Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

ANALYSIS OF DOSE-RESPONSE RELATIONSHIPS ON COMPLEX SURVEY DATA

David Judkins, Paul Zador, and Varma Nadimpalli[1]

ABSTRACT

Analysis of dose-response relationships has long been important in toxicology. More recently, this type of analysis has been employed to evaluate public education campaigns. The data that are collected in such evaluations are likely to come from standard household survey designs with all the usual complexities of multiple stages, stratification, and variable selection probabilities. To meet these challenges, a new jackknifed Jonckheere-Terpstra test for a monotone dose-response relationship is proposed. The focus of this paper is on the results of a Monte-Carlo simulation of the properties of this new test.

KEY WORDS: Jonckheere-Terpstra Test, Jackknife, Monotonicity

1. INTRODUCTION

The traditional realm for the application of dose-response relationships has been clinical and epidemiological research. In contrast, the context for our research is the evaluation of the effectiveness of the National Youth Anti-Drug Media Campaign, a Campaign that has been funded by the U. S. Congress since the beginning of 1998 with the stated goal of reducing and preventing drug use among youth, both directly, and indirectly by influencing their parents and other significant adults in their lives. The primary tool for the evaluation of the Campaign's effectiveness in achieving its objectives is the National Survey of Parents and Youth (NSPY). NSPY is a national in-home survey of complex design. NSPY collects initial and followup data from nationally representative samples of youth and the parents of these youth. Since the Campaign is national in scope, is essentially uniform in intensity by design, and was started the same time throughout the nation, conventional evaluation methods relying on before-after and/or case-control comparisons could not be adopted for this effectiveness evaluation. As a fall back option, a Westat/Annenberg evaluation team opted for assessing Campaign effectiveness by estimating exposure-outcome relationships in analogy with dose-response relationships. For more information on the Campaign, the evaluation, and NSPY, see Hornik, et al (2001).

The implementation of this strategy involves four components: (1) measuring the dose, that is the level of exposure to Campaign messages; (2) measuring the response, that is the set of outcomes the Campaign is supposed to affect; (3) controlling for potential confounders of the dose-response relationships; and (4) testing the hypothesis that a specific observed exposure-outcome relationship is in line with the hypothesis that exposure to Campaign messages is related to the outcome “the right way”, that is, in the desired direction.

We discuss here only component 4 of the evaluation strategy. As indicated earlier, our primary interest was to ascertain for each observed exposure outcome association, whether or not it was monotonic in the 'right' direction. The reason for this is that evidence that a statistical association is not merely significant, but is also monotonic, is often considered an important justification in support of claims of causality. We opted to work with the Jonckheere-Terpstra test (Terpstra, 1952; Jonckheere, 1954; Pirie, 1983)—a nonparametric test designed to reject null hypotheses of no association against ordered alternatives. An additional advantage of the JT test was that it is

[1] Westat, 1650 Research Boulevard, Rockville, MD, U.S.A. 20850

available in SAS (SAS/STAT, 1996), and could be made a part of an already very complex data processing stream that needed to be programmed in SAS.

As it was developed, the JT test was designed and tested for data from simple random samples (SRS) and, therefore, may not be valid for the NSPY data. In the next section of this paper, we present a jackknifed version of the JT test that can be used for complex samples, such as NSPY. First though, we review some recent research on tests of monotoncity.

A recent paper, Leuraud and Benichou (2001) compared several methods to test for the existence of monotone dose-response relationship for *proportions calculated from simple random samples.* In so far as our samples are not SRS—the response variables in our study can be ordinal with more than two response levels—and Leuraud and Benichou (LR) did not include the JT test among those compared, LR's results are not directly comparable to the results in the present study. Nonetheless, LR's methodology and results provide a background for our research. LR performed a Monte Carlo simulation to assess the power of four tests to reject the null hypothesis of no difference across classes against monotone, near-monotone, and strictly nonmonotone alternatives. While their findings were quite complex, it appears that a test based on adjacent contrasts devised by Dosemeci and Benichou (1998) accepted a strictly nonmonotone 'single peak' alternative as monotonic with the smallest frequency—a desirable property in the NSPY analyses. LR also reported that neither the isotonically modified chi-test test due to Barlow et al (1972), nor the modified Mantel test for overall trend (1963) performed well for this alternative.

