Biopharmaceutical Statistics for Drug Development

edited by

Karl E. Peace
G. D. Searle & Co.
Skokie, Illinois

MARCEL DEKKER, INC. New York and Basel

9 Clinical Efficacy Trials with Categorical Data

GARY G. KOCH *Biostatistics Department, School of Public Health, University of North Carolina, Chapel Hill, North Carolina*

SUZANNE EDWARDS *Clinical Statistics Department, Burroughs Wellcome Co., Research Triangle Park, North Carolina*

I. INTRODUCTION

Efficacy response measures in clinical trials often take the form of categorical variables. Such response variables may be dichotomous (e.g., healed vs. not healed), ordinal (e.g., symptom severity of none, mild, moderate, or severe), discrete counts (e.g., number of occurrences of some symptom), or grouped survival times (e.g., time interval of recurrence of symptoms). An overview of the primary methods used in the analysis of categorical efficacy response measures is shown in Table 1. Examples of both nonparametric randomization-based methods, such as Fisher's exact test, and model-based methods, such as logistic regression, appear in this table. The rationale for choosing one of these methods of inference over the other lies in the assumptions one is willing to make about the connection between the sample of patients enrolled in the trial and the target population to which results are to be generalized.

In most cases, the patients in a clinical trial are a convenience sample, chosen by their ability to satisfy the study protocol, and the investigators are a judgment sample, chosen by their expertise in the area of study. Given this sampling process, the patients in a clinical trial may not be statistically representative of any well-defined target population. If assumptions which involve random sampling from a larger target population are not tenable, then nonparametric randomization-based methods may be more appropriate than model-based methods. Nonparametric randomization-based methods require no assumptions other than the randomization of patients to treatment groups. However, these methods have limited scope in that the conclusions drawn on a statistical basis apply only to the patients who were actually randomized; generalization to a larger population requires nonstatistical arguments about the representativeness of the observed patients (see Koch et al., 1982; Koch and Gillings, 1983). Also, nonparametric randomization-based methods are useful mainly in hypothesis testing; they must be supplemented by other methods for estimation purposes.

Table 1 Overview of Statistical Methods for Categorical Efficacy Response Variables

Response[a] variable	Randomization-based nonparametric methods		Model-based methods (single or combined investigators)
	Single investigator	Combined investigators	
Dichotomous	Fisher's exact test	Mantel-Haenszel test	Maximum likelihood logistic regression, weighted least squares for correlated proportions from repeated measures
Ordinal or discrete counts	Wilcoxon rank sum or Kruskal-Wallis test	Extensions of the Mantel-Haenszel procedure	Extensions of logistic regression (e.g., the proportional odds model), weighted least squares for rank measures of association or mean scores
Grouped survival times	Logrank test	Stratified logrank test	Maximum likelihood fitting of piecewise exponential models

[a]Polytomous nominal variables are not included in the table because they are rarely used as efficacy variables in clinical trials. Such variables can be analyzed with chi-square tests, multivariate forms of the Mantel-Haenszel test, or log-linear model extensions of logistic regression.

Model-based methods are appropriate if it can be assumed that the patients selected for the trial are equivalent to a (possibly stratified) random sample of some larger population. Assumed probability models can then be used to describe the relationship of the response variable to explanatory variables such as treatment, investigator, and pretreatment patient characteristics. As noted by Koch and Sollecito (1984), model-based methods have several advantages over nonparametric randomization-based methods: (1) the capacity to assess the homogeneity of effects across strata (e.g., center-by-treatment interactions), (2) greater flexibility with respect to adjustment for demographic or pretreatment variables which are not equivalently distributed for the treatment groups, and (3) more powerful analysis through the framework of reduced variability provided by adjustment for explanatory variables which are strongly associated with the response variable. To clarify advantages (2) and (3) further, it must be noted that equivalent distributions for demographic and pretreatment variables are expected as a consequence of the randomized assignment of treatments. When noteworthy imbalances in important explanatory variables do occur, they can be adjusted for with randomization-based methods via stratification. The advantage of model-based methods for such adjustments is that a larger number of explanatory variables may be adjusted for by incorporating them in the model, and these variables may be either categorical or continuous. The disadvantages of model-based methods are that they may be more difficult to implement and interpret than randomization-based methods. Also, since the assumptions necessary for model-based methods cannot be proved, the results of such analyses are more likely to be subject to debate. A reasonable strategy is thus to use both types of methods in combination.

In this chapter, a variety of model-based and nonparametric randomization-based methods for analysis of categorical data are reviewed, using examples from clinical trials with different data structures (i.e., univariate, multivariate) and measurement scales for the response and explanatory variables.

II. ANALYSIS OF 2 × 2 TABLES

The data in Table 2 are from a randomized double-blind placebo-controlled clinical trial in patients with rheumatoid arthritis. Patients evaluated the effectiveness of treatment according to a three-point scale: no improvement, some improvement, or marked improvement. The explanatory variables of interest are treatment (test drug vs. placebo), sex, and age.

If one reduces the response variable to a dichotomous variable and considers treatment as the only explanatory variable of interest, the data may be presented as follows:

	Improvement		
Treatment	None	Some or marked	Total
Test drug	$n_{11} = 13$	$n_{12} = 28$	$n_{1+} = 41$
Placebo	$n_{21} = 29$	$n_{22} = 14$	$n_{2+} = 43$
Total	$n_{+1} = 42$	$n_{+2} = 42$	$n = 84$

(1)

Table 2 Rheumatoid Arthritis Data

Test drug treatment				Placebo treatment			
Pt. #	Sex	Age	Improvement[a]	Pt. #	Sex	Age	Improvement[a]
57	M	27	1	9	M	37	0
46	M	29	0	14	M	44	0
77	M	30	0	73	M	50	0
17	M	32	2	74	M	51	0
36	M	46	2	25	M	52	0
23	M	58	2	18	M	53	0
75	M	59	0	21	M	59	0
39	M	59	2	52	M	59	0
33	M	63	0	45	M	62	0
55	M	63	0	41	M	62	0
30	M	64	0	8	M	63	2
5	M	64	1	80	F	23	0
63	M	69	0	12	F	30	0
83	M	70	2	29	F	30	0
66	F	23	0	50	F	31	1
40	F	32	0	38	F	32	0
6	F	37	1	35	F	33	2
7	F	41	0	51	F	37	0
72	F	41	2	54	F	44	0
37	F	48	0	76	F	45	0
82	F	48	2	16	F	46	0
53	F	55	2	69	F	48	0
79	F	55	2	31	F	49	0
26	F	56	2	20	F	51	0
28	F	57	2	68	F	53	0
60	F	57	2	81	F	54	0
22	F	57	2	4	F	54	0
27	F	58	0	78	F	54	2
2	F	59	2	70	F	55	2
59	F	59	2	49	F	57	0
62	F	60	2	10	F	57	1
84	F	61	2	47	F	58	1

Table 2 (continued)

Test drug treatment				Placebo treatment			
Pt. #	Sex	Age	Improvement[a]	Pt. #	Sex	Age	Improvement[a]
64	F	62	1	44	F	59	1
34	F	62	2	24	F	59	2
58	F	66	2	48	F	61	0
13	F	67	2	19	F	63	1
61	F	68	1	3	F	64	0
65	F	68	2	67	F	65	2
11	F	69	0	32	F	66	0
56	F	69	1	42	F	66	0
43	F	70	1	15	F	66	1
				71	F	68	1
				1	F	74	2

[a]Improvement: 0, none; 1, some; 2, marked.

Although a few patients may have been excluded from each group due to protocol violations, the row marginal totals (n_{1+}, n_{2+}) can be considered fixed by the treatment allocation process. The column marginal totals (n_{+1}, n_{+2}) can also be considered as fixed under the null hypothesis H_0 of no treatment difference for each patient, n_{+1} and n_{+2} being respectively the number with no improvement and the number with at least some improvement, regardless of treatment. Given that all marginal totals are fixed, the probability model implied by randomization is given by the hypergeometric distribution, i.e.,

$$\Pr\{n_{ij} \mid H_0\} = \frac{n_{1+}!\, n_{2+}!\, n_{+1}!\, n_{+2}!}{n!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!} \tag{2}$$

Thus, the expected value of n_{11} is

$$E\{n_{11} \mid H_0\} = \frac{n_{1+}n_{+1}}{n} = m_{11} = 20.50 \tag{3}$$

and the variance is

$$V\{n_{11} \mid H_0\} = \frac{n_{1+}n_{2+}n_{+1}n_{+2}}{n^2(n-1)} = v_{11} = 5.31 \tag{4}$$

If the sample size is sufficiently large, n_{11} has approximately a normal distribution by central limit theory (Hannan and Harkness, 1963; Puri and Sen, 1971; Plackett, 1981), and so

$$Q = \frac{(n_{11} - m_{11})^2}{v_{11}} \tag{5}$$

approximately has a chi-square distribution with 1 degree of freedom (d.f.). The value of Q does not depend on which of the four cells is used in the calculations because for the ij-th cell, $(n_{ij} - m_{ij}) = \pm(n_{11} - m_{11})$ and $v_{ij} = v_{11}$, where m_{ij} and v_{ij} are the expected value and variance of n_{ij} under H_0. The statistic Q can also be written as

$$Q = \frac{\{(n_{1+}n_{2+}/n)(p_{11} - p_{21})\}^2}{v_{11}} \tag{6}$$

where $p_{i1} = (n_{i1}/n_{i+})$ denotes the proportion of patients in the ith group with no improvement. This expression shows how larger values of Q can be interpreted as indicating larger differences between treatments with respect to the proportions of patients with no improvement; in this sense, $(p_{11} - p_{21})$ describes the association between treatment and response.

The relationship between the randomization chi-square statistic, Q, and the well-known Pearson chi-square statistic, Q_P, where

$$Q_P = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}} \tag{7}$$

for 2×2 tables is $Q = [(n - 1)/n]Q_P$. Thus, in large samples the two statistics are nearly identical. In our example, $Q = 10.59$ $(p < 0.01)$ and $Q_P = 10.72$ $(p < 0.01)$.

A commonly used rule for the suitability of chi-square tests such as Q or Q_P to 2×2 contingency tables is that the expected value, $m_{ij} = n_{i+}n_{+j}/n$, for all cells should exceed 5.0. This is true in our example. A more appropriate method than chi-square tests when cell counts are small is Fisher's exact test. The p value for Fisher's (two-sided) exact test is found by enumerating all tables with the same marginal totals as the observed table, calculating the probability of each table using (2), and then summing the probabilities of those tables which are as likely as or less likely than the observed table. For the one-sided Fisher's exact test, summation is applied to the probabilities of tables with association at least as strong as that for the observed table in the direction specified by the one-sided alternative (here, tables in which the test drug has the more favorable response). In our example, the p value from a two-sided Fisher's exact test is 0.002, and the p value from a one-sided Fisher's exact test is 0.001. When the sample sizes in the two treatment groups are nearly equal, as in our example, the p value from a two-sided Fisher's exact test is approximately twice the p value from a one-sided test; when they are actually equal, this approximation becomes an identity. However, when the sample sizes differ, the

Fisher's exact test is usually nonsymmetric, and the two-sided p value can be notably less than twice the one-sided p value.

Some authors (Haber, 1986; Overall and Starbuck, 1983; Salama et al., 1984) have advocated the use of exact tests which are more powerful than Fisher's exact test. However, Fisher's exact test has the advantage of following directly from randomization, whereas other tests require that one assumes the data for each treatment have a binomial distribution. Justification of this assumption requires that the patients in each group can be viewed as equivalent to a random sample from some larger target population. For situations in which the assumption of binomial distributions is reasonable, a numerical study by Upton (1982) suggests that chi-square approximations for Q are somewhat better than those for Q_P.

A continuity correction is often applied to both the randomization and the Pearson chi-square statistics to make the results agree more closely with results of an exact test. The continuity-corrected randomization statistic is

$$Q_C = \frac{\{|n_{11} - m_{11}| - 0.5\}^2}{v_{11}} \tag{8}$$

and the continuity-corrected Pearson statistic (Yates, 1934) is

$$Q_{PC} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\{|n_{ij} - m_{ij}| - 0.5\}^2}{m_{ij}} \tag{9}$$

In our example, the continuity-corrected Pearson statistic, $Q_{PC} = 9.34$ ($p = 0.002$), does more closely approximate the two-sided Fisher's exact test than the uncorrected statistic. Fleiss (1981) recommends that the Pearson chi-square statistic always be corrected for continuity, although there is some controversy on this point (see Grizzle, 1967; Conover, 1974). One aspect of this controversy concerns whether evaluation is based on the hypergeometric distribution induced by randomization or on the assumption of binomial distributions for each treatment. Another consideration is that in nonsymmetric situations, the one-sided p value from a continuity-corrected chi-square statistic tends to approximate the one-sided Fisher's exact test p value well, but its two-sided counterpart (which is twice as large) then tends to be larger than the two-sided Fisher's exact test p value to an overly conservative extent.

In summary, for the analysis of 2 × 2 tables from clinical efficacy trials, Fisher's exact test is always applicable and easily obtained by computer. Approximations for it (i.e., all chi-square tests, continuity-corrected or not) should be used cautiously, particularly in nonsymmetric situations, unless the sample sizes are sufficiently large (e.g., all $m_{ij} \geq 10$).

III. ANALYSIS OF s × r TABLES

Another way to view the data in Table 2 is to construct a 2 × 3 contingency table for the cross-classification of treatment and the ordinal response variable.

Treatment	Improvement			
	None	Some	Marked	Total
Test drug	13	7	21	41
Placebo	29	7	7	43
Total	42	14	28	84

(10)

Let n_{ij} denote the number of patients who received the ith treatment ($i = 1, 2$) and had the jth response ($j = 1, 2, 3$). Again, the row marginals, n_{i+}, may be considered fixed by the treatment allocation process, and the column marginals, n_{+j}, may be considered fixed under the null hypothesis H_0 of no treatment effect for each patient by considering the n_{+j} as the number with the jth level of improvement, regardless of treatment. The probability model implied by randomization is

$$\Pr\{n_{ij} \mid H_0\} = \frac{\prod_{i=1}^{2} n_{i+}! \prod_{j=1}^{3} n_{+j}!}{n! \prod_{i=1}^{2}\prod_{j=1}^{3} n_{ij}!} \tag{11}$$

which is sometimes called the multivariate hypergeometric model. The frequencies n_{ij} have expected values

$$E\{n_{ij} \mid H_0\} = \frac{n_{i+}n_{+j}}{n} \tag{12}$$

and covariance structure

$$\mathrm{Cov}\{n_{ij}, n_{i'j'} \mid H_0\} = \frac{m_{ij}(n\delta_{ii'} - n_{i'+})(n\delta_{jj'} - n_{+j'})}{n(n-1)} \tag{13}$$

where $\delta_{kk'} = 1$ if $k = k'$ and $\delta_{kk'} = 0$ if $k \neq k'$.

The form of the Pearson chi-square statistic for $s \times r$ tables is the same as (7) except that the summation for i is from 1 to s, where s is the number of groups (e.g., treatments; here $s = 2$), and the summation for j is from 1 to r, where r is the number of response levels (here $r = 3$). For sufficiently large samples (e.g., all $m_{ij} \geq 5$), Q_P approximately has the chi-square distribution with $(r-1)(s-1)$ degrees of freedom. In this example, $Q_P = 13.06$, d.f. = 2, and $p < 0.01$.

The general form of the randomization chi-square statistic counterpart of Q_P is given by the quadratic form

$$Q = (\underline{n} - \underline{m})' \underline{A}' (\underline{A}\underline{V}\underline{A}')^{-1} \underline{A}(\underline{n} - \underline{m}) \tag{14}$$

where $\underline{n} = (13, 7, 21, 29, 7, 7)'$ is the compound vector of observed frequencies, $\underline{m}$ is the corresponding 6×1 vector of expected frequencies from (12), $\underline{V}$ is the 6×6 covariance matrix from (13), and $\underline{A}$ is any 2×6 matrix such that $\underline{A}\underline{V}\underline{A}'$ is nonsingular. For $\underline{A}\underline{V}\underline{A}'$ to be nonsingular, $\underline{A}$

must be linearly independent of the restrictions that sums of the $(n_{ij} - m_{ij})$ in the same row or in the same column are identically zero; thus, Rank [$\underline{A}'$, $\underline{R}'$] must equal Rank[$\underline{A}'$] + Rank[$\underline{R}'$], where $\underline{R}$ is a basis for the restrictions, e.g.,

$$\underline{R} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{15}$$

One convenient choice is $\underline{A} = [\underline{I}_2, \underline{0}_{2,4}]$, where $\underline{I}_2$ is the 2 × 2 identity matrix and $\underline{0}_{2,4}$ is the 2 × 4 matrix of zeros. As noted in Koch et al. (1982) and other references, $Q = (n - 1)Q_P/n$ for any $\underline{A}$ satisfying the above criteria, so Q has the same approximate chi-square distribution as Q_P with d.f. = $(r - 1)(s - 1)$. In this example, $Q = 12.90$, d.f. = 2, and $p < 0.01$.

