TESTS OF MONOTONE DOSE-RESPONSE IN COMPLEX SURVEYS

Varma Nadimpalli, David Judkins, and Paul Zador
Westat, 1650 Research Boulevard, Rockville, Maryland 20850

KEY WORDS: Jackknife, Association Measures, Mantel-Haenzel Test

1. Introduction

The exploration of dose-response relationships is the main focus of much clinical and epidemiological research. In recent years, dose-response analysis has been employed to evaluate public education campaigns that lacked design features required for the application of commonly used methods based on before/after comparisons and controlled dose-assignment. The data for such evaluations can come from standard household surveys with the usual complexities of multiple stages, stratification and variable selection probabilities. Judkins, Zador and Nadimpalli (2002) reported on the performance of a jackknifed Jonckheere-Terpstra (JT) for testing dose-response relationship on data from a complex survey the National Survey of Parents and Youth. The choice of statistic for analysis of data from this survey was the JT at first, but switched to Gamma as the chief analyst found it easier to interpret. Gary Simon (1978) indicated that most of the association tests for data from a simple random sample are asymptotically equivalent (but he did not study the JT). The focus of this paper is to extend our results on the jackknifed JT to tests based on other association measures and to the Cochran-Mantel-Haenzel (CMH) test (SAS Procedures Guide, 1999), and also to see if various tests are still equivalent even when the data are from a complex survey and the statistics are jackknifed. The hypothesis is that exposure to Campaign messages is monotonically related to outcomes.

We opted to work with the Jonckheere-Terpstra test, tests based on the Gamma and Kendal's Tau association measures, and the Cochran-Mantel-Haenzel test. These tests were developed for testing data from simple random samples, and therefore, may not be valid for data from complex samples; we present jackknifed versions of these tests that can be used for samples from complex surveys.

2. Modification for Complex Sample Design

Survey practitioners had already made considerable progress in determining how to analyze contingency tables based on complex sample designs. Kish and Frankel (1974) first established that although the impact of clustering on fixed parameters in models is smaller than on marginal means, it is nonnegligible for high intraclass correlation. Holt and Scott (1981) and Scott and Holt (1982) confirmed and expanded upon on that work. Rao and Scott (1981) reviewed the early work and suggested a series of three alternate adjusted chi-square statistics for two-way tables and later generalized these to multi-way tables (Rao and Scott, 1984). Fay (1985) suggested a procedure for testing for independence and various forms of conditional independence in contingency tables using a jackknifed chi-square statistic. The Rao and Scott statistics have become standard features in Wesvar (Wesvar 4.2, 2002) and Sudaan (Sudaan 8.02, 2003).

More recently, Wu, Holt and Holmes (1988) showed the seriousness of ignoring the clustering in determining an overall F statistic for clustered samples and how to correct it. Medical researchers have been slower to recognize these problems, but recent gains have been made in this field as well (Manda, 2002). We conjectured that the problems identified for other types of analyses would also impact the JT, Gamma, Tau and CMH unfavorably if we were to compute them from a weighted contingency table, even when the weights had been standardized. So we wanted to do something for these tests similar to the Fay or Rao and Scott corrections of the chi-square independence tests.

Since the JT, Gamma, Tau and CMH test statistics all have an asymptotic normal distribution under the null hypothesis of independence, it seemed like a straightforward procedure to calculate these tests on each set of replicated weights, then calculate a variance on the replicated JT, Gamma, Tau and CMH statistic, and finally use this in a z-test. More specifically, let T_0 be the standardized test statistic formed on the contingency table of Y by Z using full-sample weights and let T_r be the standardized statistic formed on the contingency table of Y by Z using the r-th set of replicate weights. Let b_r be a factor associated with the r-th replicate and the method used to create the replicate weights. Then the proposed "jackknifed association" test is:

$$JAT = \frac{T_0}{\sqrt{\sum_r b_r \left(T_r - T_0\right)^2}} \qquad (1)$$

Note that we use "jackknifed" more broadly here than to imply that the replicate weights need to be created by a jackknife method. The replicate weights can be created by balanced repeated replications, a bootstrap, or any of a variety of resampling schemes, as described in the Wesvar manual (Westat's Wesvar 4.0 Users guide), which includes further references on variance estimators of the form given in the denominator of JAT.