Simpson and Margolin (1990) investigated the asymptotic properties of a class of nonparametric tests for dose-response curves subject to what SM termed 'downturns' at high doses—defined as an initial monotone increase in the dose-response followed by a monotone decrease through the highest dose—and found that for many parametric distributions, the JT test lacks robustness to a downturn. Simpson and Margolin also found that certain related, but considerably more complicated, statistics outperformed the JT test when the alternative included a downturn.

2. MODIFICATION FOR COMPLEX SAMPLE DESIGN

The sample design for NSPY is complex. It employs a stratified multi-stage design with unequal weights. Sampling weights were prepared for the survey that reflect probabilities of selection and adjustments for nonresponse and undercoverage. Replicate weights were created to allow the consistent estimation of variance of estimated finite population quantities.

Although it might be possible to develop a parametric model for each outcome in terms of exposure, factors with differential sampling rates, and confounders, the project schedule did not allow time to develop separate models for each outcome variable. A strong advantage of the propensity scoring method is that it can be used to develop a set of counterfactual projection weights. With this single set of weights, it is possible to remove the confounding effects of all the variables that were used in the propensity scoring. The combination of propensity scoring and the Jonckheere-Terpstra test promised great speed in the production of a large number of tables that had been designed to look for Campaign effects on a variety of outcomes within a variety of population domains. In order to preserve this benefit of fast production of analytic tables while reflecting the nonignorability of the sample with respect to family composition and age of housing, as well as incorporating adjustments for differential response rates and undercoverage, it was decided to base the counterfactual projection weights on the survey sampling weights and then to calculate the Jonckheere-Terpstra test using these counterfactual projection weights. The question then was how to adjust the JT for the 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 and Sudaan.

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 (c.f., Manda, 2002). We assumed that the problems identified for other types of analysis would also impact the JT unfavorably if we were to compute it from a weighted contingency table where the weights had been standardized. So we wanted to do something for the JT similar to Fay's or Rao and Scott's corrections to chi-square tests for independence.

Since the JT has an asymptotic normal distribution under the null hypothesis of independence, it seemed like a straightforward procedure would be to replicate the JT on each of the replicated weights, then calculate a variance on the replicated JT, and finally use this in a z-test. More specifically, let J_0 be the standardized JT statistic formed on the contingency table of Y by Z using full-sample weights and let J_r be the standardized JT 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 "jackknifed" Jonckheere-Terpstra test is

$$JJT = \frac{J_0}{\sqrt{\sum_r b_r \left(J_r - J_0\right)^2}}$$

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.

A very similar formula is used in Wesvar to estimate the variance of regression coefficients, where instead of J_0 and J_r, there is β_0, the estimated regression coefficient calculated using the full sample weights, and β_r, the estimated regression coefficient calculated using the *r*-th set of replicate weights. Past research has shown that this approach generally works well, although jackknife weights work considerably better than BRR or bootstrap weights for at least some error distributions. Based on our familiarity with jackknifing of regression coefficients and the similarity of the Jonckheere-Terpstra statistic to a regression coefficient, the decision was made to use the jackknifed JT in NSPY reports. The publication schedule did not allow time for the development of theory nor even for simulation studies. Although we have still not attempted to develop theory for the test, we have conducted a simulation study which is the topic of this paper.

3. PARAMETERS FOR SIMULATION STUDY

We chose to simulate only those features of the NSPY design that seemed most likely to impact the performance of the jackknifed JT. 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 nor 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 NSPY was a late-decade survey. 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. In fact, the 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 analysis 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 effects the application of the central limit theorem to survey estimates. Basu's elephant 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 iid $n_i \sim \Gamma(7,5.7)+\Gamma(0.3,33.3)$, rounded to 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 5000, 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 set to all equal 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

$$\mu_i \sim N\left(1-\left(\frac{4}{n_i}\right)^{0.3}, \sigma_{y1}^2\right).$$

At the person level, the latent exposure Y scores were simulated as $\mu_{ij} \sim N(\mu_i, \sigma_{y2}^2)$, where σ_{y1}^2 and σ_{y2}^2 were varied to create different levels of intraclass correlation,

$$\rho_y = \frac{\sigma_{y1}^2}{\sigma_{y1}^2 + \sigma_{y2}^2} \ ,$$

while keeping the total variance $\sigma_{y1}^2 + \sigma_{y2}^2 = c$ fixed.