Although Q and Q_P are useful for detecting general types of departures from H_0, they are not as effective as other methods for alternatives involving location shifts across ordinal response levels (e.g., the more favorable response categories being more likely for some treatments than others). A class of randomization statistics which accounts for ordinal response levels can be constructed by changing the form of the $\underline{A}$ matrix in (14). Let $\underline{a} = \{a_j\} = (a_1, a_2, a_3)'$ be a set of scores reflecting response levels. Then the mean score for the test drug group is

$$\bar{f}_1 = \sum_{j=1}^{3} \frac{a_j n_{1j}}{n_{1+}} \tag{16}$$

which has expected value

$$E\{\bar{f}_1 \mid H_0\} = \sum_{j=1}^{3} \left(a_j \frac{n_{1+} n_{+j}}{n_{1+} n} \right) = \sum_{j=1}^{3} a_j \frac{n_{+j}}{n} = \mu_{\underline{a}} \tag{17}$$

and variance

$$\mathrm{Var}\{\bar{f}_1 \mid H_0\} = \sum_{j=1}^{3} \sum_{j'=1}^{3} a_j a_{j'} \frac{n_{1+}(n - n_{1+}) n_{+j} (n\delta_{jj'} - n_{+j'})}{n^2 n_{1+}^2 (n - 1)} \tag{18}$$

$$= \frac{n - n_{1+}}{n_{1+}(n - 1)} \sum_{j=1}^{3} (a_j - \mu_{\underline{a}})^2 \left(\frac{n_{+j}}{n} \right)$$

$$= \frac{(n - n_{1+}) v_{\underline{a}}}{n_{1+}(n - 1)}$$

where $\mu_{\underline{a}}$ and $v_{\underline{a}}$ are the finite population mean and variance of scores $\underline{a}$ for the 84 patients on study. In large-sample situations, $\bar{f}_1$ approximately has a normal distribution by randomization central limit theory, and so the mean score statistic

$$Q_S = \frac{(\bar{f}_1 - \mu_{\underline{a}})^2}{[(n - n_{1+})/n_{1+}(n - 1)]v_{\underline{a}}}$$

$$= \left(\frac{n - 1}{n}\right)\frac{(\bar{f}_1 - \bar{f}_2)^2}{[(1/n_{1+}) + (1/n_{2+})]v_{\underline{a}}} \qquad (19)$$

approximately has the chi-square distribution with d.f. = 1. The general expression of this statistic is

$$Q_S = (\underline{n} - \underline{m})'\underline{A}_S'(\underline{A}_S\underline{V}\underline{A}_S')^{-1}\underline{A}_S(\underline{n} - \underline{m}) \qquad (20)$$

where $\underline{A}_S = [\underline{a}', -\underline{a}']$ for the comparison of s = 2 treatments.

When the overall comparison among s > 2 treatments is of interest for a response with r levels, a convenient choice for $\underline{A}_S$ is the (s − 1) × sr matrix

$$\begin{bmatrix} \underline{a}' & \underline{0}' & \cdots & \underline{0}' & -\underline{a}' \\ \underline{0}' & \underline{a}' & \cdots & \underline{0}' & -\underline{a}' \\ & & \cdots & & \\ \underline{0}' & \underline{0}' & \cdots & \underline{a}' & -\underline{a}' \end{bmatrix} \qquad (21)$$

As noted in Koch and Bhapkar (1982) and elsewhere, Q_S in (20) can be simplified to the one-way analysis of variance form

$$Q_S = \frac{\frac{(n - 1)}{n}\sum_{i=1}^{s} n_{i+}(\bar{f}_i - \mu_{\underline{a}})^2}{v_{\underline{a}}} \qquad (22)$$

where $\bar{f}_i = (\sum_{j=1}^{r} a_j n_{ij}/n_{i+})$ is the mean score for the ith group. The statistic Q_S approximately has the chi-square distribution with d.f. = (s − 1) in large-sample situations.

A variety of possible scoring systems are described in Landis et al. (1979) and Koch et al. (1985a). In the analysis of clinical efficacy trials, the scores which may be of most interest are as follows:

1. *Integer scores*, defined as $a_j = j$ for j = 1, 2, . . ., r. In our example, $\underline{a} = (1, 2, 3)'$ and $Q_S = 12.86$. Integer scores are useful when the response levels are discrete counts or ordinal categories which can be viewed as equally spaced. A rationale for their use in more general situations lies in the fact that Q_S can be expressed as a function of

$$(\bar{f}_1 - \mu_{\underline{a}}) = \sum_{j=1}^{r} j\left(\frac{n_{1j}}{n_{1+}} - \frac{n_{+j}}{n}\right) = \frac{1}{n_{1+}} \sum_{k=2}^{r} (N_{1k} - M_{1k}) \qquad (23)$$

where $N_{1k} = \Sigma_{j=k}^{r} n_{1j}$ is the number of patients on test drug with responses at least as favorable as the kth and $M_{1k} = \Sigma_{j=k}^{r} m_{1j}$ is its corresponding expected value under H_0; thus, integer scores enable evaluation of the extent to which the $\{N_{1k}\}$ are consistently larger (or smaller) than the $\{M_{1k}\}$.

2. *Standardized midranks*, defined as

$$a_j = \frac{2[\Sigma_{k=1}^{j} n_{+k}] - n_{+j} + 1}{2(n + 1)} \qquad (24)$$

In our example, $\underline{a}$ = (43/170 = 0.253, 99/170 = 0.582, 141/170 = 0.829)' and Q_S = 12.73. Standardized midranks lie in the interval (0, 1). In situations with no ties in the sense that all $n_{+j} = 1$, they represent expected values of order statistics for the uniform distribution. The use of Q_S with standardized midranks provides the contingency table counterpart of the Wilcoxon rank sum statistic (see Koch and Bhapkar, 1982). An advantage of standardized midranks over integer scores is that they involve no scaling of the response categories other than that implied by their relative ordering. Also, they will be noted in Section V to have favorable power relative to actual midranks [i.e., $(n + 1)a_j$] in situations involving a set of 2 × r tables. In the SAS (1985) procedure FREQ, standardized midrank scores are referred to as modified ridits.

3. *Logrank scores*, defined as

$$a_j = 1 - \sum_{k=1}^{j} \left(\frac{n_{+k}}{\Sigma_{m=k}^{r} n_{+m}}\right) \qquad (25)$$

In our example, $\underline{a}$ = (0.500, 0.167, −0.833)' and Q_S = 12.61. When there are no ties in the sense that all $n_{+j} = 1$, the quantities $\{(1 - a_j)\}$ represent expected values of order statistics from the exponential distribution with unit mean. The use of Q_S with logrank scores is of interest when the data have L-shaped distributions and there is greater interest in treatment differences for higher value response categories than lower value ones; see Koch et al. (1985a, 1985b) for further discussion and examples. In the SAS procedure NPAR1WAY, the Savage statistic is based on scores which are identical to the $\{-a_j\}$ when there are no ties, but on somewhat different quantities when there are ties.

4. *Binary scores*, $a_j = 1$ for some levels of response and $a_j = 0$ for the other levels. The use of the binary scores $\underline{a}$ = (0, 1, 1)' yields a statistic identical to (5), i.e., Q_S = 10.59.

When a study is concerned with comparisons of ordinal response distributions for s treatments which are also ordinally scaled with respect to some factor like dose, potential trends in the mean scores $\{\bar{f}_i\}$ are often of

interest. A class of randomization statistics which are effective for addressing such alternative hypotheses can be constructed in terms of a linear function

$$\bar{f} = \sum_{i=1}^{s} c_i \bar{f}_i \left(\frac{n_{i+}}{n}\right) = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{c_i a_j n_{ij}}{n} \tag{26}$$

where $\underline{c} = (c_1, c_2, \ldots, cs)'$ is a vector of scores which account for the ordinal scaling of the subpopulations. Under the hypothesis H_0,

$$E\{\bar{f} \mid H_0\} = \sum_{i=1}^{s} c_i \left(\frac{n_{i+}}{n}\right) \sum_{j=1}^{r} a_j \left(\frac{n_{+j}}{n}\right) = \mu_{\underline{c}} \mu_{\underline{a}} \tag{27}$$

and

$$\mathrm{Var}\{\bar{f} \mid H_0\} = \left\{ \sum_{i=1}^{s} (c_i - \mu_{\underline{c}})^2 \left(\frac{n_{i+}}{n}\right) \sum_{j=1}^{r} \frac{(a_j - \mu_{\underline{a}})^2 (n_{+j}/n)}{(n-1)} \right\} \tag{28}$$

$$= \frac{v_{\underline{c}} v_{\underline{a}}}{(n-1)}$$

The linear function $\bar{f}$ approximately has a normal distribution in large samples, and so for these situations,

$$Q_{CS} = \frac{[\bar{f} - E\{\bar{f} \mid H_0\}]^2}{\mathrm{Var}\{\bar{f} \mid H_0\}} \tag{29}$$

$$= \frac{(n-1)[\Sigma_{i=1}^{s}\Sigma_{j=1}^{r}(c_i - \mu_{\underline{c}})(a_j - \mu_{\underline{a}})n_{ij}]^2}{[\Sigma_{i=1}^{s}(c_i - \mu_{\underline{c}})^2 n_{i+}][\Sigma_{j=1}^{r}(a_j - \mu_{\underline{a}})^2 n_{+j}]}$$

$$= (n-1) r_{ac}^2$$

approximately has the chi-square distribution with d.f. = 1. The statistic r_{ac} represents the correlation coefficient between group scores and response scores, and so Q_{CS} is often called a correlation chi-square statistic. Its use can involve any of the types of scores (1)–(4) for either the groups or the response categories.

Since Q_S and Q_{CS} are based on linear combinations of the $\{n_{ij}\}$, moderate sample sizes are usually sufficient to support chi-square approximations for their distributions (e.g., all $n_{i+} \geq 20$ for Q_S and $n \geq 25$ for Q_{CS}).

Additional discussion of the relative advantages and limitations of randomization chi-square statistics like Q, Q_S, and Q_{CS} is given in Landis et al. (1978) and Koch et al. (1982, 1985a).

In some instances (e.g., several $m_{ij} < 5$) an exact test for $s \times r$ tables may be preferable to a chi-square test. An exact test for $s \times r$ tables may be calculated using the same principle as the Fisher's exact test, but with the probabilities to be summed calculated from the multivariate hypergeometric distribution. For the example, the result for this test was $p = 0.0014$. Algorithms for calculating the exact p values are described in Pagano and Halvorsen (1981) and Mehta and Patel (1983). Also, Mehta et al. (1984) describe an algorithm for obtaining exact p values when the mean score statistic Q_S is used with standardized midranks.

IV. THE MANTEL-HAENSZEL METHOD

In the analysis of efficacy trials, one may wish to compare the treatments by combining the results from a set of strata (e.g., centers in a multicenter trial) and adjusting for the effect of the stratification variable(s). Postrandomization stratification on baseline or other explanatory variables which have at least moderately strong association with response may also be of interest. When the treatment groups are unbalanced by chance with respect to the distribution of important explanatory variables, stratification provides an adjusted framework for undertaking comparisons so that observed differences between the treatment groups can be more clearly interpreted as actually due to treatment.

In the rheumatoid arthritis example, sex has a moderately strong relationship with response (58% of the females had favorable response compared with 32% of the males). Although the treatment groups are only slightly unbalanced with respect to the distribution of sex (66% female in the test drug group versus 74% female in the placebo group), it is still of interest to use sex as a stratification variable in order to conduct an overall test of treatment effectiveness, adjusted for sex.

Let $h = 1, 2$ index the two strata (sexes) in our example, and for simplicity let us consider the response as dichotomous. The data may be presented as follows:

		Improvement		
Sex	Treatment	None	Some or marked	Total
Female	Test drug	$n_{111} = 6$	$n_{112} = 21$	$n_{11+} = 27$
Female	Placebo	$n_{121} = 19$	$n_{122} = 13$	$n_{12+} = 32$
Female total		$n_{1+1} = 25$	$n_{1+2} = 34$	$n_{1} = 59$
Male	Test drug	$n_{211} = 7$	$n_{212} = 7$	$n_{21+} = 14$
Male	Placebo	$n_{221} = 10$	$n_{222} = 1$	$n_{22+} = 11$
Male total		$n_{2+1} = 17$	$n_{2+2} = 8$	$n_{2} = 25$

(30)

Among the females, 78% of the test drug group had a favorable response, as compared to 41% of the placebo group. The males were less responsive to both treatments; 50% of the males in the test drug group had favorable response, as compared to 9% of those in the placebo group.

When a set of 2 × 2 tables arises from stratification by center in a multicenter study, the separate treatment randomizations at each center induce independent hypergeometric distributions for the within-center frequencies, and so the distribution for the full table is the product of these hypergeometric distributions. The same distribution (i.e., the product hypergeometric) applies via conditional distribution arguments to data from a single center where there is postrandomization stratification on demographic or pretreatment variables. Thus, under the null hypothesis H_0 of no treatment difference for each patient of each sex, the expected value of n_{h11} is

$$E\{n_{h11} \mid H_0\} = \frac{n_{h1+}n_{h+1}}{n_h} = m_{h11} \tag{31}$$

and its variance is

$$\text{Var}\{n_{h11} \mid H_0\} = \frac{n_{h1+}n_{h2+}n_{h+1}n_{h+2}}{n_h^2(n_h - 1)} = v_{h11} \tag{32}$$

Note that in the 2 × 2 table for males, two of the m_{hij} are less than 5.0, so a chi-square statistic is not appropriate for testing H_0 for males separately. The two-sided Fisher's exact test p value for males is 0.042, and that for females is 0.008.

A method for evaluating the overall association of treatment and response, adjusted for sex, is the Mantel-Haenszel (1959) statistic:

$$Q_{MH} = \frac{\{\Sigma_{h=1}^{2} n_{h11} - \Sigma_{h=1}^{2} m_{h11}\}^2}{\Sigma_{h=1}^{2} v_{h11}} \tag{33}$$

$$= \frac{\{\Sigma_{h=1}^{2}(n_{h1+}n_{h2+}/n_h)(p_{h11} - p_{h21})\}^2}{\Sigma_{h=1}^{2} v_{h11}}$$

$$= 12.59$$

where $p_{hi1} = (n_{hi1}/n_{hi+})$ is the proportion of patients of the hth sex who received the ith treatment and had no improvement. Even when the within-strata sample sizes are small, the Mantel-Haenszel statistic approximately has the chi-square distribution with d.f. = 1 as long as the combined strata sample sizes,

$$n_{+1+} = \sum_{h=1}^{2}\sum_{j=1}^{2} n_{h1j} = 41 \quad \text{and} \quad n_{+2+} = \sum_{h=1}^{2}\sum_{j=1}^{2} n_{h2j} = 43 \tag{34}$$

are sufficiently large. Thus, the result $Q_{MH} = 12.59$ is interpreted as significant with $p < 0.01$. Here, it is appropriate to note that some references (e.g., Fleiss, 1981; Breslow and Day, 1980) apply a continuity correction for Q_{MH} to improve the quality of the chi-square approximation.

Mantel and Fleiss (1980) proposed the following criterion for considering the chi-square approximation suitable for the distribution of the Mantel-Haenszel statistic in the general setting of q strata:

$$\min\left\{\left[\sum_{h=1}^{q} m_{h11} - \sum_{h=1}^{q} (n_{h11})_L\right], \left[\sum_{h=1}^{q} (n_{h11})_U - \sum_{h=1}^{q} m_{h11}\right]\right\} \geq 5 \tag{35}$$

where $(n_{h11})_L = \max(0, n_{h1+} - n_{h+2})$ and $(n_{h11})_U = \min(n_{h+1}, n_{h1+})$ are respectively the lowest and highest possible values for n_{h11} across all possible randomizations with the marginal frequencies $\{n_{hi+}\}$ and $\{n_{h+j}\}$ fixed. Thus, the criterion specifies that the across-strata sum of expected values for a particular cell should be at least 5.0 from both the minimum possible sum and the maximum possible sum of observed values. Any of the four cells may be used in the calculations. In our example,

$$\sum_{h=1}^{2} m_{h11} = 11.4 + 9.5 = 20.9$$

$$\sum_{h=1}^{2} (n_{h11})_L = 0 + 6 = 6 \tag{36}$$

and

$$\sum_{h=1}^{2} (n_{h11})_U = 25 + 14 = 39$$

Since $(20.9 - 6) \geq 5$ and $(39 - 20.9) \geq 5$, the Mantel-Fleiss criterion is satisfied and the use of the chi-square distribution with d.f. = 1 for the Mantel-Haenszel statistic is appropriate for this example. When the Mantel-Fleiss criterion is not met for a set of 2×2 tables, an exact test may be carried out using the algorithm of Thomas (1975). The one-sided p value provided by this method is $p = 0.0003$. A computationally more efficient algorithm for exact inference in sets of 2×2 tables is given in Mehta et al. (1985).