The rest of this paper reports on a simulation study to establish the properties of this test.

3. Parameters for the Simulation Study

We chose to simulate only those features of the design of typical household surveys that seemed most likely to impact the performance of the jackknifed tests. The features we selected were clustering at the PSU level, variation in cluster size, and 100 replicate weights for variance estimation. We elected not to simulate stratification or any sort of differential weighting. Levels of intraclass correlation at the PSU level were set as might be expected if stratification had been employed. Variation in cluster size is of interest because of the natural variation in yields in the screening procedure across PSUs and because of aging in frames between decennial censuses. By this, we mean that each Decennial Census in the U.S. is used to set probabilities of selection for the PSUs in most surveys. Sometimes, PSUs are selected early in the decade and then used for a variety of surveys over the course of the decade. The optimal set of probabilities of selection would be proportional to total eligible population at the time of the survey data collection. As the decade progresses, new construction, natural increase, immigration, and internal migration conspire to degrade the quality of the probabilities of selection. This then results in variation in cluster sample size.

Variation in cluster sample size has an impact on the variance of survey estimates that is more difficult to project in advance than the impact of intraclass correlation. It depends fairly strongly on the types of analyses being conducted with a general rule of thumb that more conditioning probably reduces its impact. However, in addition to its effect on variance, variation in sample cluster size affects the application of the central limit theorem to survey estimates. Basu's elephant (Brewer, 2002) lies at the extreme of variation in cluster size. So we decided to incorporate rather strong variation in cluster sample size into our simulation.

We chose to simulate 100 sample PSUs and 100 replicate weights because this is a common sample design and variance estimation strategy at Westat. The PSU sample sizes were generated as a mixture of two iid gamma distributions as parameterized in Encyclopedia of Statistical Sciences, Volume 3. The formula is given by: $n_i \sim \Gamma(7,5.7)+\Gamma(0.3,33.3)$, rounded to the nearest natural number. On one replicate, this produced an average of 50 units (could be either people or households) per PSU with a standard deviation of 24, a skew of 1.9, a minimum of 12, and a maximum of 202. So the total sample size was about 5,000, but was left as a random variable.

The replicate weights were generated with the jackknife method, meaning that each of the 100 PSUs was dropped in turn for one replicate. All the remaining PSUs in a replicate had their weights adjusted by a factor of 100/99. The full sample weights were all set equal to a constant. With this replication scheme, the replication factors are $b_r = 99/100$.

Exposure Variable. To simulate an ordinal exposure or dose variable *Y*, we used a double normal distribution to simulate a latent score at the person level and then scored it as 0, 1, 2, 3, where the thresholds were selected as quartiles of the latent distribution. At the PSU level, the latent exposure score was allowed to depend on the cluster size through the distribution:

$$\mu_i \sim N\left[1-\left(\frac{4}{n_i}\right)^{0.3}, \sigma^2_{y1}\right] \qquad (2)$$

The latent exposure score for person *j* in cluster *i* was simulated as $\mu_{ij} \sim N\left(\mu_i, \sigma^2_{y2}\right)$, where σ^2_{y1} and σ^2_{y2} were varied to create different levels of intraclass correlation, $\rho_y = \dfrac{\sigma^2_{y1}}{\sigma^2_{y1}+\sigma^2_{y2}}$, while keeping the total variance $\sigma^2_{y1}+\sigma^2_{y2} = c$ fixed.

These latent scores for *Y* were then translated to manifest ordinal scores, y_{ij}, representing the quartiles of μ.

Outcome Variable. For the outcome variable *Z*, we constructed a model in terms of a PSU-level perturbation in the marginal mean, a PSU-level perturbation in the strength of the exposure-outcome relationship, and a person-level relationship between exposure and outcome. The PSU level perturbation in the marginal mean was simulated as $\xi_i \sim N\left(0, \sigma^2_{z1}\right)$, and the PSU-level perturbation in the strength of the exposure-outcome relationship was simulated as $\theta_i \sim N\left(\alpha, \sigma^2_{\theta}\right)$, where α can be thought of as an average slope and σ^2_{θ} as a measure of the between-PSU variance in the strength of the dose-response relationship.

