Statistical Package for Improved Analysis of Hillslope Monitoring Data Collected as Part of the Board of Forestry's Long-Term Monitoring Program

Agreement No. PSW-96-CL-032, CDF #8CA95056

Final Report

Jack Lewis

Jim Baldwin

May 1997

Table of Contents

1. Introduction

The State of California has embarked upon a Long-Term Monitoring Program whose primary goal is to assess the effectiveness of the Forest Practice Rules and Review Process in protecting the beneficial uses of waters from the impacts of timber operations on private timberlands. The Board of Forestry's Monitoring Study Group concluded that hillslope monitoring should be used to assess Rule implementation and effectiveness at preventing erosion and maintaining canopy cover, while instream monitoring would be used to assess conditions and trends in water quality and stream channels.

Hillslope monitoring objectives include evaluating how well the Rules: 1) are implemented, 2) keep soil on hillslopes away from stream channels, and 3) maintain riparian zones, including canopy cover along watercourses. A plan has been drafted for achieving these objectives and it is discussed and described in detail in a report by Andrea Tuttle (1995). To briefly summarize, evaluation forms and procedures were developed for five transect types: 1) roads, 2) landings, 3) skid trails, 4) watercourse crossings, and 5) water and lake protection zones (WLPZs). The physical characteristics and logging activities on these transects are described in detail and erosion problems are identified. Rule implementation is evaluated for three types of sampling units: 1) individual points where problems have been identified, 2) entire transects, and 3) entire Timber Harvest Plans. Rule effectiveness is defined as the absence of certain types of problems for individual points and entire transects, but is defined in terms of impacts on the beneficial uses of water for entire Timber Harvest Plans (THPs).

Under a Pilot Monitoring Program, hillslope monitoring data were collected on 17 THPs from 1993 to 1994 and were available for this investigation. Data recorded during the summer and fall of 1996 from an additional 50 THPs were not available at the time this report was prepared.

The objectives and tasks of this report (distilled from the Study Plan) are to:

1. Evaluate and propose various statistical techniques, including nonparametric statistics, that can be used to analyze the kind of data collected on the 17 THPs during the Pilot Monitoring Program.
2. Provide guidelines for determining necessary sample sizes to achieve a specified degree of certainty in hypothesis testing and estimation.
3. Evaluate the sampling design and provide advice on how it might be improved. Assess the scope of the inferences that can be made.
4. Discuss the statistical compromises that may be necessary and recommend ways to address identified weaknesses.

This Final Report consists of six sections including this introduction. Section 2 proposes a variety of specific techniques for evaluating implementation and effectiveness and presents methods for estimating required sample sizes. Section 3 describes methods for validating the repeatability of subjective evaluations. Section 4 is a general discussion of some of the statistical compromises that might be necessary including issues related to sampling design. Section 5 addresses specific weaknesses identified in the study. Section 6 summarizes our conclusions. In addition, the Appendices list source code from S-Plus, SAS, and BASIC to perform the recommended analyses.

2. Specific Techniques

Measures of association for ordinal variables

The basic data for evaluating rule implementation and effectiveness was presented as a 2x2 contingency table by Tuttle (1995), reflecting two categories for each variable. The two categories of effectiveness are presence or absence of a problem, defined as "an observable erosion or sedimentation feature, or failed erosion control structure which is not due to some unrelated cause." However, seven implementation categories have been defined in the pilot phase. Three of these categories (rule cannot be evaluated, criteria not applicable at this site, cannot determine because evidence is masked) are uninformative, leaving four substantive categories:

1. Exceeds forest practice rule or THP requirements
2. Meets forest practice rule or THP requirements
3. Minor departure from forest practice rule or THP requirements
4. Major departure from forest practice rule or THP requirements

Implementation is recorded as an ordinal categorical variable, *i.e.,* the categories have a natural ordering, but the relative distances between the categories are unknown. The analysis should reflect the fact that implementation is recorded as an ordinal categorical variable. If these categories are lumped into two categories, *e.g.,* "adequate" and "inadequate", or if the ordering of the categories is ignored, the analysis will have less power for detecting alternatives to null hypotheses such as that of independence. Methods for describing and analyzing ordinal data are more akin to those (*e.g.,* correlation and regression) for continuous (interval) data than to methods for nominal categorical data.

Dichotomous variables, such as the effectiveness measure employed in this study, can be considered either nominal or ordinal because there are only two possible orderings and these orderings are opposite. If such a variable is treated with ordinal methods, reversing the categories changes the direction but not the magnitude of ordinal association and does not affect the conclusions.

Concordance and discordance

For ordinal data with two variables such as implementation and effectiveness one wants to assess the question "Does level of effectiveness increase as level of implementation increases?" For interval-scale data a common measure of association is the Pearson correlation coefficient. But most of the commonly used measures of association for ordinal variables are based on the numbers of concordant and discordant pairs in a sample among all possible pairings of observations. A concordant pair is one in which one member of the pair ranks higher on both variables. A discordant pair is one in which one member ranks higher on variable X and the other observation ranks higher on variable Y. Pairs in which a tie occurs on variable X or Y or both are considered neither concordant nor discordant. Positive association is indicated when the number of concordant pairs is large relative to the number of discordant pairs. To illustrate, consider the following example in Table 2.1.

The number of concordant pairs is

$$C = \sum_{i<k} \sum_{j<l} n_{ij} n_{kl} = 3(7 + 4 + 2) + 6(4 + 2) + 2(2) = 79$$

Table 2.1. Classification of 26 evaluations for rule implementation and problem occurrence.

	Exceeds Rule	Meets Rule	Minor Departure	Major Departure	**Total**
No Problem	3	6	2	1	**12**
Problem	1	7	4	2	**14**
Total	**4**	**13**	**6**	**3**	**26**

The first summation in each equation is over all pairs of rows satisfying $i<k$ and the second summation is over all pairs of columns satisfying $j<l$. The number of discordant pairs is

$$D = \sum_{i<k} \sum_{j<l} n_{ij} n_{kl} = 6(1) + 2(1 + 7) + 1(1 + 7 + 4) = 34$$

The number of pairs tied on the row variable (X) is the total of the numbers of pairs in the same row:

$$T_X = \sum_i \frac{n_{i+}(n_{i+} - 1)}{2} = \frac{12(11)}{2} + \frac{14(13)}{2} = 157$$

in which n_{i+} is the number of observations in row i. Similarly the number of pairs tied on the column variable (Y) is the total of the numbers of pairs in the same column:

$$T_Y = \sum_i \frac{n_{j+}(n_{j+} - 1)}{2} = \frac{4(3)}{2} + \frac{13(12)}{2} + \frac{6(5)}{2} + \frac{3(2)}{2} = 102$$

Pairs of observations from the same cell are tied on both X and Y. The number of these pairs is

$$T_X = \sum_i \sum_j \frac{n_{ij}(n_{ij} - 1)}{2} = \frac{3(2)}{2} + \frac{6(5)}{2} + \frac{2(1)}{2} + \frac{1(0)}{2} + \frac{1(0)}{2} + \frac{7(6)}{2} + \frac{4(3)}{2} + \frac{2(1)}{2} = 47$$

To verify these counts, the total number of pairs can be calculated two ways

$$\frac{n(n - 1)}{2} = \frac{26(25)}{2} = 325$$
$$C + D + T_X + T_Y - T_{XY} = 79 + 34 + 157 + 102 - 47 = 325$$

in which n is the total number of observations. Several measures of association based on the quantity $C - D$ have been proposed. The direction of the association is determined by the sign of $C - D$.

Gamma

Among untied pairs, the proportion of concordant pairs is $C / (C + D)$ and the proportion of discordant pairs is $D / (C + D)$. The difference between these two proportions is the measure *gamma* proposed by Goodman and Kruskal (1954), estimated by

$$\hat{\gamma} = \frac{C - D}{C + D}$$

The range of *gamma* is $-1 \leq \gamma \leq 1$, with positive values indicating positive association and negative values indicating negative association. For the example in Table 2.1,

$$\hat{\gamma} = \frac{79 - 34}{79 + 34} = 0.398$$

Somers' *d*

A similar measure to *gamma* was proposed by Somers (1962). It differs in that pairs untied on X serve as the base rather than only those untied on both X and Y:

$$d_{YX} = \frac{C - D}{(n(n - 1) / 2) - T_X}$$

Somers' d_{YX} is an asymmetric measure intended for use when Y is a response variable. When X is the response variable, the appropriate measure is based on pairs untied on Y:

$$d_{XY} = \frac{C - D}{(n(n - 1) / 2) - T_Y}$$

For the example in Table 2.1, the response variable is X (effectiveness), so d_{XY} would be used:

$$d_{XY} = \frac{79 - 34}{((26)(25) / 2) - 102} = 0.202$$

Kendall's *tau* and *tau-b*

This coefficient of rank correlation was introduced by Kendall (1938) for continuous variables. It differs from *gamma* and Somers' *d* only in the denominator, which is the total number of pairs of observations,

regardless of the number of ties. The sample estimate is

$$\hat{\tau} = \frac{C - D}{(n(n-1)/2)} = \frac{79 - 34}{(26)(25)/2} = 0.139$$

Because a large proportion of the pairs are tied when the measurements are categorical, this measure cannot attain very large values, and Kendall (1945) proposed another measure, *tau-b*, estimated by:

$$\hat{\tau}_b = \frac{C - D}{\sqrt{\left[(n(n-1)/2 - T_X\right]\left[n(n-1)/2 - T_Y\right]}} = \sqrt{d_{YX} d_{XY}}$$

which is the geometric mean of Somers' d_{YX} and d_{XY}. This relationship is analogous to the relationship $r^2 = b_{YX} b_{XY}$ between the Pearson correlation r and the least-squares slopes for regressing Y on X and X on Y. In fact, Kendall's *tau-b* and Somers' *d* are special cases of the correlation and regression slope for a model in which the observations are defined as signs of the X and Y differences between pairs of observations. For our example data, $\hat{\tau}_b = \sqrt{d_{YX} d_{XY}} = \sqrt{(0.268)(0.202)} = 0.232$.

Ridit analysis

Because effectiveness has been recorded as a dichotomous variable, methods designed for one nominal and one ordinal variable may be just as effective as those designed for two ordinal variables. One such method is ridit analysis. The jth ridit is the proportion of observations below category j of the ordinal variable (Y) plus half the proportion in category j. In mathematical terms, the jth ridit is

$$r_j = \sum_{k=1}^{j-1} p_{+k} + \tfrac{1}{2}(p_{+j})$$

where p_{+k} is the proportion of observations in column k of Y, summed over all levels of X (rows). The ridits are ordered $r_1 < r_2 < \cdots < r_c$. Mean ridits are defined for each row as

$$R_i = \sum_j r_j p_{j(i)}$$

where $p_{j(i)}$is the proportion of observations in row i belonging to column j. The mean ridit approximates the probability that a random observation from row i ranks higher *in the underlying continuous distribution of Y* than a random observation from the overall population. In the underlying continuous distribution, it is assumed there are no ties, so the ties in the sample are split between the two possible inequalities. The quantity $R_1 - R_2 + 0.5$ approximates the probability that a random observation

from row 1 ranks higher than a random observation from row 2. Also, the quantity $2(R_2 - R_1)$ is Somers' d_{YX}. In our example (Table 2.1), the ridits are

$$r_1 = \tfrac{1}{2}(4) / 26 = 0.077$$
$$r_2 = [4 + \tfrac{1}{2}(13)] / 26 = 0.404$$
$$r_3 = [17 + \tfrac{1}{2}(6)] / 26 = 0.769$$
$$r_4 = [23 + \tfrac{1}{2}(3)] / 26 = 0.942$$

and the mean ridits are

$$R_1 = [(.077)(3) + (.404)(6) + (.769)(2) + (.942)(1)] / 12 = 0.428$$
$$R_2 = [(.077)(1) + (.404)(7) + (.769)(4) + (.942)(2)] / 14 = 0.562$$

The probability that a random observation from row 1 ranks higher in the underlying continuum for Y than a random observation from row 2 is approximately 0.428-0.562+0.5=0.366.

Inference for ordinal measures of association

If the population proportions of concordant and discordant pairs are denoted by Π_c and Π_d, respectively, the null hypothesis of independence can expressed as $H_0: \Pi_c = \Pi_d$, with alternatives corresponding to the possible inequalities. In evaluating the association between implementation and effectiveness of forest practice rules, the logical alternative hypothesis is $H_A: \Pi_c > \Pi_d$. That is, a positive association is expected. The measures of association based on the quantity $C - D$ do not require separate tests, because their population values always have the same sign as $\Pi_c - \Pi_d$ and Simon (1978) showed that all measures having numerator $C - D$ have the same local power for testing independence. For sample sizes greater than 10, the normal distribution satisfactorily approximates the distribution of $C - D$ (Kendall, 1970). In the case where there are ties in both rankings and one of the rankings is a dichotomy, the variance of the normal approximation for $C - D$ is estimated by

$$\hat{\sigma}^2_{C-D} = \frac{n_{1+}n_{2+}}{3n(n-1)}\left(n^3 - n - \sum_j (n^3_{+j} - n_{+j})\right)$$

The distribution of $C - D$ is discontinuous and the possible values are generally separated by several units. To improve the normal approximation, we regard the frequency at $C - D$, instead of being concentrated at one point, as being spread uniformly over an interval equal to the average distance d between the discrete points of the distribution. This distance is equal to twice the average distance between midranks. Thus

$$d = 2\left(\frac{m_c - m_1}{c - 1}\right)$$

where c is the number of columns or categories of the ordinal variable Y, and m_i are the average ranks (midranks) for the levels of Y. Since all observations in a given column are tied with respect to Y, there

are c distinct midranks. The midranks are related to ridits and can be computed as

$$m_j = nr_j + \tfrac{1}{2} = \sum_{k=1}^{j-1} n_{+k} + \tfrac{1}{2}(n_{+j} + 1)$$

If $C - D$ is positive, *0.5d* should be subtracted from $C - D$ to correct for continuity, thus the test statistic to be compared with the standard normal distribution is

$$z = \frac{C - D - \left(\frac{m_c - m_1}{c - 1}\right)}{\hat{\sigma}_{C-D}}$$

For the example in Table 2.1,

$$\hat{\sigma}^2_{C-D} = \frac{14(12)}{3(26)(25)}\left(26^3 - 26 - (4^3 - 4) - (13^3 - 13) - (6^3 - 6) - (3^3 - 3)\right) \cong 1299$$

$$m_1 = nr_1 + 0.5 = 26(0.077) + 0.5 = 2.5$$

$$m_c = nr_4 + 0.5 = 26(0.942) + 0.5 = 25$$

$$z = \frac{79 - 34 - \left(\frac{25 - 2.5}{3}\right)}{\sqrt{1299}} = 1.041$$

$$\mathrm{P}\left[C - D > 1.041 \mid \mathrm{H}_0\right] \cong 0.149$$

We can avoid resorting to approximations if we assume that sampling was conducted in such a way as to produce tables with fixed marginal totals. Under that assumption, the probability distribution of the cell counts is the multiple hypergeometric distribution, and exact significance tests can be computed. For $\mathrm{H_A}: \Pi_c > \Pi_d$, the procedure consists of determining the probability that the observed value of $C - D$ would be exceeded among random samples matching the observed set of marginal totals. If every possible sample were enumerated, then the proportion of samples in which $C - D$ exceeds the observed value would provide a p-value for the one-sided test of $\mathrm{H}_0: \Pi_c = \Pi_d$. For a simple example, see Agresti (1990, p. 64).

Mann-Whitney-Wilcoxon test

It happens that the exact test for $2 \times c$ contingency tables is equivalent to a Mann-Whitney-Wilcoxon test using midranks for the levels of the ordinal variable (Agresti, 1990, p. 70). This test is one of the most frequently used and best known distribution-free tests. It is used to test whether two independent distributions *F(x)* and *G(x)* are equal for all *x* against the alternative that *F(x)≥G(x)* for all *x* with at least one strict inequality. Another way of stating the alternative is that, if *X1* and *X2* are two independent random variables with respective distributions *F(x)* and *G(x)*, then *X1* is *stochastically larger* than *X2*.

To illustrate how the Mann-Whitney-Wilcoxon test is applied to contingency table data, the example data is shown once again (Table 2.2), with a few additional rows. The Wilcoxon rank-sum test is based on the sum of the midranks, $W_i = \sum_j W_{ij}$ where $\mathrm{W_{ij}} = n_{ij} m_j$. It is equivalent to the Mann-Whitney U test, described next.

Table 2.2. Calculations for Mann-Whitney and Wilcoxon statistics

	Exceeds Rule	Meets Rule	Minor Departure	Major Departure	**Total**
No Problem	3	6	2	1	12
Problem	1	7	4	2	14
Total	4	13	6	3	26
m_j	2.5	11	20.5	25	**NA**
W_{1j}	7.5	66	41	25	139.5
W_{2j}	2.5	77	82	50	211.5
U_{1j}	1.5	27	20	13	61.5
U_{2j}	1.5	42	40	23	106.5

The cells in the row labelled U_{1j} give the number of pairs involving column *j* of row 1 in which implementation exceeds that of an observation in row 2. Similarly, the cells in the row labeled U_{2j} give the number of pairs involving column *j* of row 2 in which implementation exceeds that of an observation in row 1. Ties are counted as one-half. The formulas for these counts are

$$U_{1j} = n_{1j}(\sum_{k=1}^{j-1} n_{2k} + \tfrac{1}{2}n_{2j})$$

$$U_{2j} = n_{2j}(\sum_{k=1}^{j-1} n_{1k} + \tfrac{1}{2}n_{1j})$$

The Mann-Whitney *U* statistics are $U_i = \sum_j U_{ij}$ and are linearly related to the Wilcoxon statistics:

$$U_i = W_i - \frac{n_{i\cdot}}{2}(n_{i+} + 1)$$

A simple check on the validity of the calculations is

$$U_1 + U_2 = n_{1+}n_{2+}$$

Most published tables for the *U* statistic are not valid for situations involving large proportions of tied observations. However, programs are available which compute the exact probabilities for the *U* statistic in cases with many ties. Jerome Klotz's wilcox program, available as freeware, (http://justzaam.stat.wisc.edu/~klotz/Wilcox.zip), is one example. To run wilcox, one enters the *U* statistic and row total for the smaller group, the number of columns, and the column sums (referred to as

tie groups by the program). The U statistic for a $2 \times k$ table can be obtained using the S-Plus function U2k (Appendix A). For the example data, the program gives the result, under the null hypothesis, that

$$P[U_1 \leq 61.5] = 0.1288$$

Other programs which compute these probabilities are Jerry Dallal's PC-Pitman shareware program (http://www.simtel.net/pub/simtelnet/msdos/statstcs/pitman5.zip), or the commercial program StatXact (http://www.cytel.com/statxact.html).