The Mantel-Haenszel statistic Q_{MH} is effective for detecting patterns of treatment differences across the respective strata when there is a strong tendency for the $\{(p_{h11} - p_{h21})\}$ to have the same sign. For this reason, it is sometimes called a test of average partial association in order to distinguish it from the total partial association statistic

$$Q_T = \sum_{h=1}^{q} \frac{(n_{h11} - m_{h11})^2}{v_{h11}} \tag{37}$$

When there are sufficiently large within-stratum sample sizes (e.g., all $m_{hij} \geq 5$), Q_T has approximately the chi-square distribution with d.f. = q. For our example, $Q_T = 12.69$, d.f. = 2, but as noted previously, the sample sizes for males may not be large enough to support the use of the chi-square approximation for Q_T.

Because Q_{MH} algebraically cannot exceed Q_T and because Q_{MH} tends to increase with greater similarity among strata with respect to strength and direction of the association of treatment and response, the difference between Q_T and Q_{MH} has been used as a statistic for assessing the homogeneity of the strata with respect to the association of treatment and response. Given that the sample size conditions stated for Q_T are met, $(Q_T - Q_{MH})$ has approximately a chi-square distribution with d.f. = $(q - 1)$ under H_0. However, this statistic must be interpreted cautiously because H_0 involves a framework of no association rather than one of homogeneous nonnull association. For this reason, $(Q_T - Q_{MH}) = Q_{PH}$ needs to be viewed as a pseudohomogeneity statistic for detecting departures from H_0 which involve substantial variation across the respective strata in the magnitude and direction of treatment differences. For the example, $Q_{PH} = 12.69 - 12.59 = 0.10$ is a relatively small component of Q_T, and so the pattern of treatment differences is sufficiently consistent across strata for the Mantel-Haenszel statistic to provide the principal basis for contradicting H_0. Additional discussion of the interpretation of the pseudohomogeneity statistic Q_{PH} is given in Koch et al. (1985a). In Section VI, an appropriate method for assessing the across-strata homogeneity of nonnull association of treatment and response is presented in the setting of a logistic regression model.

V. EXTENSIONS OF THE MANTEL-HAENSZEL METHOD

Mantel (1963) proposed an extension of the Mantel-Haenszel procedure to the analysis of a set of $(2 \times r)$ tables with ordinal response categories. This method involves a combination of the principles underlying the mean score statistic Q_S in (19) and the Mantel-Haenszel statistic Q_{MH} in (33). Its application to the rheumatoid arthritis data is described here.

After stratification by sex, the frequencies for the cross-classification of treatment with the ordinal response variable are as follows:

		Improvement			
Sex	Treatment	None	Some	Marked	Total
Female	Test drug	6	5	16	27
Female	Placebo	19	7	6	32
Female total		25	12	22	59

Sex	Treatment	Improvement			Total
		None	Some	Marked	
Male	Test drug	7	2	5	14
Male	Placebo	10	0	1	11
Male total		17	2	6	25

(38)

Let n_{hij} denote the number of patients of the hth sex (h = 1, 2) who received the ith treatment (i = 1, 2) and had the jth response (j = 1, 2, 3). Under the null hypothesis of no treatment difference for each patient, the applicable probability model is the product multivariate hypergeometric distribution:

$$P\{n_{hij} \mid H_0\} = \prod_{h=1}^{2} \frac{[\Pi_{i=1}^{2} n_{hi+}! \Pi_{j=1}^{3} n_{h+j}!]}{[n_h! \Pi_{i=1}^{2} \Pi_{j=1}^{3} n_{hij}!]} \tag{39}$$

Let $\{a_{hj}\}$ be a set of scores for the response levels in the hth stratum. The sum of across-strata scores for the test drug group is

$$f_{+1+} = \sum_{h=1}^{2} \sum_{j=1}^{3} a_{hj} n_{hlj} = \sum_{h=1}^{2} n_{h1+} \bar{f}_{h1} \tag{40}$$

where $\bar{f}_{h1} = \Sigma_{j=1}^{3}(a_{hj} n_{hlj}/n_{hl+})$ is the mean score for the test drug group in the hth stratum. Under H_0, f_{+1+} has expected value

$$E\{f_{+1+} \mid H_0\} = \sum_{h=1}^{2} n_{h1+} \mu_h = \mu_* \tag{41}$$

and variance

$$\text{Var}\{f_{+1+} \mid H_0\} = \sum_{h=1}^{2} \frac{n_{h1+}(n_h - n_{h1+})}{(n_h - 1)} v_h = v_* \tag{42}$$

where $\mu_h = \Sigma_{j=1}^{3}(a_{hj} n_{h+j}/n_h)$ and $v_h = \Sigma_{j=1}^{3}(a_{hj} - \mu_h)^2 (n_{h+j}/n_h)$ are the finite subpopulation mean and variance of scores for the hth stratum.

When the across-strata sample sizes $\{n_{+1+}\}$ are large, f_{+1+} approximately has a normal distribution by central limit theory, and so

$$Q_{EMH} = \frac{\{f_{+1+} - \mu_*\}^2}{v_*} \tag{43}$$

approximately has a chi-square distribution with d.f. = (s − 1) = 1. In the case of two treatments, Q_{EMH} can be shown to be a linear function of the differences in the mean scores of the two treatments for the respective strata, i.e.,

$$Q_{EMH} = \frac{[\Sigma_{h=1}^{2} n_{h1+}(\bar{f}_{h1} - \mu_h)]^2}{[\Sigma_{h=1}^{2} n_{h1+}n_{h2+}v_h/(n_h - 1)]}$$

$$= \frac{[\Sigma_{h=1}^{2}(n_{h1+}n_{h2+}/n_h)(\bar{f}_{h1} - \bar{f}_{h2})]^2}{\Sigma_{h=1}^{2}(n_{h1+}n_{h2+}/n_h)^2\bar{v}_h} \tag{44}$$

where the $\{\bar{v}_h = (1/n_{h1+} + 1/n_{h2+})[n_h/(n_h - 1)]v_h\}$ are the variances of the mean score differences $\{(\bar{f}_{h1} - \bar{f}_{h2})\}$ for the respective strata. From the structure of (44), it is apparent that Q_{EMH} is effective for detecting consistent patterns of treatment differences where the $(\bar{f}_{h1} - \bar{f}_{h2})$ strongly tend to have the same sign.

In our example, the results of the extended Mantel-Haneszel procedure for the scores defined in 1-4 of Section III were as follows:

1. Integer scores: $Q_{EMH} = 14.63$
2. Within-stratum standardized midrank scores: $Q_{EMH} = 15.00$
3. Within-stratum logrank scores: $Q_{EMH} = 13.89$
4. Binary scores for at least some improvement: $Q_{EMH} = 12.59$

Relative to the chi-square distribution with d.f. = 1, all of these results are significant with $p < 0.01$. It is of interest to note that Q_{EMH} with standardized midranks represents the categorical data counterpart of the procedure proposed by van Elteren (1960) for combining Wilcoxon rank sum tests across a set of strata. As discussed in Lehmann (1975), the van Elteren statistic has a locally most powerful property for continuous data situations where treatment effects are similar for the respective strata.

Extensions of the Mantel-Haenszel procedure to sets of (s × r) tables can also be constructed. These have the general form

$$Q_{EMH} = \left\{\sum_{h=1}^{q} (\underline{n}_h - \underline{m}_h)'\underline{A}'_h\right\}\left\{\sum_{h=1}^{q} \underline{A}_h\underline{V}_h\underline{A}'_h\right\}^{-1}\left\{\sum_{h=1}^{q} \underline{A}_h(\underline{n}_h - \underline{m}_h)\right\} \tag{45}$$

where $\underline{n}_h$, $\underline{m}_h$, and $\underline{V}_h$ are the observed frequency vector, the expected frequency vector, and the covariance matrix for the hth stratum and have definitions analogous to $\underline{n}$, $\underline{m}$, and $\underline{V}$ in (14); the $\{\underline{A}_h\}$ are (u × sr) matrices with full rank $u \leq (s - 1)(r - 1)$ and structure such that

$\{\Sigma_{h=1}^{q} \underline{A}_h \underline{V}_h \underline{A}'_h\}$ is nonsingular. For purposes of interpretation, the $\{\underline{A}_h\}$ specify the linear functions of the $\{(\underline{n}_h - \underline{m}_h)\}$ at which the test statistic is directed. Alternative choices of the $\{\underline{A}_h\}$ along the lines discussed in Section III provide the stratification-adjusted counterparts to the overall randomization chi-square statistic Q, the mean score statistic Q_S, and the correlation statistic Q_{CS}. These test statistics can be computed with the SAS (1985) procedure FREQ or the program PARCAT documented in Landis et al. (1979).

It is also possible to construct counterparts to the total partial association statistic Q_T and the pseudohomogeneity statistic Q_{PH} for sets of s × r tables. Issues concerning their application are similar to those outlined in Section IV for sets of 2 × 2 tables. For additional discussion and references concerning the construction and properties of alternative types of randomization-based test statistics for sets of s × r tables, see Landis et al. (1978) and Koch et al. (1982, 1985a).

VI. LOGISTIC REGRESSION

Logistic regression can be used to describe the relationship between a dichotomous response variable and a set of explanatory variables (e.g., treatment, age, sex, pretreatment status). Such analysis is of interest for the evaluation of treatment effects as well as interaction effects of treatment with other explanatory variables. Confirmation of no interaction between treatment and any particular explanatory variable is desirable because treatment effects can then be interpreted as homogeneous across the levels of that variable, thus supporting the generalizability of treatment effects.

Let us return to the data structure of Section IV as shown in Table 3. If we assume that the patients of each sex are statistically representative of some larger target population (i.e., the sample of patients enrolled is equivalent to a stratified simple random sample), then the overall sex-by-treatment-by-response cross-classification has the product binomial distribution:

$$P\{n_{hij}\} = \prod_{h=1}^{2} \prod_{i=1}^{2} \frac{n_{hi+}!}{(n_{hi1}! n_{hi2}!)} (1 - \theta_{hi})^{n_{hi1}} (\theta_{hi})^{n_{hi2}} \tag{46}$$

where θ_{hi} is the probability that a patient of sex h who receives treatment i will have some or marked improvement and n_{hi1} and n_{hi2} are the numbers of patients of the hth sex and ith treatment who had no improvement or improvement, respectively. The logistic model for describing the variation among the $\{\theta_{hi}\}$ has the specification:

$$\theta_{hi} = \{1 + \exp[-(\alpha + \underline{x}'_{hi}\underline{\beta})]\}^{-1} \tag{47}$$

where α is the intercept parameter, $\underline{\beta}$ is a vector of regression parameters, and $\underline{x}'_{hi}$ is a row vector of explanatory variables corresponding to the (h, i)th subpopulation. If we let $\underline{x}'_{hiA} = [1, \underline{x}'_{hi}]$, then the $\underline{x}'_{hiA}$ are the rows of a model specification matrix $\underline{X}_A$. An important property of the

Table 3 Observed Frequencies, Observed Percentages, and Predicted Percentages from Logistic Regression Analysis of Rheumatoid Arthritis Data

		Observed frequencies		Percentage of patients with some or marked improvement	
Sex	Treatment	No improvement	Some or marked improvement	Observed % (s.e.)	Model-predicted % (s.e.)
Female	Test drug	6	21	77.8 (8.0)	79.4 (7.2)
	Placebo	19	13	40.6 (8.7)	39.3 (8.2)
Male	Test drug	7	7	50.0 (13.4)	47.0 (11.9)
	Placebo	10	1	9.1 (8.7)	13.0 (6.8)

logistic model is that all possible values of $(\alpha + \underline{x}'_{hi}\underline{\beta})$ in $(-\infty, +\infty)$ correspond to values of θ_{hi} in (0, 1). With a logistic transformation, the model becomes

$$\text{logit}(\theta_{hi}) = \log_e \frac{\theta_{hi}}{(1 - \theta_{hi})} = \alpha + \underline{x}'_{hi}\underline{\beta} \tag{48}$$

The logit of θ_{hi} is thus the logarithm of the odds of some or marked improvement to no improvement for the (h, i)th subpopulation.

The parameters α and $\underline{\beta}$ are usually estimated by maximum likelihood. By replacing θ_{hi} in the likelihood equation (46) with its model expression in (47), then differentiating $\log_e$ of (46) with respect to α and $\underline{\beta}$, and setting the results to 0, one obtains the following equation:

$$\sum_{h=1}^{2}\sum_{i=1}^{2}(n_{hi1} - n_{hi+}\hat{\theta}_{hi})\underline{x}'_{hiA} = 0 \tag{49}$$

where $\hat{\theta}_{hi} = \{1 + \exp[-(\hat{\alpha} + \underline{x}'_{hi}\hat{\underline{\beta}})]\}^{-1}$ is the model predicted maximum likelihood estimate of θ_{hi} based on $\underline{x}_{hi}$, and $\hat{\alpha}$ and $\hat{\underline{\beta}}$ are the maximum likelihood estimates of α and $\underline{\beta}$. Since this equation is nonlinear, $\hat{\alpha}$ and $\hat{\underline{\beta}}$ must usually be calculated with an iterative procedure such as the Newton-Raphson method [see Cox (1970), Koch and Edwards (1985), or McCullagh and Nelder (1983) for further discussion]. The estimates $\hat{\alpha}$ and $\hat{\underline{\beta}}$ are approximately normal when $\Sigma_{h=1}^{2}\Sigma_{i=1}^{2}n_{hi1}\underline{x}'_{hiA}$ is approximately normal. If we let $\hat{\underline{\beta}}_A = (\hat{\alpha}, \hat{\underline{\beta}}')'$, then a consistent estimate for the covariance matrix of $\hat{\underline{\beta}}_A$ is

$$\underline{V}(\hat{\underline{\beta}}_A) = \left[\sum_{h=1}^{2}\sum_{i=1}^{2}\{n_{hi+}\hat{\theta}_{hi}(1 - \hat{\theta}_{hi})\underline{x}_{hiA}\underline{x}'_{hiA}\}\right]^{-1} \tag{50}$$

A model of interest for the data in Table 3 is the main effects model

$$\underline{X}_A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \tag{51}$$

for which the parameter estimates, their standard errors, and p values for tests of zero values are as follows:

Parameter	Estimate	Standard error	p value	Interpretation
α	−1.904	0.598	<0.010	Log_e odds of improvement for males receiving placebo

Parameter	Estimate	Standard error	p value	Interpretation
β_1	1.469	0.576	0.011	Increment in $\log_e$ odds due to female sex
β_2	1.782	0.519	<0.010	Increment to $\log_e$ odds due to test drug

(52)

The p values are obtained from the Wald statistics based on the squares of the ratios of the parameter estimates to their estimated standard errors, i.e., $\{\text{est.}/\text{se(est.)}\}^2$. These test statistics approximately have chi-square distributions with d.f. = 1.

The goodness of fit of model $\underline{X}_A$ can be evaluated either by the Pearson chi-square statistic

$$Q_P = \sum_{h=1}^{2}\sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{hij} - \hat{m}_{hij})^2}{\hat{m}_{hij}} = 0.26 \qquad (53)$$

where $\hat{m}_{hi1} = n_{hi+}(1 - \hat{\theta}_{hi})$ and $\hat{m}_{hi2} = n_{hi+}\hat{\theta}_{hi}$, or by the log-likelihood ratio chi-square statistic

$$Q_L = \sum_{h=1}^{2}\sum_{i=1}^{2}\sum_{j=1}^{2} 2n_{hij} \log_e \frac{n_{hij}}{\hat{m}_{hij}} = 0.28 \qquad (54)$$

[in which, for $n_{hij} = 0$, $n_{hij} \log_e(n_{hij}/\hat{m}_{hij})$ is defined to be 0]. For large-sample situations (e.g., all $m_{hij} \geq 5$), the statistics Q_P and Q_L are asymptotically equivalent, and both approximately have chi-square distributions with d.f. = (qs − t) = 1, where q = 2 is the number of sexes, s = 2 the number of treatments, and t = 3 the rank of $\underline{X}_A$.

