Rao-Scott corrections and their impact

Alastair Scott
Department of Statistics, University of Auckland,
Auckland, New Zealand, 1011

Abstract

We review the basic ideas underlying the Rao-Scott corrections to chi-squared tests for contingency tables when the estimated cell proportions are derived from survey data (Rao & Scott, 1981,1984), and look briefly at the impact of this work in the 25 or so years since it was first published. We also look at a variant of the corrections that gives improved results when testing for homogeneity and a useful spin-off of this work to the analysis of clustered binary data.

KEY WORDS: Survey data; contingency tables; chi-squared tests.

1. Introduction

This talk is about Rao-Scott corrections to chi-squared tests for cross-classified categorical data in applications where the data comes from a complex sample survey, rather than from a simple random sample. This is something that Jon Rao and I have been working on, off and on, for a long time, starting way back in 1978 when we were both on sabbatical leave at the University of Southampton, working on a joint program organized by Fred Smith and Tim Holt. At various times, that program also included Gad Nathan, Wayne Fuller, Graham Kalton, Chris Skinner, and (a very young) Danny Pfefferman so that it has had a big influence on the development of survey methodology over the years. I have learnt a tremendous amount from Jon over the intervening years and it is a great honour to be asked to speak in this session honoring Jon's contributions to Statistics in the year of his 70th birthday. Jon has made many fundamental contributions to Statistics over the past 50 years and, as a glance at the current literature will show immediately, he shows very little sign of slowing down. I have no doubt that there are many more fundamental contributions to come from him yet.

We start this paper with a brief review of basic methods for the analysis of cross-classified categorical data in the simple case when the data comes from a simple random sample and then look at the complications that result when the data actually come from a survey with more complex structure. In Section 4, we outline the methods developed in Rao & Scott (1981, 1984) to adapt the standard methods to handle this complexity and trace the growth in the use of these methods over the past 25 years.

In Section 5, we look at another, much less well-known paper where we introduce an alternative correction for tests of homogeneity and look at a surprising application of that work to clustered binary data. Finally, we give a brief glimpse at the large body of more recent work that has been influenced by those original Rao-Scott papers.

2. Background Review

Suppose that $\boldsymbol{\pi} = (\pi_t)$ denotes the $T \times 1$ vector of population cell proportions when the cells of a multiway table are ordered in some way. A log-linear model for $\boldsymbol{\pi}$ takes the form

$$\boldsymbol{\mu} = u(\boldsymbol{\theta})\mathbf{e} + \mathbf{X}\boldsymbol{\theta} \tag{1}$$

where $\boldsymbol{\mu}$ is the $T \times 1$ vector with components $\mu_t = \log(\pi_t)$, $\boldsymbol{\theta}$ is a p-vector of unknown parameters, $\mathbf{X}$ is a $T \times p$ matrix of known constants with $r(\mathbf{X}) = p \leq T - 1$ and $\mathbf{X^T e} = \mathbf{0}$. Here $\mathbf{e}$ denotes a T-vector of ones and $u(\boldsymbol{\theta})$ is a normalizing constant chosen so that $\sum_t \pi_t = 1$.

We want a representation for $\boldsymbol{\pi}$ that is as parsimonious as possible. Thus we are led to check the fit of nested models of the form

$$\boldsymbol{\mu} = u_1(\boldsymbol{\theta}_1)\mathbf{e} + \mathbf{X}_1\boldsymbol{\theta}_1 \tag{2}$$

with $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix}$ and $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$. Clearly, checking the fit of this model is equivalent to testing the null hypothesis $H_0 : \boldsymbol{\theta}_2 = 0$ where $\boldsymbol{\theta}_2$ is $k \times 1$.

In the simplest case, and the one for which most of the standard theory was developed, a random sample of n observations is drawn from the population and the results classified according to the cells of the table. Let $\widehat{\mathbf{p}}$ denote the resulting vector of observed proportions. Then the maximum likelihood estimator, $\widehat{\boldsymbol{\pi}}$, of $\boldsymbol{\pi}$ in model (1) is the (unique) solution to

$$\mathbf{X^T}\widehat{\boldsymbol{\pi}} = \mathbf{X^T}\widehat{\mathbf{p}} \tag{3}$$

satisfying (1). For some special models (including, of course, the important case of independence in a two-way table), $\widehat{\boldsymbol{\pi}}$ can be found explicitly. When this is not the case, very efficient algorithms are available for calculating the solution to (1) and (2) in general.