Confidence intervals for ordinal measures of association

The measures of association discussed above (*gamma*, Somers' *d*, and Kendall's *tau-b*) have asymptotic normal distributions (under full multinomial sampling), which can be used to obtain variances via the delta method (Bishop, *et. al.,* 1975). Large-sample confidence intervals for the population parameter values may be constructed from the asymptotic standard errors (Agresti, 1984, p. 185-188). For *gamma*, the asymptotic standard error is estimated by

$$ASE[\hat{\gamma}] = \left[\frac{\sum \sum p_{ij} \phi_{ij}^2}{n(P_c + P_d)^4} \right]^{1/2}$$

where $p_{ij} = n_{ij} / n$, $P_c = 2C / n^2$, $P_d = 2D / n^2$, and

$$\phi_{ij} = 4[P_d p_{ij}^{(c)} - P_c p_{ij}^{(d)}]$$

The terms $np_{ij}^{(c)}$ and $np_{ij}^{(d)}$ represent the number of observations that are concordant and discordant, respectively, when matched with cell (*i,j)*.

For Somers' *d*, the asymptotic standard error is estimated by

$$ASE[d_{YX}] = 2\left[\frac{\sum \sum p_{ij} \phi_{ij}^2 - (P_c - P_d)^2}{n\delta^4} \right]^{1/2}$$

where

$$\delta = 1 - \sum_i p_{i+}^2$$

and

$$\phi_{ij} = p_{i+}(P_c - P_d) + \delta(p_{ij}^{(c)} - p_{ij}^{(d)})$$

For Kendall's *tau-b*, the asymptotic standard error is estimated by

$$ASE[\hat{\tau}_b] = \left[\frac{\sum\sum p_{ij}\phi_{ij}^2 - \left(\sum\sum p_{ij}\phi_{ij}\right)^2}{n(\delta_1\delta_2)^4} \right]^{1/2}$$

where

$$\delta_1 = (1 - \textstyle\sum_i p_{i+}^2)^{1/2}$$

$$\delta_2 = (1 - \textstyle\sum_j p_{+j}^2)^{1/2}$$

and

$$\phi_{ij} = \delta_1\delta_2\left[2(p_{ij}^{(c)} - p_{ij}^{(d)}) + (P_c - P_d)p_{+j}\right] + \frac{\delta_2}{\delta_1}(P_c - P_d)p_{i+}$$

S-Plus code is provided in Appendix B for calculating gamma, Somers' *d*, Kendall's *tau-b*, and their asymptotic standard errors. Appendix C lists the equivalent code for SAS. The example (Tables 2.1 and 2.2) may be a bit small for accurate use of the asymptotic approximations, but will be used for the sake of illustration. For the example, $\hat{\gamma} = 0.398$ and $ASE[\hat{\gamma}] = 0.286$, leading to the 95% confidence interval [-0.16,0.96] for gamma. Similarly, $d_{XY} = 0.202$ and $ASE[d_{XY}] = 0.149$, leading to the 95% confidence interval [-0.09,0.49] for Somers' *d*. Finally, $\hat{\tau}_b = 0.232$ and $ASE[\hat{\tau}_b] = 0.173$, leading to the 95% confidence interval [-0.11,0.57] for $\hat{\tau}_b$. Note that all confidence intervals include 0, which is in concordance with the results of the Mann-Whitney-Wilcoxon test.

Partial association and conditional independence

In a cross-sectional study such as this one, there are a myriad of uncontrolled variables which could be influencing the results. Two-dimensional contingency tables are essentially collapsed over all the values of the uncontrolled variables. If an association is observed, there is always the possibility that it is due to the association between each of the two variables and a third, perhaps unmeasured, variable. For example, among a group of children of diverse ages, an association would likely be found between vocabulary size and height. There is probably no functional relation; the association simply results from the fact that both vocabulary size and height increase with age. Vocabulary size and height are *conditionally independent* for any given value of age. That is, there is no partial association (holding age fixed), even though the marginal table (which ignores age) would indicate an association.

In fact marginal tables can indicate a positive association even when the partial associations are *negative*. This situation, called Simpson's paradox, is illustrated by the following example, presented first by Radelet (1981). At issue was whether race (white, black) of a defendant affects the decision whether to impose the death penalty. The basic (marginal) table shows a slight association between the defendant's

Defendant's race	Death	No Death	Proportion Death
White	19	141	0.119
Black	17	149	0.102

race and imposition of the death penalty. The association is not statistically significant, but the proportion of white defendants receiving the death penalty was 0.017 higher than that of black defendants. Now, if the victim's race is considered, two tables indicating the partial associations can be

Victim's race	Defendant's race	Death	No Death	Proportion Death
White	**White**	19	132	0.126
White	**Black**	11	52	0.175
Black	**White**	0	9	0.000
Black	**Black**	6	97	0.058

constructed. When the victim was white, the proportion of white defendants receiving the death penalty was 0.049 lower than that of black defendants; and when the victim was black, the proportion of white defendants receiving the death penalty was 0.058 lower than that of black defendants. The apparent contradiction can be understood if it is recognized that:

1. When the victim is white, the death penalty was imposed in 14% of cases, compared to 5% of cases with black victims
2. 94% of the victims of white killers are white, but only 38% of the victims of black killers are white.

So white killers kill whites much more than black killers do, hence they get the death penalty more often. But black killers get the death penalty more often when only black victims are considered or only white victims are considered.

Figure 2.1 (adapted from Agresti, 1990, p. 139 with permission from John Wiley and Sons) provides further insight into the apparent paradox. The letters inside the circles indicate the color of the victim and the sizes of the circles are proportional to the number of observations at that combination of the victim's and defendant's races. The graph clearly shows four conditions.

1. Most white victims are killed by whites and most black victims are killed by blacks.
2. The death penalty is applied more when whites are murdered than when blacks are murdered, regardless of the defendant's race.
3. For either white or black victims, the death penalty is imposed more when the defendant is black.
4. The marginal association between defendant's race and death penalty is reversed from the partial associations.

The graph also illuminates the mechanism behind the reversal of the marginal association. When collapsing over the victim's race, the death penalty rates are not halfway between those for cases involving black and white victims. The percent of white defendant's receiving the death penalty is dominated by cases in which the victim is white and the percent of black defendant's receiving the death penalty is dominated by cases with black victims. Thus, when ignoring the victim's race, the white defendants' death penalty rate is elevated and the black defendants' rate is depressed, reversing the direction of the marginal association relative to the partial associations.

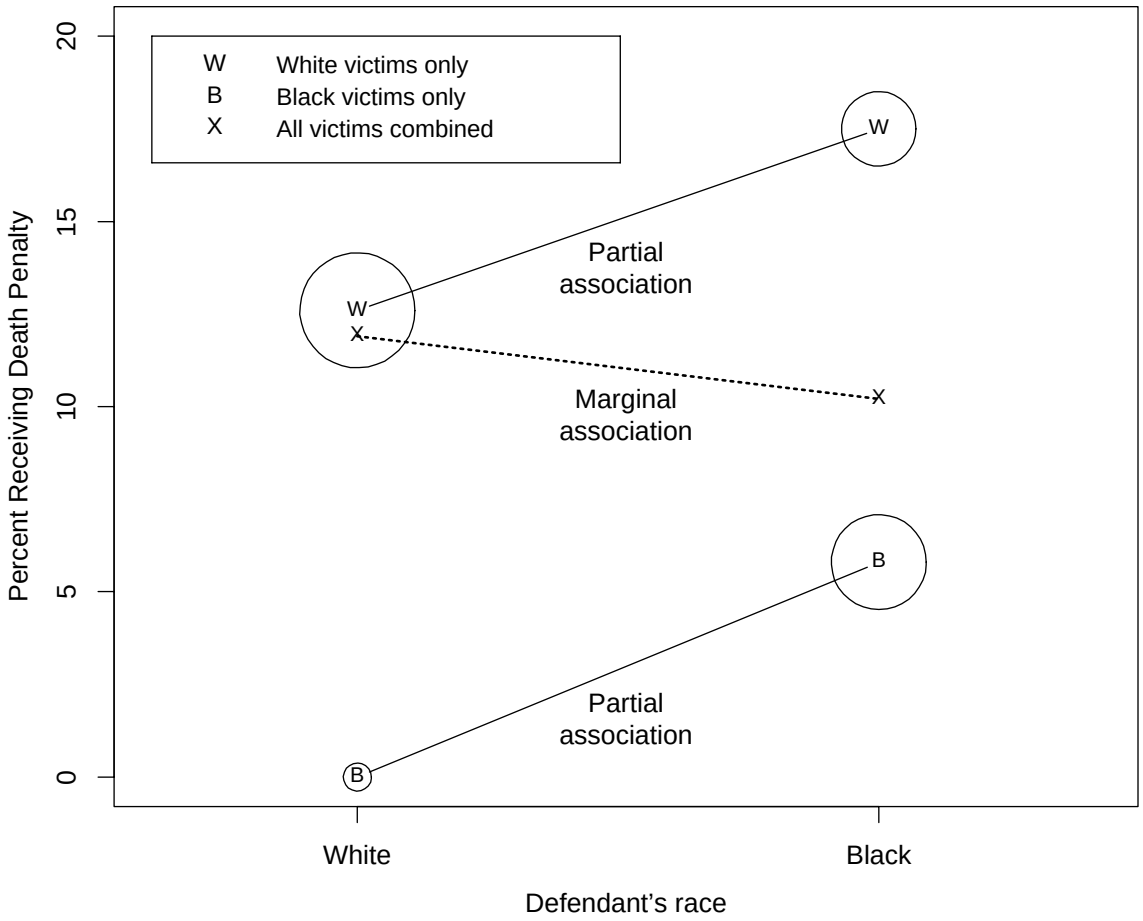

Figure 2.1. Marginal and partial associations between defendant's race and death penalty.

In general, the partial associations between *X* and *Y* for a given level of *Z* are the same as the marginal association *only* when one of the following conditions is true:

1. *X* and *Z* are conditionally independent at all levels of *Y*, or
2. *Y* and *Z* are conditionally independent at all levels of *X*.

If either of these conditions is met, then it is reasonable to collapse the partial tables into a single marginal table to describe the association of *X* and *Y*. On the other hand, if *X* is dependent on *Z* at one or more levels of *Y*, *and Y* is dependent on *Z* at one or more levels of *X*, there is the potential that the observed marginal association between *X* and *Y* is induced by their associations with *Z*, and differs from their partial associations at fixed levels of *Z*. This was the case in the death penalty example where both the defendant's race and imposition of the death penalty were associated with the victim's race (*Z*).

To be confident that an observed marginal association between *X* and *Y* is not the result of both *X* and *Y* being associated with some other variable *Z*, all potential correlates should be measured and examined in relation to *X* and *Y*. The two collapsibility conditions listed above could then be tested. If either condition holds, the marginal table of *X* and *Y* is adequate to describe their association. Otherwise, the partial associations should be computed. If the partial associations are similar at all levels of *Z*, then it is useful to employ a measure of association that pools the data within levels of *Z* to arrive at a single measure of partial association.

Mantel score test for conditional independence

The Mantel Score Test is a generalized statistic designed to detect association between ordinal variables. Suppose a monotone conditional relation is expected between X and Y, with the same direction at each level of Z. With the Mantel Score Test, we can test the conditional independence of X and Y against such an alternative. Sets of monotone scores $\{u_i\}$ and $\{v_j\}$ must be assigned to the levels of X and Y, respectively. There is evidence of a positive association between X and Y if, within each level k of Z, the covariance measure $Q_k = \sum_i \sum_j u_i v_j n_{ijk}$ is greater than its expectation under the null hypothesis of independence. The Mantel's M^2 statistic summarizes the covariance measures over all levels of Z:

$$M^2 = \frac{\left(\sum_k [Q_k - \mathrm{E}(Q_k)] \right)^2}{\sum_k \mathrm{Var}(Q_k)}$$

in which

$$\mathrm{E}(Q_k) = \frac{\left(\sum_i u_i n_{i+k} \right) \left(\sum_j v_j n_{+jk} \right)}{n_{++k}}$$

and

$$\mathrm{Var}(Q_k) = \frac{1}{n_{++k}} \left[\sum_i u_i^2 n_{i+k} - \frac{\left(\sum_i u_i n_{i+k} \right)^2}{n_{++k}} \right] \left[\sum_j v_j^2 n_{+jk} - \frac{\left(\sum_j v_j n_{+jk} \right)^2}{n_{++k}} \right]$$

Under the null hypothesis, M^2 has an approximately chi-squared distribution with one degree of freedom.

By swapping variables, the Mantel score test can be used to test the conditional independence of X and Z or Y and Z (these are the collapsibility conditions). However, if the test is not rejected, one still cannot be sure the variables are conditionally independent because an interaction might be present in which partial associations oppose one another. To verify that there is no such interaction, the partial associations need to be examined. Conditional testing methods using log-linear models provide a more rigorous method for identifying the underlying structure of multidimensional tables (Agresti, 1990, pp. 283-286; Bishop, Feinberg, and Holland, pp. 146-149).

Kendall's partial rank-order correlation coefficient

As stated above, if the partial associations are similar at all levels of Z (but differ from the marginal association), then it is useful to employ a measure of association that summarizes the partial association over all levels of Z. There exist versions of both *gamma* and Kendall's *tau* that describe partial associations with a single measure. The *gamma* statistic is rather sensitive to the choice of categories for both X and Y, therefore only the Kendall's partial rank-order correlation coefficient $\tau_{xy.z}$ will be

described. To compute it, one only needs the values of Kendall's *tau-b* (τ_{xy}, τ_{xz}, and τ_{yz}), computed from each of the 3 marginal tables:

$$\tau_{xy.z} = \frac{\tau_{xy} - \tau_{xz}\tau_{yz}}{\sqrt{(1 - \tau_{xz}^2)(1 - \tau_{yz}^2)}}$$

If the partial associations of X and Y are opposite in sign, $\tau_{xy.z}$ may be close to zero and therefore is not an appropriate measure of association in that case. However, when, they are similar, $\tau_{xy.z}$ provides a single measure of the association between X and Y for fixed Z, based on all the data. For the death penalty example, the various values of *tau* to be computed are:

Variables associated	Type of association	Tau subscript	Value
Defendant's race, victim's race	Marginal	*xz*	0.594
Death penalty, victim's race	Marginal	*yz*	0.131
Death penalty, defendant's race	Marginal	*xy*	0.026
Death penalty, defendant's race	Partial, black victims only	*xy.1*	-0.070
Death penalty, defendant's race	Partial, white victims only	*xy.2*	-0.064
Death penalty, defendant's race	Partial, black and white victims	*xy.z*	-0.065

Note that $\tau_{xy.1}$ (partial association for black victims) and $\tau_{xy.2}$(partial association for white victims) are similar and Kendall's partial $\tau_{xy.z}$ lies between the two.

Loglinear Models

Loglinear models have come into wide use in recent years for analyzing categorical data. Consider a contingency table with r rows and c columns, from a population with relative frequencies π_{ij} for each cell (i,j). The row and column membership indicates levels of the variables X and Y. In a loglinear model the expected cell frequencies m_{ij} are modelled as functions of the row and column membership. In a model where row and column identities are thought to be independent variables, the logarithms of the expected frequencies are modelled as an additive relationship:

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y$$

The number of linearly independent parameters in this model is $1 + (r - 1) + (c - 1) = r + c - 1$. A more general model includes a term for the interaction of level i of X and level j of Y:

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

In this model the number of linearly independent parameters is
$1 + (r - 1) + (c - 1) + (r - 1)(c - 1) = rc$, equal to the number of cells in the table. Since there are as many parameters as cells, it is called a saturated model. A saturated model has 0 degrees of freedom and provides a perfect fit for any observed data.

To test the independence hypothesis, the loglinear model for independence is fit to the data using maximum likelihood estimation, producing expected frequencies for each cell. In a 2-way table, the estimated expected frequencies are simply

$$\hat{m}_{ij} = \frac{n_{i+}n_{+j}}{n}$$

The expected frequencies $\hat{m}_{ij}$ may then be compared to the observed frequencies n_{ij} using the chi-squared statistic

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

or the likelihood ratio statistic

$$G^2 = 2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right)$$

both of which, asymptotically, have a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom.

Row effects model

The goal here is to specify a model that is more complex than the independence model but not saturated. This is done by defining scores $\{v_j\}$ for the levels of Y and adding a linear term with a coefficient τ_i for each row defining the linear dependence on the scores.

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \tau_i(v_j - \bar{v})$$

This model has an extra $r-1$ additional linearly independent parameters compared to the independence model, hence the degrees of freedom are $(r-1)(c-1) - (r-1) = (r-1)(c-2)$ and the model is unsaturated when $c > 2$. The maximum likelihood estimates $\hat{m}_{ij}$ do not have a closed form expression, but can be estimated using an iterative scaling procedure (Darroch and Ratcliff, 1972; Fienberg, 1980, p. 63; Agresti, 1984, p. 86, implemented in the Model Selection Loglinear Analysis procedure of SPSS Advanced Statistics™ 6.1) or a Newton-Raphson method (implemented in the computer package GLIM). With the $\hat{m}_{ij}$, the likelihood ratio statistic G^2 can be used to assess the fit of the model. The statistical significance of the association between X and Y is based on the null hypothesis $H_0: \tau_i = 0, \quad i = 1, \cdots, r$ using a test statistic $G^2(I|R)$ equal to the difference between $G^2(R)$ for the row effects model and $G^2(I)$ for the independence model. The test uses the chi-squared distribution, with degrees of freedom $r-1$. Given that the row effects model holds, this test is asymptotically more powerful at detecting an association than the $G^2(I)$ test, in which the alternative hypothesis ignores the ordinal nature of Y.

For example 1, the expected frequencies for the independence model are

	Exceeds Rule	Meets Rule	Minor Departure	Major Departure	**Total**
No Problem	1.85	6	2.77	1.38	**12**
Problem	2.15	7	3.23	1.62	**14**
Total	**4**	**13**	**6**	**3**	**26**

and for the row effects model they are

	Exceeds Rule	Meets Rule	Minor Departure	Major Departure	**Total**
No Problem	2.58	6.54	2.16	0.71	**12**
Problem	1.42	6.46	3.84	2.29	**14**
Total	**4**	**13**	**6**	**3**	**26**

The effects in the estimated row effects model, reconstructed from the logarithms of the above frequencies and using integer column scores $\{v_j = j\}$, are

$$\hat{\mu} = 0.9556$$
$$\hat{\lambda}^X = (-0.1407,\ \ 0.1407)$$
$$\hat{\lambda}^Y = (-0.3072,\ \ 0.9162,\ \ 0.1021,\ \ -0.7111)$$
$$\hat{\tau} = (-0.2947,\ \ 0.2947)$$

The likelihood ratio statistics are

Model	G^2	df	P-value
Independence, I	1.99	3	0.57
Row effects, R	0.45	2	0.80
Conditional independence, given row effects, I|R	1.54	1	0.21

Here, the increase in p-value from model I to model R shows an improvement in fit, but the difference would not normally be considered significant at p=0.21.