A third equivalent way to test the goodness of fit of model $\underline{X}_A$ is to fit an expanded model which includes a parameter for sex-by-treatment interaction and then to verify that the Wald statistic for this parameter is nonsignificant. If we fit the model $[\underline{X}_A, \underline{W}]$ where $\underline{W} = (1\ 0\ 0\ 0)'$, the Wald statistic for this interaction term is $Q = 0.26$. All three of the statistics Q_P, Q_L, and Q_W are nonsignificant with $p > 0.50$, so the goodness of fit of the model is well supported. In this regard, one can note that numerical studies (e.g., Larntz, 1978) tend to support the use of Q_P because of the applicability of chi-square approximations to its distribution for many types of situations with small or moderate sample sizes (e.g., most $\hat{m}_{hij} > 2$ and few < 1).

Values of the logits predicted by model $\underline{X}_A$ for each of the four subpopulations are the respective elements of $\underline{X}_A\underline{\hat{\beta}}_A$, as shown below. Their estimated standard errors are the square roots of the diagonal elements of $\underline{X}_A\underline{V}(\underline{\hat{\beta}}_A)\underline{X}'_A$. The predicted odds of improvement for each subpopulation are found from exponentiating the predicted logit.

		Estimates from model $\underline{X}_A$		
Sex	Treatment	logit	s.e. (logit)	Odds of improvement
F	Test drug	$\hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 = 1.347$	0.437	exp(1.347) = 3.846
F	Placebo	$\hat{\alpha} + \hat{\beta}_1 = -0.435$	0.345	exp(−0.435) = 0.647
M	Test drug	$\hat{\alpha} + \hat{\beta}_2 = -0.122$	0.480	exp(−0.122) = 0.885
M	Placebo	$\hat{\alpha} = -1.904$	0.598	exp(−1.904) = 0.149

(55)

From the model-estimated odds of improvement, one can calculate the model-estimated probability of improvement. For example, the estimated probability of improvement for females receiving test drug ($\hat{\theta}_{11}$) is found from $\hat{\theta}_{11}/(1 - \hat{\theta}_{11}) = 3.846$, i.e., $\hat{\theta}_{11} = 0.794$. The standard error of $\hat{\theta}_{11}$ is found from $(\hat{\theta}_{11})(1 - \hat{\theta}_{11})[\text{s.e.}\ (\underline{x}'_{11A}\hat{\underline{\beta}}_A)] = (0.794)(0.206)(0.437) = 0.072$. The estimated probabilities of improvement for the four groups are shown alongside the observed probabilities in Table 3. Note that the standard errors for the model-estimated probabilities are somewhat smaller than those for the observed proportions. This gain in precision is due to the elimination of extraneous variability for effects not in the model $\underline{X}_A$ (i.e., the variability due to sex-by-treatment interaction). The results from the maximum likelihood logistic regression analysis of this example were calculated with the SAS (1985) procedure CATMOD; they could also be obtained with the SAS procedure LOGIST documented in Harrell (1986) or the BMDP procedure PLR documented in Engelman (1983).

Since the fit of the model $\underline{X}_A$ in (51) is supported by the nonsignificance of Q_P, Q_L, and the Wald statistic Q_W for the interaction term in the model $[\underline{X}_A, \underline{W}]$, the association of treatment and response can be interpreted as being relatively similar for the two sexes. A measure of the association between treatment and response which can be estimated directly from the logistic model is the odds ratio, i.e., the ratio of the odds of improvement for the test drug to the odds of improvement for the placebo treatment. Since $\underline{X}_A$ is a main effects model, the estimated odds ratio for males, $\exp(\hat{\alpha} + \hat{\beta}_2)/\exp(\hat{\alpha}) = \exp(\hat{\beta}_2)$, is equal to the estimated odds ratio for remales, $\exp(\hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2)/\exp(\hat{\alpha} + \hat{\beta}_1) = \exp(\hat{\beta}_2) = \exp(1.782) = 5.94$. A 95% normal approximation confidence interval for the common odds ratio, $\exp(\beta_2)$, is

$$\exp\{\hat{\beta}_2 \pm 1.96[\text{s.e.}(\hat{\beta}_2)]\} = (2.15,\ 16.43) \qquad (56)$$

An exact confidence interval for the common odds ratio can be obtained through the product noncentral hypergeometric distribution (see Breslow and Day, 1980; Gart, 1971; Kleinbaum et al., 1982) from the computing procedure of Thomas (1975). The exact 95% confidence interval obtained by this method is (1.94, 18.78). A more efficient algorithm for obtaining this type of exact confidence interval is described in Mehta et al. (1985).

Logistic regression, as previously described, can be generalized to include continuous as well as categorical explanatory variables. The main difference between a logistic model with a small number of categorical explanatory variables and a model with both categorical and continuous explanatory variables is in the evaluation of goodness of fit. For example, in the previously described model, there were four subpopulations corresponding to the cross-classification of sex and treatment. If we expand this model to include age also, so that the number of subpopulations corresponds to the cross-classification of sex by treatment by age, then methods [such as the Pearson chi-square (53) or the log-likelihood ratio chi-square (54)] which depend on having a sufficiently large number of patients per subpopulation (e.g., $n_{hi+} \geq 10$) are no longer applicable.

The strategy of fitting an expanded model and then verifying the nonsignificance of effects not in the original model is still applicable, however. If the specification matrix $\underline{X}_A$ for the original model has rank t, then the expanded model $[\underline{X}_A, \underline{W}]$ must have rank t + w, where w is the rank of $\underline{W}$. The significance of the contribution of $\underline{W}$ may be evaluated by the difference of the log-likelihood ratio chi-square statistics for the models $\underline{X}_A$ and $[\underline{X}_A, \underline{W}]$, i.e.,

$$Q_{LR} = \sum_{i=1}^{s} \sum_{j=1}^{2} 2n_{ij} \left[\log_e \left(\frac{\hat{m}_{ij,w}}{\hat{m}_{ij}} \right) \right] \tag{57}$$

where s is the total number of subpopulations with at least one subject for the cross-classification of explanatory variables, n_{ij} the number of patients in the ith subpopulation with the jth response, $\hat{m}_{ij}$ the predicted value of n_{ij} for model $\underline{X}_A$, and $\hat{m}_{ij,w}$ the predicted value of n_{ij} for model $[\underline{X}_A, \underline{W}]$. Q_{LR} has an approximate chi-square distribution with d.f. = w.

An alternative statistic which does not require fitting the expanded model is the Rao score statistic for assessing the association of the residuals $(\underline{n}_{*1} - \hat{\underline{m}}_{*1})$ with $\underline{W}$ through the linear functions $\underline{g} = \underline{W}'(\underline{n}_{*1} - \hat{\underline{m}}_{*1})$. A computational expression for this criterion is

$$Q_{RS} = \underline{g}' \{ \underline{W}' [\underline{D}_{\hat{\underline{v}}}^{-1} - \underline{D}_{\hat{\underline{v}}}^{-1} \underline{X}_A (\underline{X}'_A \underline{D}_{\hat{\underline{v}}}^{-1} \underline{X}_A)^{-1} \underline{X}'_A \underline{D}_{\hat{\underline{v}}}^{-1}] \underline{W} \}^{-1} \underline{g} \tag{58}$$

where $\underline{n}_{*1} = (n_{11}, n_{21}, \ldots, n_{s1})'$, $\hat{\underline{m}}_{*1} = (\hat{m}_{11}, \hat{m}_{21}, \ldots, \hat{m}_{s1})'$, and $D_{\hat{v}}$ is a diagonal matrix with diagonal elements $\hat{v}_i = [n_{i+}\hat{\theta}_i(1 - \hat{\theta}_i)]^{-1}$. Both Q_{LR} and Q_{RS} approximately have the chi-square distribution with d.f. = w when the overall sample size is sufficiently large to support an approximately multivariate normal distribution for the linear functions $[\underline{X}'_A, \underline{W}']\underline{n}_{*1}$; see Imrey et al. (1981, 1982) and Koch et al. (1985a) for further discussion.

A logistic regression model with explanatory variables of sex, treatment, and age was fit to the data of Table 2 with the SAS procedure LOGIST of Harrell (1986). The estimated model parameters and their standard errors were as follows:

Parameter	Estimate	Standard error	p value	Interpretation
α	−4.503	1.307	<0.010	Log_e odds of improvement for a hypothetical male of age 0 who received placebo
β_1	1.488	0.595	0.012	Increment due to female sex
β_2	1.760	0.536	<0.010	Increment due to test drug
β_3	0.049	0.021	0.018	Increment per year of age

(59)

Thus, sex, treatment, and age are all significant predictors of improvement. It is important to note that the model is useful for predicting the relationship between age and response only for patients with ages similar to the ages of patients in the study (i.e., ages 23 to 74). Also, the model assumes a linear relationship between age and response, which may not be the case. One way of assessing potential departures from linearity is through the contribution of the square of age to the model.

The goodness of fit of the model was tested by evaluating the potential contribution of the following four additional variables: the sex × treatment interaction, the age × treatment interaction, the sex × age interaction, and the square of age (where interaction variables are defined as products of their components). The Rao score statistic (58) for the joint contribution of these four variables was $Q_{RS} = 4.03$, which is nonsignificant when compared to the chi-square distribution with d.f. = 4 ($p = 0.402$). However, when attention was directed at the individual components of this overall test, the sex × age interaction was found to be nearly significant ($Q_{RS} = 3.69$, d.f. = 1, $p = 0.055$). This result can be interpreted either as a chance event for the situation with no sex × age interaction or as an indicator that the model needs to be expanded to include the sex × age interaction. The former interpretation is supported by the sex × age interaction not seeming strongly evident in view of the multiplicity of goodness-of-fit assessments involved in the overall Q_{RS} and its separate components. Alternatively, the latter interpretation is compatible with the theme of confirming conclusions about treatment effects by considering settings which account for all other possibly relevant sources of variation. When the expanded model with age, sex, age × sex interaction, and treatment effects was applied, the test drug was still found to be associated with significantly ($p < 0.010$) more favorable response, and (treatment × sex) and (treatment × age) interactions were found to be nonsignificant. The sex × age interaction corresponded to the disparity between a significant linear relationship with age for females and an essentially null relationship for males. For additional discussion and references concerning logistic regression methods, see Breslow and Day (1980), Cox (1970), Kleinbaum et al. (1982),

Koch et al. (1982, 1985a), Koch and Edwards (1985), and McCullagh and Nelder (1983).

VII. EXTENSIONS OF LOGISTIC REGRESSION

For ordinal variables with three levels (e.g., no improvement, some, or marked), two logits for more favorable versus less favorable response can be constructed:

1. The logit for some or marked improvement versus no improvement. If, for the ith subpopulation, we let π_{i1} denote the probability of no improvement, π_{i2} denote the probability of some improvement, and π_{i3} denote the probability of marked improvement, this logit can be expressed as $\log_e\{(\pi_{i2} + \pi_{i3})/(\pi_{i1})\} = \text{logit}(\theta_{i1})$. This is the logit considered in Section VI.
2. The logit for marked improvement versus some or no improvement, i.e., $\log_e\{(\pi_{i3})/(\pi_{i1} + \pi_{i2})\} = \text{logit}(\theta_{i2})$.

The assumed probability structure for the frequencies $\{n_{ij}\}$ of response outcomes $j = 1, 2, \ldots, r$ (here, $r = 3$) for the samples of $\{n_{i+}\}$ subjects from the $i = 1, 2, \ldots, s$ subpopulations is the product multinomial distribution:

$$P\{n_{ij}\} = \prod_{i=1}^{s}\left[n_{i+}! \prod_{j=1}^{r} \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!}\right] \tag{60}$$

As noted in Section VI, a supportive basis for the structure (60) is the perspective that the subjects under study are conceptually representative of the respective subpopulations in a sense equivalent to stratified simple random sampling.

We can consider a model for both of the logits in 1 and 2 simultaneously with the specification

$$\text{logit}(\theta_{ik}) = \alpha_k + \underline{x}_i'\underline{\beta}_k \tag{61}$$

where $k = 1, 2$ indexes the two logits. With this model, there are separate intercept parameters $\{\alpha_k\}$ and vectors of regression parameters $\{\underline{\beta}_k\}$ for the two types of logits. This model does not account for the ordinal nature of the response because comparisons between the ith and i'th subpopulations with respect to it have $(r - 1) = 2$ components

$$\{\text{logit}(\theta_{ik}) - \text{logit}(\theta_{i'k})\} = (\underline{x}_i - \underline{x}_{i'})'\underline{\beta}_k \qquad \text{for } k = 1, 2 \tag{62}$$

which express general association. However, under the condition that the respective logits have the same regression parameter vector $\underline{\beta}$ (i.e., all $\underline{\beta}_k = \underline{\beta}$), the structure (62) simplifies to

$$\{\text{logit}(\theta_{ik}) - \text{logit}(\theta_{i'k})\} = (\underline{x}_i - \underline{x}_{i'})'\underline{\beta} \tag{63}$$

which involves only one component. Since $\exp\{(\underline{x}_i - \underline{x}_{i'})'\underline{\beta}\}$ expresses the extent to which more favorable response categories are more likely in the ith subpopulation than the i'th subpopulation, it is potentially indicative of location shifts. Thus, the corresponding model

$$\text{logit}(\theta_{ik}) = \alpha_k + \underline{x}_i'\underline{\beta} \tag{64}$$

provides a framework which accounts for ordinal response categories; it is called the proportional odds model by McCullagh (1980) as well as other authors. Maximum likelihood estimation for the parameters of the proportional odds model (64) is discussed in McCullagh (1980), McCullagh and Nelder (1983), and Walker and Duncan (1967). Computations can be readily undertaken with the SAS procedure LOGIST of Harrell (1986).

For the rheumatoid arthritis data, the model with sex and treatment as explanatory variables in a form analogous to the last two columns of (51) was fit. The resulting maximum likelihood estimates, their standard errors, and p values were as follows:

Parameter	Estimate	Standard error	p value	Interpretation
α_1	−1.813	0.565	<0.010	Log_e odds of some or marked improvement versus no improvement for males receiving placebo
α_2	−2.667	0.606	<0.010	Log_e odds of marked improvement versus some or no improvement for males receiving placebo
β_1	1.319	0.538	0.014	Increment for both types of $\log_e$ odds due to female sex
β_2	1.797	0.472	<0.010	Increment for both types of $\log_e$ odds due to test drug

(65)

For situations with categorical explanatory variables and contingency table data structure, like that shown in Section V for this example, the goodness of fit of the model (64) can be evaluated with counterparts to the Pearson chi-square Q_P in (53) or the log-likelihood ratio chi-square Q_L in (54). The model-predicted frequencies for these methods are obtained as follows:

$$\hat{m}_{i1} = n_{i+}[1 + \exp(\hat{\alpha}_1 + \underline{x}_i'\hat{\underline{\beta}})]^{-1} \tag{66}$$

$$\hat{m}_{ij} = n_{i+}\{[1 + \exp(\hat{\alpha}_j + \underline{x}_i'\hat{\underline{\beta}})]^{-1} - [1 + \exp(\hat{\alpha}_{(j-1)} + \underline{x}_i'\hat{\underline{\beta}})]^{-1}\}$$

$$\text{for } j = 2, \ldots, (r-1)$$

$$\hat{m}_{ir} = n_{i+}[1 + \exp(-\hat{\alpha}_{(r-1)} - \underline{x}_i'\hat{\underline{\beta}})]^{-1}$$

When the sample sizes for each subpopulation are sufficiently large (e.g., most $\hat{m}_{ij} > 2$ and few < 1), Q_P approximately has the chi-square distribution with d.f. = $\{(r-1)s - t\}$, where t is the dimension of $\underline{\beta}_A = (\alpha_1, \alpha_2, \underline{\beta}')'$. The result $Q_P = 4.80$ supports the use of the model illustrated here by its nonsignificance ($p = 0.308$) with respect to the chi-square distribution with d.f. = 4.