The latent outcome score for person *j* in cluster *i* was simulated as:

$$\xi_{ij} \sim N\left[\xi_i + \theta_i f_\lambda\left(y_{ij}\right), \sigma^2_{z2}\right] \qquad (3)$$

where 8 shapes were selected for the relationship, f_λ, of the latent outcome score to the manifest latent exposure score and σ^2_{z1} and σ^2_{z2} were again varied to create a variety of intraclass correlations.

Figure 1 shows the relationship between the mean latent *Z* score (vertical axis) and the manifest *Y* score (horizontal axis) for each of the 8 patterns. Patterns Linear, Square Root, Fourth Power, Early Jump and Late Jump reflect monotone relationships. Flat pattern reflects independence, and Patterns Central Butte and Early Spike

reflect nonmonotone dependence. We varied the level of α to create patterns that were more or less easy to detect.

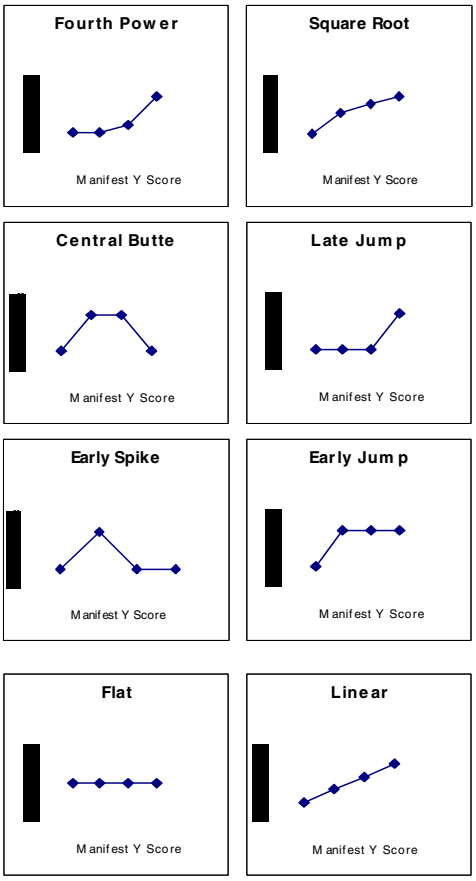

Figure 1. Eight shapes tested in simulation

Note that the intraclass correlation for *Z* is fairly complex. For the flat pattern, the intraclass correlation on *Z* is

$$\rho_z = \frac{\sigma_{z1}^2 + \sigma_\theta^2/4}{\sigma_{z1}^2 + \sigma_\theta^2/4 + \sigma_{z2}^2} \qquad (4)$$

We then converted the person-level latent *Z* scores into the manifest ordinal variable *Z* by splitting ξ at its quartiles.

Sample. We generated a sample of 2,000 draws from this distribution for each dependence pattern, f_λ, and level of intraclass correlation, where we kept $\rho_y = \rho_z$. We hoped that the null hypothesis of independence would be rejected only five percent of the time for the flat, central butte, and early spike patterns, and that power to reject the null hypothesis would be reasonably strong for the other patterns. In this simulation, we also compared the jackknifed test against the ordinary test.

4. Results

The primary statistic of interest is the true size of the test, the percent of draws for which the null hypothesis was rejected when it is true. Of secondary interest was the power of the test under various alternate hypotheses and other hypotheses that are properly neither null nor alternate hypotheses.

Figure 2 shows the true size of the test as a function of the intraclass correlation. A two-sided test with nominal size 0.05 was used. Any values greater than 0.05 indicate that the test is overly aggressive (i.e., rejects at a rate above the nominal size). The same intraclass correlation was used for both the exposure and outcome variables. Note that ordinary gamma performs well for small to moderately large intraclass correlation. It is only when the intraclass correlation is about 0.05 or more the ordinary gamma no longer provides the nominal significance level. If the intraclass correlation is about 0.1, then the ordinary gamma is much too aggressive. The jackknifed gamma protects the significance level at all levels of intraclass correlation. Note that with 2000 draws, the 95 percent confidence interval on the estimated power is about plus or minus one percentage point. We obtained similar graphs for other statistics.