Logit models

For loglinear models, it is unnecessary to distinguish between response variables and explanatory variables. For a binary response, logit models, also known as logistic regression models, express the logit (logarithm of the odds of success), as a linear function of explanatory variables. When both variables are ordinal, it is called the uniform association model. When the response is binary and the explanatory variable is ordinal, the uniform association model takes a very simple form:

$$L_j = \log\left[\frac{m_{1j}}{m_{2j}}\right] = \alpha + \beta(v_j - \bar{v}), \qquad j = 1, \cdots, c$$

The $\{v_j\}$ are, again, arbitrary scores assigned to the levels of the ordinal variable. This is a model for c logits. There are 2 parameters, α and β, therefore the residual degrees of freedom for fitting the model are $c - 2$. Usually the main parameter of interest is the association parameter β, which describes how the logits are related to the level of the explanatory variable. The difference between logits for any two columns is

$$L_a - L_b = \log\left[\frac{m_{1a}}{m_{2a}}\right] - \log\left[\frac{m_{1b}}{m_{2b}}\right] = \log\left[\frac{m_{1a}}{m_{2a}} \div \frac{m_{1b}}{m_{2b}}\right]$$

which is the log of the odds ratio for the 2×2 table formed by columns a and b. For adjacent columns, if integer column scores $\{v_j = j\}$ are used, then

$$L_{j+1} - L_j = \alpha + \beta(j + 1) - (\alpha + \beta j) = \beta$$

and

$$\exp(\beta) = \frac{m_{1,j+1}m_{2,j}}{m_{2,j+1}m_{1,j}}$$

thus the odds ratio is constant and equal to $\exp(\beta)$ for all adjacent pairs of columns. Parameters for the logit model are estimated by many commercially available statistical packages. Expected cell frequencies can be computed from the parameters and the likelihood ratio statistic G^2 can be used to assess the fit of the model. In a fashion completely analogous to log-linear models with 2 rows, a one degree-of-freedom conditional test of independence can be performed using the difference between G^2 for the uniform association model and a model of independence in which $\beta = 0$. Note that the logit independence model is simply

$$L_j = \log\left[\frac{m_{1j}}{m_{2j}}\right] = \alpha, \qquad j = 1, \cdots, c$$

implying that the ratio between row cell frequencies (odds) is the same for all columns, hence all odds ratios are equal to unity. It is equivalent to the log-linear independence model. In addition, the logit model with binary response X is equivalent to the log-linear $2 \times c$ row effects model, yielding identical expected frequencies. For the example above, with integer column scores $\{v_j = j\}$, the estimated logit model is

$$L_j = \log\left[\frac{m_{1j}}{m_{2j}}\right] = -1.192 + 0.589(j - 2.5), \qquad j = 1, \cdots, c$$

The constant value of the expected odds ratio is $\exp(0.589) = 1.80$.

Cumulative logit model

When there are more than two response levels, one can model the cumulative logits. Let $L_{i(j)}$ denote the ith cumulative logit within column j; that is,

$$L_{i(j)} = \log\left[\frac{m_{i+1,j} + \cdots + m_{r,j}}{m_{1,j} + \cdots + m_{i,j}}\right]$$

In words, $L_{i(j)}$ is the log odds of an observation from column j falling in a row with index greater than i. As before, rows (i) will represent levels of the response and columns (j) represent levels of an explanatory variable. The cumulative logit model for uniform association assumes the cumulative logits are a linear function of the column index, with the same slope for all columns:

$$L_{i(j)} = \alpha_i + \beta(v_j - \bar{v}), \qquad i = 1, \cdots, r\text{-}1 \qquad j = 1, \cdots, c$$

(Models such as this, in which β is independent of the rows, are also known as proportional odds models.) This is a model for $r - 1$ logits in each of c columns, a total of $(r - 1)c$ logits. It has one association parameter (β) and $r - 1$ α's pertaining to the cutpoints for forming the logits. The residual degrees of freedom for fitting the model are therefore $df = (r - 1)c - r = rc - r - c$. If $\beta = 0$, then the ith logit is the same in each column, implying that the column and row variables are independent. Note that the difference between logits for adjacent columns

$$L_{i(j+1)} - L_{i(j)} = \log\left[\frac{m_{i+1,j+1} + \cdots + m_{r,j+1}}{m_{1,j+1} + \cdots + m_{i,j+1}} \div \frac{m_{i+1,j} + \cdots + m_{r,j}}{m_{1,j} + \cdots + m_{i,j}}\right]$$

is the log odds ratio for the 2×2 table formed by taking adjacent columns and collapsing the rows into 2 levels with cutpoint at i. Applying the logit model with integer column scores $\{v_j = j\}$, the log odds can also be expressed as

$$L_{i(j+1)} - L_{i(j)} = \alpha_i + \beta(j + 1 - \bar{v}) - \alpha_i + \beta(j - \bar{v}) = \beta$$

Hence

$$\exp(L_{i(j+1)} - L_{i(j)}) = \frac{m_{i+1,j+1} + \cdots + m_{r,j+1}}{m_{1,j+1} + \cdots + m_{i,j+1}} \div \frac{m_{i+1,j} + \cdots + m_{r,j}}{m_{1,j} + \cdots + m_{i,j}} = \exp(\beta)$$

Therefore the model implies that the odds ratios are constant for all 2×2 tables formed from adjacent columns with dichotomous collapsings of the rows.

Additional covariates, such as other rule implementation ratings and environmental variables, can be added to this model. They need not be ordinal variables. For example, a model with two rules and slope steepness as explanatory variables might be formulated as:

$$L_{i(X)} = \alpha_i + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3, \qquad i = 1, \ldots, 3$$

where X_1 and X_2 are implementation ratings (1-4) for the two rules, X_3 is slope steepness and $L_{i(X)}$ is the log odds, at the given level of the covariates (X_1, X_2, and X_3), of observing effectiveness greater than i. The model includes $r - 1 = 3$ parameters α (as before) and 6 parameters β for the effects of rules, slope and all order 2 interactions between them. If the mean score is not subtracted from each covariate as in this example, certain precautions must be taken when the coefficients and statistics are being interpreted (see Appendix D). To estimate the parameters and their significance, a program is needed that performs logistic regression for ordinal responses. The program should handle both continuous and categorical explanatory variables, and interaction terms. One such program is "lrm", found in Frank Harrell's (1995) Design library of functions for S-Plus, available at the Carnegie Mellon "statlib" web site, under lib.stat.cmu.edu/S. An example analysis using "lrm" is shown in Appendix D. A simple ordinal logistic regression example using SAS is found in Appendix E. Ordinal explanatory variables are generally assigned scores in these analyses and treated as if they were continuous. The appropriateness of assigned scores can be tested in S-Plus by coding the variables as ordered factors or by looking at dummy variable coefficients in which the ordinal ratings are treated as nominal categorical variables.

Comparison of ordinal methods

When working with large tables or multidimensional tables, loglinear and logit models have some clear advantages over measures of association. With the correct model they can more fully describe the structural pattern of multidimensional tables, permitting identification of interactions and conditional dependencies among variables. Expected cell frequencies and odds ratios can be obtained from model parameters. Measures of association, on the other hand, generally reduce a table to a single number. On the other hand, measures of association have simple interpretations. For small tables, they can be as illuminating as models. And, they provide a simple method for summarizing results from many tables for which numerous different models may be required. For our example, and for many types of data, conclusions will not differ substantively among these various methods (Agresti, 1984, p. 222). Each of the measures of association have slightly different meanings, but all three that we have presented are based on the difference between the number of concordant and discordant pairs, and if one measure indicates a significant association, then all do. Since implementation is naturally an explanatory variable and effectiveness a response, perhaps the most appropriate measure of association is Somers' d_{XY}. Figure 2.2 summarizes the values of Somers' d_{XY} for all tables in the pilot data set that have at least 10 rule ratings in categories 1 to 4.

Required sample sizes

Wilcoxon-Mann-Whitney test

A formula for the sample size required to achieve a given power, $1 - \beta$, for a one-sided test at significance level α, was derived by Noether (1987). The alternative hypothesis is stated in terms of the

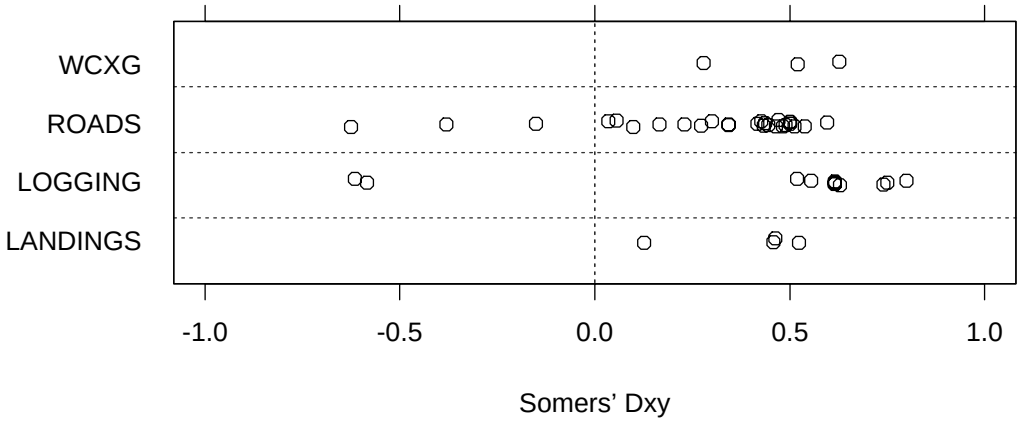

Figure 2.2. Association between problem occurrence (the response) and implementation (based on rules with at least 10 rated sites).

probability, p, that a random observation from row 1 ranks higher than a random observation from row 2. Recall that this probability can be estimated from a contingency table by adding 0.5 to the difference in mean ridits for the 2 rows. An equivalent way to estimate this probability is from the Mann-Whitney statistics,

$$p = P[X1 > X2] \cong U_1 / (U_1 + U_2)$$

Noether's formula for the required sample size is

$$N = \frac{(z_\alpha + z_\beta)^2}{12c(1-c)(p - \frac{1}{2})^2}$$

where $c = n_{1+} / n$, and z_α and z_β are the values in a standard normal table corresponding to the areas α and β. That is, if Z is a standard normal random variable, z_α and z_β are chosen so that $P[Z < z_\alpha] = \alpha$ and $P[Z < z_\beta] = \beta$. It may be useful to express N in terms of the odds, $r = p / (1 - p) = P[X1 > X2] / P[X2 > X1]$, which can be estimated by the ratio U_1 / U_2. In this case, p may be replaced by $r / (1 + r)$ in the above formula. Figure 2.3 shows the distribution of the odds (calculated by the U2k function in Appendix A) for the different categories of rules in the pilot study. For the validity of these calculations, one must assume that the variance of the Mann-Whitney statistic under the alternative hypothesis is equal to its variance under the null hypothesis. For alternatives that do not differ much from p=0.5, this will often be appropriate. Figure 2.4 shows the required sample sizes for a variety of odds and significance levels as a function of the test's power (probability of rejecting the null hypothesis for the given odds ratio).

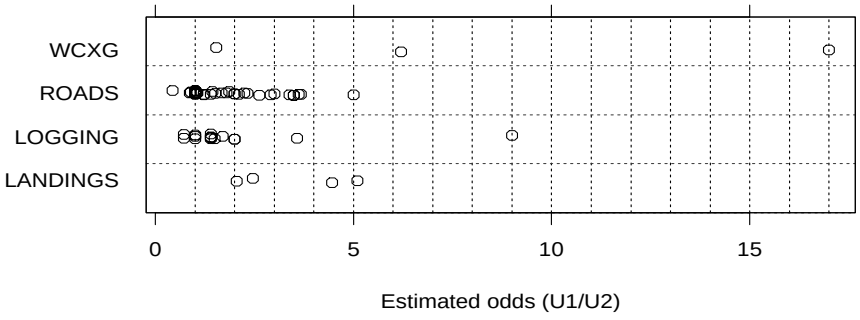

Figure 2.3. Odds of a random non-problem site having better implementation than a random problem site (based on rules with at least 10 rated sites)

Logistic regression

Sample size requirements for logistic regression have been calculated using the S-Plus function mvlogistic.design(), from a library of S-Plus routines for asymptotic power calculations obtained from the 'statlib' archives at Carnegie Mellon University. The calculations presented here are for hypothesis tests using the likelihood ratio method (Cox and Hinkley, 1974). To determine the necessary sample size, the following information must be supplied:

1. The values of the coefficients in the alternative hypothesis.
2. The design points (combinations of covariate values). For example, there are 16 possible design points in a model for effectiveness with 2 predictors rated on a scale of 1-4.
3. The relative sample size at each design point.
4. The significance level and desired power of the test.

Obviously, it is impossible to give just one number for the required sample size. In fact, it is very difficult to obtain appropriate answers at all because the values of the coefficients in the alternative hypothesis and the relative sample sizes at the design points are unknown in advance. The relative sample sizes can be estimated from the pilot data but they vary from rule to rule, so there are different answers for each rule. For the coefficients, a reasonable approach is generally to specify an alternative hypothesis coinciding with the minimum effects that are considered important. Since the coefficients can be interpreted in terms of odds ratios, we can specify an important effect in terms of odds ratios, and translate those to coefficient values. However, in testing a particular coefficient, the values of the intercept and all other coefficients involving that variable being considered affect the required sample size, so it is not enough to specify the size of important effects. You must actually have an estimate of the coefficients to get an accurate answer.

Despite these difficulties, it is possible to get a feeling for required sample sizes and what factors influence them by running various alternatives. In this section, we will look at a hierarchy of logistic regression models, and the power of tests for a main effect and interaction effects up to third order. In an attempt to make this analysis relevant, models for effectiveness on road transects were fit to 3 rules

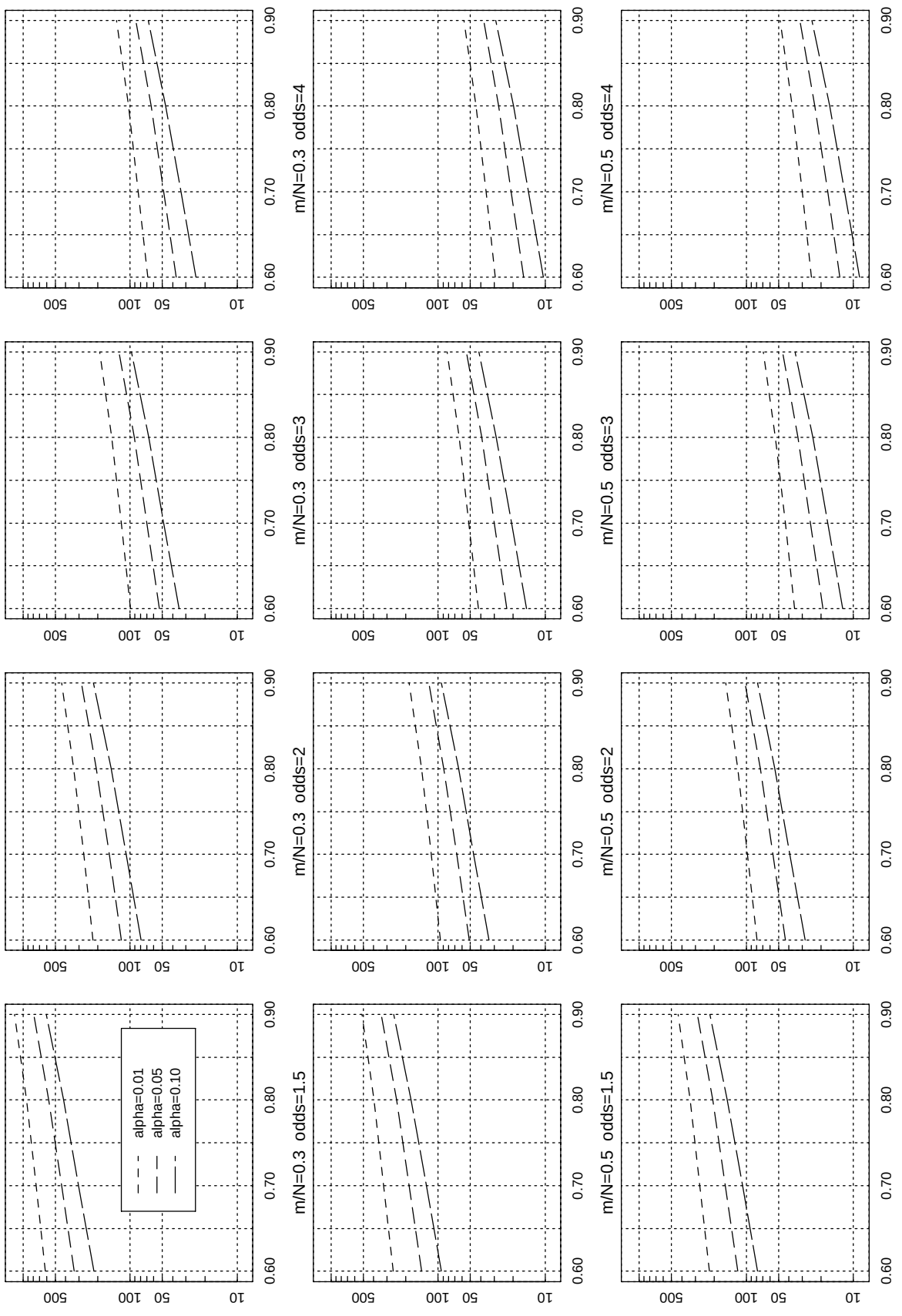

Figure 2.4. Required sample sizes for one-sided Mann-Whitney-Wilcoxon test (m/N is fraction in smaller group)

relating to roads. These rules were selected mainly because they had more ratings in the substantive categories 1-4 and a better spread of ratings than most other rules.

> Rule 1. Are drainage structures and facilities of sufficient size, number, and location to minimize erosion from roadbeds, landings, and fill slopes? [923.2(h); 19.C.2b]
>
> Rule 2. Is the road located on natural benches, flatter slopes, and areas of stable soils to minimize effects on watercourses? [923(f); 18.A.05]
>
> Rule 3. Have all roadside berms been removed or breached except as needed to facilitate erosion control? [923.4(e); 22.C.3]

First, consider a model for effectiveness, based on X1, the implementation of Rule 1. The marginal contingency table for Rule 1 looks like this

	Exceeds Rule	Meets Rule	Minor Departure	Major Departure	**Total**
No Problem	0	7	1	0	**8**
Problem	0	5	6	3	**14**
Total	**0**	**12**	**7**	**3**	**22**

and the logistic regression fit for Rule 1 by itself is

$$Y = 1.61 + 0.33(X1 - 2.5)$$

The coefficient 0.33 implies an odds ratio of $\exp(0.33) \cong 1.4$ between implementation ratings differing by 1 unit. For example, predicted odds of a problem are 1.4 times higher for an implementation rating of 3 than for an implementation rating of 2. A likelihood ratio test for the significance of the coefficient 0.33 yields a chi-square probability of 0.27, so this coefficient would not be considered to be different from zero. (None of the coefficients in the models discussed in this section were significant at the 0.05 level). Let's suppose the minimum important effect is an odds ratio of 2 between adjacent ratings. Given the relative numbers of the ratings 1 to 4 are 0, 12, 7, and 3, respectively, and the intercept is really 1.61, the following sample sizes would be required to detect an odds ratio of 2 for various values of α and β.