Maximum likelihood estimation for the parameters of the proportional odds model is also applicable to situations with categorical and continuous explanatory variables. An example of such analysis is the fit of the model with age, sex, and treatment explanatory variables to the rheumatoid arthritis data in Table 2. Results from the SAS procedure LOGIST of Harrell (1986) are as follows:

Parameter	Estimate	Standard error	p value	Interpretation
α_1	−3.784	1.144	<0.010	Log_e odds of some or marked improvement versus no improvement for males of age o receiving placebo
α_2	−4.683	1.187	<0.010	Log_e odds of marked improvement versus some or no improvement for males of age o receiving placebo
β_1	1.252	0.546	0.022	Increment for both types of log_e odds due to female sex
β_2	1.745	0.476	<0.010	Increment for both types of log_e odds due to test drug
β_3	0.038	0.018	0.038	Increment for both types of log_e odds per year of age

(67)

As discussed for logistic models in Section V, the evaluation of goodness of fit of proportional odds models with categorical and continuous variables needs to be undertaken in terms of the potential contribution of additional explanatory variables. For this purpose, either extensions of the log-likelihood ratio chi-square statistic (57) or the Rao score statistic (58) can be used. For the model expansion consisting of sex × treatment interaction, age × treatment interaction, sex × age interaction, and the square of age, the Rao score statistic $Q_{RS} = 3.53$ was nonsignificant ($p = 0.473$) relative to the chi-square distribution with d.f. = 4. On this basis, the model with age, sex, and treatment as the explanatory variables is found to be satisfactory. Treatment effects are thereby interpreted as homogeneous across sex and age. The nature of treatment effects is expressed through $\exp(\hat{\underline{\beta}}_2) = \exp(1.745) = 5.73$. This quantity represents the model-predicted ratio of test drug to placebo for both the odds of some or marked improvement versus no improvement and the odds of marked improvement versus some or no improvement in the setting where sex and age are held constant; i.e., it is an odds ratio which describes the multiplicative extent to which more favorable response is more likely for test drug patients than placebo patients. Thus, the odds of some or marked

improvement versus no improvement and the odds of marked improvement versus some or no improvement are 5.73 times greater for test drug patients than placebo patients. Additional discussion of the proportional odds model and methods to assess its goodness of fit are given in Harrell (1986), Koch et al. (1985c), McCullagh and Nelder (1983), and Peterson (1986).

Another extension of the logistic model which is useful for the analysis of data for responses with $r > 2$ levels is the log-linear model with the structure

$$\pi_{ij} = \frac{\exp(\alpha_j + \underline{x}_i'\underline{\beta}_j)}{\{1 + \sum_{j=1}^{(r-1)} \exp(\alpha_j + \underline{x}_i'\underline{\beta}_j)\}} \tag{68}$$

where $j = 1, 2, \ldots, (r-1)$ for a set of explanatory variables $\{\underline{x}_i\}$ for the respective subpopulations; for $j = r$, the model is $\pi_{ir} = 1 - \sum_{j=1}^{(r-1)} \pi_{ij}$. Since (68) implies that

$$\log_e\left\{\frac{\pi_{ij}}{\pi_{ir}}\right\} = \alpha_j + \underline{x}_i'\underline{\beta}_j \tag{69}$$

for $j = 1, 2, \ldots, (r-1)$, the $\{\alpha_j\}$ represent intercept parameters and the $\{\underline{\beta}_j\}$ represent vectors of regression parameters. As was the case for the model (61), the model (68) describes general association between subpopulations and response in the sense that comparisons between the ith and i'th subpopulations have $(r-1)$ components

$$\log_e\left\{\frac{\pi_{ij}\pi_{i'r}}{\pi_{ir}\pi_{i'j}}\right\} = (\underline{x}_i - \underline{x}_{i'})'\underline{\beta}_j \tag{70}$$

for $j = 1, 2, \ldots, (r-1)$. When response outcomes are nominal rather than ordinal, the evaluation of such general association would be of interest. Maximum likelihood methods for estimating model parameters and assessing goodness of fit for the log-linear model (68) in these situations are discussed in such references as Andersen (1980), Bishop et al. (1975), Fienberg (1980), Haberman (1978), and Imrey et al. (1981, 1982). Computations for this type of analysis can be undertaken with the SAS (1985) procedure CATMOD or the BMDP procedure P4F documented in Brown (1983).

A refinement of the log-linear model (68) which can account for ordinal response categories is the additional condition of equality for odds ratios involving adjacent response categories, i.e.,

$$\left(\frac{\pi_{ij}\pi_{i',(j+1)}}{\pi_{i,(j+1)}\pi_{i'j}}\right) = (\underline{x}_i - \underline{x}_{i'})'\underline{\beta} \tag{71}$$

for $j = 1, 2, \ldots, (r-1)$. Since (71) also implies that

$$\log_e\left\{\frac{\pi_{ij}\pi_{i'r}}{\pi_{ir}\pi_{i'j}}\right\} = (r-j)(\underline{x}_i - \underline{x}_{i'})'\underline{\beta} \tag{72}$$

the comparison of the ith and i'th subpopulations with respect to the log-linear model with equal adjacent odds ratio structure has only one component. A specification for this model is

$$\pi_{ij} = \left\{ \frac{\exp\{\alpha_j + (r-j)\underline{x}_i'\underline{\beta}\}}{\{1 + \sum_{j=1}^{(r-1)} \exp\{\alpha_j + (r-j)\underline{x}_i'\underline{\beta}\}} \right\} \quad \text{for } j = 1, 2, \ldots, (r-1)$$

$$\pi_{ir} = 1 - \sum_{j=1}^{(r-1)} \pi_{ij} \qquad (73)$$

for i = 1, 2, . . ., s. Aspects of the application of maximum likelihood methods to the model (73) are similar to those for the general model (68) and are discussed in the references cited previously for it. Results for the rheumatoid arthritis data from the fit of the equal adjacent odds ratio log-linear model with sex and treatment as explanatory variables are as follows:

Parameter	Estimate	Standard error	p value	Interpretation
α_1	2.607	0.707	<0.010	Log_e odds of no improvement versus marked improvement for males receiving placebo
α_2	0.566	0.515	0.271	Log_e odds of some improvement versus marked improvement for males receiving placebo
β_1	−0.741	0.325	0.023	Increment in adjacent odds of less favorable versus more favorable response due to female sex
β_2	−1.076	0.293	<0.010	Increment in adjacent odds of less favorable versus more favorable response due to test drug

(74)

The goodness of fit of this model can be assessed by application of counterparts to the Pearson chi-square statistic Q_P in (53) or the log-likelihood ratio chi-square statistic Q_L in (54) to the contingency table data structure shown in Section V. For these test statistics, the model-predicted frequencies are given by $\hat{m}_{ij} = n_{i+}\hat{\pi}_{ij}$, where $\hat{\pi}_{ij}$ is the maximum likelihood estimate for π_{ij} as obtained from substitution of the maximum likelihood estimates $\{\hat{\alpha}_j\}$ and $\hat{\underline{\beta}}$ for model parameters into (73). Since $Q_P = 2.22$ and $Q_L = 3.36$ are both nonsignificant (p = 0.695 and p = 0.499, respectively) relative to their approximately chi-square distribution for which d.f. = $\{(r-1)s - t\} = 4$, where t is the dimension of $\underline{\beta}_A = (\alpha_1, \alpha_2, \underline{\beta}')'$, the use

of this model is supported. Accordingly, $\exp(\underline{\hat{\beta}}_2) = \exp(-1.076) = 0.341$ expresses the nature of treatment effects in the sense of the extent to which both the odds of no response versus some response and the odds of some response versus marked response are less likely for test drug patients than placebo patients. The corresponding reciprocal, $\exp(-\underline{\hat{\beta}}_2) = 2.93$, expresses the extent to which the odds of more favorable response in adjacent pairs versus less favorable response are greater for test drug patients than placebo patients.

For applications such as the rheumatoid arthritis data analyzed in this chapter, the choice between the proportional odds model and the equal adjacent odds model mainly involves philosophical and computational considerations since both had satisfactory goodness of fit. The proportional odds model has two noteworthy advantages. One is that the pooling of adjacent response categories only influences its specification through reduction of the number of logit functions which it encompasses, whereas such pooling necessitates complete respecification of the equal adjacent odds ratio model due to the creation of a new set of adjacent outcomes. A second advantage of the proportional odds model is the availability of the SAS procedure LOGIST of Harrell (1986), which can be used to implement models with both categorical and continuous explanatory variables. Readily available computing procedures for the equal adjacent odds ratio model are currently limited to situations with categorical explanatory variables. However, the equal adjacent odds ratio model has the important advantage of being in the general class of log-linear models, and therefore the methods for the evaluation of its goodness of fit are more straightforward than those for the proportional odds model. Also, the equal adjacent odds model has somewhat greater flexibility for extensions which account for potential lack of fit. Of course, in some situations, only one of these two models might provide a satisfactory fit, so the choice of model would be based on practical considerations. If neither of these types of models was appropriate, the applicability of other reasonable structures would need to be explored. References which deal with alternative methods for the analysis of ordinal data or provide illustrative examples include Agresti (1983, 1984), Andrich (1979), Clogg (1982), Cox and Chuang (1984), Goodman (1979), Koch et al. (1982, 1985a, 1985c), McCullagh (1980), and McCullagh and Nelder (1983).

VIII. RANK MEASURES OF ASSOCIATION

The application of the extended Mantel-Haenszel statistic to a set of $2 \times r$ tables with ordinal response categories was described in Section V. A relevant consideration for the effectiveness of that method was the consistency across strata for the patients with one treatment to have more favorable responses than those with the other treatment. One way to address this issue of homogeneity of treatment effects is through tests of treatment × strata interaction effects in a statistical model for ordinal data such as the proportional odds model or the equal adjacent odds ratio model discussed in Section VII. A limitation of this approach, however, is that these types of models may not have sufficiently satisfactory goodness of fit for their use to be appropriate. In such situations, an alternative approach involving the across-strata comparison of measures of association between treatment and response is of interest.

One useful rank measure of association between a dichotomous treatment classification and an ordinal response variable for the hth stratum is

$$g_h = \left\{ \sum_{j=1}^{3} p_{h1j} \left[\left(\sum_{k=1}^{j} p_{h2k} \right) - 0.5 p_{h2j} \right] \right\} \tag{75}$$

where $p_{hij} = (n_{hij}/n_{hi+})$ denotes the proportion of subjects with the jth response for the hth stratum and the ith treatment. The function g_h is related to the Mann-Whitney (1947) statistic in the sense of being an hth stratum estimator for the probability ξ_h with which a randomly selected test drug patient has more favorable response than a randomly selected placebo patient when ties are randomly broken with probability 1/2. Since the hypothesis of no treatment effects on the responses of subjects in the hth stratum implies $\xi_h = 0.5$, the values of $(\xi_h - 0.5)$ describe the extent of treatment differences for the respective strata.

Given that the frequencies $\{n_{hij}\}$ have a product multinomial distribution like that shown in (60), a consistent estimator $\underline{V}_{\underline{g}}$ for the covariance matrix of $\underline{g} = (g_1, g_2)'$ can be constructed by linear Taylor series methods. This estimator has the form

$$\underline{V}_{\underline{g}} = [\underline{H}(\underline{p})]\underline{V}_{\underline{p}}[\underline{H}(\underline{p})]' \tag{76}$$

where $\underline{p} = (\underline{p}'_{11}, \underline{p}'_{12}, \underline{p}'_{21}, \underline{p}'_{22})'$ is the compound vector with components $\underline{p}'_{hi} = (p_{hi1}, p_{hi2}, p_{hi3})$ for the respective subpopulations; $\underline{V}_{\underline{p}}$ is the block diagonal matrix with diagonal blocks $[\underline{D}_{\underline{p}_{hi}} - \underline{p}_{hi}\underline{p}'_{hi}]/n_{hi+}$ and corresponds to the estimated covariance matrix of $\underline{p}$; and $\underline{H}(\underline{p}) = [\partial \underline{g}/\partial \underline{p}]$. Also, if $\underline{g}$ is expressed as the compound function

$$\underline{g} = \underline{A}_3\{\underline{\exp}[\underline{A}_2 \underline{\log}_e(\underline{A}_1\underline{p})]\} \tag{77}$$

with $\underline{\exp}$ and $\underline{\log}_e$ being operations that respectively exponentiate and compute natural logarithms of the elements of a vector and

$$\underline{A}_1 = \begin{bmatrix} 1 & 0 & 0 & & & & & & & & & \\ 0 & 1 & 0 & & & & & & & & & \\ 0 & 0 & 1 & & & & & & & & & \\ & & & .5 & 0 & 0 & & & & & & \\ & & & 1 & .5 & 0 & & & & & & \\ & & & 1 & 1 & .5 & & & & & & \\ & & & & & & 1 & 0 & 0 & & & \\ & & & & & & 0 & 1 & 0 & & & \\ & & & & & & 0 & 0 & 1 & & & \\ & & & & & & & & & .5 & 0 & 0 \\ & & & & & & & & & 1 & .5 & 0 \\ & & & & & & & & & 1 & 1 & .5 \end{bmatrix} \tag{78}$$

$$\underline{A}_2 = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & & & & & & \\ 0 & 1 & 0 & 0 & 1 & 0 & & & & & & \\ 0 & 0 & 1 & 0 & 0 & 1 & & & & & & \\ & & & & & & 1 & 0 & 0 & 1 & 0 & 0 \\ & & & & & & 0 & 1 & 0 & 0 & 1 & 0 \\ & & & & & & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \underline{A}_3 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

then $\underline{H}(\underline{p}) = \underline{A}_3\underline{D}_{\underline{a}_2}\underline{A}_2\underline{D}_{\underline{a}_1}{}^{-1}\underline{A}_1$ with $\underline{a}_1 = \underline{A}_1\underline{p}$ and $\underline{a}_2 = \underline{exp}\{\underline{A}_2\ \underline{log}_e\underline{a}_1\}$. Additional discussion concerning the use of compound functions to represent rank measures of association $\underline{g}$ and the construction of their estimated covariance matrix $\underline{V}_{\underline{g}}$ is given in Forthofer and Lehnen (1981), Koch et al. (1982, 1985a), and Semenya et al. (1983).

For the example,

$$\underline{g} = \begin{bmatrix} 0.733 \\ 0.698 \end{bmatrix} \quad \text{and} \quad \underline{V}_{\underline{g}} = \begin{bmatrix} 0.3814 & 0 \\ 0 & 0.6704 \end{bmatrix} \times 10^{-2} \tag{79}$$

As a consequence of central limit theory for Mann-Whitney statistics (see Puri and Sen, 1971), the estimators $\underline{g}$ approximately have a multivariate normal distribution when the sample sizes for the respective (sex × treatment) subpopulations are sufficiently large (e.g., $n_{hi+} \geq 10$). Thus, the variation among the $\{g_h\}$ can be analyzed by weighted least-squares methods and Wald statistics as discussed in Grizzle et al. (1969) and Koch et al. (1977, 1985a). The Wald statistic for comparing the $\{g_h\}$ for the two strata is

$$Q_W = \underline{g}'\underline{W}'[\underline{W}\underline{V}_{\underline{g}}\underline{W}']^{-1}\underline{W}_{\underline{g}} = 0.12 \tag{80}$$

where $\underline{W} = [1, -1]$. This result is nonsignificant with $p = 0.732$ relative to the approximately chi-square distribution with d.f. = 1 for Q_W, and so the two sexes are interpreted as having similar patterns of association between treatment and response relative to the measures $\{\xi_h\}$. On this basis, the $\{g_h\}$ can be described with the linear model

$$\underline{E}_{\underline{A}}\{\underline{g}\} = \underline{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}\beta = \underline{X}\beta \tag{81}$$

where $E_A\{\ \}$ denotes expected value for large samples, $\underline{X}$ is the model specification matrix, and β is the parameter for the common value of the $\{\xi_h\}$ for the two strata under the model. The weighted least-squares estimator b for β and a consistent estimator V_b for its variance are given by

$$b = (\underline{X}'\underline{V}_{\underline{g}}{}^{-1}\underline{X})^{-1}\underline{X}'\underline{V}_{\underline{g}}{}^{-1}\underline{g} = 0.720 \tag{82}$$

and

$$V_b = (\underline{X}'\underline{V}_{\underline{g}}{}^{-1}\underline{X})^{-1} = 0.002431 \tag{83}$$

The goodness of fit of the model specified by $\underline{X}$ is supported by the statistic

$$Q_W = (\underline{g} - \underline{X}b)'\underline{V}_{\underline{g}}{}^{-1}(\underline{g} - \underline{X}b) = 0.12 \tag{84}$$

which approximately has the chi-square distribution with d.f. = (u − t) = 1, where u denotes the number of elements in $\underline{g}$ and t denotes the number of elements in b. Q_W is identical to the Wald statistic shown in (80) when

$\underline{X}$ and $\underline{W}'$ are orthocomplements of one another (i.e., $\underline{W}\underline{X} = \underline{0}_{(u-t),t}$ and $\text{Rank}[\underline{X}, \underline{W}'] = u$).