These latent scores for *Y* were then translated to manifest ordinal scores 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 $\xi_i \sim N\left(0, \sigma_{z1}^2\right)$, and the PSU-level perturbation in the strength of the exposure-outcome relationship was simulated as $\theta_i \sim N\left(\alpha, \sigma_\theta^2\right)$, where α can be thought of as a slope and σ_θ^2 as a measure of the between-PSU variance in the strength of the dose-response relationship.

At the person level, the latent Z scores were simulated as

$\xi_{ij} \sim N(\xi_i + f_\lambda(y_{ij}), \sigma^2_{z2})$, where 8 shapes were selected for the dependence of the latent *Z* score on the manifest *Y* score and σ^2_{z1} and σ^2_{z2} were again varied to create variety of intraclass correlations. The eight patterns were defined as follows:

Flat:	$f_A(y_{ij}) = 1.5\theta_i / 3$
Linear:	$f_B(y_{ij}) = \theta_i y_{ij} / 3$
Square root:	$f_C(y_{ij}) = \theta_i \sqrt{y_{ij} / 3}$
Fourth Power:	$f_D(y_{ij}) = \theta_i \left(y_{ij} / 3 \right)^4$
Early jump:	$f_E(y_{ij}) = \theta_i \left[\text{if } y_{ij} > 0 \right]$ and 0 otherwise
Late jump:	$f_F(y_{ij}) = \theta_i \left[\text{if } y_{ij} = 3 \right]$ and 0 otherwise
Central Butte:	$f_G(y_{ij}) = \theta_i \left[\text{if } y_{ij} = 1 \text{ or } 2 \right]$ and 0 otherwise
Early spike:	$f_H(y_{ij}) = \theta_i \left[\text{if } y_{ij} = 1 \right]$ and 0 otherwise

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 for the average of 10 random draws from the joint distribution with α=0.66. and σ^2_θ=0. Patterns B through F reflect a monotone relationship. Pattern A reflects independence, and Patterns G and H reflect nonmonotone dependence. We varied the level of α to create patterns that were more or less easy to detect.

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

$$\rho_z = \frac{\sigma^2_{z1} + \sigma^2_\theta / 4}{\sigma^2_{z1} + \sigma^2_\theta / 4 + \sigma^2_{z2}} .$$

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 2000 draws from this distribution for each dependence pattern and level of intraclass correlation. We hoped that the null hypothesis of independence would be rejected only 5 percent of the time for Patterns A, G, and H, and that power to reject the null hypothesis would be reasonably strong for patterns B through F. In this simulation, we only compared the jackknifed JT against the ordinary JT. It would be interesting to compare the performance of other tests such as that proposed by Leuraud and Benichou, but we did not attempt it. We expected that the ordinary JT would have higher power for all patterns, which of course, is only good for patterns B through F.