	Power of the test			
Significance Level	0.60	0.70	0.80	0.90
0.05	209	263	334	448
0.10	153	200	263	365

Some ratings of 1 would probably occur in a larger sample. If we assume the ratings proportions would be 15, 120, 70, and 30, the sample size requirements become

	Power of the test			
Significance Level	0.60	0.70	0.80	0.90
0.05	144	182	231	309
0.10	106	138	182	252

The numbers dropped by about 31% for a small change in proportions, so the power of the test is sensitive to the proportions in the different ratings groups. If the ratings proportions are all the same, as they would be in a controlled experiment with a balanced design, these numbers decrease another 52%:

	Power of the test			
Significance Level	0.60	0.70	0.80	0.90
0.05	70	88	112	149
0.10	51	67	88	122

The test is even more sensitive to the magnitude of effect to be detected. In order to detect a smaller odds ratio of 1.4, even with a balanced design, the sample sizes are quite a bit larger:

	Power of the test			
Significance Level	0.60	0.70	0.80	0.90
0.05	270	341	433	580
0.10	199	260	341	473

Next, suppose the implementation of Rule 2 is to be included in the model, and we wish to test for an interaction between Rule 1 and Rule 2. Because both rules were not rated 1-4 on every transect, the sample size in the pilot data drops from 22 to 19. A logistic regression model fit to these data is

$$Y = 1.68 + 0.24(X1 - 2.5) + 0.16(X2 - 2.5) - 0.12(X1 - 2.5)(X2 - 2.5)$$

Let's assume the true intercept and coefficients for the main effects are as given in the above equation, and we would like to detect any interaction that could change the odds by a factor of 2 for a unit change in the interaction term. That is, we would like to detect an interaction whose coefficient is log(2)=0.619 or greater in absolute value. It happens that only 5 of the 16 possible design points have non-zero values in the pilot data. Assuming the relative ratings proportions from the pilot study are typical, the sample sizes needed are

	Power of the test			
Significance Level	0.60	0.70	0.80	0.90
0.05	750	945	1201	1608
0.10	551	720	946	1311

If a balanced design is used, with equal numbers of ratings for each of the 16 possible design points, the required sample sizes again drop dramatically:

	Power of the test			
Significance Level	0.60	0.70	0.80	0.90
0.05	59	75	95	127
0.10	44	57	75	104

Tests for interaction between Rules 1 and 3 require slightly larger sample sizes than the above tables for Rules 1 and 2. An interaction between Rules 2 and 3 cannot be tested for at all because the design space in the pilot data is too sparse, leading to computational singularities.

Finally, consider testing for a 3-way interaction in a model that includes all main and second order effects (except the interaction between Rules 2 and 3, which cannot be fit in the pilot data). Because all rules must be rated 1-4 to be included in the data set, the sample size drops to 18 in the pilot data set for this analysis. In a fashion analogous to that followed for the order 2 interactions, a model was fit and the parameters for main and second order effects were set equal to their estimated values. The fitted model was

$$\begin{aligned} Y = {} & 1.89 + 0.017(X1 - 2.5) + 0.37(X2 - 2.5) + 0.32(X3 - 2.5) \\ & -0.12(X1 - 2.5)(X2 - 2.5) - 0.27(X1 - 2.5)(X3 - 2.5) \\ & +0.08(X1 - 2.5)(X2 - 2.5)(X3 - 2.5) \end{aligned}$$

We will again assume the minimum effect of interest is a coefficient of log(2)=0.619 for the 3-way interaction. Of the 64 possible design points, the pilot study includes only 8. Assuming the ratings proportions would be the same as that in the pilot study, the required sample sizes are:

	Power of the test			
Significance Level	0.60	0.70	0.80	0.90
0.05	1641	2067	2629	3519
0.10	1206	1576	2071	2868

If, on the other hand, a balanced experimental design were used with equal numbers of observations in all 64 possible combinations of ratings for rules 1, 2, and 3, the sample sizes needed are only

	Power of the test			
Significance Level	0.60	0.70	0.80	0.90
0.05	63	79	101	135
0.10	46	60	79	110

The parameter estimates used in these calculations are very poor due to the small sample sizes used to derive them. Furthermore, the parameters will differ from rule to rule and from model to model. Also the ratings proportions will vary. The rules selected for this analysis had better balanced ratings proportions than most of the rules not selected. Nevertheless, these calculations seem to imply that it will be very difficult to detect effects using logistic regression unless a balanced sample is obtained with about equal numbers of observations in each of the ratings categories. Logistic regression, however, is

probably the best tool available for detecting interactions that could lead to incorrect conclusions in analyses of two-dimensional contingency tables. In theory, the best way to overcome these difficulties is by using a controlled experimental design in which the levels of nuisance variables are fixed and the various treatments (rule implementation levels) are applied to equal numbers of experimental units.

Proportions from independent samples

When comparisons of proportions are of interest from independent samples, the necessary sample size can be calculated after stating values for the significance level, power, and proportions in the alternative hypothesis. If the true proportions for the independent samples are labeled P_1 and P_2, respectively, then the null hypothesis of equal proportions is $H_0: P_1 = P_2$ with an alternative hypothesis of $H_A: P_1 \neq P_2$, or $H_A: P_1 > P_2$, or $H_A: P_1 < P_2$ depending on what is known. Consider setting the significance level to α for the two-sided alternative and the power to $1 - \beta$ for two specified values of P_1 and P_2. If equal sample sizes are desired, then the necessary sample size for each sample is given by

$$n = n' \cdot \left(\frac{1}{2} + \frac{1}{2} \sqrt{1 + \frac{4}{n' \left| P_2 - P_1 \right|}} \right)^2$$

where

$$n' = \frac{\left(z_{1-\alpha/2} \sqrt{2\bar{P}(1 - \bar{P})} - z_{\beta} \sqrt{P_1(1 - P_1) + P_2(1 - P_2)} \right)^2}{\left(P_1 - P_2 \right)^2}$$

and $\bar{P} = (P_1 + P_2) / 2$, $\Phi\left(z_{1-\alpha/2}\right) = 1 - \alpha / 2$, and $\Phi\left(z_{\beta}\right) = \beta$ with $\Phi(\cdot)$ being the standard normal distribution function.

Sometimes costs for one sample type might be much larger than another. In such a case we can consider determining the necessary sample size of *m* for one sample type and *rm* for the other where *r* might represent the ratio of the costs. Once we fix the ratio, *r*, we have

$$m = m' + \frac{r + 1}{r \left| P_2 - P_1 \right|}$$

where

$$m' = \frac{\left(z_{1-\alpha/2} \sqrt{(r + 1)\bar{P}(1 - \bar{P})} - z_{\beta} \sqrt{rP_1(1 - P_1) + P_2(1 - P_2)} \right)^2}{r\left(P_1 - P_2 \right)^2}$$

and *r* is a specified ratio of sample sizes and the other parameters are the same as before.

3. Quality assurance

Quality assurance is a set of procedures to help with consistency and assessment of the data gathering procedures. These procedures vary widely in subject matter and scope and can cover measurement error, training of observers, calibration of instruments, checks on the consistency of observers, etc. For studies considered here we mention the tests for marginal homogeneity, symmetry, and overall agreement.

Marginal homogeneity and symmetry

Sometimes measurements are repeated on the same unit. When two measurements of a continuous variable are made on the same unit, the standard paired *t*-test is commonly used to compare the means at the two sampling times. The pairing of the continuous measurements on the same unit is often for reasons of convenience but the pairing also tends to reduce the variability of the difference in the two sample means. This results in increased precision and gives us more power for the same sample size, or the same power for a smaller sample size, than if there were no pairing. Similar paired sampling designs exist for categorical data and result in increased precision for estimates of differences.

Generally there are two characteristics that are examined. If we are interested in determining if there is a change in the overall proportions for each category from one time period to another, then we examine *marginal homogeneity*. But if we are interested in determining if the proportions of, say, changing from category A to B are the same as changing from category B to A, then we examine *symmetry*.

Tests and estimation procedures for evaluating both marginal homogeneity and symmetry are described in the following sections.

McNemar's test

Consider a categorical variable that takes on values of "inadequate" and "adequate" and that we take measurements at two different time periods. If we associate "0" with "inadequate" and "1" with "adequate", then we have 4 possible outcomes when sample units are measured: "00" when the unit is rated "inadequate" at both time periods, "01" when the unit is rated "inadequate" for the first time period but "adequate" for the second, "10" when the unit is rated "adequate" for the first time period but "inadequate" for the second time period, and "11" when the unit is rated "adequate" at both time periods.

The probabilities of each type of rating are shown in the following table:

		Time Period 2		
		Inadequate	Adequate	*Totals*
Time Period 1	Inadequate	p_{00}	p_{01}	p_{0+}
	Adequate	p_{10}	p_{11}	p_{1+}
	Totals	p_{+0}	p_{+1}	1

From such a table we might want to test that the probability of being "adequate" is the same for both time periods. A test of marginal homogeneity would test that the marginal probabilities are equal: $p_{+1} = p_{1+}$ and $p_{+0} = p_{0+}$. A test of symmetry would test that the off-diagonal elements are equal: $p_{01} = p_{10}$.

Because $p_{+0} + p_{+1} = 1$ and $p_{0+} + p_{1+} = 1$, we can write the null hypothesis of marginal homogeneity as

$$H_0: p_{1+} = p_{+1}$$

This is equivalent to

$$H_0: p_{10} + p_{11} = p_{01} + p_{11}$$

and

$$H_0: p_{10} = p_{01}$$

So in the case of a 2-by-2 table marginal homogeneity and symmetry are equivalent.

Now consider a random sample of n locations. The data will be displayed as counts in a table structured similar to the table of probabilities:

		Time Period 2		
		Inadequate	Adequate	*Totals*
Time Period 1	Inadequate	x_{00}	x_{01}	x_{0+}
	Adequate	x_{10}	x_{11}	x_{1+}
	Totals	x_{+0}	x_{+1}	n

with n representing the total sample size and $n = x_{00} + x_{01} + x_{10} + x_{11}$.

The equivalent paired t-test for categorical data is McNemar's Test and the test statistic for McNemar's Test is relatively simple:

$$T = \frac{(x_{10} - x_{01})^2}{x_{10} + x_{01}}$$

Under the null hypotheses of no difference in marginal proportions, T has approximately a χ^2 distribution with one degree of freedom. If the nominal significance level of the test is set to 0.05, the null hypothesis is rejected when $T > 3.84$.

For example, consider the following hypothetical example. At each of 40 sites implementation level was determined at two separate times. With McNemar's test we could see if the proportion of sites determined to have adequate implementation changed over time (*i.e.,* marginal homogeneity). Table 3.1 lists those hypothetical counts.

Table 3.1. Hypothetical example of agreement of implementation ratings at two different time periods.

		Time Period 2		
		Inadequate	Adequate	*Totals*
Time Period 1	Inadequate	18	7	25
	Adequate	10	5	15
	Totals	28	12	40

The test statistic is

$$T = \frac{(x_{10} - x_{01})^2}{x_{10} + x_{01}} = \frac{(10 - 7)^2}{10 + 7} = 0.5294$$

and because $T \leq 3.84$, the null hypothesis of equal proportions between time periods 1 and 2 is not rejected. (See Appendix F and Appendix G, respectively, for examples of S-Plus and SAS code using this data.)

Calculating power for McNemar's test

Power for McNemar's test is relatively simple to calculate exactly but processing time becomes prohibitive for large sample sizes. When the probability of rejecting the null hypothesis of no difference in proportions is nominally set to 0.05, the null hypothesis is rejected when the test statistic is greater than 3.84. For a given set of cell probabilities, we simply sum the probabilities of all of the possible tables for a sample of size n where the test statistic is larger than 3.84. The probability of each possible table is computed from the sample size and the cell probabilities under the alternative hypothesis, using the multinomial distribution for the cell counts. A simple BASIC program to perform these computations is found in Appendix H.

Besides setting n, we must also set all four cell probabilities to calculate power. We could set all four probabilities by first setting the marginal probabilities $p_{1+} = p_{10} + p_{11}$ and $p_{+1} = p_{01} + p_{11}$ and then just one of the cells, say, p_{11}. (The values for all of the other cells would be predetermined at that point.) One might not have too hard of a time setting the marginal probabilities but setting p_{11} is much more difficult to do.

We might choose $p_{11} = p_{1+} \cdot p_{+1}$ which is what is expected if the responses are independent between time periods. But this will probably result in underestimating power as the pairing is expected to show a positive dependence between time periods. We want to take advantage of the

positive dependence to increase the power over non-paired sampling techniques. Of course, by assuming independence we don't see the efficiencies expected from the positive dependence. Below are two examples where we first assume independence and then assume some dependence.

Suppose we have n=40, $p_{1+} = p_{10} + p_{11} = 0.3$, and $p_{+1} = p_{01} + p_{11} = 0.6$. If we assume independence, then we have $p_{11} = p_{1+} \cdot p_{+1} = 0.3 \cdot 0.6 = 0.18$,
$p_{10} = p_{1+} - p_{11} = 0.3 - 0.18 = 0.12$, $p_{01} = p_{+1} - p_{11} = 0.6 - 0.18 = 0.42$, and
$p_{00} = 1 - p_{01} - p_{10} - p_{11} = 1 - 0.42 - 0.12 - 0.18 = 0.28$.
The power calculated from the BASIC program in Appendix H, is 0.78053.

		Time Period 2		
		Inadequate	Adequate	*Totals*
Time Period 1	Inadequate	0.28	0.42	0.70
	Adequate	0.12	0.18	0.30
	Totals	0.40	0.60	1

If there is a positive association with $p_{11} = 0.26$, then $p_{10} = p_{1+} - p_{11} = 0.3 - 0.26 = 0.04$,
$p_{01} = p_{+1} - p_{11} = 0.6 - 0.26 = 0.34$, and
$p_{00} = 1 - p_{01} - p_{10} - p_{11} = 1 - 0.34 - 0.04 - 0.26 = 0.36$. The power is 0.92828. The positive association gives us a higher power than when the cell probabilities exhibit independence.

		Time Period 2		
		Inadequate	Adequate	*Totals*
Time Period 1	Inadequate	0.36	0.34	0.70
	Adequate	0.04	0.26	0.30
	Totals	0.40	0.60	1

Calculating the significance level for McNemar's test

As a short aside, there needs to a discussion of the actual significance level. For large enough sample sizes the test statistic, T, has approximately a χ_1^2 distribution. This suggests the use of $T > 3.84$ as the rule for the rejection of the null hypothesis which results in a nominal significance level of 0.05. But the distribution of T is only approximated by a χ_1^2 and the goodness of the approximation depends on the values of all four cell probabilities.

For a specific sample size and set of cell probabilities satisfying the null hypothesis, we define the significance level as the probability of rejecting the null hypothesis. Where there are many possible parameter values (sets of cell probabilities) satisfying the null hypothesis, the "size" of the test is a useful index. It is defined to be the largest significance level over all possible sets of cell probabilities that satisfy the null hypothesis for a given sample size. Since, it is unknown which combinations of cell probabilities give the largest significance levels, the size of the test must be approximated by trial and error.

After setting the marginal cell probabilities and specifying p_{11} one can determine the significance level for the particular cell values. For example, if $p_{+1} = p_{1+} = 0.8$ and $p_{11} = 0.64$, then using the program in Appendix H (same program used to compute power) results in a significance level of 4.4412% rather than the nominal 5% significance level.

We state that the significance level is 5% for McNemar's test. But in practice the size of the test will be not be exactly 5% and unfortunately there is no way of determining the actual size of the test from the sample data.

Bowker's test for symmetry in larger tables

The previous sections described McNemar's test for 2x2 tables which tested the equality of the marginal distributions of the response variable on two separate measurement occasions. Another characterization mentioned was that of symmetry of the cell probabilities about the main diagonal. In other words, the null hypothesis can be stated as $H_0: p_{01} = p_{10}$. As mentioned earlier symmetry and equal marginal distributions are one and the same for 2-by-2 tables.

But for tables with more than two categories for the response variable, symmetry implies marginal homogeneity but marginal homogeneity does not imply symmetry. So not rejecting a hypothesis of symmetry would imply not rejecting a hypothesis of marginal homogeneity.

One test statistic of symmetry for a variable with *K* nominal classifications is the usual Pearson statistic with the expected counts (E_{ij}) being the mean of the corresponding diagonal cell counts:

$$X^2 = \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{\left(n_{ij} - \left(n_{ij} + n_{ji}\right)/2\right)^2}{\left(n_{ij} + n_{ji}\right)/2} = \sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{\left(n_{ij} - n_{ji}\right)^2}{\left(n_{ij} + n_{ji}\right)}$$

When the null hypothesis of symmetry is true, then X^2 will have a χ^2 distribution with $K(K-1)$ degrees of freedom. This is known as Bowker's test. (If $K = 2$, then X^2 simplifies to the McNemar's test statistic.) A SAS example is found in Appendix I.

If the question of interest is about equality of the marginal distributions (*i.e.,* marginal homogeneity), then the above test will be somewhat liberal. Again, this is because symmetry implies marginal homogeneity but lack of symmetry does not imply a lack of marginal homogeneity.

Marginal homogeneity for nominal variables (Bhapkar's test)

If we have a square table for a variable with *K* categories for each of the two sample times, the sample difference in marginal probabilities for category *i* is $\hat{d}_i = \left(n_{i+} - n_{+i}\right)/n$. Let $\hat{\mathbf{d}} = \left(\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_{K-1}\right)$. (Note that $\hat{d}_K$ is not used because knowing the first $K-1$ values determines $\hat{d}_K$.) The sample covariance matrix $\hat{\mathbf{V}}$ of $\sqrt{n} \cdot \hat{\mathbf{d}}$ has elements

$$\hat{v}_{ij} = \begin{cases} -(\hat{p}_{ij} + \hat{p}_{ji}) - (\hat{p}_{i+} - \hat{p}_{+i})(\hat{p}_{j+} - \hat{p}_{+j}), & i \neq j \\ \hat{p}_{i+} + \hat{p}_{+i} - 2\hat{p}_{ii} - (\hat{p}_{i+} - \hat{p}_{+i})^2, & i = j \end{cases}$$

The Bhapkar test uses the test statistic

$$W = n\hat{\mathbf{d}}'\hat{\mathbf{V}}^{-1}\hat{\mathbf{d}}$$

which has asymptotically a chi-square distribution with $K - 1$ degrees of freedom. Appendix J lists a sample SAS program to perform Bhapkar's test.

Symmetry and marginal homogeneity for ordinal variables

With ordinal variables one can attempt to fit more parsimonious models to assess (among other things) symmetry and marginal homogeneity. One logit model is

$$\log\left(m_{ij} / m_{ji}\right) = \tau \qquad \text{for} \quad i < j$$

which says that the log of the ratio of expected counts of the corresponding off-diagonal cells is a constant. The estimates are

$$\hat{\tau} = \log \frac{\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} n_{ij}}{\sum_{j=1}^{K-1} \sum_{i=j+1}^{K} n_{ij}}$$

$$\hat{m}_{ij} = \begin{cases} \dfrac{\exp(\hat{\tau})\left(n_{ij} + n_{ji}\right)}{\exp(\hat{\tau}) + 1}, & i < j \\ \dfrac{n_{ij} + n_{ji}}{\exp(\hat{\tau}) + 1}, & i > j \\ n_{ii}, & i = 1,2, \ldots, K \end{cases}$$

The usual Pearson chi-square statistic will have $(K + 1)(K - 1) / 2$ degrees of freedom. Such a model is called a conditional symmetry model and $\tau = 0$ implies marginal homogeneity.