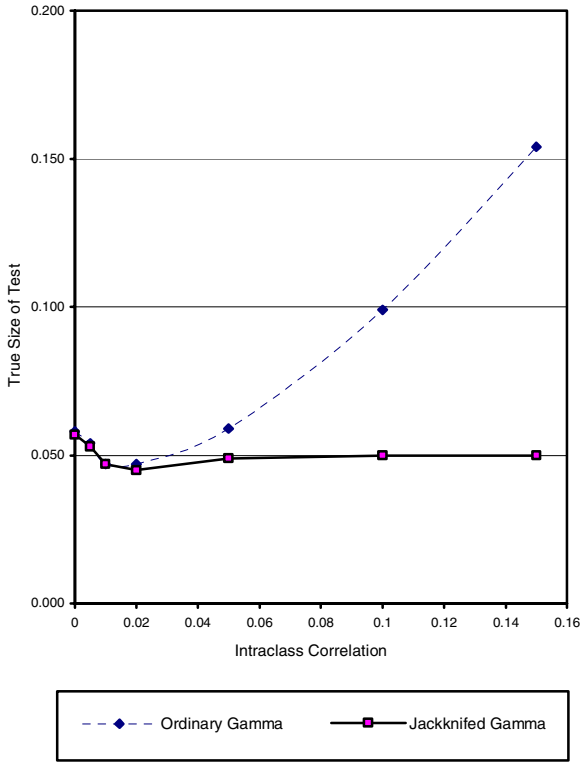

Figure 2. Probability of rejecting the null hypothesis of a flat relationship when the relationship is indeed flat (same intraclass correlation on *Y* and *Z*, total variance =35 on each)

Figure 3 displays power for the jackknifed JT, Gamma, Tau and CMH for a linear pattern with $\alpha = 0.66$ and $\sigma_\theta^2 = 0$. Note that power decreases as intraclass correlation increases, as would be expected. We produced similar graphs for other monotone patterns. We can also see that the jackknifed statistics are essentially equivalent.

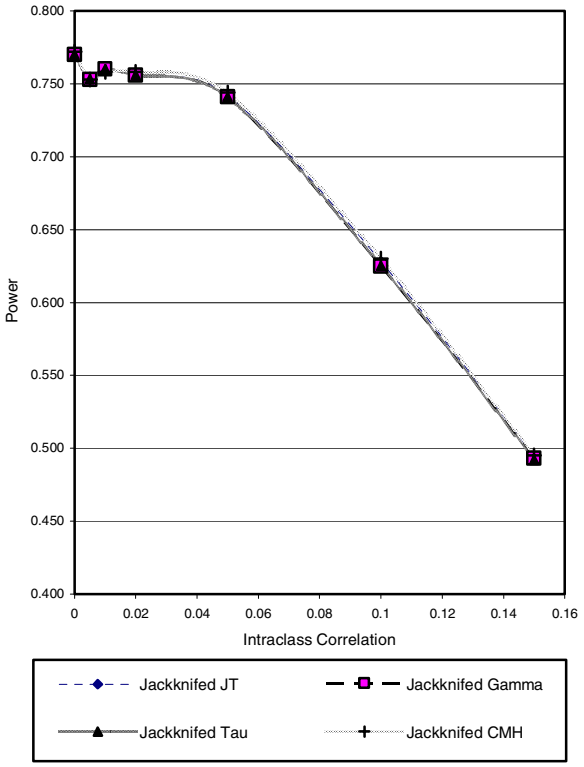

Figure 3. Power with the jackknifed statistics for a linear pattern

Figure 4 focuses on the relative power loss with jackknifed JT, Gamma, Tau and CMH for a linear pattern with $\alpha = 0.66$ and $\sigma_\theta^2 = 0$. Note that for small to moderately large intraclass correlation, the loss of power with jackknifed statistics is minimal. Not surprisingly this range of intraclass correlation coincides with the range shown in Figure 2, where the ordinary test is valid. Figure 5 shows test size for dependent but nonmonotone patterns.