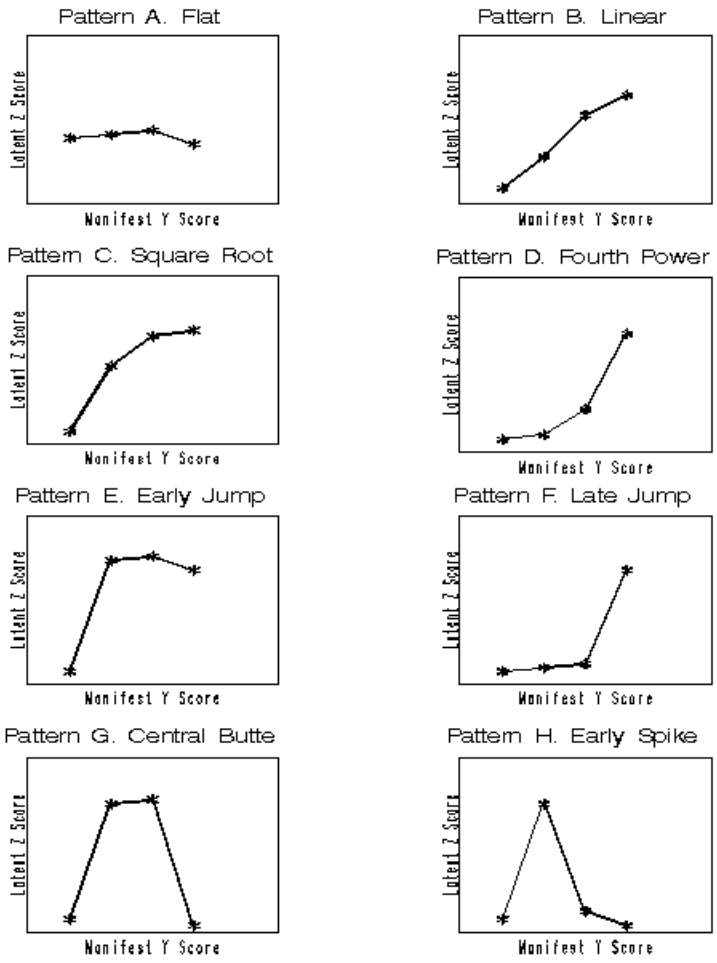

Figure 1. Eight Shapes Tested in Simulation

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 1 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 the ordinary JT performs very well for small to moderately large intraclass correlation. In most surveys, an intraclass correlation at the PSU level no larger than 0.005 would be expected. It is only when intraclass correlation is about 0.05 or more that the ordinary JT no longer provides the nominal significance level. If the intraclass correlation is about 0.01, then the ordinary JT is much too aggressive. The jackknifed JT 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.

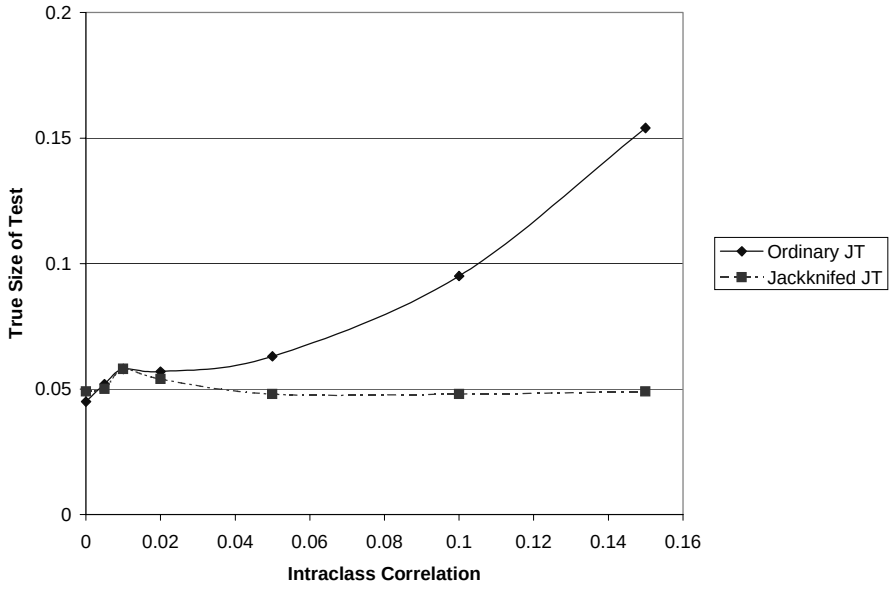

Figure 2. Probability of rejecting the null hypothesis of a flat relationship when the relationship is indeed flat (same intraclass correlation on *Y* and *Z*, total variance =35 on each)

Figure 3 focuses on power of the jackknifed JT for monotone patterns with α=0.66 and σ_θ^2=0. Here we see that power is always best for a linear pattern. The power for smooth nonlinear patterns is next best. Power is worst for nonsmooth nonlinear patterns where all the variation in the outcome is related to just one of the exposure transitions. We also see that power decreases for all patterns as intraclass correlation increases, as we had expected.