One test is to take the difference between the chi-square values of the above Pearson chi-square statistic and that of Bhapkar's test of symmetry. The resulting test statistic has approximately a chi-square distribution with one degree of freedom when conditional symmetry holds. Significant values of the chi-square statistic mean that the hypothesis of marginal homogeneity is rejected. A short BASIC program for performing the calculations is given in Appendix K.

Another test of marginal homogeneity when a conditional symmetry model holds and for large enough samples is to examine the standardized value of $\hat{\tau}$ which is

$$z = \frac{\hat{\tau}}{\hat{\sigma}(\hat{\tau})}$$

where

$$\hat{\sigma}^2(\hat{\tau}) = \left(\sum_{i=1}^{K-1} \sum_{j=i+1}^{K} n_{ij} \right)^{-1} + \left(\sum_{j=1}^{K-1} \sum_{i=j+1}^{K} n_{ij} \right)^{-1}$$

and compare that statistic with a standard normal distribution. Very large values of z, say larger than 1.96 and above, (and very small values, say -1.96 and below) suggest that there is no marginal homogeneity.

Cochran's Q Test

If we observe a binary response (say, effective and not effective) at n locations on T occasions, we might be interested in testing if the effectiveness rate changes over time. If $T = 2$, then we can perform McNemar's test described earlier. When $T > 2$, then Cochran's Q test is used.

Let $b_{it} = 1$ if a success is observed at location i on occasion t and let $b_{it} = 0$ otherwise. Define the means for each location, each time period, and overall as

$$\bar{x}_{i+} = \frac{1}{T} \sum_{t=1}^{T} x_{it}, \; \bar{x}_{+t} = \frac{1}{n} \sum_{i=1}^{n} x_{it}, \text{ and } \bar{x}_{++} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it}$$

Cochran's Q statistic is

$$Q = \frac{n^2(T-1)}{T} \cdot \frac{\sum_{t=1}^{T} \left(\bar{x}_{+t} - \bar{x}_{++} \right)^2}{\sum_{i=1}^{n} \bar{x}_{i+}(1 - \bar{x}_{i+})}$$

and has an approximate chi-square distribution with $T - 1$ degrees of freedom.

Consider the example in Table 3.2, in which there are $n = 8$ locations assessed for effectiveness on $T = 4$ occasions. The value "1" indicates effective and "0" indicates not effective. We have

$$Q = \frac{8^2(4-1)}{4} \cdot \frac{0.04296875}{0.5625} \cong 3.6667$$

with 3 degrees of freedom. The 5% cut-off point for a chi-square distribution with 3 degrees of freedom is 7.815. Because $Q < 7.815$ we would not reject the null hypothesis of no change in effectiveness over the 4 time periods. Appendixes L, M, and N, respectively, contain BASIC, S-Plus, and SAS code for calculating Cochran's Q.

Table 3.2. Hypothetical example of assessing effectiveness (0 = no and 1=yes) at 4 times at 8 locations. The row and column means ($\bar{x}_{i+}$ and $\bar{x}_{+t}$, respectively) are used in calculating Cochran's *Q*.

	Time (*t*)				
Location (*i*)	**1**	**2**	**3**	**4**	$\bar{x}_{i+}$
1	1	1	0	1	0.75
2	1	1	1	1	1.00
3	0	0	0	0	0.00
4	0	0	0	0	0.00
5	1	1	1	1	1.00
6	1	0	0	0	0.25
7	1	1	1	1	1.00
8	1	1	1	0	0.75
$\bar{x}_{+t}$	0.750	0.625	0.50	0.50	$\bar{x}_{++}$=0.59375

Table 3.3. Hypothetical example of assigned implementation levels performed independently at 26 sites by two different raters.

		Rater 2				
		Exceeds Rule	Meets Rule	Minor Departure	Major Departure	**Total**
Rater 1	Exceeds Rule	3	2	0	1	**6**
	Meets Rule	1	6	1	0	**8**
	Minor Departure	0	1	4	3	**8**
	Major Departure	0	0	1	3	**4**
	Total	**4**	**9**	**6**	**7**	**26**

Measuring agreement

Measuring agreement between raters is part of a good quality assurance plan. But assessing agreement requires more than examining a measure of association. Strong agreement requires a strong positive association but with just a small proportion of sites in the cells where there is disagreement. For example, if one rater tended to assign evaluations one level lower than another other rater, we would find a strong association but not strong agreement.

Consider Table 3.3, which lists a hypothetical example of a table of counts of assigned implementation levels performed independently at 26 sites by two different raters. If there were perfect agreement, we would see all of the counts along the main diagonal where the raters give

the same rating. Instead we see that most of the counts are along the main diagonal with just a single discrepancy of more than two categories.

We find that $\hat{\gamma} = \frac{C - D}{C + D} = 0.801$ which suggests a strong positive association between the two raters. However, we notice that when there is disagreement, rater 2 tends to give sites a lower implementation rating.

Symmetry and marginal homogeneity would be desirable conditions for strong agreement but both of those characteristics could occur even when we have no observations along the main diagonal! Clearly, to assess agreement there needs to be a summary statistic that attempts to combine all of the desirable conditions for agreement.

One measure of agreement for nominal variables is Cohen's *kappa* which is constructed as follows. For a randomly selected site let π_{ij} be the probability that observer 1 classifies the site as level i and observer 2 classifies the site as level j with $i = 1,2, \ldots, K$ and $j = 1,2, \ldots, K$. The probability that the two observers agree is

$$\Pi_O = \sum_{i=1}^{K} \pi_{ii}$$

probability that the two observers agree is

$$\Pi_O = \sum_{i=1}^{K} \pi_{ii}$$

If there is independence between observers (which implies that agreement in ratings is just by chance), the probability of agreement is

$$\Pi_e = \sum_{i=1}^{K} \pi_{i+}\pi_{+i}$$

The difference $\Pi_O - \Pi_e$ represents the additional agreement between observers over and above what is expected by chance. To obtain a measure with a maximum value of 1 Cohen's *kappa* is used:

$$\kappa = \frac{\Pi_O - \Pi_e}{1 - \Pi_e}$$

For a particular sample we use the observed cell proportions to estimate κ:

$$\hat{\kappa} = \frac{\sum_{i=1}^{K} n_{ii} / n - \sum_{i=1}^{K} n_{i+}n_{+i} / n^2}{1 - \sum_{i=1}^{K} n_{i+}n_{+i} / n^2} = \frac{n\sum_{i=1}^{K} n_{ii} - \sum_{i=1}^{K} n_{i+}n_{+i}}{n^2 - \sum_{i=1}^{K} n_{i+}n_{+i}}$$

For the above table we have $k = 4$, $n = 26$, $\sum_{i=1}^{K} n_{ii} = 3 + 6 + 4 + 3 = 16$, and which results in. $\sum_{i=1}^{K} n_{i+}n_{+i} = 6 \cdot 4 + 8 \cdot 9 + 8 \cdot 6 + 4 \cdot 7 = 172$ which results in

$$\hat{\kappa} = \frac{26 \cdot 16 - 172}{26^2 - 172} = \frac{244}{504} \cong 0.484.$$

Confidence intervals can be calculated for κ with the following estimate of the asymptotic variance:

$$\hat{\sigma}^2(\hat{\kappa}) = \frac{1}{n}\left\{\frac{P_O(1 - P_O)}{(1 - P_e)^2} + \frac{2(1 - P_O)\left[2P_OP_e - \sum_{i=1}^{K} p_{ii}(p_{i+} + p_{+i})\right]}{(1 - P_e)^3} + \frac{(1 - P_O)^2\left[\sum_{i=1}^{K}\sum_{j=1}^{K} p_{ij}(p_{j+} + p_{+i})^2 - 4P_e^2\right]}{(1 - P_e)^4}\right\}$$

where $p_{ij} = n_{ij} / n$, $p_{+j} = \sum_{i=1}^{K} p_{ij}$, $p_{i+} = \sum_{j=1}^{K} p_{ij}$, $P_O = \sum_{i=1}^{K} p_{ii}$, and $P_e = \sum_{i=1}^{K} p_{i+}p_{+i}$. The 95% confidence limits would be $\hat{\kappa} \pm 1.96 \cdot \hat{\sigma}(\hat{\kappa})$. For the example data, the calculated confidence limits are (0.239, 0.729).

The above summary statistic is for nominal variables and does not give a penalty for being far off of the main diagonal. The generalized *kappa* measure allows one to use different weights for each level of agreement. If we let w_{ij} be a symmetric weight (*i.e.,* $w_{ij} = w_{ji}$) taking values from zero to one with larger values indicating greater agreement, a "weighted" Cohen's *kappa* can be constructed:

$$\kappa_w = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij}\pi_{ij} - \sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij}\pi_{i+}\pi_{+j}}{1 - \sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij}\pi_{i+}\pi_{+j}}$$

One suggestion by Fleiss and Cohen (mentioned in Agresti, 1990) is to have $w_{ij} = 1 - (i - j)^2 / (K - 1)^2$. This gives cells along the diagonal an agreement measure of 1 and the cells farthest from the diagonal an agreement measure of 0. In a 4×4 table, this suggestion would give the weights 1 to the cells along the main diagonal, 8/9 and 5/9 to cells in the two off-diagonals, and 0 to the two cells of most extreme disagreement. An S-Plus function, submitted by S. Hilsenbeck to S-news and listed in Appendix O, computes weighted or unweighted *kappa* and its standard error. We have not fully tested this function for correctness in the weighted case but the above table results in $\hat{\kappa} = 0.677$ with a estimated standard error of 0.191. SAS produces a weighted and unweighted *kappa* with associated standard errors in PROC FREQ. But SAS uses a slightly different weighting function: $w_{ij} = 1 - |i - j| / (K - 1)$. This results in $\hat{\kappa} = 0.600$ and an estimated standard error of 0.112. The SAS code is listed in Appendix I.

4. Evaluation of statistical compromises that may be necessary

The very fact that we have to sample rather than simply describe the true state of nature means that we've already started to compromise. Listed below are a variety of issues that generally require some compromise and included are some suggestions on how to deal with those issues.

Clear goals vs. massive data collection

A somewhat facetious definition of monitoring is "the collection of data. PERIOD." Not often enough is there any mention as to what should be done with the data, how it will be reported, how it will be stored and archived, and how the consequences of the resulting management actions will be assessed.

In a study where the costs of traveling to each site are high, it usually makes sense to record additional information at each site that is not directly required in proposed analyses if it might help with the interpretation or might be used in an expanded analysis. This philosophy in its extreme would dictate that we measure everything we know how to measure in anticipation of any conceivable analysis that might be proposed. Of course this would result in greatly increased time spent at each site at the expense of seriously reducing the sample size and increasing data management costs. Most of the measurements would never be used but would cost a great deal to organize, enter into a database, and archive. In addition, it may be a false perception that the scope of analyses is greatly expanded by recording a lot of extra information, because the analyses are limited primarily by the sampling design (see next section). The analyses may not be worth doing if they are not powerful enough to detect important effects because of the small sample size obtained following this strategy. Data collection is expensive and massive data collection is not a good substitute for a study design with clear objectives, limited in scope, but supported by well-defined analytical procedures and appropriate field methods.

Types of sampling designs: retrospective (case-control), prospective, and cross-sectional studies

Sampling designs can be categorized into several types and it is important to distinguish between them as each can answer a different set of very specific questions. More importantly is to recognize the kinds of questions that a particular sampling design cannot answer.

Consider looking at the association between the level of implementation of a rule and the level of effectiveness of a rule. Implementation would be considered an explanatory variable and effectiveness a response variable. We would like to know how implementation affects effectiveness. Three basic sampling designs are described:

> A *retrospective* (meaning "looking into the past") sampling design would take a sample of sites for each level of effectiveness. For each of the samples of effectiveness the levels of implementation would be determined.
>
> A *prospective* (meaning "looking to the future") sampling design would take a sample of sites at each level of implementation and then the level of effectiveness would be determined.

A *cross-sectional* sampling design would take a sample from all sites in the population and implementation and effectiveness would be assessed at the same time.

To show some of the differences in the sampling designs consider the following table of numbers of sites that are classified by level of implementation and level of effectiveness:

		Level of Implementation				
		Exceeds Rule	Meets Rule	Minor Departure	Major Departure	**Total**
Level of Effectiveness	No Problem	3	6	2	1	**12**
	Problem	1	7	4	2	**14**
	Total	**4**	**13**	**6**	**3**	**26**

A retrospective study would fix the totals for the response variable, which would be the row totals in this example. A prospective study would fix the totals for the explanatory variable, which would be the column totals here. Finally, the cross-sectional study would simply fix the total sample size; the row and column totals would fall as they may.

The important difference is in the kind of questions that can be addressed by each sampling design. If the totals for the level of effectiveness are fixed one cannot estimate the proportion of problem sites for each level of implementation without additional information about the population. Prospective and cross-sectional sampling designs are not limited in this way. For the question of interest here the prospective design is preferred. This is because the cross-sectional design cannot remove the effect of some sites being less amenable to some levels of implementation. The result is that site characteristics differ among implementation categories and inferences about the association between implementation and effectiveness are likely to be influenced by such confounding effects.

Two types of prospective studies are considered: clinical trials and cohort studies. The basic difference between the two is that clinical trials use a random assignment of sites to levels of implementation and cohort studies use the assignment that just happened to occur. Clinical trials are controlled experiments in which the influences of confounding variables are removed by the random assignment process. The intent is to obtain groups that differ only in the variable whose effect is being studied--groups that are approximately homogeneous with respect to all other variables. Randomized assigments accomplish this without having to specifically identify potential confounding variables.

We strongly recommend the use of clinical trials as the most defensible and most likely to address questions of "cause-and-effect" such as "Do sites where rules are better implemented tend to result in sites with fewer problems?"

Limitations of sample surveys

As alluded to in the section on sample design, sampling that assigns a site to a level of an explanatory variable without some sort of random process is difficult to defend. Not assigning the level randomly allows other unmeasured variables to add bias to the interpretation of effects.

In practice we might find that a lower level of implementation might be observed because the rule is difficult to apply at a particular site. Such sites might be more susceptible to problems regardless of the level of implementation. In the same manner, sites where a rule is easy to apply might be less likely to show a problem. That makes such sites different from the general population of sites and, therefore, inferences might be induced by the population differences rather than by any inherent association between implementation and effectiveness.

The only way to effectively eliminate such a confounding factor is to assign the level of implementation using some random process. The random assignment of implementation levels to sites will ensure that implementation level is unrelated to site characteristics that might be associated with erosion problems. This doesn't mean that there needs to be an equal number of sites for each implementation level. What is important is to have the probability of being assigned an implementation level be known.

Stratification

Stratification is the act of dividing a population into smaller, non-overlapping subpopulations. The reasons for stratification vary and can be for convenience or for obtaining more homogeneous subpopulations. When the stratification does result in more homogeneous subpopulations, estimates of population parameters have more precision (*i.e.,* less variability). When stratification does not result in more homogeneous subpopulations, inferences are still valid (with some form of random sampling) but parameter estimates might not be as precise as desired.

In practice, because of the investigator's knowledge of the population, stratification rarely reduces precision and almost always increases precision. With appropriate sample design and execution, the ability to make assessments of population parameters should be improved by stratification. Subpopulation parameters can usually be estimated more efficiently after stratification.

Most of the methods described in this report concern categorical data. And such data require sample sizes far larger than what one typically needs for data of a more continuous nature. The current amount of data available is not considered large despite the costs of the collection from 17 THPs in 1995 and 50 THPs in 1996.

Consider the assessment of a single rule with 4 levels of implementation and just 2 levels of effectiveness. Such a design results in 8 cells. For the 17 THPs, that's an average of only about 2 observations per cell with the possibility of many cells being empty. General guidelines state that the expected number of counts per cell be at least 4 or 5 just so that the distributional approximations hold let alone enough to detect differences and changes of interest.

Because of the sparseness of the cell counts that is expected to continue, only stratification that is known to result in just a few but far more homogenous subpopulations can be recommended.

Random vs. systematic vs. opportunistic sampling

Short of sampling the complete population, making valid statistical inferences requires the use of randomness in some fashion (*i.e.,* probability sampling). While inferences can be made with non-random or "opportunistic" samples, there is no statistical justification for such inferences. And part of making

inferences is not just convincing oneself about the inferences. By using randomness appropriately, one provides the basis that can be used to convince others of the validity of the inferences.

Using randomness does not mean just "simple random sampling". One can use any form of probability sampling. Systematic sampling (which is generally used for purposes of convenience) can be considered a form of probability sampling when there is a random starting location.

When opportunistic samples have been collected, they should be analyzed separately from the samples collected using probability sampling. Besides not being selected based on a probabilistic rule, use of non-random samples suggests that the population of interest is not very well defined. Not having a well-defined population is far more disconcerting than having non-random samples.

Defining categories

Four implementation levels have been considered: exceeds rule, meets rule, minor departure from rule, and major departure from rule. The investigator needs to decide if such a characterization is appropriate for the needs of the study. If only "meets rule" and "doesn't meet rule" is all that is needed, then that is a decision appropriately made by the investigator.

What is lost when collapsing categories is the ability to make inferences below the level of the analyzed categories. It might very well be that contrasting "major departures" with a category created by combining the categories "exceeds", "meets", and "minor departures" is more appropriate than a contrast of "meets rule" and "doesn't meet rule."

So the investigator needs to consider how one creates the categories as well as the number of categories. However, as mentioned earlier, collapsing categories is discouraged because it can result in the loss of statistical power.

Stating desired levels of precision

When summary statistics are used one must provide estimates of precision. Otherwise, the summary statistics are at best of unknown value. And prior to setting the sample size one needs to state the desired level of precision. This presupposes that one knows the statistic that will characterize the feature of interest and that the level of desired precision can be agreed to by others.

If the sample size is inadequate to achieve the desired levels of precision, then one must either lower one's standards or not do the data collection. If the sampling effort is mandated by law, then the latter choice is not possible (without dire consequences upon one's career). But avenues other than lowered expectations can be considered. One way is to limit the scope of the experiment. Another is to use more efficient measurement methods that would result in higher precision measurements or lower precision measurements at many more sites. In either case the overall result will be increased precision of the population estimates at the same or lower cost.

Levels of desired precision should be tempered with the consequences of not achieving that precision. And selecting those levels is not always an easy task. One should always ask "What decisions would I make if all of the data possible were available?" Then, because we know that we'll be sampling only a fraction of that data, we ask how much uncertainty can we accept and still expect to make good decisions.

