratio structure in ordinal contingency tables. Environmetrics **24**, 281–288 (2013)
18. Lang, J.B.: Multinomial-Poisson homogeneous models for contingency tables. Ann. Stat. **32**, 340–383 (2004)
19. Lang, J.B.: Homogeneous linear predictor models for contingency tables. J. Am. Stat. Assoc. **100**, 121–134 (2005)
20. Lang, J.B., Agresti, A.: Simultaneously modeling joint and marginal distributions of multivariate categorical responses. J. Am. Stat. Assoc. **89**, 625–632 (1994)
21. Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**, 13–22 (1986)
22. Masarotto, G., Varin, C.: Gaussian copula marginal regression. Electron. J. Statist. **6**, 1517–1549 (2012)
23. Molenberghs, G., Lesaffre, E.: Marginal modelling of multivariate categorical data. Stat. Med. **18**, 2237–2255 (1999)
24. O'Brien, S.M., Dunson, D.B.: Bayesian multivariate logistic regression. Biometrics **60**, 739–746 (2004)

25. Park, Y., Kalbfleisch, J., Taylor, J.: Confidence intervals under order restrictions. Stat. Sinica **24**, 429–445 (2014)
26. Pauli, F., Racugno, W., Ventura, L.: Bayesian composite marginal likelihoods. Stat. Sinica **21**, 149–164 (2011)
27. Robertson, T., Wright, F.T., Dykstra, R.: Order Restricted Statistical Inference. Wiley, New York (1988)
28. Shapiro, A.: Towards a unified theory of inequality constrained testing in multivariate analysis. Int. Stat. Rev. **56**, 49–62 (1988)
29. Silvapulle, M.J., Sen, P.K.: Constrained Statistical Inference. Wiley, New York (2005)
30. Touloumis, A., Agresti, A., Kateri, M.: GEE for multinomial responses using a local odds ratios parameterization. Biometrics **69**, 633–640 (2013)
31. Varin, C., Reid, N., Firth, D.: An overview of composite likelihood methods. Stat. Sinica **21**, 5–42 (2011)
32. Williamson, J., Kim, K.: A global odds ratio regression model for bivariate ordered categorical data from opthalmologic studies. Stat. Med. **15**, 1507–1518 (1996)
33. Wolak, F.: Local and global testing of linear and nonlinear inequality constraints in nonlinear econometric models. Econom. Theor. **5**, 1–35 (1989)