The standard tests of $H_0 : \boldsymbol{\theta}_2 = 0$ are based either on the Pearson chi-squared statistic

$$\mathrm{X}_P^2 = n \sum_t \frac{(\widehat{\pi}_t - \widehat{\pi}_t^*)^2}{\widehat{\pi}_t^*}$$

or on the likelihood-ratio statistic

$$G^2 = 2n \sum_t \widehat{\pi}_t log \left(\frac{\widehat{\pi}_t}{\widehat{\pi}_t^*} \right),$$

where $\widehat{\pi}_t^*$ is the MLE of π under the restricted model (2). X_P^2 and G^2 are asymptotically equivalent and both have asymptotic χ_k^2 distributions under H_0. Details of this standard theory can be found in many places. Good accounts are given in Bishop, Fienberg & Holland (1975) or Agresti (2007)

Methods for choosing a parsimonious representation based on these test statistics have proved very successful for making sense of complex interrelationships among categorical variables. However, the data for many studies, particularly in the social sciences, are drawn using more complicated sampling schemes and there is a natural desire to carry the methods over to data collected using such more complex sampling methods. This turns out to be by no means straightforward. We look at this in more detail in the next section

3. Survey data

Almost all large-scale surveys involve multi-stage sampling with units in the same PSU positively correlated, stratification and variable selection probabilities leading to differential weighting among units, and many other such complications. This makes the assumption of independent and identically distributed observations underlying the standard multinomial theory far from true.

Suppose instead that our sample of n units is actually drawn from a survey with a more complex design. As in the simple case above, we let $\widehat{\mathbf{p}}$ denote the vector of estimated cell proportions but now $\widehat{\mathbf{p}}$ might be extremely complicated in general, incorporating design weights, for example, and involving ratio estimation or post-stratification. All we assume is that $\widehat{\mathbf{p}}$ is a consistent estimator of the population cell proportions, $\boldsymbol{\pi}$, and that a central limit theorem is available for the combination of design and estimator so that $\sqrt{n}(\widehat{\mathbf{p}} - \boldsymbol{\pi})$ converges in distribution to a T-variate normal with mean vector $\underset{\sim}{0}$ and covariance matrix $\mathbf{V}_p$, say. In the case of simple random sampling without replacement, $\mathbf{V}_p$ is equal to the multinomial covariance matrix, $\mathbf{V}_M = diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T$.

In Rao & Scott (1981, 1984) we showed that, under such a general sampling scheme, X_P^2 and G^2 are still asymptotically equivalent but now both are asymptotically distributed as $\sum_1^k \delta_i Z_i^2$ under H_0, where the Z_is are independent $N(0,1)$ random variables and the δ_is are the eigenvalues of

$$\left(\widetilde{\mathbf{X}}_2^T \mathbf{V}_M \widetilde{\mathbf{X}}_2 \right)^{-1} \left(\widetilde{\mathbf{X}}_2^T V_p \widetilde{\mathbf{X}}_2 \right)$$

where

$$\widetilde{\mathbf{X}}_2 = \mathbf{X}_2 - \mathbf{X}_1 \left(\mathbf{X}_1^T \mathbf{V}_M \mathbf{X}_1 \right)^{-1} \left(\mathbf{X}_1^T \mathbf{V}_M \mathbf{X}_2 \right).$$

The δ_is are the design effects of particular linear combinations of the estimated cell proportions and are often known as the "generalized design effects".

Provided that we have an estimate, $\widehat{\mathbf{V}}_p$ say, of $Cov\{\widehat{\mathbf{p}}\}$, available, we can calculate estimates of the δ_is and hence get approximate percentage points for the asymptotic null distribution of X_P^2 and G^2.