Relative to the model $\underline{X}$ in (81), the hypothesis of no treatment effects can be expressed as H_0: $\beta = 0.5$. Since b approximately has a normal distribution,

$$Q_b = \frac{(b - 0.5)^2}{V_b} = 19.91 \qquad (85)$$

approximately has the chi-square distribution with d.f. = 1. This result is significant with $p < 0.01$, and so a difference between treatments is evident. The estimator $b = 0.720$ enables this difference to be interpreted as involving 0.72 probability for a randomly selected test drug patient having more favorable response than a randomly selected placebo patient. The computations for obtaining $\underline{g}$ and $\underline{V}_g$ and carrying out weighted least-squares analysis were undertaken with the program GENCAT documented in Landis et al. (1976); with some modifications in the operations for specifying g, the SAS (1985) procedure CATMOD could have been used.

Since the statistic Q_b in (85) is based on the combination of the rank measures of association $\{g_h\}$, it is analogous to the extended Mantel-Haenszel statistic Q_{EMH} with standardized midrank scores in Section V. The main advantage it has over Q_{EMH} is its determination from a framework through which the across-strata homogeneity of measures of association between treatment and response can be evaluated. On the other hand, the application of Q_b is limited to situations where the $\{g_h\}$ are homogeneous because this condition is an assumption of its underlying model (81), whereas Q_{EMH} can be used more broadly. Also, since Q_{EMH} is based on linear statistics, chi-square approximations for its distribution tend to be reasonable for somewhat smaller sample sizes than the methods for the nonlinear statistics considered in this section. Further discussion of analyses of rank measures of association and additional examples illustrating their comparison with other methods are given in Forthofer and Lehnen (1981), Koch et al. (1982, 1985a, 1986b), and Semenya et al. (1983).

IX. A GROUPED SURVIVAL TIMES EXAMPLE

The data in Table 4 are from a multicenter trial to compare a test drug to placebo with respect to time to healing of a gastrointestinal condition. Patients were examined at 2 weeks to determine whether healing had occurred. If healing had not occurred, they were reexamined at 4 weeks. All patients were considered to have completed the study at 4 weeks regardless of whether their condition had been healed.

A variety of useful preliminary analyses can be performed for these data without giving specific attention to their time to healing (or ulcer survival) structure. For example, the Mantel-Haenszel method (Section IV) was applied to each of two binary response variables: healed at 2 weeks versus not healed at 2 weeks, and healed by 4 weeks (i.e., healed at 2 weeks or healed at 4 weeks) versus not healed. There was no significant difference between the treatments with respect to healing by 2 weeks, controlling for center ($Q_{MH} = 1.94$, d.f. = 1, $p = 0.164$). Also, no lack of homogeneity between centers is suggested with respect to this variable

Table 4 Data from a Multicenter Trial for the Comparison of Treatments with Respect to Healing of a Gastrointestinal Condition

		Contingency table format			
Center	Treatment	Number healed at 2 weeks	Number healed at 4 weeks but not at 2 weeks	Number not healed	Total
1	Test drug	15	17	2	34
	Placebo	15	17	7	39
2	Test drug	17	17	10	44
	Placebo	12	13	15	40
3	Test drug	7	17	16	40
	Placebo	3	17	18	38

($Q_{PH} = Q_T - Q_{MH} = 2.51 - 1.94 = 0.57$, d.f. = 2, $p = 0.752$). However, there was a significant treatment difference with respect to healing by 4 weeks, controlling for center ($Q_{MH} = 4.01$, d.f. = 1, $p = 0.045$), although none of the treatment differences at the individual centers was significant at the 0.05 level. The pseudohomogeneity statistic was nonsignificant for this response variable ($Q_{PH} = 0.99$, d.f. = 2, $p = 0.610$).

Logistic regression analyses were also undertaken for each of the previously specified dichotomous response variables in order to compare the results with those from the Mantel-Haenszel method and to obtain a more convincing assessment of the homogeneity of the association between treatment and response at the three centers. Computations were performed with the SAS procedure LOGIST. Two indicator variables were formed to represent the effect of center (1 if center 1, 0 otherwise; and 1 if center 2, 0 otherwise). The treatment variable was coded as 1 if test drug, 0 if placebo. Two indicator variables were also formed to represent center-by-treatment interaction (1 if center 1 and test drug, 0 otherwise; and 1 if center 2 and test drug, 0 otherwise). The indicator variables for center and treatment were forced to be included in the model, and the two interaction indicator variables were considered as candidate variables for entry. The Rao score statistic for testing the combined contribution of the interaction variables was nonsignificant for both response variables, so the models with main effects for center and treatment were considered to have satisfactory goodness of fit.

In the main effects logistic regression model for healing by 2 weeks, the indicator variables for center 1 and center 2 were both significant ($p < 0.01$ for both Wald statistics), and the parameter estimates were both positive, indicating that healing was significantly faster at centers 1 and 2 than at center 3. The Wald statistic for treatment effect was nonsignificant ($Q_W = 1.95$, $p = 0.163$) and was very similar to the corresponding Mantel-Haenszel statistic ($Q_{MH} = 1.94$). In the logistic regression for healing by 4 weeks, the indicator variable for center 1 was significant

($p < 0.001$) and the indicator variable for center 2 was nearly significant ($p = 0.069$), with both parameter estimates being positive. The Wald statistic for treatment effect ($Q_W = 4.01$, $p = 0.045$) was again virtually identical to the corresponding Mantel-Haenszel statistic ($Q_{MH} = 4.01$).

Other possible methods which could be applied to these data include extensions of the Mantel-Haenszel method (Section V) and extensions of logistic regression (Section VII). In these analyses, the three responses (healed at 2 weeks, healed between 2 and 4 weeks, and not healed) would be considered as an ordinal response variable with three levels.

A survival data extension of the Mantel-Haenszel procedure was suggested by Mantel (1966) and later derived by Cox (1972) from likelihood theory under the Cox regression model for essentially continuous time to event response variables. The Mantel-Cox test Q_{MC} involves restructuring the contingency table data as a set of 2×2 tables with a life table format for each center, and then applying the Mantel-Haenszel procedure to this set of tables. The data in Table 4 would be restructured as shown in Table 5. The frequencies in this table satisfy the Mantel-Fleiss (1980) criterion (Section IV), so a chi-square approximation is appropriate for the evaluation of the Mantel-Haenszel (or Mantel-Cox) statistic. There is a significant difference between the treatments according to the Mantel-Cox statistic, $Q_{MC} = 4.25$ (d.f. = 1, $p = 0.039$). If the Mantel-Fleiss criterion had not been satisfied, the computing procedure of Thomas (1975) could have been used to provide an exact probability for Q_{MC}.

The Mantel-Cox statistic is closely related to the mean score statistic Q_S in (20) with logrank scores or its stratified extension Q_{EMH} in (44). For this reason, tests based on Q_{MC} are often referred to as logrank tests. Both Q_S and Q_{MC} have the same numerator. The denominator of Q_{MC} is like that shown in (33) for sets of 2×2 tables, whereas the denominator of the Q_S or Q_{EMH} is like that shown in (44) for sets of $2 \times r$ tables [see Lee (1980) and Koch et al. (1985b) for additional discussion of the difference between the two statistics].

One could also approach the data in Table 5 from a model-based survival analysis standpoint. One model which can be used to describe the relationship between a grouped survival time response variable and a set of categorical explanatory variables is the piecewise exponential model. This model requires the assumption that within each of the intervals, 0 to 2 weeks and 2 to 4 weeks, the events of healing have independent exponential distributions. The piecewise exponential likelihood function for our example may be written as

$$L_{PE} = \prod_{i=1}^{6} \prod_{j=1}^{2} \lambda_{ij}^{n_{ij}} [\exp(-\lambda_{ij} N_{ij})] \tag{86}$$

where i indexes the six subpopulations formed from the cross-classification of center and treatment; j indexes the two time intervals; and n_{ij}, N_{ij}, and λ_{ij} are respectively the number of patients with healing, the number of person-weeks at risk to heal, and the hazard for the ith subpopulation in the jth interval. The quantities n_{ij} and N_{ij} are shown in the fourth and last columns, respectively, of Table 5. The N_{ij} were calculated under the assumption that healing occurred uniformly throughout the interval and

Table 5 Life Table Format for Data from a Multicenter Trial for the Comparison of Treatments with Respect to Healing of a Gastrointestinal Condition

Center	Time interval (weeks)	Treatment	Life table format: Number healed during interval	Life table format: Number not healed during interval	Total	Estimated person-weeks at risk
1	0–2	Test drug	15	19	34	53
		Placebo	15	24	39	63
	2–4	Test drug	17	2	19	21
		Placebo	17	7	24	31
2	0–2	Test drug	17	27	44	71
		Placebo	12	28	40	68
	2–4	Test drug	17	10	27	37
		Placebo	13	15	28	43
3	0–2	Test drug	7	33	40	73
		Placebo	3	35	38	73
	2–4	Test drug	17	16	33	49
		Placebo	17	18	35	53

therefore can be viewed as occurring at the midpoint of the interval; thus, N_{ij} = (n_{ij} × 1) + [(number in subpopulation i not healed in interval j) × 2].

We can express the hazard rates λ_{ij} in terms of a log-linear model with specification matrix $\underline{X}$ and parameter vector $\underline{\beta}$ as follows:

$$\lambda_{ij} = \exp(\underline{x}'_{ij}\underline{\beta}) \tag{87}$$

where $\underline{x}'_{ij}$ is the row of $\underline{X}$ corresponding to the ith subpopulation and jth interval. The matrix $\underline{X}$ is required to have full rank $t \leqslant sd$, where $s = 6$ is the number of subpopulations and $d = 2$ the number of time intervals. Maximum likelihood estimation of the regression parameters for this model can be undertaken using log-linear Poisson-regression computing procedures because, as shown by Holford (1980) and Laird and Olivier (1981), the likelihoods for the piecewise exponential and Poisson frameworks are proportional. If we consider the numbers healed $\{n_{ij}\}$, conditional on their exposures to treatment $\{N_{ij}\}$, as having independent Poisson distributions with means $\mu_{ij} = N_{ij}\lambda_{ij}$, then the Poisson counterpart to (86) is

$$L_{PO} = \prod_{i=1}^{6}\prod_{j=1}^{2} \frac{(N_{ij}\lambda_{ij})^{n_{ij}}\,[\exp(-N_{ij}\lambda_{ij})]}{n_{ij}!} \tag{88}$$

$$= L_{PE}\left[\prod_{i=1}^{6}\prod_{j=1}^{2} \frac{N_{ij}^{\,n_{ij}}}{n_{ij}!}\right]$$

It is not necessary to actually assume that the $\{n_{ij}\}$ have Poisson distributions. The Poisson likelihood (88) is presented only to substantiate the proportionality of L_{PO} and L_{PE}. Since the two likelihood functions are proportional, the $\underline{\beta}$ which maximize L_{PO} also maximize L_{PE}, and thus maximum likelihood Poisson regression computing procedures may be used for estimation of piecewise exponential model parameters. This is of practical as well as theoretical importance because computing procedures for Poisson regression are more widely available than those for piecewise exponential modeling.

The maximum likelihood Poisson regression approach to parameter estimation involves substituting $\exp(\underline{x}'_{ij}\underline{\beta})$ for λ_{ij} in (88), differentiating the $\log_e$ of (88) with respect to $\underline{\beta}$, and equating the result to zero. The set of nonlinear equations resulting from this process has the form

$$\underline{X}'\underline{n} = \underline{X}'\hat{\underline{\mu}} = \underline{X}'\underline{D}_{\underline{N}}[\exp(\underline{X}\,\hat{\underline{\beta}})] \tag{89}$$

where $\underline{n} = \{n_{ij}\} = (n_{11}, n_{12}, \ldots, n_{61}, n_{62})'$, $\hat{\underline{\mu}}$ is the corresponding vector of predicted values (where $\hat{\mu}_{ij} = N_{ij}\hat{\lambda}_{ij}$), and $\underline{D}_{\underline{N}}$ is a diagonal matrix with the elements of $\underline{N} = (N_{11}, N_{12}, \ldots, N_{61}, N_{62})'$ down the main diagonal. Since these equations usually do not have an explicit solution, an iterative method, such as the Newton-Raphson procedure, is necessary to solve for $\underline{\beta}$. The maximum likelihood estimator $\hat{\underline{\beta}}$ approximately has a multivariate normal distribution with a covariance matrix which can be consistently estimated by

$$\underline{V}(\hat{\underline{\beta}}) = (\underline{X}'\underline{D}_{\hat{\underline{\mu}}}\underline{X})^{-1} \tag{90}$$

when the observed n_{ij} are sufficiently large for the linear functions $\underline{X}'\underline{n}$ to have approximately a multivariate normal distribution from central limit theory; a relatively conservative guideline for this purpose is all $n_{ij} \geqslant 5$.

Two measures of goodness of fit for the model $\underline{X}$ are the Pearson chi-square criterion

$$Q_P = \sum_{i=1}^{6}\sum_{j=1}^{2} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \tag{91}$$

and the log-likelihood ratio chi-square criterion

$$Q_L = \sum_{i=1}^{6}\sum_{j=1}^{2} 2n_{ij} \log_e\left[\frac{n_{ij}}{\hat{\mu}_{ij}}\right] \tag{92}$$

Both of these statistics have d.f. equal to the number of rows of $\underline{X}$ minus the number of columns of $\underline{X}$; in this example, d.f. = (sd − t) = (12 − t).

As in Section VI, the goodness of fit of the model may also be evaluated by assessing the contribution of columns $\underline{W}$ (a matrix with rank w) as an expansion to the model $\underline{X}$ via a log-likelihood ratio statistic Q_{LR} with structure analogous to (57) or by the Rao score statistic

$$Q_{RS} = (\underline{n} - \hat{\underline{\mu}})'\underline{W}[\underline{W}'\{\underline{D}_{\hat{\underline{\mu}}} - \underline{D}_{\hat{\underline{\mu}}}\underline{X}(\underline{X}'\underline{D}_{\hat{\underline{\mu}}}\underline{X})^{-1}\underline{X}'\underline{D}_{\hat{\underline{\mu}}}\}\underline{W}]^{-1}\underline{W}'(\underline{n} - \hat{\underline{\mu}}) \tag{93}$$

Both Q_{LR} and Q_{RS} approximately have chi-square distributions with w degrees of freedom.

A model of interest for the data in Table 5 is

$$\underline{X} = \begin{bmatrix} 1&1&1&1&1&1&1&1&1&1&1&1 \\ 0&0&0&0&1&1&1&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&1&1&1 \\ 1&0&1&0&1&0&1&0&1&0&1&0 \\ 0&0&1&1&0&0&1&1&0&0&1&1 \end{bmatrix}' \tag{94}$$

for which the parameter estimates and their standard errors are as follows:

Parameter	Estimate	Standard error	p value	Interpretation
β_1	−1.515	0.172	<0.010	Reference value of the $\log_e$ hazard rate for placebo group patients in center 1 during weeks 0–2
β_2	−0.421	0.181	0.020	Increment due to center 2

Parameter	Estimate	Standard error	p value	Interpretation
β_3	−0.887	0.197	<0.010	Increment due to center 3
β_4	0.306	0.156	0.049	Increment due to test drug
β_5	0.964	0.158	<0.010	Increment for interval 2–4 weeks

(95)

Computations for this example were undertaken with the SAS macro CATMAX documented in Stokes and Koch (1983).

The goodness of fit of model $\underline{X}$ is supported by the nonsignificance of the Pearson and log-likelihood criteria with respect to the chi-square distribution with d.f. = 7 (Q_P = 6.29, p = 0.506 and Q_L = 6.66, p = 0.466). A Rao score statistic was used to explicitly test the effect of center-by-treatment interaction; the result was nonsignificant (Q_{RS} = 0.14, d.f. = 2, p = 0.931).

The model $\underline{X}$ is said to have a proportional hazards structure because the effects of the explanatory variables are specified to be the same in both time intervals. This proportionality assumption was also tested with Rao score statistics. The center-by-time interaction was assessed by the Rao score statistic for the contribution of the matrix

$$\underline{W}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}' \tag{96}$$

Its result approached significance (Q_{RS} = 5.08, d.f. = 2, p = 0.078). The contribution of the treatment-by-time interaction

$$\underline{W}_2 = [0\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ 0]' \tag{97}$$

was nonsignificant (Q_{RS} = 0.01, d.f. = 1, p = 0.931). Also, the joint contribution of the center-by-time and treatment-by-time interactions, $[\underline{W}_1, \underline{W}_2]$, was nonsignificant ($Q_{RS}$ = 5.09, d.f. = 3, p = 0.165). Thus, the overall interpretation is that the effects of center and treatment during the first 2 weeks of the study are similar to their effects during the second 2 weeks, although it is recognized that there is some suggestion of possible departure of center effects from this structure. If this were considered important, the model could be expanded to include center-by-time interaction; such analysis is not illustrated here. For further discussion of the application of Poisson regression methods to the analysis of grouped survival data like those shown in Table 4, see Frome (1983), Holford (1980), Koch et al. (1985a, 1986a), and Laird and Olivier (1981).