In the framework of the Gamma, these are patterns that have been ruled out a priori as not sensible.

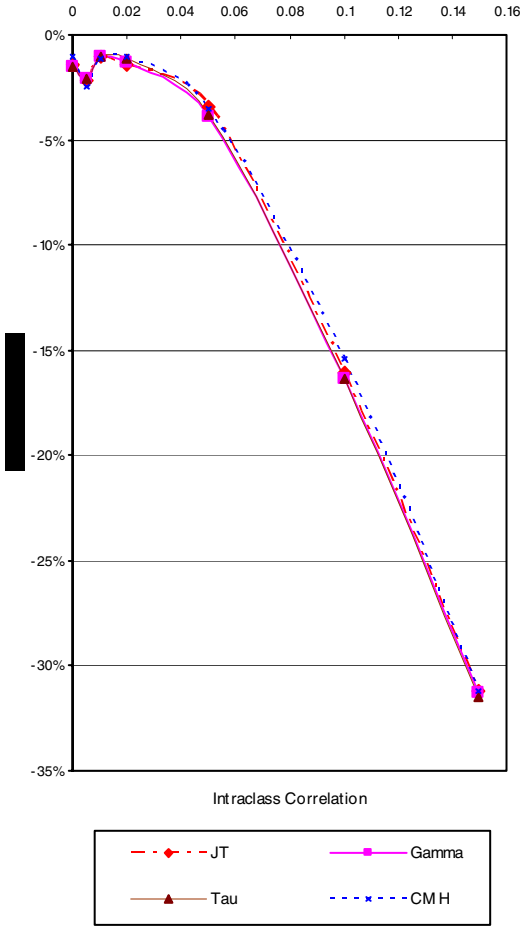

Figure 4. Relative power loss with the jackknifed statistics for a linear pattern

As discussed above, the null hypothesis is that the two variables are independent while the alternative hypothesis is that there is a monotone dose-response relationship. Nonmonotone patterns are outside the parameter space.Nonetheless, there might be situations where a nonmonotone patterns exists for complex reasons.We used the Gamma in the hope that it would reject the null hypothesis no more than five percent of the time if such a nonmonotone pattern was found. This hope was fulfilled for the Central Butte pattern but partially disappointed for the Early Spike pattern. Nonetheless, because the jackknifed Gamma generally has lower size than the ordinary Gamma, using the jackknifed Gamma does result in fewer false claims of monotone trends even when there is an Early Spike pattern. We obtained similar graphs for the other statistics.

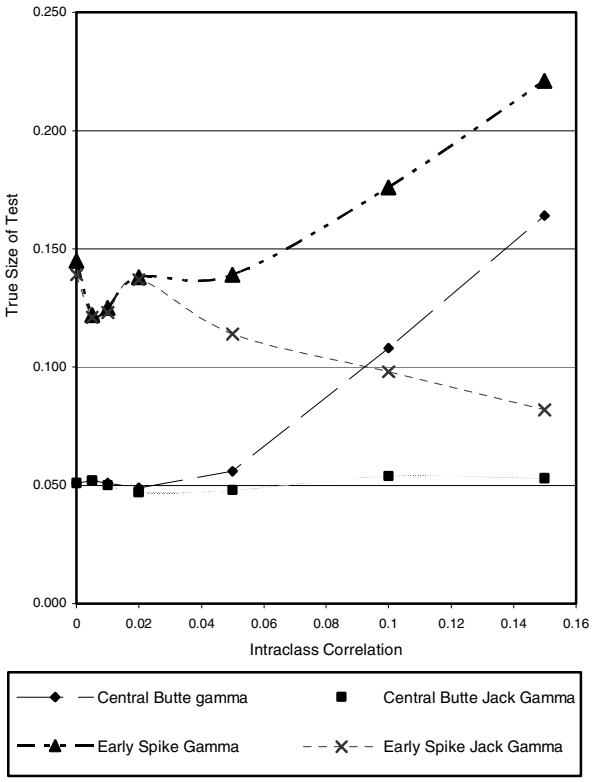

Figure 5. Size with the ordinary and jackknifed gamma for two nonmonotonically dependent patterns