Consequences of Type I and Type II errors

Typically Type I and Type II errors are only addressed through their respective rates of occurrence. For example, Type I error rates are typically set to 5% and, although Type II error rates are less frequently mentioned, desired rates are many times in the range of 5% to 30%. But determining necessary sample size based on selections of these two percentages alone ignores the relative consequences (costs) of .committing Type I and II errors.

To fix ideas we repeat the definitions of the two types of errors:

Type I Error: Rejection of the null hypothesis when the null hypothesis is true.
Type II Error: Acceptance of the null hypothesis when a specified alternative hypothesis is true.

Three other common terms are associated with hypothesis testing and Type I and Type II errors:

Significance Level: The probability of a Type I error.
Power: One minus the probability of a Type II error. The probability of rejecting the null hypothesis when a specified alternative hypothesis is true.
***P*-value**: The probability of rejecting the null hypothesis with a test statistic at least as extreme as that observed *when the null hypothesis is true*.

The significance level is denoted by the symbol α and power with $1 - \beta$. The *P*-value is generally represented by *P*.

The premise of this section is that the consequences of each type of error need to be considered when setting α and $1 - \beta$. Consider testing hypotheses about the mean number of road failures associated with two levels of culverts per 1,000 feet of road. Let the null hypothesis state that the current rule's requirement is adequate and the alternative hypothesis be that four times the number of culverts are necessary. (And for this example, we'll assume that only these two hypotheses are of possible interest.)

Clearly there is a great deal of difference in the cost of installing culverts at a 1X rate or a 4X rate. When a Type I error occurs, the alternative hypothesis is accepted and far too many culverts will be installed.

When a Type II error occurs, we have chosen the 1X rate when we should have chosen the 4X rate. This will result in more road failures, possibly exceeding the cost of installing the additional culverts required by the alternative hypothesis.

A statistical solution requires reasonable estimates of costs but such estimates are generally hard to give with general agreement as to the precision of those estimates. Depending on the costs of the two types of errors one might treat Type I and Type II errors with very different levels of importance.

As an extreme example, if the cost of a Type II error is much higher than that of a Type I error, one might always reject the null hypothesis irrespective of the data! But suppose there is more than an adequate level of power for detecting a particular alternative hypothesis. Then one might consider reducing α to lower the Type I error rate. While this will lower the power (and raise the Type II error rate) the resulting power might still be adequate.

If cost considerations are important, then deciding the significance level and power in isolation of such costs is neither a wise nor justifiable decision. While we are not presenting any particular formula for determining the optimal balance between the two types of error, we do recommend at least a contemplation of the consequences of each type of error.

5. Study specific concerns

Problem and non-problem points

Two stated goals (Tuttle, 1995; p. 3) of the Hillslope Component are to collect information on

1. how well the Rules are being implemented in the field; and
2. where, when, to what degree, and in what situations problems occur – and don't occur – under proper implementation

Problems are defined (Tuttle, 1995, p. 7) as "observable erosional or sediment deposition features, or failed erosion control features such as waterbars or culverts".

The current study design states that no implementation ratings will be given to non-problem sites. Without a sampling protocol for non-problem sites, the first goal can only be met with regard to problem sites and the second goal cannot be met at the site level at all. If it is desired to perform a statistical analysis contrasting the populations of problem and non-problem sites, it is necessary that both populations be well-defined, and that both populations be sampled or censused. If implementation is to be contrasted between the two populations, implementation needs to be evaluated in both populations. Under the current design, it will thus be impossible to evaluate the relation between implementation and effectiveness at the individual site level. To address this gap, the study plan includes implementation evaluations for whole transects, whether or not problems were encountered. This compromise has the following drawbacks:

1. Sample size is effectively reduced to a fraction of what it would be if individual points or plots along the transect were used as sampling units.
2. There are no well-defined rules for integrating evaluations of implementation over whole transects. Instructions are to evaluate implementation based upon overall, professional impressions.
3. It would seem impossible to objectively evaluate implementation over an entire transect if only problem sites have been evaluated. Non-problem sites must be observed as well. If those observations were recorded for a sample of non-problem sites, a more powerful analysis could be performed using individual sites as basic units.

If non-problem sites were to be sampled, it is not obvious what the sampling units of the non-problem population should be. The problem site population is rather unorthodox in that the sampling units consists of a variety of feature types. For roads, the defined types have been broken into 10 categories (Appendix B, Code Sheet 1: Road and Skid Trail Features), consisting of between 5 and 19 types each. Each identified feature is assigned a starting and ending distance on the road, which may or not be the same, and there is no limit on the number of features that might be assigned to a particular segment of road. When an erosion feature is identified, there are no formal rules for deciding which of these overlapping features it should be assigned to. Nevertheless, waterbars, culverts, cut slopes, and inboard ditches are a few obvious examples of possible problem sites. To be comparable to the problem site population as defined, it would be necessary to construct a different sampling frame for each defined feature type. Such an undertaking is clearly impractical.

To permit an analysis contrasting the populations of problem and non-problem sites, a population of non-problem plots needs to be defined upon which a well-defined sampling design can be carried out. For example, plots could have some fixed size and shape. (They should be large enough to have a high probability of containing one or more waterbars and culverts. Rules could be defined for selecting among drainage features when more than one of a given type falls in a plot). The critical sites erosion study (CSES) (Lewis and Rice, 1989) took such an approach. CSES was a case-control study, i.e. a separate sampling design was employed on the two contrasting sub-populations in order to obtain roughly equal numbers of plots from both sub-populations. A similar approach could be employed here. Within randomly-selected transects, all problem sites and a systematic sample of non-problem sites could be evaluated. Because a systematic sample is equivalent to a cluster sample, the design could be viewed as a one-stage cluster sample of problem sites and a two-stage cluster sample of non-problem sites.

There is one drawback of using problem plots and non-problem plots as sampling units, and that is the issue of pseudo-replication. Since plots are likely to be more similar within than between transects, *i.e.* plots are not independent observations, then the sampling assumptions behind the ordinal methods discussed earlier will not be satisfied. The effect this will have upon the results depends on the degree of non-independence among plots, but would be difficult to assess.

Rule identification

When a site is identified as a problem, an evaluator needs to know which rules apply to that site. Without clear guidelines, there could be a tendency to only evaluate those rules that were thought to have been inadequately implemented and to have caused the problem. It is important that a consistent set of rules is evaluated for each feature type. In the pilot data, the set of rules evaluated for whole transects was generally fixed, but the set of rules evaluated for transect points varied. Currently the only categorization of rules is by transect type (road, landing, logging, watercourse crossing, or WLPZ). There are not separate lists of rules designated for fill failures, waterbar failures, etc. To be certain that all applicable rules have been considered, all the rules listed for the type of transect being sampled should be evaluated for each problem point, utilizing N/A for those rules that do not apply.

Effectiveness evaluations

As Tuttle (1995) pointed out, an effectiveness evaluation will be meaningless unless stressing events have occurred between the time of the logging activity and the time of the evaluation. For example, if no storms have occurred with at least a one-year return interval, it may make sense to postpone an evaluation because it is unlikely that any problems would be found, regardless of rule implementation. One approach to the issue might be to specify a minimum event size that must occur before effectiveness would be evaluated. An additional requirement might be to wait some minimum length of time. The justification for the latter proposal is that it may take some time after disturbance before a logged site becomes vulnerable to certain types of stress (e.g. shear stress after root strength declines).

Specifying a minimum stressful event raises another issue. Once it has been specified, if we do not observe any problems can we conclude that the rules are effective? Suppose we specify the recurrence interval of the minimum stressful event as 3 years. And suppose the rules are designed to prevent erosion problems up to the 50-year event. It would be optimistic to declare that the rules are effective after observing that no problems occurred after a mere 3-year event. They were not tested at the level of stress they were designed for. Yet it is clearly impractical to wait for the 50-year event. Thus we must employ a test that does not measure performance to our standard of concern.

A solution to this dilemma is to include the magnitude of the stressful event(s) that have occurred as part of the analysis. For example, if the data are being modelled using logistic regression, it is simple to include maximum storm event magnitude as an explanatory variable.

Whole THP evaluations

The goals quoted at the beginning of this section from Tuttle (1995, p. 3) *can* be addressed at the whole transect and whole THP level under the current design, since both rule implementation and problem occurrence or effectiveness are recorded for these entities. The main disadvantages with these analyses, as stated before, are:

1. Sample size is reduced by aggregating the data.
2. Without well-defined rules for aggregating evaluations over many sites into a single value, there is a greater degree of subjectivity in the data.

For whole THP evaluations, effectiveness is defined differently than on transects. Instead of being defined as problem occurrence, effectiveness on whole THPs is a numerical rating similar to that used for implementation. The ratings for effectiveness are defined in terms of beneficial uses of water:

1. Improved protection of beneficial uses of water over pre-project conditions
2. Adequate protection of beneficial uses of water
3. Minor effects (short term or low magnitude negative effect on beneficial uses of water)
4. Major effects (long term or substantial negative effect on beneficial uses of water)

Protection of beneficial uses can be assessed either by looking at how well the harvest plan was executed (*i.e.,* evaluating implementation), or by looking at the actual impacts on beneficial uses. If the association between implementation and effectiveness is to be analyzed, then clearly the evaluations need to be based on actual impacts, not on implementation, which is evaluated separately. Therefore, it is recommended that the definitions of ratings 1 and 2 be modified, replacing "improved protection of" with words such as "no deleterious effects on" or "no observable effects on" and replacing "adequate protection of" with words such as "enhancement of".

Another consideration in the definition of effectiveness is the lack of information about effects on beneficial uses of water. As stated by Tuttle (page 17), the hillslope component of the Long Term Monitoring Program "does not answer the question of how well the Rules protect beneficial uses". Perhaps it would be better to rephrase these definitions in terms of observed erosion and evidence of sediment delivery to stream channels, rather than assumed effects on beneficial uses of water.

Because effectiveness is rated on an ordinal scale from 1 to 4, any of the ordinal measures of association discussed earlier in this report may be used. Loglinear and logit uniform association models (Agresti, 1984) may also be used, but they are not equivalent when the response has more than 2 categories. The loglinear row effects model described earlier is applicable only to a situation where the rows represent levels of a nominal (possibly binary) variable. The uniform association models provide little advantage over simple measures of association when considering only 2 variables, e.g. effectiveness and implementation of some rule. However, if multiple rules and environmental variables are being analyzed, the effort in applying these models may be worthwhile. The (cumulative) logit model is perhaps preferred because effectiveness is treated as a response, results are independent of the groupings

of the response categories, scores need not be assigned to the levels of the response, and it is generally easier to formulate than the loglinear model.

Outcome-oriented rules

Implementation evaluation for some rules (e.g. "Were temporary crossings removed?" [916.4c(3); 14.D.27.c]) is quite straightforward. However, for many rules (e.g. "Have the rules and mitigation measures described in the THP provided protection for upslope stability?" [916.4(b); 14.D.5.d]), evaluating implementation can be a rather subjective process. One could argue about what constitutes "protection". A further difficulty is that presence of a slope failure is an obvious evaluation criterion. If evaluations are based upon whether the slope has failed, then a perfect correspondence between problem occurrence and poor implementation (and between lack of problems and good implementation) is guaranteed. That the rule "works" when implemented becomes a foregone conclusion, because implementation is judged by the erosional outcome. To avoid that trap, the evaluator must clearly distinguish between the implementation of protective measures described in the THP from whether those measures succeeded in preventing a failure. Such an evaluation, while subjective, should be possible.

However, there are a great many rules with variations on this problem. "Are drainage structures and facilities of sufficient size, number, and location to minimize erosion from roadbeds, landings, and fill slopes? [923.2h; 19.C.2.b]" What is sufficient to minimize erosion? It is natural to use observations of erosion as evidence for such a judgment. Yet doing so guarantees that erosion problems (i.e. poor effectiveness) are equated with poor implementation. Implementation must be judged by what was done to comply, *not* by the end result. That is the only fair method to both the landowner (when looking at possible violations) and to the rule evaluation process. Unfortunately, the wording of such rules makes it extremely difficult to judge implementation fairly.

Some rules are worded in such a way that it is indeed impossible to evaluate their implementation without looking at the outcome. For example, "Was excess material from road construction and reconstruction deposited and stabilized in a manner or in areas that did not adversely affect beneficial uses of water?" [923.2m; 19.B.4]. While there is a great deal of subjectivity in judging adverse effects on beneficial uses, a perhaps greater difficulty is that the proper implementation of this rule can *only* be determined by the subsequent occurrence (or non-occurrence) of problems. By definition, to be implemented properly, no problems can occur. Since we have defined effectiveness as the prevention of problems, implementation and effectiveness are one and the same. The rule must prevent problems if it is followed, hence it automatically works! (Unfortunately, the landowner can never know if he has complied with such a rule until a stressful event occurs.)

The essential difference between rules like [923.2h; 19.C.2.b] and [923.2m; 19.B.4] is the use of the words "was sufficient to minimize [adverse effects on]" instead of "did not adversely affect". This may seem like a subtle difference, but vigilantly distinguishing between them is the only way the study can conceivably hope to draw any valid conclusions regarding the effectiveness of the former type rule. Critics will question whether the evaluators can successively maintain such a distinction, unless some evidence is provided to support the contention. However, the issue can be avoided by *evaluating implementation before any stressing events occur, and evaluating effectiveness after stressing events have occurred.* If logistics prevent implementation evaluations being done at some sites until after stressing events have occurred, then at least a sample of other sites should be evaluated independently both before and after stressful events. These paired evaluations can provide a test of repeatability for implementation evaluations.

Multiple rule interactions

Tuttle (1995) pointed out that the following four cases pertaining to a rule R1 can be inconclusive:

Implementation	Effectiveness	
	No Problem	Problem
Proper	Case 1	Case 2
Improper	Case 3	Case 4

When classifying individual sites, most cases are in fact inconclusive without further information.

Case 1: Rule may have prevented a problem if stressful storms occurred, or it may not have been needed.
Case 2: Rule may not have worked, or some other deficiency may have caused problem.
Case 3: Rule is not needed if stressful storms occurred.
Case 4: Problem may have resulted from non-compliance or from some other deficiency.

Only in case 3, could we draw conclusions (assuming stressful storms have occurred) without knowing something about the cause of the problem and how other rules were implemented. If, instead of a single site, a sample of sites is classified, the difficulties persist. Consider 3 possible outcomes:

A. Positive association: Ratio of cases 1 and 3 exceeds ratio of cases 2 and 4.
B. Negative association: Ratio of cases 2 and 4 exceeds ratio of cases 1 and 3.
C. No association: Cases 1 and 3 occur in the same ratio as cases 2 and 4.

Simplistic interpretations of these outcomes are, respectively:

A. Rule tends to work (has intended effect)
B. Rule tends to have the opposite of its intended effect.
C. Rule has no effect.

However, the problems with interpreting both positive and negative associations were pointed out in the discussion on partial associations and conditional independence. Other variables can be responsible for the observed associations. These other variables might be the implementation of other rules. For example,

Trap A. Positive association can be caused by the fact that another rule R2, which is usually implemented similarly to R1, causes problems when it is poorly implemented.
Trap B. Negative association can be caused when some rule R3, which is usually implemented opposite that of R1, causes problems when it is poorly implemented.
Trap C. No association can be an artifact of R1 working well when R2 is implemented, but having a reverse effect when R2 is poorly implemented. If the implementations of R1 and R2 are uncorrelated, summing over both cases of R2 could result in a table with no association.

Trap C describes a situation in which two rules both must be implemented properly in order to be effective. It is perhaps an unusual situation that, given poor implementation of R2, good implementation of R1 is more likely to result in problems than poor implementation. However, it is probably quite common that a number of rules must be jointly implemented in order to be effective. Lack of attention to

any rule in the set can result in a problem. Therefore, it may be useful to consider an implementation rating for a group of rules, before considering the individual rule ratings. For example, one could consider the set of all rules governing waterbar construction. Some summary measure such as the maximum (the worst implementation of the group) could be used for a group rating. This approach would avoid having to sort out the various rule interactions that might be expected within the group, and it would still allow conclusions to be drawn about the set of rules as a whole. If the group of rules were found to be ineffective, the member rules could then be examined individually. Or if some member rule was felt to be unnecessary, it could be examined individually.

Rules have already been categorized into groups on the implementation evaluation forms. For example, logging operations have been broken into 7 groups. Six rules are listed under "timber operations", 11 rules under "waterbreaks", and 5 rules under "permitted activities". Just 1 rule each is listed under "felling practices", "cable yarding", "winter period operations", and "compliance with the act and rules". A contingency table for the the maximum implementation rating of the waterbreak group indicates that

	Exceeds Rule	Meets Rule	Minor Departure	Major Departure	**Total**
No Problem	0	9	0	0	**9**
Problem	0	0	5	1	**6**
Total	**0**	**9**	**5**	**1**	**15**

.
the group as a whole, when implemented properly, seems to be very effective at preventing problems. No problems were found on transects where all the rules were followed, but every transect with even a minor departure from a waterbreak rule had problems.

Logistic regression for multiple rules

Groups of rules can be modelled using logistic regression to provide information about independence and interactions. Let p be the probability of the event Y (a problem) at a site, within some specified time after logging. Let $X_1, X_2, \cdots, X_k$ be the values of implementation on a scale of 1 to 4 for the k rules in a group. Let

$$\log\left(\frac{p}{1-p}\right) = (X_1 + X_2 + \cdots + X_k)^k$$

denote a linear logistic model with k main effects and all interactions. For example, the order 2 interactions are of the form $X_i X_j$ for all pairs of rules (i, j). There is only one order k interaction $(X_1 X_2 \cdots X_k)$. In the complete model there is one coefficient for each term and one intercept, for a total of 2^k parameters. We will also consider all hierarchical sub-models. Hierarchical sub-models are those in which each term may be included only when all lower order terms composed of the variables in that term are also in the model. For example, $X_2 X_4 X_5$ may be included only if X_2, X_4, X_5, $X_2 X_4$, $X_2 X_5$, and $X_4 X_5$ are included. Non-hierarchical models are sensible in very few applications (Agresti, 1990, p. 144). We thus drop all rules whose estimated main effects are not significantly different from zero.

These rules (with non-significant main effects) are unlikely to cause any of the traps (A, B, and C) discussed above. We can essentially collapse our contingency tables across the levels of these variables, because they are conditionally independent of problem occurrence.

Among those terms with significant main effects, the next step is to look at order 2 interactions. Since we are only considering hierarchical models, if the estimated effect $X_i X_j$ is non-significant, we can drop all higher order interactions involving rules *i* and *j* . In that case, the relation between Y and X_i is unaffected by X_j, so partial measures of association such as Kendall's partial *tau* are meaningful for describing the relation between Y and X_i given X_j. If the estimated effect $X_i X_j$ is significant, then the partial associations between Y and X_i will vary depending on the level of X_j and should be examined individually.