X. A LONGITUDINAL STUDY EXAMPLE

Many clinical trials have longitudinal designs in which the response status of each patient in the respective treatment groups is determined at two or

more visits. For the resulting repeated measurements data structure of these studies, the objectives of statistical analysis are as follows:

1. Comparisons among groups for the response distributions at each visit separately and for their average across visits
2. Comparisons among visits for the response distributions within each group and for their average across subpopulations
3. Comparisons among groups for differences between visits in response distributions (i.e., group × visit interaction)
4. Description of the relationship of response distributions with group and visit

Some methods for addressing these objectives are discussed here in the context of an example involving two treatments (test drug and placebo) in a randomized clinical trial for a skin disorder. The response variables are the dichotomous response status, satisfactory (S) or unsatisfactory (U), at each of two visits (7 days and 14 days). The data for this example are summarized in Table 6 in terms of a 2 × 4 contingency table. The rows of this table correspond to the $s = 2$ treatments, and the columns correspond to the $r = 2^2 = 4$ possible response profiles [i.e., (S, S), (S, U), (U, S), and (U, U)].

For the data in Table 6, the proportion of patients with satisfactory classifications for each (group × visit) describe the corresponding response distributions. These quantities indicate that 0.50 of the patients with test drug had satisfactory response at 7 days, and 0.73 of them had this status at 14 days. In contrast, the proportions of placebo patients with satisfactory response were 0.28 at 7 days and 0.23 at 14 days. Thus, the differences between treatments in the proportions of patients with satisfactory response status favor test drug at both visits, and this tendency appears much stronger at 14 days than at 7 days. Subsequent analysis provides more formal confirmation of these conclusions.

The comparison of the two treatments at each visit separately can be undertaken by applying the methods in Section II to the corresponding 2 × 2 marginal tables. More specifically, let n_{ijk} denote the number of patients in the ith treatment group with the jth response outcome at 7 days and the kth response outcome at 14 days, where $i = 1, 2$ for test drug and placebo, and $j, k = 1, 2$ for satisfactory and unsatisfactory. Then,

Table 6 Data from a Longitudinal Clinical Trial for the Comparison of Treatments for a Skin Disorder

Treatment	Frequencies of response at 7 days and 14 days[a]				Number of patients
	(S, S)	(S, U)	(U, S)	(U, U)	
Test drug	24	6	20	10	60
Placebo	8	9	6	37	60

[a] S, Satisfactory; U, unsatisfactory.

the $\{n_{ij+} = \Sigma_{k=1}^{2} n_{ijk}\}$ and the $\{n_{i+k} = \Sigma_{j=1}^{2} n_{ijk}\}$ represent the response distributions for the two treatments at 7 and 14 days, respectively. For these 2 × 2 tables, the two-sided Fisher's exact tests yield $p = 0.024$ for the treatment comparison at 7 days and $p < 0.001$ for that at 14 days, and so patients with test drug have significantly more favorable experience than their placebo counterparts at both visits.

Since the proportions of patients with satisfactory response at 7 days and at 14 days are given by $p_{i1+} = (n_{i11} + n_{i12})/n_{i++}$ and $p_{i+1} = (n_{i11} + n_{i21})/n_{i++}$, respectively, their average is given by the mean score

$$\bar{f}_i = \frac{p_{i1+} + p_{i+1}}{2} = \frac{2n_{i11} + n_{i12} + n_{i21}}{2n_{i++}} = \sum_{j=1}^{2} \sum_{k=1}^{2} \frac{a_{jk} n_{ijk}}{n_{i++}} \tag{98}$$

where $(a_{11}, a_{12}, a_{21}, a_{22}) = (1, 0.5, 0.5, 0) = \underline{a}'$. It follows from Section III that the mean score statistic Q_S in (19) is applicable for the comparison of the $\{\bar{f}_i\}$ for the two treatment groups. A convenient computational strategy to obtain Q_S is to apply the integer scores (1, 2, 3) to the following transformation of Table 6:

Treatment	(U, U)	{(U, S) or (S, U)}	(S, S)	Total
Test drug	10	26	24	60
Placebo	37	15	8	60

(99)

The justification for this approach is the invariance of Q_S with respect to linear transformations of $\underline{a}$. The result $Q_S = 23.77$ has $p < 0.01$ relative to the chi-square distribution with d.f. = 1, and so the test drug is interpreted as significantly better than placebo with respect to the average proportion of patients with satisfactory response across the two visits.

Information concerning variation between visits is provided by the differences

$$\bar{g}_i = (p_{i+1} - p_{i1+}) = \frac{n_{i21} - n_{i12}}{n_{i++}} = \sum_{j=1}^{2} \sum_{k=1}^{2} \frac{\bar{a}_{jk} n_{ijk}}{n_{i++}} \tag{100}$$

where $(\bar{a}_{11}, \bar{a}_{12}, \bar{a}_{21}, \bar{a}_{22}) = (0, -1, 1, 0) = \underline{\bar{a}}'$. Their comparison between groups through Q_S in (19) provides an assessment of group × visit interaction. A convenient way to obtain Q_S for this situation is to apply the integer scores (1, 2, 3) to the following transformation of Table 6:

Treatment	(S, U)	{(S, S) or (U, U)}	(U, S)	Total
Test drug	6	34	20	60
Placebo	9	45	6	60

(101)

Such computation yields $Q_S = 7.17$, which is significant ($p < 0.01$) relative to the chi-square distribution with d.f. = 1. This significant group × visit interaction corresponds to the much larger difference in favor of test drug at 14 days than at 7 days.

The comparison of visits within each group separately is addressed in terms of the extent to which the $\{\bar{g}_i\}$ are different from 0, or equivalently the extent to which the ratios $\{(n_{i21}/(n_{i21} + n_{i12})\}$ are different from 0.5. This latter type of hypothesis can be evaluated exactly with the binomial distribution in a spirit similar to the sign test. Such analysis is usually known as McNemar's test. For test drug, $n_{121} = 20$ and $(n_{112} + n_{121}) = 26$. The two-sided exact probability of outcomes at least this extreme for the binomial distribution with 26 trials and probability parameter 0.5 is $p = 0.0094$. Similarly, for placebo, where $n_{221} = 6$ and $(n_{212} + n_{221}) = 15$, the two-sided exact probability for outcomes at least as extreme is $p = 0.607$. Thus, for active treatment patients, the proportion of patients with satisfactory response increases significantly between 7 days and 14 days, whereas for placebo it tends to remain unchanged. This substantial difference in the pattern of change across visits for the two treatment groups corresponds to an alternative description of the group × visit interaction. Its nature implies that any comparison between visits for the pooled groups would need to be viewed cautiously. The application of McNemar's test to this setting would involve outcomes at least as extreme as $(n_{121} + n_{221}) = 26$ relative to $(n_{112} + n_{121} + n_{212} + n_{221}) = 41$ binomial trials. Since the number of binomial trials under consideration exceeds 40, approximation with the chi-square distribution with d.f. = 1 is applied to the continuity-corrected statistic

$$Q_M = \frac{\{|\,26 - (41/2)\,| - 0.5\}^2}{(41/4)} = 2.44 \qquad (102)$$

Since the two-sided $p = 0.118$, it appears that the change between visits for the pooled groups is nonsignificant. As noted previously, the nature of the significant group × visit interaction tends to imply that the separate groups provide a better framework for interpreting comparisons between visits than the combined groups.

Although the previously considered methods for longitudinal studies are useful for specific statistical tests concerning groups, visits, and group × visit interaction, they do not provide an overall description of the variation among groups and visits for the proportions of patients with satisfactory response. One way to address this latter objective of analysis is to fit linear regression models by the weighted least-squares (WLS) procedures discussed in Grizzle et al. (1969). This general methodology has three stages: (1) construction of appropriate functions of the responses and a corresponding consistent estimate of their covariance structure, (2) estimation of model parameters, and (3) tests for hypotheses concerning model parameters. Its principal assumption is the availability of sufficient sample size for the functions in (1) to have an approximately multivariate normal distribution for which the covariance matrix can be viewed as known. This requirement tends to limit the use of WLS methods to situations with categorical explanatory variables for which each of the cross-classified subpopulations has at least moderate sample size (e.g., ≥ 25). A noteworthy advantage of WLS methods is their applicability to other types of functions than those involved in logistic regression or its extensions (as discussed in Sections VI and VII). These include linear functions as well as more

complex functions like the rank measures of association discussed in Section VIII. This capability of WLS methods is particularly relevant to longitudinal studies in which linear functions like p_{i1+} and p_{i+1} for describing the marginal distributions of response at each visit are often of interest.

In the subsequent discussion of the example summarized in Table 6, the patients in each treatment group are assumed to be conceptually representative of corresponding large target subpopulations in a sense equivalent to simple random sampling. On this basis, the $\{n_{ijk}\}$ have the product multinomial distribution

$$P\{n_{ijk}\} = \prod_{i=1}^{2}\left[n_{i++}! \prod_{j=1}^{2}\prod_{k=1}^{2}\left(\pi_{ijk}^{n_{ijk}} \Big/ n_{ijk}!\right)\right] \tag{103}$$

where the $\{\pi_{ijk}\}$ denote the probabilities of the jth response outcome at 7 days and the kth response outcome at 14 days for randomly selected patients who received the ith treatment. Let $\underline{\pi}_i = (\pi_{i11}, \pi_{i12}, \pi_{i21}, \pi_{i22})'$, and let $\underline{p}_i = (p_{i11}, p_{i12}, p_{i21}, p_{i22})'$, where $p_{ijk} = (n_{ijk}/n_{i++})$ denotes the sample estimator of π_{ijk}. Also, define the compound vectors $\underline{\pi} = (\underline{\pi}_1', \underline{\pi}_2')'$ and $\underline{p} = (\underline{p}_1', \underline{p}_2')'$. A set of linear functions of $\underline{p}$ can be expressed in the form $\underline{F} = \underline{A}\underline{p}$. For the proportions p_{i1+} and p_{i+1} of patients with satisfactory response at 7 days and 14 days,

$$\underline{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \tag{104}$$

A consistent estimate for the covariance matrix of the linear functions $\underline{F}$ is given by

$$\underline{V}_{\underline{F}} = \underline{A}\underline{V}_{\underline{p}}\underline{A}' \tag{105}$$

where $\underline{V}_{\underline{p}}$ is the estimated covariance matrix for $\underline{p}$ and has the block diagonal structure

$$\underline{V}_{\underline{p}} = \begin{bmatrix} (\underline{D}_{\underline{p}_1} - \underline{p}_1\underline{p}_1')/n_{1++} & \underline{0}_{4,4} \\ \underline{0}_{4,4} & (\underline{D}_{\underline{p}_2} - \underline{p}_2\underline{p}_2')/n_{2++} \end{bmatrix} \tag{106}$$

In (106), $\underline{D}_{\underline{y}}$ denotes a diagonal matrix with elements of $\underline{y}$ on the diagonal, and $\underline{0}_{4,4}$ denotes the 4 × 4 matrix of 0's. The specific forms of $\underline{F}$ and $\underline{V}_{\underline{F}}$ for the data in Table 6 are as follows:

$$\underline{F} = \begin{bmatrix} p_{11+} \\ p_{1+1} \\ p_{21+} \\ p_{2+1} \end{bmatrix} = \begin{bmatrix} 0.500 \\ 0.733 \\ 0.283 \\ 0.233 \end{bmatrix} \quad \text{and} \quad \underline{V}_{\underline{F}} = \begin{bmatrix} 41.67 & 5.56 & 0.00 & 0.00 \\ 5.56 & 32.59 & 0.00 & 0.00 \\ 0.00 & 0.00 & 33.84 & 11.20 \\ 0.00 & 0.00 & 11.20 & 29.82 \end{bmatrix} \times 10^{-4} \tag{107}$$

The functions $\underline{F}$ are viewed here as approximately having a multivariate normal distribution since the sample sizes for the two treatment groups are moderately large (i.e., $n_{1++} = n_{2++} = 60$ and all $n_{ijk} \geq 5$). On this basis, the application of weighted least-squares methods to fit a linear model to $\underline{F}$ is appropriate. Linear models for a $(u \times 1)$ vector $\underline{F}$ are expressed as

$$\underline{E}_A\{\underline{F}\} = \underline{X}\,\underline{\beta} \tag{108}$$

where $\underline{E}_A\{\ \}$ denotes expected value for large samples, $\underline{X}$ is the $(u \times t)$ model specification matrix with full rank $t \leq u$, and $\underline{\beta}$ is the $(t \times 1)$ vector of unknown parameters. The weighted least-squares estimators $\underline{b}$ for $\underline{\beta}$ and their estimated covariance matrix are given by

$$\underline{b} = (\underline{X}'\underline{V}_{\underline{F}}^{-1}\underline{X})^{-1}\underline{X}'\underline{V}_{\underline{F}}^{-1}\underline{F} \qquad \text{and} \qquad \underline{V}_{\underline{b}} = (\underline{X}'\underline{V}_{\underline{F}}^{-1}\underline{X})^{-1} \tag{109}$$

Goodness of fit of the model (108) can be assessed with the Wald statistic

$$Q_W = (\underline{F} - \underline{X}\,\underline{b})'\underline{V}_{\underline{F}}^{-1}(\underline{F} - \underline{X}\,\underline{b}) = \underline{F}'\underline{W}'(\underline{W}\,\underline{V}_{\underline{F}}\underline{W}')^{-1}\underline{W}\,\underline{F} \tag{110}$$

where $\underline{X}$ and $\underline{W}'$ are orthocomplements to one another (i.e., $\underline{W}\,\underline{X} = \underline{0}$ and Rank $[\underline{X}, \underline{W}'] = u$). When the model applies, Q_W approximately has the chi-square distribution with d.f. $= (u - t)$. For models which are considered to have satisfactory fit, linear hypotheses H_0: $\underline{C}\,\underline{\beta} = \underline{0}$ (where $\underline{C}$ is a $c \times t$ specification matrix) can be tested with the Wald statistic

$$Q_C = \underline{b}'\underline{C}'(\underline{C}\,\underline{V}_{\underline{b}}\underline{C}')^{-1}\underline{C}\,\underline{b} \tag{111}$$

Q_C approximately has the chi-square distribution with d.f. $= c$. A useful preliminary framework for assessing the sources of variation that pertain to the functions $\underline{F}$ in (107) for the example is the cell mean (or identity model). It has the structure

$$E_A\{\underline{F}\} = \underline{X}\,\underline{\beta} = \underline{I}_4\underline{\beta} \tag{112}$$

where $\underline{I}_4$ is the (4×4) identity matrix. For this model, $\underline{b} = \underline{F}$ and $\underline{V}_{\underline{b}} = \underline{V}_{\underline{F}}$. The goodness-of-fit statistic Q_W is not applicable for the cell mean model because the model involves no reduction in dimension (i.e., $u = t$, so d.f. $= 0$ for Q_W). For this reason, the principal use of this model is for testing hypotheses concerning $\underline{F}$ with the Q_C statistic in (111). The specifications and results for some hypotheses of interest concerning group effects, visit effects, and group × visit interaction are as follows:

Comparison of groups	$\underline{C} = [1\ \ 1\ -1\ -1]$	$Q_C = 29.96$	$p < 0.010$
Comparison of visits	$\underline{C} = [1\ -1\ \ 1\ -1]$	$Q_C = 3.22$	$p = 0.073$
Group × visit interaction	$\underline{C} = [1\ -1\ -1\ \ 1]$	$Q_C = 7.69$	$p < 0.010$

(113)

The conclusions from this analysis agree with those from the previously described methods involving explicit construction of statistical tests. Both analyses indicate that the difference between groups (in the sense of an average over visits) and the group × visit interaction are significant.

From the elements of $\underline{F}$ in (107), one can see that the group × visit interaction is due to the disparity between the test drug group having a substantial increase in the proportion of patients with satisfactory response over the time from 7 days to 14 days and the placebo group having essentially no change. This interpretation can be expressed with the reduced model

$$E\{\underline{F}\} = \underline{X}_R\underline{\beta}_R = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_{R,1} \\ \beta_{R,2} \\ \beta_{R,3} \end{bmatrix} \tag{114}$$

For this model, $\beta_{R,1}$ corresponds to the probability of satisfactory response for placebo patients at either 7 days or 14 days, $\beta_{R,2}$ is the effect for test drug at 7 days, and $\beta_{R,3}$ is the interaction increase in the effect of test drug between 7 days and 14 days.