4. Rao-Scott corrections

Let $\bar{\delta} = \sum_1^k \widehat{\delta}_i / k$ be the average of the estimated δ_is. It is convenient to define the equivalent sample size, $\widetilde{n}$, by setting $\widetilde{n} = n/\bar{\delta}$. The corresponding "corrected" test statistics are then

$$X_{RS}^2 = \widetilde{n} \sum_t \frac{(\widehat{\pi}_t - \widehat{\pi}_t^*)^2}{\widehat{\pi}_t^*} \quad \text{or} \quad G_{RS}^2 = 2\widetilde{n} \sum_t \widehat{\pi}_t log \left(\frac{\widehat{\pi}_t}{\widehat{\pi}_t^*} \right).$$

There are several ways of approximating the asymptotic null distribution of X_{RS}^2 or G_{RS}^2. In particular, we could approximate it by:

- χ_k^2 (the 1st-order Rao-Scott (RS) correction)
- $c\chi_{k_1^*}^2$, where $c = \frac{\sum \delta_i^2}{k\bar{\delta}^2} > 1$ and $k_1^* = k/c$ (the 2nd-order RS correction).
- $kF_{k_1^*, k_2^*}$, where $k_2^* = k_1^* \nu$ with $\nu = rank(\widehat{V}_p)$. (Note that typically $\nu = \#\ of\ P.S.U.s - \#\ of\ strata$ which may be relatively small even in big surveys).

The first-order approximation matches the first moments of the distributions, ignoring sampling variation in the estimated covariance matrix, $\widehat{V}_p$. The second-order approximation matches the first two moments, again ignoring sampling variation in $\widehat{V}_p$. The third approximation makes an allowance for this sampling variation and is the most accurate in general. In fact, it is good enough for almost all practical purposes (see Thomas & Rao, 1987, Rao & Thomas, 1989, 2003, Thomas, Singh and Roberts, 1996, Servy, Hachuel & Wojdyla, 1998).

Why use the first order approximation at all? One reason is that the other approximations require information on the full $(T-1) \times (T-1)$ covariance matrix V_p and this is not always available, especially when doing secondary analysis from published tables. It turns out that, for many models, $\bar{\delta}$ (and hence the first-order correction which requires no additional information) can be calculated using only information on the standard errors of cell proportions and appropriate marginal proportions, information which should be available as a matter of course in any well-run survey. (More specifically, this happens whenever $\widehat{\boldsymbol{\pi}}^*$ has an explicit form - see Bedrick, 1983, Rao, 1982, Gross, 1984, Rao & Scott, 1984.) The first-order correction is slightly conservative, in general, but much less conservative than the alternative of doing nothing at all.

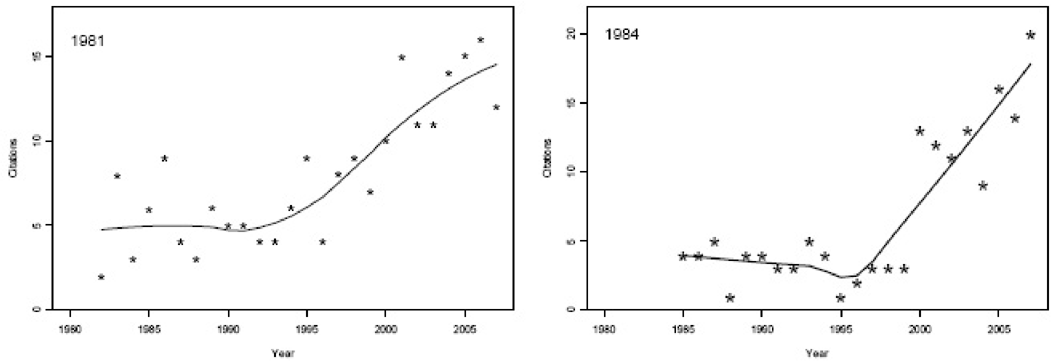

Figure 1: Citation counts for Rao & Scott (1981, 1984) by year.

There has been a fair amount of work trying to improve on the first order correction, while still using only information on cell and marginal standard errors. For example, see Holt, Scott & Ewing (1980), Scott & Styan (1985), Rao & Scott (1987), Shin (1994). However none of this work seems to had much impact in practice.

There are a number of viable alternatives to the RS tests (the Wald, Fay's Jackknife,and Bonferroni inequalities) that all perform well in appropriate circumstances but the Rao-Scott tests have the virtues of simplicity and familiarity, in the sense that experience gained in the standard i.i.d. case can be carried over directly, and they have turned out to work pretty well in practice.