Although we are concerned only with main effects and order 2 interactions, it may be necessary to consider a hierarchical model with higher order terms in order to adequately to fit the data. Since the number of parameters is a power function of the number of rules, this is likely to be impractical. However, it is unlikely that terms higher than order 3 will contribute much to the fit and even less likely that they will contribute to our understanding. A reasonable approach to model-building might be to start by fitting a main effects model. The main effects model could include implementation ratings for all rules in a rule group, as well as variables such as erosion hazard rating that might be related to problem occurrence. The second step would be to fit a second order model with those variables that were significant in the main model. Finally, a third order model could be built from those terms that were significant in the second order model. The primary purpose of this last step would be to validate the results of the second order model.

Confounding variables

What are some of the variables, besides implementation of associated rules, that might affect apparent associations between rule implementation and effectiveness? Any variables that might be associated with both implementation and effectiveness need to be considered.

Erosion Hazard Rating

Environment variables associated with problem occurrence need to be considered. A reliable erosion hazard rating (EHR) is an obvious candidate, and has the attractive property that it may integrate a number of important variables into a single measure. If implementation is also systematically related to EHR, then this variable could be very important to understand. For example, if landowners are more diligent where EHR is high and problems are expected, it will inflate the count of problem sites with good implementation in the marginal table. The effect is to induce a marginal association between implementation and problem occurrence while there may be no association for given levels of EHR.

Other erosion-related environment variables

Variables related to both implementation and effectiveness may or may not be part of the EHR. If a variable is directly related to the ease of implementing some rule, it may be more informative to consider it individually, even if it is part of the EHR. For example, the ability to stabilize fill slopes on abandoned

roads is closely related to slope steepness. Rules governing fill slope stabilization [923.8(b); 23.B.2] will be more difficult to implement on steep slopes, and those are the places where problems are most likely to occur. For another example, in areas with high drainage density, it may be difficult to locate roads and landings to avoid watercourses. Implementation of Rule 923(d); 18.A.06c, which directs that such locations are to be avoided, is likely to violated more frequently when drainage density is high than when it is low. If high drainage density is a geomorphic expression of high erosion rates, then it could induce a spurious marginal association between implementation and effectiveness for this rule.

Ground cover and canopy cover measurements

The instructions specify procedures for measuring ground cover and canopy cover on WLPZ transects, presumably to help evaluate implementation of rules governing the retention of surface, overstory, and understory canopy. The rules differentiate between overstory and understory canopy, and between conifer and non-conifer overstory, so these measurements may not be sufficient.

On midzone transects, both ground cover and canopy cover are to be measured. On streambank transects, ground cover is not measured and canopy cover is only measured on Class 1 streambanks. Transects are 1,000 feet in length. Ground cover is measured at 100-foot intervals using the toe point method, in which ground cover is estimated for approximately 1-foot diameter plots. Canopy cover is measured at 200-foot intervals, each measurement being integrated over a 200-foot distance centered at the measurement point. The method of canopy cover measurement and its integration over 200 feet have not yet been defined.

Ground cover measurements represent a systematic sample of the transect, but canopy cover measurements, consisting of 200-foot plots every 200 feet, represent a complete census. There is no sampling error in a complete census; however, if the integrated measurements are less accurate than individual canopy cover measurements, then increased measurement error is effectively being substituted for sampling error. This may not be a desirable substitution because sampling error is easily estimated from the sample data while additional investigations would be required to estimate measurement errors.

6. Conclusions

Sampling design

Clinical trials are the most effective way to evaluate the causal relations between specific rules and subsequent problems. In such a prospective design, implementation levels would be randomly assigned to sites in the populations or subpopulations of interest and effectiveness would be evaluated after some specified time period or some level of stress had been applied. Such a design ensures that implementation level is unrelated to site characteristics that tend to cause erosion.

Time of assessing implementation

The next best design to clinical trials is the cohort study, also a prospective design. In a cohort study, the treatments are not randomly assigned, but they are identified prior to observing the response. In terms of quality assurance, evaluating the level of implementation prior to stressing events is critical. If this is not done and site damage is observed, it might be much more likely that a rater would judge that a rule was not properly implemented. This is particularly true for many of the rules that are defined in terms of their erosional outcome. It is recommended that, in spite of increased costs, the implementation assessment be done prior to any stressing events.

Time of assessing effectiveness

Stressing events need to be defined and effectiveness should be evaluated only after stressing events have occurred. Some measure of the magnitude of the stressing events that occurred should be included in the analysis.

Quality assurance

Many of the rules are written in subjective language that will make it difficult to assign implementation ratings reliably. If the study is to be credible, it must be shown that the implementation ratings are repeatable. It is straightforward to do so by repeating the assessments with different observers, and measuring their agreement with statistics such as Cohen's *kappa* and others described in Section 3 of this report. At least one of the assessments should be done *before* stressing events have occurred.

A study of observer agreement might result in a decision that some rules are too subjective to bother with in future evaluations. Such a study could even have implications for rule changes.

Marginal associations

Any of the ordinal measures of association (gamma, Somers' d, Kendall's tau-b) presented in Section 2 are appropriate for characterizing the marginal associations between rule implementation and effectiveness. Somers' d_{XY} is perhaps the most suitable because of its asymmetry. Confidence intervals may be computed for these measures and hypothesis testing may be done using a version of the Mann-

Whitney-Wilcoxon test that accounts for a large proportion of tied observations. Mean ridits may be useful because of their probability interpretations. In a case-control or cohort study, these measures must be interpreted very cautiously because they are easily influenced by uncontrolled variables.

Confounding variables

If two variables are each conditionally dependent with one or more additional variables, their marginal association will be different than their partial associations at fixed levels of these confounding variables. In this case, the marginal association is misleading and the partial associations are needed to describe the relationships unambiguously. The partial associations may vary according to the levels of the confounding variables. When considering a single confounding variable, the Mantel score test of conditional independence and Kendall's partial rank-order correlation coefficient may be useful for identifying these relationships.

Loglinear and logit (logistic regression) models are powerful tools for identifying multidimensional structure. Conditional independence can be tested by examining association terms in loglinear models or interaction terms in logit models. Logit models are probably preferred in this application because they model effectiveness as a response, their interpretation is more straightforward than loglinear models, and they can be generalized to handle ordinal responses with more than 2 categories. In addition, most programs for logistic regression permit the inclusion of continuous explanatory variables that need to be considered such as slope steepness and storm recurrence interval.

Sample sizes

Methods for determining samples size for some of the analyses are presented in Section 2. Tests involving categorical data generally require larger samples than analogous tests involving continuous data. This is particularly true when the cell counts are unbalanced. Sample sizes much larger than those obtained in the Pilot Monitoring Program and the first year of the Long Term Monitoring Program will probably be required to detect minimally important effects. Given the high costs of data collection, it may be impractical to obtain the sample sizes and statistical power necessary to disentangle the web of interactions that is likely to be operating. In a test with low power, if the null hypothesis cannot be rejected, the result is inconclusive.

Sample sizes could be increased without increasing costs by spending less time at each site collecting information that is not likely to be used.

Another strategy for increasing sample size is to focus effort on an analysis using individual sites rather than transects or THP's as the units of replication. This would require developing a sampling protocol for non-problem sites, but could greatly increase the sample size.

Software

To perform many of the recommended analyses a high-level statistical package is required. This package needs to have numeric reliability, a scripting language, and the necessary nonparametric statistical procedures. We recommend S-Plus as the statistical package of choice for such analyses. SAS would be a distant second recommendation.

Data analyst

The analysis of the hillslope monitoring database is not a trivial undertaking that can be accomplished using a mechanical approach. To maximize the probability of success, we recommend that the analysis be conducted by a person with training beyond a bachelor's degree and a strong knowledge of statistics.

Other suggestions

1. The definitions of effectiveness for whole THP evaluations should be rephrased in terms of observed erosion (or perhaps evidence of sediment delivery to stream channels), rather than protection of beneficial uses of water. The existing definitions seem difficult to apply and may invite confusion between effectiveness and implementation.

2. A set of rules should be identified in advance that would be evaluated at each site of a given type.

3. Considering implementation evaluations for groups of rules could reduce the complexity of the analysis by reducing the number of interactions that need to be considered.

4. WLPZ canopy measurements need to be more detailed in order to verify implementation of WLPZ rules that distinguish between overstory and understory cover, conifer and non-conifer cover.

References

Agresti, A. 1984. Analysis of Ordinal Categorical Data. New York: Wiley.

Agresti, A. 1990. Categorical Data Analysis. New York: Wiley.

Bishop, M. M., S. E. Fienberg, and P. W. Holland. 1975. Discrete Multivariate Analysis: Theory and Practice. 5th printing. Cambridge: MIT Press.

Brown, B. W., J. Lovato, and K. Russell. In Preparation. Asymptotic power calculations and the art of the merely reasonable. Available at ftp://odin.mdacc.tmc.edu/pub/S/asypow/powpap.ps

Cox, D. R. and D. V. Hinkley. 1974. Thoretical Statistics. London: Chapman and Hall.

Darroch, J. N. and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. Ann. Math. Statist. 43: 1470-1480.

Fienberg, S. 1980. The analysis of cross-classified categorical data. 2nd ed. Cambridge: MIT Press.

Fleiss, J. L. 1981. Statistical Methods for Rates and Proportions. 2nd ed. New York: Wiley.

Goodman, L. A. 1978. Analyzing Qualitative/Categorical Data: Log-Linear Models and Latent Structure Analysis. Cambridge, MA: Abt Books.

Goodman, L. A., and W. H. Kruskal. 1954. Measures of association for cross classifications. J. Amer. Statist. Assoc. 49: 732-764.

Kendall, M. G. 1938. A new measure of rank correlation. Biometrika 30: 81-93.

Kendall, M. G. 1945. The treatment of ties in rank problems. Biometrika 33: 239-251.

Kendall, M. G. 1970. Rank Correlation Methods. 4th ed. London: Griffin.

Lewis, J. and R. M. Rice. 1989. Critical sites erosion study. Technical Report Volume II: Site conditions related to erosion on private timberlands in Northern California. Cooperative investigation by CDF and USFS Pacific Southwest Research Station. Arcata, CA, 95p.

McCullagh, P. and J. A. Nelder. 1989. Generalized Linear Models. 2nd ed. London: Chapman and Hall.

Noether, G. E. 1987. Sample size determination for some common nonparametric tests. J. Amer. Statist. Assoc. 82(398): 645-647.

Plackett, R. L. 1981. The analysis of categorical data. 2nd ed. London: Griffin.

Radelet, M. 1981. Racial characteristics and the imposition of the death penalty. Amer. Sociol. Rev. 46: 918-927.

SAS Institute Inc. 1989. SAS/STAT® User's Guide, Version 6, Fourth Edition, Volume 1. Cary, NC: SAS Institue Inc.

Somers, R. H. 1962. A new asymmetric measure of association for ordinal variables. Amer. Sociol. Rev. 27: 799-811.

Spector, P. 1994. An Introduction to S and S-Plus. Belmont, CA: Duxbury Press.

SPSS Inc. 1994. SPSS Advanced Statistics™ 6.1. Chicago, IL. SPSS Inc.

Tuttle, 1995. Board of Forestry Pilot Monitoring Program: Hillslope Component. Final Report. Contract #9CA38120. California Department of Forestry and Fire Protection.

Appendices

Appendix A. Mann-Whitney U-statistic. (S-Plus)

```
"U2k"<-
function(x)
{
#  Computes Mann-Whitney U-statistic for a 2xk contingency table
#  with the columns representing an ordered categorical variable.
#  Specifically, if X represents the row with the smaller row sum m,
#  and Y represents the row with the larger row sum n, the
#  statistic returned is the sum, over each X observation, of the
#  number of smaller Y values.  Ties count as 1/2.
#  The program also returns the ratio of U12:U21, that is the
#  odds of P(1<2):P(1>2) where 1 is a random observation in
#  row 1 and 2 is a random observation in row 2.  In other words,
#  for the HMDB data, the odds that a random observation from a
#  problem site will have a higher rating than a random observation
#  from a non-problem site.
     sum1 <- sum(x[1,  ])
     sum2 <- sum(x[2,  ])
     if(sum1 >= sum2) {
          maxrow <- 1
          r1 <- x[2,  ]
          r2 <- x[1,  ]
     }
     else {
          maxrow <- 2
          r1 <- x[1,  ]
          r2 <- x[2,  ]
     }
     tmp1 <- cumsum(r1) - 0.5 * r1
     tmp2 <- cumsum(r2) - 0.5 * r2
     Uyx <- sum(r1 * tmp2)
     Uxy <- sum(r2 * tmp1)
     if(maxrow == 1)
          odds <- Uyx/Uxy
     else odds <- Uxy/Uyx
     list(Uyx = Uyx, odds = odds)
}
```

Appendix B. Measures of association. (S-Plus)

```
"assoc.ord"<-
function(x)
{
# Gamma, Somers' dXY, Kendall's tau-b and asymptotic standard errors.
# X is the row variable, Y is the column variable.
# X is expected to be the response, so the matrix is transposed,
# resulting in appropriate computation of dXY rather than dYX.
# Variance is according to Agresti (1984), p. 185-7.
     x <- t(x)  # transposes matrix
     ntot <- sum(x)
     pi <- x/ntot
     nr <- dim(x)[1]
     nc <- dim(x)[2]
     Pcmat <- Pdmat <- pc <- pd <- phi <- matrix(NA, nr, nc)
     for(i in 1:nr)
          for(j in 1:nc) {
               Mplus <- sum(pi[row(x) < i & col(x) < j])
               Mminus <- sum(pi[row(x) < i & col(x) > j])
               Nminus <- sum(pi[row(x) > i & col(x) < j])
               Nplus <- sum(pi[row(x) > i & col(x) > j])
               pc[i, j] <- Mplus + Nplus
               pd[i, j] <- Mminus + Nminus
               Pcmat[i, j] <- pi[i, j] * Nplus
               Pdmat[i, j] <- pi[i, j] * Nminus
          }
#
# pc and pd are matrixes of Agresti's pi^(c) and pi^(d), i.e.
proportion
# of concordant and discordant cells, when matched with cell (i,j)
     pdiff <- pc - pd
     Pc <- 2 * sum(Pcmat)  # proportion of concordant pairs
     Pd <- 2 * sum(Pdmat)  # proportion of discordant pairs
     Pdiff <- Pc - Pd
     Psum <- Pc + Pd  #
# compute gamma and its asymptotic standard error
     gamma <- Pdiff/Psum
     gammaphi <- 4 * (Pd * pc - Pc * pd)
     gammavar <- sum(pi * gammaphi^2)/Psum^4
     ase.gamma <- sqrt(gammavar/ntot) #
# compute somersd and taub
     rowsum <- apply(pi, 1, sum)
     colsum <- apply(pi, 2, sum)
     somersdelta <- 1 - sum(rowsum^2)
     delta1 <- sqrt(somersdelta)
     delta2 <- sqrt(1 - sum(colsum^2))
     taudelta <- delta1 * delta2
```

```
      dXY <- Pdiff/somersdelta
      taub <- Pdiff/taudelta     #
# compute somersphi and tauphi
      rowmat <- matrix(rep(rowsum, nc), ncol = nc)
      colmat <- matrix(rep(colsum, nr), nrow = nr, byrow = T)
      somersphi <- rowmat * Pdiff + somersdelta * pdiff
      tauphi <- (2 * pdiff + Pdiff * colmat) * delta2 * delta1 + (Pdiff
*
           rowmat * delta2)/delta1    #
# Note sum(pi*somersphi) = Pdiff verifies somersd computations
# Compute asymptotic standard errors of somersd and taub
      numer1 <- 4 * sum(pi * somersphi^2)
      numer2 <- 4 * Pdiff^2
      somersvar <- (numer1 - numer2)/somersdelta^4
      ase.dXY <- sqrt(somersvar/ntot)
      numer1 <- sum(pi * tauphi^2)
      numer2 <- sum(pi * tauphi)^2
      tauvar <- (numer1 - numer2)/taudelta^4
      ase.taub <- sqrt(tauvar/ntot)
      list(gamma = gamma, dXY = dXY, taub = taub, ase.gamma =
ase.gamma,
           ase.dXY = ase.dXY, ase.taub = ase.taub)
}
```

Appendix C. Measures of association. (SAS)

```
* Measures of association;

* Read in hypothetical implementation data from Table 1-1;
data;
  input problem $ rule $ count;
  cards;
  no  exceeds 3
  yes exceeds 1
  no  meets   6
  yes meets   7
  no  minor   2
  yes minor   4
  no  major   1
  yes major   2
;

* Calculate summary statistics of association;
proc freq order=data;
  table problem*rule / all;
  weight count;
run;
```

Appendix D. Example interpretation of ordinal logistic regression output from lrm(). (S-Plus)

This example uses artificial data and is designed to show some of the simpler types of output that are possible using Frank Harrell's "Design" library of functions for S-Plus. For more information about these functions, see the online help and Harrell's (1996) book, "Predicting Outcomes: Applied Survival Analysis and Logistic Regression", which can be obtained from the University of Virginia bookstore. Harrell's book shows how to derive much more about a model than we show here, including a variety of graphical methods. This section is intended only as a brief introduction.

In the following, commands are shown in bold, preceded by the ">" prompt symbol, and S-Plus output is shown in `Courier font like this`.

A data frame called example.data contains 240 observations of 3 variables, each rated on a 1-4 scale:

(1) whole THP effectiveness
(2) implementation of rule 1
(3) implementation of rule 2

The lrm() function performs logistic regression for an ordinal response. The fitted model is known as the cumulative logit or proportional odds model. The expression used as the first argument to lrm() below requests a model for effectiveness as a function of rule 1, rule 2, and the interaction of rules 1 and 2.

```
> lrmfit <- lrm(effectiveness ~ rule1 * rule2, data = example.data)
> lrmfit

lrm(formula = effectiveness ~ rule1 * rule2, data = example.data)

Frequencies of Responses
  1  2  3  4
 90 95 30 25

 Obs Max Deriv Model L.R. d.f. P     C   Dxy Gamma Tau-a    R2 Brier
 240     4e-08      49.45   3  0 0.678 0.356  0.42 0.241 0.204 0.206

                     Coef   S.E. Wald Z      P
            y>=2 -2.7242 1.2827 -2.12  0.0337
            y>=3 -4.7070 1.3014 -3.62  0.0003
            y>=4 -5.7727 1.3089 -4.41  0.0000
           rule1  1.0098 0.5898  1.71  0.0869
           rule2  0.7126 0.5072  1.41  0.1600
    rule1 * rule2 -0.1134 0.1925 -0.59  0.5556
```

Among the statistics listed are Somers' d_{XY}, Goodman-Kruskal *gamma*, and Kendall's *tau* for the association between the ranked predicted probabilities and the observed response.

The first three coefficients are the intercepts for the three logits based on the possible cutpoints between the four ordinal effectiveness scores. They may be interpreted as distance measures, indicative of the spacing between categories. The last three coefficients are the effect of each rule and their interaction. If

there were no interaction in the model, each coefficient could be interpreted as the log of the constant odds ratio for a change in 1 unit of that variable. With an interaction in the model, the interpretation is more complicated. With an interaction, the effect of rule 1 is different for each possible value of rule 2. The summary() function, shown further below, helps in interpreting the coefficients. The standard errors of the coefficients are based on a large-sample normal approximation, and the Wald statistics are just the coefficients divided by their standard errors. The p-values are for the null hypothesis that the coefficient is zero. These p-values are not of great interest unless the mean score (2.5) has been subtracted from each rating in advance. Both the coefficients and the p-values will change if a constant is added to or subtracted from the ratings. To determine whether each term contributes significantly to the fit, the anova() function should be used.