Figure 6 shows how the power of the jackknifed Gamma depends on the source of the intraclass correlation of the outcome variable. More specifically, it shows the effect of intraclass correlation caused partially or completely by variability in the strength of the relationship across clusters as opposed to be due to variability in the underlying outcome tendencies. In Figure 3, the entries for $\rho_z = 0.1$ were simulated by setting $\sigma^2_{z1} = 0.1 \times 35$, $\sigma^2_{z2} = 0.9 \times 35$ and $\sigma^2_{\theta} = 0$. When we simulated $\rho_z = 0.1$ by setting $\sigma^2_{z1} = \frac{0.1 \times 35}{2}$, $\sigma^2_{z2} = 0.9 \times 35$ and $\sigma^2_{\theta} = 7$, so that half of the intraclass correlation on the outcome was due to variable slopes, power fell from the range of 54 to 63 percent down to the range of 32 to 37 percent. When we went further and simulated $\rho_z = 0.1$ by setting $\sigma^2_{z1} = 0$, $\sigma^2_{z2} = 0.9 \times 35$ and $\sigma^2_{\theta} = 14$, so that all of the intraclass correlation on the outcome was due to variable slopes, power fell down to the range of just 26 to 28 percent. This power loss may not be a bad thing in the sense that the Gamma is supposed to be testing for a monotone dose-response relationship that is universal. If the strength of the relationship varies substantially because of interactions with unknown covariates, then one might not want to conclude that there is a universal monotone dose-response relationship.

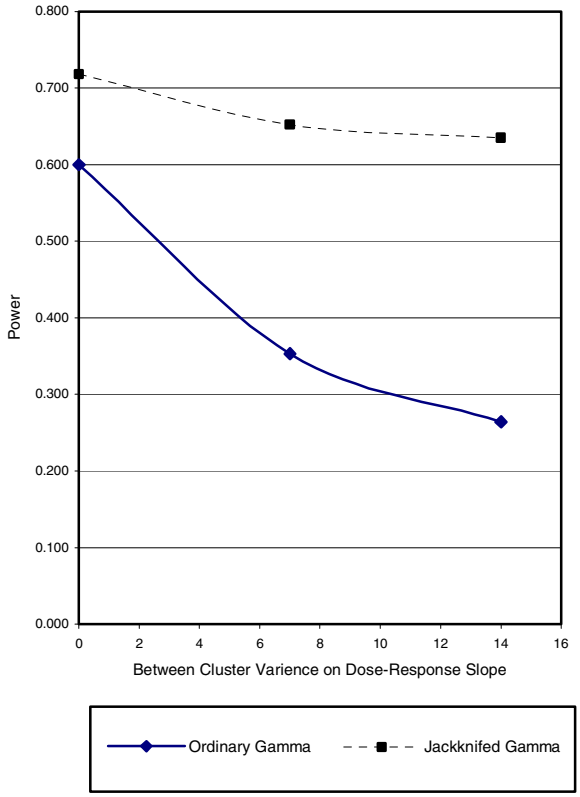

Figure 6. Power of the tests given variability in the slope of square root dose-response relationship

5. Recommendations

The jackknifed tests have been shown to be reasonable tests for monotone dose-response relationships on clustered data as might be expected in a complex sample survey, repeated measures design or randomized cluster design. It accepts the null hypothesis at the desired rate when the true pattern is flat or symmetric (as in the Central Butte pattern). It rejects the null hypothesis with only slightly worse power than the ordinary statistic for the monotone patterns when the true size of the ordinary test is close to its nominal size. The jackknifed tests accepts the alternate hypothesis of a monotone dose-response relationship more often than desired for a true asymmetric nonmonotone dependent pattern (as in the Early Spike pattern), but it does reject this hypothesis more often than not and does much better than the ordinary statistic when there is strong clustering.

With cluster sample sizes of about 50, an intraclass correlation of 0.05, (the level where we found jackknifing to be important), translates to a design effect of 3.45. This is larger than is typically seen in national household surveys in the U.S. because the PSUs in such surveys are

typically whole metropolitan areas and groups of nonmetropolitan counties, areas large enough to encompass great diversity. However, since the power loss is minimal in situations where the correction is unnecessary, we recommend that the procedure always be used on clustered data, regardless of the level of intraclass correlation expected.