4.1 Impact

The plots in Figure 1 show citation counts (taken from the Web of Science) for the two basic papers, with lowess curves superimposed. It seems that the corrections are being used increasingly in recent years.

Basically, the plots indicate the importance of usable software. We see that citations of both papers meandered along at a low level until they took off in the mid- to late-90s. This is presumably because of their inclusion in major software packages and in Stata and SAS in particular.

A closer look at recent individual citations is instructive. The first thing of note is that almost all the applications are in medicine, health, and biometrics. There are very few in social science, which was the main motivation for the original work. This may be partly due to selection bias, since the coverage of the Web of Science is better for Science and Medicine than for Social Science (Survey Methodology is not among the journals covered, for example). However, a more important reason may be that, as far as I can see, Rao-Scott corrections are not in SPSS, which is the main package used in Social Science applications. (It is possible that it is included under another name.)

The second point of note is that I can find no recent applications at all that use the first-order correction calculated just from standard errors for cells and margins. All that clever work referred to in Section 4 appears to have had very little lasting impact. I suspect that this means that a lot of analysis of published data is still being carried by simply ignoring the effect of the sampling scheme.

5. Another Rao-Scott chi-squared statistic

There is actually another first-order RS correction (Scott & Rao, 1981), which I want to discuss briefly. Apart from one small (but practically important) special case, this one has more or less disappeared into oblivion. I would like to use this forum to attempt a bit of PR on its behalf.

One reason for its obscurity is that it only applies to the special case of testing homogeneity. However, tests of homogeneity turn up reasonably often in a survey context, for example when we want to compare:

- different regions in a regionally stratified survey;
- different national surveys;
- different surveys, supposedly from the same population;
- agreement between interviewers based on Mahalanobis' interpenetrating subsamples.

In ordinary multinomial theory, of course, testing homogeneity and testing independence in a two-way table are identical. This comes about because the multinomial has the nice property that, when we condition on the marginal row totals, cell counts in different rows become independent. This does not happen with more general survey designs. Formally, we could just calculate the equivalent sample size $\widetilde{n}$, form X^2_{RS} and proceed as in the general case. However, this makes no use of the block diagonal structure of the covariance matrix and we can exploit this structure to get improved results for testing homogeneity that do not carry across to testing independence.

It is convenient to change notation slightly for this section. Suppose that we have an $r \times c$ table with the rows representing independent samples. Let $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ic})$ be the vector of category proportions for the ith row and let $\widehat{\mathbf{p}}_i$ be the corresponding vector of sample estimates.

We assume that the $\widehat{\mathbf{p}}_i$s are independent and that $\sqrt{n_i}(\widehat{\mathbf{p}}_i - \boldsymbol{\pi}_i)$ converges in distribution to a c-variate normal with mean vector $\underline{0}$ and (singular) covariance matrix V_i.

The hypothesis of interest is

$$H_0 : \boldsymbol{\pi}_i = \boldsymbol{\pi} \text{ for } i = 1, \ldots, r$$

and the usual multinomial Pearson chi-squared statistic has the form

$$\mathrm{X}^2_{HP} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_i(\widehat{p}_{ij} - \widehat{p}_{+j})^2}{\widehat{p}_{+j}},$$

with $\widehat{p}_{+j} = \frac{\sum_i n_i \widehat{p}_{ij}}{\sum_i n_i}$.

To get the alternative Rao-Scott chi-squared statistic, we simply replace n_i by $\widetilde{n}_i = n_i/\bar{\delta}_i$, where $\bar{\delta}_i$ is the average generalized design effect for the ith row:

$$\mathrm{X}^2_{HRS} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\widetilde{n}_i(\widehat{p}_{ij} - \widetilde{p}_{+j})^2}{p_{+j}},$$

where $\widetilde{p}_{+j} = \frac{\sum_i \widetilde{n}_i \widehat{p}_{ij}}{\sum_i \widetilde{n}_i}$.