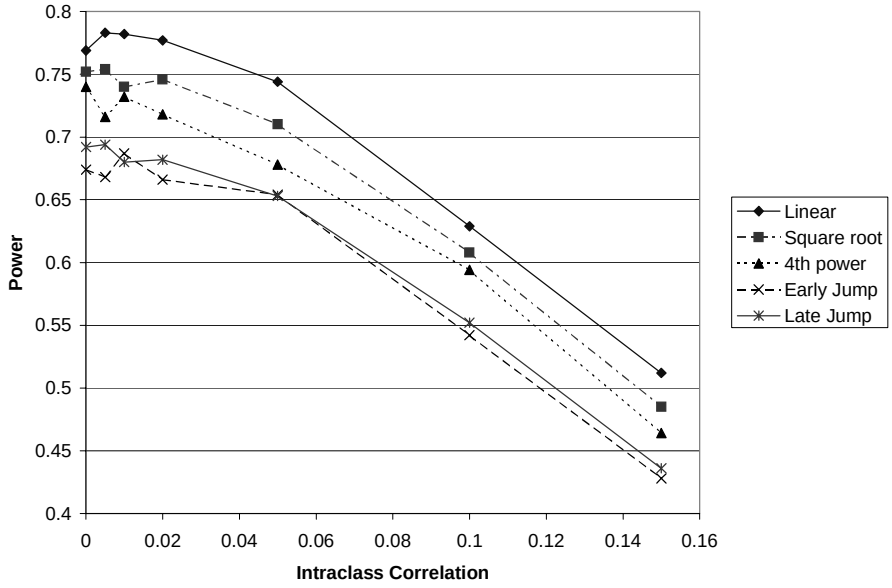

Figure 3. Power with the jackknifed JT for various monotone patterns and slopes

Figure 4 shows power for dependent but nonmonotone patterns. In the framework of the JT, these are patterns that have been ruled out a priori as not sensible. 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 JT in the hope that it would reject the null hypothesis no more than 5 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 JT generally has lower power than the ordinary JT, using the jackknifed JT does result in fewer false claims of monotone trends even when there is an Early Spike pattern.

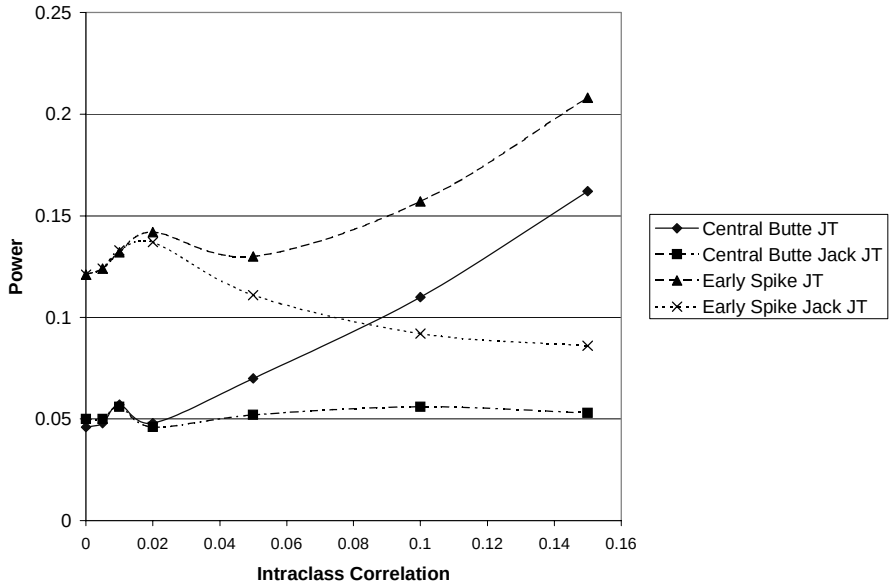

Figure 4. Power with the ordinary and jackknifed JT for various dependent but nonmonotone patterns and slopes

Figure 5 shows how the power of the jackknifed JT depends on the source of the intraclass correlation on 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% down to the range of 32 to 37%. 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%. This power loss may not be a bad thing in the sense that the JT 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.