Also intraclass correlation considerably larger than 0.05 can be seen in surveys of individuals associated with institutions such as schools or prisons. An example of such survey is the multinational surveys of Education Achievement (Rust and Ross, 1993). The intraclass at the level of primary school and secondary school population for various countries ranges from 0.05 to 0.88 for reading literacy. Another example of surveys with high intraclass correlation and large cluster sizes are third world surveys (Le and Verma, 1997). In such surveys, whole urban neighborhoods and rural villages are often canvassed and intraclass correlation for these levels ranges as high as 0.15 for measures of child health and knowledge of contraception. Taking these into consideration, we recommend that the procedure be used on all the clustered data, regardless of the level of intraclass correlation.

6. References

Brewer, K. (2002). Combined Survey Sampling Inference, Arnold Publications, London.

Fay, R.E. (1985). A jackknifed chi-squared test for complex samples, *Journal of the American Statistical Association*, 80, pp. 148-157.

Holt, D. and Scott, A.J. (1981). Regression analysis using survey data, *The Statistician*, 30, pp. 169-178.

Jonckheere, A.R. (1954). Distribution-free k-sample test against ordered alternatives, *Biometrika,* 7, pp. 93-100.

Judkins, D., Zador, P., and Nadimpalli, V. (2002). Analysis of Dose-Response Relationships on Complex *Survey Data, Proceedings of Statistics Canada Symposium.*

Kish, L. and Frankel, M.R. (1974). Inference from complex samples (with discussion), *Journal of the Royal Statistical Society.* Series B, 36, pp. 1-37.

Lê, T.N. and Verma, V.K. (1997). *An Analysis of Sample Designs and Sampling Errors of the Demographic and Health Surveys*. Demographic and Health Surveys, Analytic Reports No. 3. Calverton, MD. Macro International.

Lucas, W.F. (1983). Gamma distribution, in Katz, S. and Johnson, N.L, (eds) *Encyclopedia of Statistical Sciences*, New York: Wiley, Vol. 3, pp. 292-298.

Manda, S.O.M. (2002). A Bayesian ordinal model for heterogeneity in a multi-centre myocardial infarction clinical trial, *Statistics in Medicine*, 21, pp. 3011-3022.

Mantel N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure, *Journal of the American Statistical Association*, 58, pp. 690-700.

Pirie, W. (1983). Jonckheere tests for ordered alternatives, in Kotz, S., and Johnson, N. L (eds.) *Encyclopedia of Statistical Sciences*, New York:Wiley, vol. 4, pp. 315-318.

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, pp. 221-230.

Rao, J.N.K. and Scott, A.J. (1984). On Chi-squared tests for multiway contingency tables with cell proportions estimated from survey data, *The Annals of Statistics*, 12, pp. 46-60.

Rust, K.F. and Ross, K.N. (1993). Multinational Survey of Educational Achievement, presented at the 49th *Session of the International Statistical Institute*, Florence Italy.

SAS Procedures Guide, Version 8 (1999). SAS Institute, Inc., Carey, NC.

Scott, A.J. and Holt, D. (1982). The effects of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, pp. 848-854.

Simon, G. (1978). Effects of Measures of Association for Ordinal Contingency Tables, *Journal of the American Statistical Association*, September 1978, Volume 73.

Sudaan 8.02 Software. (2003). RTI, Research Triangle Park, NC.

Terpstra, T.J. (1952). The asymptotic normality and consistency of Kendall's test against trend when ties are present in one ranking, *Indag. Mat*., 14, pp. 327-333.

Wesvar 4.0 User's Guide. (2000). Westat, Inc. Rockville, MD.

Wesvar 4.2 Software. (2002). Westat, Inc. Rockville, MD.

Wu, C.F.J., Holt, D., and Holmes, D.J. (1988). The effect of two-stage sampling on the F statistic, *Journal of the American Statistical Association*, 83, pp. 150-159.