Note that we can write $\bar{\delta}_i = \sum_{j=1}^{c} \frac{var(\widehat{p}_{ij})}{p_{ij}}$. Thus estimating $\bar{\delta}_i$ only needs the standard errors of the cell proportions which should always be available in a well-run survey (see Scott & Rao, 1981, for details). (Note also that the standard RS chi-squared, X^2_{RS}, can be written in exactly the same form as X^2_{HRS} but with $\widetilde{n}_i$ replaced by $\widetilde{\widetilde{n}}_i = n_i/\bar{\delta}$ with $\bar{\delta} = \sum_i n_i \bar{\delta}_i / \sum_i n_i$. Obviously this also only needs cell standard errors.)

The asymptotic null distribution of X^2_{HRS} is a linear combination of $k = (r-1)(c-1)$ χ^2_1 random variables under H_0 in general. Equating first moments leads to treating X^2_{HRS} as a χ^2_k random variable as a first order approximation. If we have estimates of the full covariance matrices V_i for $i = 1, \ldots, r$, then we can get improved second-order and $F-$based approximations as with the usual RS statistic. However, this is often not necessary since the first order χ^2_k correction is uniformly better for X^2_{HRS} than for the standard X^2_{RS}. Both statistics have the same asymptotic mean of k but $Var(\mathrm{X}^2_{RS}) \geq Var(\mathrm{X}^2_{HRS}) \geq Var(\chi^2_k) = 2k$. We are still doing more empirical work here but preliminary results suggest that the second order correction will often add very little value.

Finally we note that, when $c = 2$ (i.e. when we are comparing proportions from independent surveys), the first order chi-squared approximation for X^2_{HRS} is actually asymptotically correct. This is not true of the usual X^2_{RS} statistic.

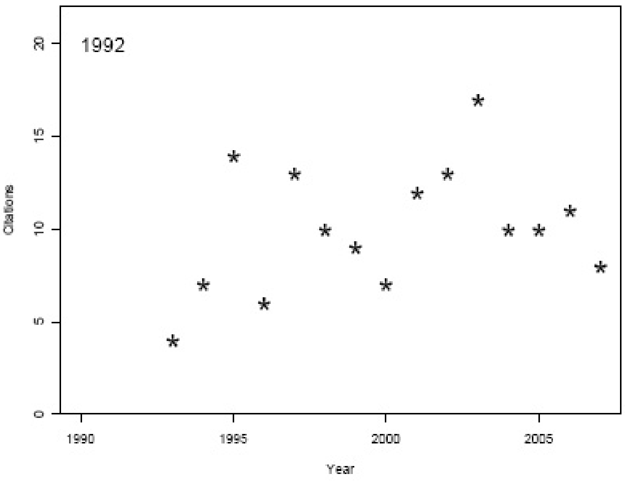

Figure 2: Citation counts for Rao & Scott (1992) by year.

5.1 A special case

The fact that X^2_{HRS} gives asymptotically correct results w when comparing proportions from independent surveys is equivalent to saying that we can treat the "equivalent totals" $\{\widetilde{t}_i = \widetilde{n}_i \widehat{p}_i\}$ as if they were independent $Bin(\widetilde{n}_i, p_i)$ random variables. This trick works with any procedure for handling independent binomial proportions - logistic regression, Mantel-Haenszel tests, Cochran-Armitage tests, and so on. Details are given in Rao & Scott (1992).

Such applications are of limited interest in survey work but have turned out to be very useful in other fields where clustered binary data arise, particularly medicine and biology (litters of animals in toxicity studies, repeated measures on the same individual, twin studies, etc). No new computer programs are needed to implement this technique and as the plot in Figure 2 shows, the take-up was much more immediate.

6. Extensions

The basic idea of the RS-corrections has been applied in a number of other contexts, particularly by Jon Rao and people working with him. For example, Roberts, Rao & Kumar (1987) applied similar techniques to handle logistic regression with survey data. Rotnitzky & Jewell (1990) treated generalized linear models with clustered data and Rao, Scott & Skinner (2000) extended this to handle more general survey data. Rao & Thomas (1990) tackled problems arising with classification errors in categorical data. Bellhouse & Rao (2002) developed corrections to tests for domain means. Problems with the analysis of multiple response table, where a respondent can answer yes to more than one response category, have been discussed by Bilder & Loughin (2002) and Decady & Thomas (2000) and in several other papers by both pairs of authors. In Wang & Rao (2002) and continuing work by the same authors, similar methods to develop empirical likelihood tests with missing data. It seems that there is a lot of life left yet in the basic ideas.