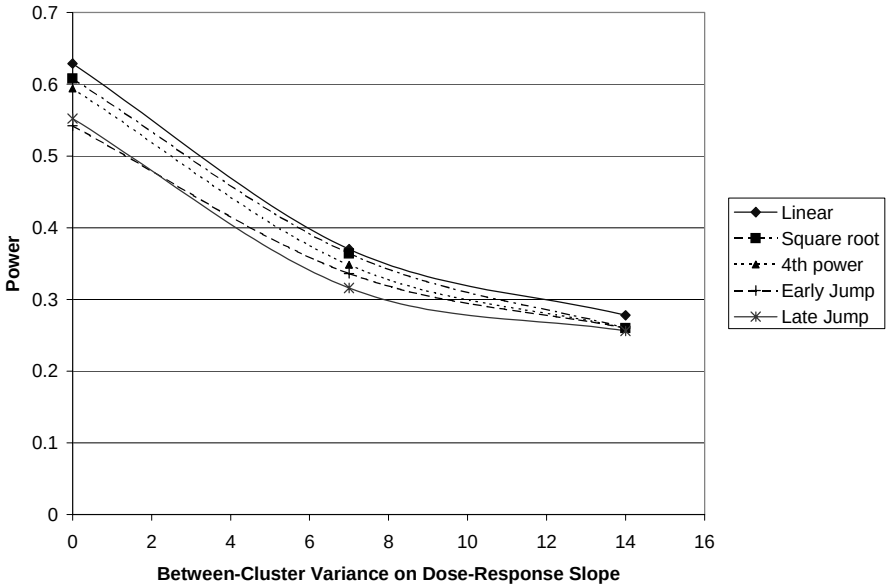

Figure 5. Power of the jackknifed JT by source of intraclass correlation on outcome

5. RECOMMENDATIONS

The jackknifed JT has been shown to be a reasonable test 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 JT when the true size of the nominal JT is close to its nominal size. The jackknifed JT 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 that the ordinary JT when there is strong clustering.

The advantages of the jackknifed JT are most apparent at levels of intraclass correlation that may not often be achieved. 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.

REFERENCES

Barlow, R. E., Bartholmew, D. J., Bremner, J. M., and Brunk, H. D. (1972), *Statistical Inference under Order Restrictions. The Theory and Application of Isotonic Regression*, New York:Wiley.

Collings, B. J., Margolin, B. H., and Oehlert, G. W. (1981), "Analyses for Binomial Data, with Applications to the Fluctuation Test for Mutagenicity", Biometrics 37, pp. 775-794.

Dosemeci M., and Benichou J. (1998), "An alternative test for trend in exposure-response analysis", *Journal of Exposure Analysis and Environmental Epidemiology*, 8 (1), pp. 9-15.

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.

Hornik, R., Maklan, D., Orwin, R., Cadell, D., Judkins, D., Barmada, C., Yanovitzky, I., Moser, M., Zador, P., Southwell, B., Baskin, R., Morin, C., Jacobsohn, L., Prado, A., and Steele, D., (2001). Evaluation of the National Youth Anti-Drug Media Campaign: Third Semi-Annual Report of Findings - October 2001. Rockville, Maryland: Westat.

Imbens, G. W. (1999), "The role of propensity score in estimating dose-response functions". Technical Working Paper 237 (http ://www.nber.org/papers/T0237). Cambridge, MA:NBER.

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

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.

Leraud, K., and Benichou J. (2001), "A comparison of several methods to test for the existence of monotone dose-response relationship in clinical and epidemiological studies", *Statist. Med*., 20, pp. 3335-3351.

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.

Rosenbaum P. R. (1995), *Observational Studies*, New York:Springer-Verlag.

Rosenbaum P., and Rubin, R. (1983), "The central role of propensity score in observational studies for causal effects", *Biometrika*, 70, pp. 41-55.

SAS/STAT Software: Changes and Enhancements for Release 6.12. (1996), Carey, N.C.:SAS Institute, Inc.

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.

Simpson D. G., and Margolin B. H. (1990), "Nonparametric testing for dose-response curves subject to downturns: asymptotic power considerations", *The Annals of Statistics*, 18, No 1, pp. 372-390.

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.

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.