The anova() function tests most meaningful hypotheses in a design. If the rating scores had been centered on zero, then the p-values above would have matched those given below for the individual terms.

```
> anova(lrmfit,main.effect=T)
                Wald Statistics          Response: effectiveness

                                     Factor     Chi-Square d.f.      P
 rule1                                           2.93      1    0.0869
 rule1  (Factor+Higher Order Factors)            9.46      2    0.0088
  All Interactions                               0.35      1    0.5556
 rule2                                           1.97      1    0.1600
 rule2  (Factor+Higher Order Factors)            5.20      2    0.0742
  All Interactions                               0.35      1    0.5556
 rule1 * rule2                                   0.35      1    0.5556
 rule1 * rule2  (Factor+Higher Order Factors)    0.35      1    0.5556
 TOTAL                                          44.59      3    0.0000
```

First, the TOTAL Chi-Square statistic indicates there is strong evidence for some associations with y. The p-values for the rule 1 and rule 2 main effects were included only for illustration and are not meaningful when higher order terms containing these variables are in the model. For example, if the interaction is dropped from this model, the main effect for rule 1 becomes significant. From line 2 of the above table, we conclude that rule 1 should be retained, and line 3 says not to keep the interaction. Terms involving rule 2 do not contribute significantly to the model fit.

The summary() command helps to interpret the coefficients. The first 2 commands below are prerequisites to running summary() on an object created by the lrm() function.

```
> dd <- datadist(lrmfit, data=example.data)
> options(datadist="dd")
> summary(lrmfit)

           Effects              Response : effectiveness

 Factor     Low High Diff. Effect S.E. Lower 0.95 Upper 0.95
 rule1       2   3    1     0.78   0.28 0.24       1.33
  Odds Ratio 2   3    1     2.19     NA 1.27       3.77
 rule2       2   3    1     0.49   0.21 0.06       0.91
  Odds Ratio 2   3    1     1.63     NA 1.07       2.48
(Adjusted to: rule1=2 rule2=2)
```

The summary(lrmfit) output shows the effect of a unit change in rule1 when rule2 is fixed at 2, and the effect of a unit change in rule2 when rule1 is fixed at 2. These values are simply computed from the coefficients:

1.0098 + 2(-.1134) = 0.783 odds ratio = exp(0.783) = 2.19
0.7126 + 2(-.1134) = 0.486 odds ratio = exp(0.486) = 1.63

Confidence intervals are also shown for the effects and odds ratios.

Next, let us look at the probabilities predicted by the model for each possible combination of the predictors:

```
> newdat <- expand.grid(rule1=1:4, rule2=1:4)
> predvals <- predict(lrmfit, newdata=newdat, type="fitted")
> cbind(newdat, round(predvals, 3))

     rule1   rule2  y>=2  y>=3  y>=4
 1       1       1 0.247 0.043 0.015
 2       2       1 0.445 0.100 0.037
 3       3       1 0.663 0.213 0.085
 4       4       1 0.828 0.399 0.186
 5       1       2 0.374 0.076 0.028
 6       2       2 0.566 0.152 0.058
 7       3       2 0.741 0.282 0.119
 8       4       2 0.862 0.463 0.229
 9       1       3 0.521 0.130 0.049
10       2       3 0.680 0.226 0.091
11       3       3 0.806 0.363 0.164
12       4       3 0.890 0.527 0.277
13       1       4 0.664 0.214 0.086
14       2       4 0.775 0.322 0.141
15       3       4 0.857 0.453 0.222
16       4       4 0.913 0.591 0.332
```

Each row of the output shows the predicted probabilities that the effectiveness rating for that configuration of rule implementation equals or exceeds 2, 3, and 4, respectively. Thus, we have the modelled distribution of effectiveness for each combination of rule implementation.

The results of lrm() can be sensitive to the rating scale. To see this, suppose rule 1 and rule 2 are treated as categorical variables. Let us first look at the results of fitting a model with no interaction.

```
> lrm(effectiveness ~ factor(implem1)+factor(implem2), data=x4.df)

Logistic Regression Model

lrm(formula = effectiveness ~ factor(rule1) + factor(rule2), data =
x4.df)

 Obs Max Deriv Model L.R. d.f. P     C   Dxy Gamma Tau-a    R2 Brier
 240     4e-07      57.39    6 0 0.683 0.366 0.434 0.248 0.233 0.202
```

```
             Coef   S.E. Wald Z      P
    y>=2 -0.29655 0.4615 -0.64  0.5205
    y>=3 -2.34475 0.4891 -4.79  0.0000
    y>=4 -3.43119 0.5187 -6.62  0.0000
  rule1=2  0.09411 0.4507  0.21  0.8346
  rule1=3  1.53456 0.5916  2.59  0.0095
  rule1=4  1.57397 0.7331  2.15  0.0318
  rule2=2  0.17925 0.3498  0.51  0.6083
  rule2=3  0.79748 0.4631  1.72  0.0850
  rule2=4  0.83688 0.6617  1.26  0.2059
```

Although the standard errors are rather high, the similarity of coefficients for ratings 3 and 4 (on both rules) suggests that these ratings could be treated as one category. To demonstrate the effect that can have on the results, let's look at the anova() table after combining these categories.

```
> attach(example.data)
> recoded1 <- rule1
> recoded2 <- rule2
> recoded1[recoded1==4] <- 3
> recoded2[recoded2==4] <- 4
> anova(lrm(effectiveness ~ recoded1 * recoded2))

 Wald Statistics          Response: effectiveness

                                  Factor     Chi-Square d.f.      P
 recoded1  (Factor+Higher Order Factors)    18.46      2    0.0001
  All Interactions                           5.48      1    0.0192
 recoded2  (Factor+Higher Order Factors)    10.87      2    0.0044
  All Interactions                           5.48      1    0.0192
 recoded1 * recoded2  (Factor+Higher Order)  5.48      1    0.0192
 TOTAL                                      50.15      3    0.0000
```

Now the interaction as well as both rules appear to be significant. One must be conservative, however, when fitting and refitting different models suggested by the data, because iterative model fitting has the effect of increasing Type I errors. In other words, the *P*-values given by the program are too low, because they ignore the fact that these tests are conditional upon the results of earlier analyses.

Appendix E. Ordinal logistic regression. (SAS)

```
* Mental impairment by SES and Life Events example;
* From Agresti (1990), page 325;

data mental;
  input mental $ SES events;
  cards;
  Well 1 1
  Well 1 9
  Well 1 4
  Well 1 3
  Well 0 2
  Well 1 0
  Well 0 1
  Well 1 3
  Well 1 3
  Well 1 7
  Well 0 1
  Well 0 2
  Mild 1 5
  Mild 0 6
  Mild 1 3
  Mild 0 1
  Mild 1 8
  Mild 1 2
  Mild 0 5
  Mild 1 5
  Mild 1 9
  Mild 0 3
  Mild 1 3
  Mild 1 1
  Moderate 0 0
  Moderate 1 4
  Moderate 0 3
  Moderate 0 9
  Moderate 1 6
  Moderate 0 4
  Moderate 0 3
  Impaired 1 8
  Impaired 1 2
  Impaired 1 7
  Impaired 0 5
  Impaired 0 4
  Impaired 0 4
  Impaired 1 8
  Impaired 0 8
  Impaired 0 9
```

```
;

proc logistic data=mental order=data;
  model mental = ses events;
run;
```

Appendix F. McNemar's Test. (S-Plus)

```
y <- matrix(c(18,10,7,5), nrow=2, ncol=2)
mcnemar.test(y, correct=F)
```

Appendix G. McNemar's Test. (SAS)

```
* McNemar's test;

* Read in data from Table 2-1;
data table2_1;
  input time1 $ time2 $ count;
  cards;
    adequate   adequate 18
    adequate inadequate  7
  inadequate   adequate 10
  inadequate inadequate  5
;

* Calculate statistics of agreement;
proc freq data=table2_1;
  table time1*time2 / agree;
  weight count;
run;
```

Appendix H. Programming the exact power for McNemar's Test. (BASIC)

Below is a simple BASIC program that will calculate power given the sample size and the cell probabilities. No attempt is made to make the program user-friendly or check for values the sample size that might be too large to run in a reasonable amount of time.

```
DIM logF(50)

REM Set sample size and cell probabilities
n=40
p00=0.36
p01=0.34
p10=0.04
p11=0.26

REM Set some constants
logp00=LOG(p00)
logp01=LOG(p01)
logp10=LOG(p10)
logp11=LOG(p11)

REM Calculate log factorials
logF(0)=0
logF(1)=0
FOR i=2 TO 50
  logF(i)=logF(i-1)+LOG(i)
NEXT i

REM Examine all possible tables
totalp=0
power=0
FOR x00=0 TO n
  FOR x01=0 TO n-x00
    FOR x10=0 TO n-x00-x01
      x11=n-x00-x10-x01
      REM Calculate multinomial probability of table
      p=logF(n)-logF(x00)-logF(x10)-logF(x01)-logF(x11)
      p=p + x00*logp00 + x10*logp10 + x01*logp01 + x11*logp11
      p=EXP(p)
      totalp=totalp + p
      REM Calculate test statistic
      IF x01=x10 THEN
        t=0
      ELSE
        t=(x01-x10)^2/(x01+x10)
      END IF
      IF t>3.84 THEN power=power + p
    NEXT x10
```

```
  NEXT x01
NEXT x00
PRINT power;totalp
```

Appendix I. Kappa statistics and Bowker's test of symmetry. (SAS)

```
* Test of symmetry and kappa statistics;

* Read in hypothetical implementation data from Table 3-1;
* (Note that zero counts are entered as a very small
   number.  This is because SAS will ignore cells where
   the weight is zero.  And keeping the counts zero but just
   adding a statement "if count=0 then count=0.00001"
   results in a different ordering of the variable values.)
;
data;
  input r1 $ r2 $ count;
  cards;
  exceeds exceeds 3
  exceeds meets   2
  exceeds minor   0.000001
  exceeds major   1
  meets   exceeds 1
  meets   meets   6
  meets   minor   1
  meets   major   0.000001
  minor   exceeds 0.000001
  minor   meets   1
  minor   minor   4
  minor   major   3
  major   exceeds 0.000001
  major   meets   0.000001
  major   minor   1
  major   major   3
;

* Calculate statistics of agreement;
proc freq order=data;
  table r1*r2 / agree;
  weight count;
run;
```

Appendix J. Testing marginal homogeneity for square tables using Bhapkar's test. (SAS)

```
* Bhapkar's test of marginal homogeneity;

* Read in hypothetical implementation data from Table 3-1;
* (Note that zero counts are entered as a very small
   number.  This is because some SAS procedures will
   ignore cells where the weight is zero.  And keeping
   the counts zero but just adding a statement
   "if count=0 then count=0.00001" results in a different
   ordering of the variable values which can affect the
   ordinal tests.);
data rdata;
  input r1 $ r2 $ count;
  cards;
  exceeds exceeds 3
  exceeds meets   2
  exceeds minor   0.000001
  exceeds major   1
  meets   exceeds 1
  meets   meets   6
  meets   minor   1
  meets   major   0.000001
  minor   exceeds 0.000001
  minor   meets   1
  minor   minor   4
  minor   major   3
  major   exceeds 0.000001
  major   meets   0.000001
  major   minor   1
  major   major   3
;

* Calculate Bhapkar's test of marginal homogeneity;
* The desired output will be the row labeled "RATERS"
  in the "Analysis of Variance" section of the output.;
proc catmod data=rdata;
  weight count;
  response marginals;
  model r1*r2= _response_ / freq;
  repeated RATERS 2;
run;
```

Appendix K. Testing marginal homogeneity for square tables using conditional symmetry. (BASIC)

```
DIM x(4,4)
REM Set an example (Agresti, 1990, Page 364)
DATA 31,12,14,6
DATA 5,1,1,1
DATA 5,0,2,1
DATA 0,0,1,0
I=4
FOR i=1 TO I
  FOR j=1 TO I
    READ x(i,j)
  NEXT j
NEXT i

CLS
REM Symmetry
s=0
gs=0
dfS=0
FOR i=1 TO I-1
  FOR j=i+1 TO I
    dfS=dfS + 1
    mij=(x(i,j)+x(j,i))/2
    gs=gs + 2*x(i,j)*LOG(x(i,j)/mij)
    gs=gs + 2*x(j,i)*LOG(x(j,i)/mij)
    s=s + (x(i,j)-x(j,i))^2/(x(i,j)+x(j,i))
  NEXT j
NEXT i
PRINT "            Symmetry:  X2 = ";s;"  df = ";dfS
PRINT "            Symmetry:  G2 = ";gs;"  df = ";dfS

REM Conditional symmetry
t1=0
t2=0
cs=0
FOR i=1 TO I-1
  FOR j=i+1 TO I
    t1=t1 + x(i,j)
    t2=t2 + x(j,i)
  NEXT j
NEXT i
t=LOG(t1/t2)
PRINT "tau = ";t
expt=EXP(t)
gcs=0
FOR i=1 TO I-1
```

```
  FOR j=i+1 TO I
    mij=expt*(x(i,j)+x(j,i))/(expt+1)
    mji=(x(i,j)+x(j,i))/(expt+1)
    cs=cs + (x(i,j)-mij)^2/mij
    cs=cs + (x(j,i)-mji)^2/mji
    gcs=gcs + 2*x(i,j)*LOG(x(i,j)/mij)
    gcs=gcs + 2*x(j,i)*LOG(x(j,i)/mji)
  NEXT j
NEXT i
dfCS=((I+2)*(I-1))/2
PRINT "Conditional symmetry:  X2 = ";cs;"  df = ";dfCS
PRINT "Conditional symmetry:  G2 = ";gcs;"  df = ";dfCS%
PRINT "Marginal homogeneity:  X2 = ";s-cs;"  df = ";dfS-dfCS
PRINT "Marginal homogeneity:  G2 = ";gs-gcs;"  df = ";dfS-dfCS
```

Appendix L. Cochran's Q. (BASIC)

```
DIM x(10,100)

REM Sample data
REM number of observations (n) and number of
REM time periods
DATA 8,4
REM Binary response with each row representing
REM the T time periods for a particular location
DATA 1,1,0,1
DATA 1,1,1,1
DATA 0,0,0,0
DATA 0,0,0,0
DATA 1,1,1,1
DATA 1,0,0,0
DATA 1,1,1,1
DATA 1,1,1,0

REM Read in data
READ n,T
FOR j=1 TO n
  FOR i=1 TO T
    READ x(i,j)
  NEXT i
NEXT j

REM Calculate necessary means
REM Overall mean (x(0,0)) and time period means
x(0,0)=0
FOR i=1 TO T
  x(i,0)=0
  FOR j=1 TO n
    x(i,0)=x(i,0)+x(i,j)
  NEXT j
  x(0,0)=x(0,0) + x(i,0)
  x(i,0)=x(i,0)/n
NEXT i
x(0,0)=x(0,0)/(n*T)

REM Location means
FOR j=1 TO n
  x(0,j)=0
  FOR i=1 TO T
    x(0,j)=x(0,j) + x(i,j)
  NEXT i
  x(0,j)=x(0,j)/T
NEXT j
```

```
REM Now construct the Q statistic
a=0
FOR i=1 TO T
  a=a+(x(i,0)-x(0,0))^2
NEXT i
b=0
FOR j=1 TO n
  b=b+x(0,j)*(1-x(0,j))
NEXT j
q=n^2*(T-1)*a/(T*b)

REM Print results
CLS
PRINT "Cochran's Q = ";q
PRINT "Degrees of freedom are ";T-1
```

Appendix M. Cochran's Q. (S-Plus)

```
"cochran.test"<-
function(x)
{
#  Cochran's Q test for change over time in a binary variable.
#  x is a matrix with one column for each observation time.
#  Rows correspond to subjects being observed.
     n <- dim(x)[[1]]
     t <- dim(x)[[2]]
     rowmean <- apply(x, 1, mean)
     colmean <- apply(x, 2, mean)
     numer <- sum((colmean - mean(x))^2)
     denom <- sum(rowmean * (1 - rowmean))
     Q <- (n * n * (t - 1) * numer)/(t * denom)
     p <- 1 - pchisq(Q, t - 1)
     list(Q = Q, prob = p)
}
```

Appendix N. Cochran's Q. (SAS)

```
* Cochran's Q-test;

* Read in hypothetical data from Table 2-2;
data timedata;
  input location t1-t4;
  cards;
  1 1 1 0 1
  2 1 1 1 1
  3 0 0 0 0
  4 0 0 0 0
```

```
  5 1 1 1 1
  6 1 0 0 0
  7 1 1 1 1
  8 1 1 1 0
;

* Calculate Cochran's Q;
proc freq data=timedata;
  table t1*t2*t3*t4 / agree;
run;
```

Appendix O. Weighted and unweighted kappa. (S-Plus)

```
"kappa.stat"<-
function(nij, wij = diag(rep(1, ncol(nij))))
{
#
# SGH: 05/02/95
# Susan Galloway Hilsenbeck          Internet:
sue@oncology.uthscsa.edu
# UTHSCSA                              MaBell:    (210) 567-4749
# 7703 Floyd Curl Drive                   Fax:    (210) 567-6687
# San Antonio, TX 78284-7884  USA
#
# After Fleiss JL "Statistical methods for rates and proportions",2nd
ed.
# J Wiley. p223-224, 1981.
#
# nij is a square matrix of counts
# wij is a symmetric matrix of weights (0-1).
#    default weight matrix wij is the identity matrix, and counts only
#    perfect matches.
#
     nij <- as.matrix(nij)
     wij <- as.matrix(wij)
     if(sum(wij != t(wij))) {
          stop(message = "Weight matrix must be symmetric")
     }
     if(sum(dim(nij) != dim(wij))) {
          stop(message = "Dimensions of data and weight matrices must
match"
               )
     }
     pij <- nij/sum(nij)   # convert count data to proportions
     po <- sum(wij * pij)  # compute observed agreement
     pi <- apply(pij, 1, sum)
     pj <- apply(pij, 2, sum)
     pe <- sum(wij * pi %o% pj) # compute agreement expected by chance
     k <- (po - pe)/(1 - pe)    # compute kappa
     wbi <- pj %*% wij
     wbj <- pi %*% wij
     wb <- matrix(rep(wbi, ncol(nij)), ncol = ncol(nij), byrow = T) +
matrix(
          rep(wbj, ncol(nij)), ncol = ncol(nij))
     # compute standard error of weighted kappa
     se.k <- 1/((1 - pe) * sqrt(sum(nij))) * sqrt(sum(pi %o% pj *
t(wij - wb
          )^2) - pe^2)
     z <- k/se.k     # test of h0: k=0
     p.value <- 1 - pnorm(z)
     return(data.frame(po, pe, k, se.k, z, p.value))
}
```