REFERENCES

Agresti, A. (2007), *An Introduction to Categorical data Analysis, 2nd ed*, Wiley-Interscience, Hoboken, NJ.

Bedrick, E.J. (1983), “Adjusted chi-squared tests for cross-classified tables of survey data,” *Biometrika*, **70**, 591-595.

Bellhouse, D.R. and Rao, J.N.K. (2002), “The analysis of domain means in complex surveys,” *J. Statist. Planning & Inference*, **17**, 601-606.

Bilder, C.M., and Loughin, T.M. (2002), “Testing for marginal independence between two categorical variables with multiple responses,” *Biometrics*, **58**, 200-208.

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

Decady, Y.J., and Thomas, D.R. (2000), “A simple test of association for contingency tables with multiple column responses,” *Biometrics*, **56**, 893–896.

Gross, W.F. (1984), “A note on “Chi-squared tests with survey data,” *J. R. Statist. Soc. B*, **46**, 270–272.

Holt, D., Scott, A.J., and Ewing, P.D. (1984), “Chi-squared tests with survey data,” *J. R. Statist. Soc. A*, **143**, 303–320.

Rao, J.N.K., and Scott, A.J. (1981), “The analysis of categorical data from complex sample surveys: chi-squared tests for goodness-of-fit and independence in two-way tables” *Journal of the American Statistical Association*, **76**, 221–230.

Rao, J.N.K., and Scott, A.J. (1984), “On chi-squared tests for multi-way tables with cell proportions estimated from survey data” *Annals of Statistics*, **12**, 46–60.

Rao, J.N.K., and Scott, A.J. (1987), “On simple adjustments to chi-squared tests with survey data” *Annals of Statistics*, **15**, 385–397.

Rao, J.N.K., and Scott, A.J. (1992), “A simple method for the analysis of clustered binary data” *Biometrics*, **15**, 385–397.

Rao, J.N.K., Scott, A.J., and Skinner, C.J. (1998), “Quasi-score tests with survey data” *Statistica Sinica*, **8**, 1059–1070.

Rao, J.N.K., and Thomas, D.R. (1989), “ Chi-squared tests for contingency tables ” in *Analysis of Complex Surveys*, eds D.Holt,C.J. Skinner and T.M.F. Smith. Wiley, New York.

Rao, J.N.K., and Thomas, D.R. (1993), “ Analysis of Categorical Response Data from Complex Surveys: An Appraisal and Update ” in *Analysis of Survey Data*, eds R.L. Chambers and C.J. Skinner. Wiley-Interscience, New York.

Roberts,G., Rao, J.N.K., and Kumar,S. (1987), “Logistic regression analysis of sample survey data” *Biometrika*, **74**, 1–12.

Rotnitzky, A. and Jewell, N. (1990), “Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data” *Biometrika*, **77**, 385–397.

Scott, A.J. and Rao, J.N.K. (1981), “Chi-squared tests for homogeneity with proportions estimated from survey data”, in *Current Topics in Survey Sampling*, D. Krewski, R. Platek, and J.N.K. rao (eds), 209–224. New York; Academic Press.

Scott, A.J. and Styan, G.P.H. (1985), “On a separation theorem for generalized eigenvalues and a problem in the analysis of sample surveys” *Linear Algebra and its Applications* , ,

Servy, E., Hachuel, L., and Wojdyla, D. (1997),“A simulation study for analyzing the perfomance of tests of independence under cluster sampling” *Bull. Int. Statist. Inst*, **51**, 295–311.

Thomas, D.R. and Rao, J.N.K.(1987), “Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics Under Cluster Sampling ” *Journal of the American Statistical Association*, **82**, 630–636.

Thomas, D.R., Singh, A.C. and Roberts, G.R.(1996), “A simple method for the analysis of clustered data” *Int. Statist. Rev*, **64**, 295–311.

Wang, Q., and Rao, J.N.K. (2002), “Empirical likelihood-based inference under imputation for missing response data ” *Annals of Statistics*, **30**, 896–924.