The goodness of fit of the model $\underline{X}_R$ is evaluated with the Wald statistic in (110). Since $Q_W = 0.61$ is nonsignificant ($p = 0.436$) with respect to the chi-square distribution with d.f. = 1, use of the model is supported. The weighted least-squares estimates $\underline{b}_R$ of the parameters $\underline{\beta}_R$ and their estimated covariance matrix $\underline{V}_{\underline{b}_R}$ are

$$\underline{b}_R = \begin{bmatrix} 0.256 \\ 0.244 \\ 0.233 \end{bmatrix} \quad \text{and} \quad \underline{V}_{\underline{b}_R} = \begin{bmatrix} 21.42 & -21.42 & 0.00 \\ -21.42 & 63.08 & -36.11 \\ 0.00 & -36.11 & 63.15 \end{bmatrix} \times 10^{-4} \tag{115}$$

Model-predicted values for the proportions of patients with satisfactory response are given by $\hat{\underline{F}} = \underline{X}_R\underline{b}_R$. Their estimated standard errors are obtainable as square roots of the diagonal elements of $\underline{V}_{\hat{\underline{F}}} = \underline{X}_R\underline{V}_{\underline{b}_R}\underline{X}'_R$. These results were as follows:

	Test drug		Placebo	
	7 days	14 days	7 days	14 days
Predicted proportion with satisfactory response	0.500	0.733	0.256	0.256
Estimated standard error	0.065	0.057	0.046	0.046

(116)

Thus, the use of models like $\underline{X}_R$ provides an effective way of describing the relationship of response distributions with group and visit.

The methods described in this section can be extended to more complex studies with longitudinal or repeated measurements designs, e.g., those involving more than two visits, ordinal data, or several groups which encompass both treatments as well as other explanatory variables. Discussion

of some analysis strategies for such situations is given in Koch et al. (1977, 1985a, 1986b, 1987), Landis and Koch (1978, 1979), Landis et al. (1987), and Wei et al. (1985).

XI. A CHANGEOVER STUDY EXAMPLE

Clinical trials pertaining to the relief of symptoms of chronic or recurrent conditions often make use of changeover designs. In such studies, each patient is randomly assigned to a sequence of treatments which are to be received during successive evaluation periods. The resulting data have a repeated measurements structure for which the methods of analysis are similar to those discussed for the longitudinal study example in Section X. Some aspects of their application are considered here.

The data in Table 7 are from a two-period changeover study concerned with the comparison of a test drug and placebo for relief of a recurrent pain condition. There are two sequence groups, to each of which 50 patients were randomly assigned. The patients in the T:P group received test drug for the first episode of the pain condition during the course of the study and placebo for the second episode; the P:T group received the opposite regimen. Also, the two evaluation periods for the two episodes were separated by several weeks so that the treatment for the first would tend not to influence the response during the second (i.e., to minimize treatment × period interaction due to potential carryover effects). For each evaluation period, the response of each patient was classified as either favorable (F) or unfavorable (U), so for periods jointly, there are $r = 2^2 = 4$ possible response profiles [i.e., (F, F), (F, U), (U, F), (U, U)].

As was the case for the longitudinal study example in Section X, the response distributions for the data in Table 7 for the respective (group × period) combinations are described by the corresponding proportions with favorable response. These quantities are the following linear functions of the frequencies $\{n_{ijk}\}$ for the joint occurrence of the jth response outcome during evaluation period 1 and the kth response outcome during evaluation period 2 for the $n_{i++} = \sum_{j=1}^{2} \sum_{k=1}^{2} n_{ijk} = 50$ patients in the ith sequence group:

Table 7 Data from a Changeover Clinical Trial for the Comparison of Treatments for Relief of a Recurrent Pain Condition

Treatment sequence	Frequencies of response at evaluation periods 1 and 2[a]				Number of patients
	(F, F)	(F, U)	(U, F)	(U, U)	
Test drug : placebo	20	16	5	9	50
Placebo : test drug	16	6	18	10	50

[a]F, Favorable; U, unfavorable.

Sequence group	Period	Treatment	Proportion favorable
T:P	1	T	$p_{11+} = (n_{111} + n_{112})/n_{1++} = 0.720$
T:P	2	P	$p_{1+1} = (n_{111} + n_{121})/n_{1++} = 0.500$
P:T	1	P	$p_{21+} = (n_{211} + n_{212})/n_{2++} = 0.440$
P:T	2	T	$p_{2+1} = (n_{211} + n_{221})/n_{2++} = 0.680$

(117)

Thus, about 70% of the patients have favorable response during the period with test drug, whereas about 50% have favorable response during the period with placebo. Also, the tendency of about 20% more patients to have favorable response during test drug treatment than during placebo is relatively homogeneous across both periods and sequence groups. Results which support these general conclusions are presented in the following discussion.

A two-sided Fisher's exact test is used to compare the treatments for each of the two periods separately. In both periods the proportion of patients with favorable response is higher in the test drug group. In period 1 the result of Fisher's exact test is significant ($p = 0.008$), and in period 2 it is suggestive ($p = 0.103$). As discussed in Koch et al. (1983), the extent of carryover effects (i.e., treatment × period interaction) can be assessed through the comparison of the sequence groups with respect to the averages $\{\bar{f}_i = (p_{i1+} + p_{i+1})/2\}$. From computations like those for the analogous $\{\bar{f}_i\}$ in (98) for the longitudinal study example in Section X, the test statistic $Q_S = 0.475$ is obtained. This result is nonsignificant ($p = 0.491$) relative to the chi-square distribution with d.f. = 1. Similarly, the comparison of sequence groups with respect to the differences $\{\bar{g}_i = (p_{i+1} - p_{i1+})\}$ provides a test for treatment effects (in an average sense across periods). The result from computations like those for the analogous $\{\bar{g}_i\}$ in (100) of Section X is $Q_S = 11.64$, which is significant ($p < 0.010$) with respect to the chi-square distribution with d.f. = 1. Thus, when the combined information for both periods is taken into account, a strong difference in favor of test drug is detected.

For changeover studies in which carryover effects are expected to be negligible (on the basis of knowledge about the clinical condition under investigation) and are confirmed to be nonsignificant by a method like that described previously, an exact test is available for the assessment of treatment effects. It is undertaken by applying Fisher's exact test to the two sequence groups for the subtable of Table 7 corresponding to the (F, U) and (U, F) response profiles; the rationale for this method is discussed in Gart (1969). A relevant consideration is that the (F, F) and (U, U) response profiles do not provide information useful for discriminating between the two treatments since the patients with either of these profiles had the same response to both treatments. The result from this application of Fisher's two-sided exact test is $p = 0.001$; it significantly favors test drug because (F, U) was more prevalent for the T:P sequence (where the F corresponded to test drug) and (U, F) was more prevalent for the P:T sequence (where the F also corresponded to test drug).

The fitting of linear regression models by weighted least-squares methods provides a way of describing the variation among the proportions of patients with favorable response for the respective (treatment × period) combinations. The considerations involved in applying this analysis strategy are essentially the same as outlined in (103)–(111) for the longitudinal data example in Section X. A model of interest for the data in Table 7 has the specification

$$\underline{E}_A\{\underline{F}\} = \underline{E}_A \begin{bmatrix} p_{11+} \\ p_{1+1} \\ p_{21+} \\ p_{2+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \underline{X}\,\underline{\beta} \qquad (118)$$

for which β_1 represents the probability of favorable response for placebo during period 1, β_2 represents the effect of period 2, and β_3 represents the effect of test drug. The goodness of fit of this model is assessed with the Wald statistic (110). The result, $Q_W = 0.48$, is nonsignificant ($p = 0.489$) with respect to the chi-square distribution with d.f. = 1; thus, use of the model is supported. Moreover, this finding can be interpreted as indicating that carryover effects are negligible since the structure of $\underline{X}$ expresses additive effects for treatment and period. The weighted least-squares estimates $\underline{b}$ of the parameters $\underline{\beta}$, their estimated standard errors [from square roots of the diagonal elements of $\underline{V}_b$ in (109)], and p-values from the test statistics Q_C in (111) for hypotheses H_0: $\beta_h = 0$ are as follows:

Parameter	Estimate	Standard error	p value
β_1	0.466	0.060	<0.010
β_2	0.006	0.063	0.920
β_3	0.231	0.063	<0.010

(119)

These results indicate that the proportion of patients with favorable response is significantly greater by 0.231 for test drug than for placebo. Also, for each treatment, there is essentially no difference between the two periods for the proportions of patients with favorable response. A description of the pattern of variation corresponding to these conclusions is provided by the following model-predicted values $\hat{\underline{F}} = \underline{X}\underline{b}$ and their estimated standard errors (from square roots of diagonal elements of $\underline{X}\underline{V}_b\underline{X}'$):

	Test drug Period 1	Placebo Period 2	Placebo Period 1	Test drug Period 2
Predicted proportion with favorable response	0.697	0.472	0.466	0.703
Estimated standard error	0.054	0.058	0.060	0.057

(120)

Since neither period effects nor carryover effects are evident for this example, one additional test for treatment effects is of interest. It is based on the application of McNemar's test to determine whether the favorable components of (F, U) or (U, F) responses are equally split between test drug and placebo. For the data in Table 7, (16 + 18) = 34 of the (16 + 5 + 6 + 18) = 45 responses of this type favored test drug. The continuity-corrected McNemar statistic for this is $Q_M = \{|34 - 22.5| - 0.5\}^2/(45/4) = 10.76$, which is significant ($p < 0.010$) with respect to the chi-square distribution with d.f. = 1.

Other types of repeated measurements data structures which have research designs analogous to those of change-over studies can be analyzed by the methods described here. A noteworthy example is the bilateral study where one treatment is applied to the right side of the body (e.g., skin on arms or legs, eye, mouth) and the other to the left side. Also, extensions to deal with greater complexity, such as more than two evaluation periods or more than two treatments, are available or feasible to develop, particularly for situations where the sample size for each sequence group is moderately large. For additional discussion of statistical methods for the analysis of categorical data from change-over design studies, see Fleiss (1981), Gart (1969), and Koch et al. (1977, 1983, 1985a).

XII. SOME COMMENTS ON POWER AND SAMPLE SIZE ISSUES

This chapter has been primarily concerned with describing statistical methods for the analysis of clinical efficacy trials with categorical data. However, two related considerations which often require attention are the power of statistical tests for treatment comparisons and the sample size needed to achieve desired levels of power. With model-based methods, such as logistic regression or its extensions (Sections VI and VII), piecewise exponential modeling (Section IX), or weighted least-squares procedures (Section VIII, X, and XI), an indication of the a posteriori power for an already completed study can be obtained by making use of the approximately normal distribution of the estimated model parameter $\hat{\beta}_T$ for treatment effects. More specifically, for the null hypothesis H_0: $\beta_T = 0$, an approximation of the power to detect the alternative, H_A: $\beta_T = \delta$, through the test statistic $Z = \hat{\beta}_T/\{\text{s.e.}(\hat{\beta}_T)\}$ in the setting with two-sided type I error α is given by

$$\text{Power} = 1 - \Phi\left\{ Z_{(\alpha/2)} - \left[\frac{\delta}{\text{s.e.}(\hat{\beta}_T)} \right] \right\} \tag{121}$$

where Φ is the cumulative distribution function for the standard normal distribution with expected value 0 and variance 1 and $Z_{(\alpha/2)}$ is the corresponding $100\{1 - (\alpha/2)\}$th percentile.

Methods for determining sample sizes to have specific levels of power in studies which are to be initiated also often involve the use of normal approximations. For the comparison of two treatments with respect to a dichotomous response, Fleiss (1981) provides extensive tables of sample sizes. Further discussion of sample size issues which is relevant to this and more general situations with categorical data is given in Donner (1984) and Lachin (1981).

ACKNOWLEDGMENTS

The research for this chapter was supported in part by the U.S. Bureau of the Census through Joint Statistical Agreement JSA-84-5. The authors would also like to thank Ingrid Amara, Susan Atkinson, Gregory Carr, and Amanda Sullivan for assistance with computations.

REFERENCES

Agresti, A. (1983). A Survey of Strategies for Modeling Classifications Having Ordinal Variables, *J. Am. Stat. Assoc.*, 78: 184–197.

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*, Wiley, New York.

Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications*, North Holland, Amsterdam.

Andrich, D. (1979). A Model for Contingency Tables Having an Ordered Response Classification, *Biometrics*, 35: 403–415.

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

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research*, vol. I: *The Analysis of Case-Control Studies*, International Agency for Research on Cancer, Lyon.

Brown, M. B. (1983). P4F: Frequency Tables, *BMDP Statistical Software* (W. J. Dixon et al., eds.), University of California Press, Los Angeles, pp. 143–206.

Clogg, C. C. (1982). Some Models for the Analysis of Association in Multiway Cross-Classifications Having Ordered Categories, *J. Am. Stat. Assoc.*, 77: 803–815.

Conover, W. J. (1974). Some Reasons for Not Using the Yates Continuity Correction on 2 × 2 Contingency Tables, *J. Am. Stat. Assoc.*, 69: 374–382.

Cox, D. R. (1970). *The Analysis of Binary Data*, Methuen, London.

Cox, D. R. (1972). Regression Models and Life Tables, *J. R. Stat. Soc. Ser. B*, 34: 187–220.

Cox, C. and Chuang, C. (1984). A Comparison of Chi-Square Partitioning and Two Logit Analyses of Ordinal Pain Data from a Pharmaceutical Study, *Stat. Med.*, 3: 273–285.

Donner, A. (1984). Approaches to Sample Size Estimation in the Design of Clinical Trials—a Review, *Stat. Med.*, 3: 199–214.

Engelman, L. (1983). PLR: Stepwise Logistic Regression, *BMDP Statistical Software* (W. J. Dixon et al., eds.), University of California Press, Los Angeles, pp. 330–344.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd edition, MIT Press, Cambridge, Mass.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Edition, Wiley, New York.

Forthofer, R. N. and Lehnen, R. G. (1981). *Public Program Analysis: A New Categorical Data Approach*, Wadsworth, Belmont, Calif.

Frome, E. L. (1983). The Analysis of Rates Using Poisson Regression Models, *Biometrics*, 39: 665–674.

Gart, J. J. (1969). An Exact Test for Comparing Matched Proportions in Crossover Designs, *Biometrika*, 56: 75–80.

Gart, J. J. (1971). The Comparison of Proportions: A Review of Significance Tests, Confidence Intervals, and Adjustments for Stratification, *Int. Stat. Rev.*, 39: 148–169.

Goodman, L. A. (1979). Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories, *J. Am. Stat. Assoc.*, 74: 537–552.

Grizzle, J. E. (1967). Continuity Correction in the χ^2-Test for 2 × 2 Tables, *Am. Stat.*, 21: 28–32.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of Categorical Data by Linear Models, *Biometrics*, 25: 489–504.

Haber, M. (1986). An Exact Unconditional Test for the 2 × 2 Comparative Trial, *Psychol. Bull.*, 99: 129–132.

Haberman, S. J. (1978). *Analysis of Qualitative Data: Vol. 1 - Introductory Topics; Vol. 2 - New Developments*, Academic Press, New York.

Hannan, J. and Harkness, W. L. (1963). Normal Approximation to the Distribution of Two Independent Binomials, Conditional on Fixed Sum, *Ann. Math. Stat.*, 34: 1593–1595.

Harrell, F. E. (1986). The LOGIST Procedure, *SUGI Supplemental Library User's Guide, Version 5 Edition*, SAS Institute, Cary, N.C., pp. 269–293.

Holford, T. R. (1980). The Analysis of Rates and Survivorship Using Log-Linear Models, *Biometrics*, 36: 299–305.

Imrey, P. B., Koch, G. G., Stokes, M. E., Darroch, J. N., Freeman, D. H., and Tolley, H. D. (1981). Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression, Part I, *Int. Stat. Rev.*, 49: 265–283.

Imrey, P. B., Koch, G. G., Stokes, M. E., Darroch, J. N., Freeman, D. H., and Tolley, H. D. (1982). Categorical Data Analysis: Some Reflections on the Log Linear Model and Logistic Regression, Part II, *Int. Stat. Rev.*, 50: 35–64.

Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*, Lifetime Learning Publications, Belmont, Calif.

Koch, G. G. and Bhapkar, V. P. (1982). Chi-Square Tests, *Encyclopedia of Statistical Sciences*, vol. 1 (N. L. Johnson and S. Kotz, eds.), Wiley, New York, pp. 442–457.

Koch, G. G. and Edwards, S. (1985). Logistic Regression, *Encyclopedia of Statistical Sciences*, vol. 5 (N. L. Johnson and S. Kotz, eds.), Wiley, New York, pp. 128–133.

Koch, G. G. and Gillings, D. B. (1983). Inference, Design Based vs. Model Based, *Encyclopedia of Statistical Sciences*, vol. 4 (N. L. Johnson and S. Kotz, eds.), Wiley, New York, pp. 84–88.

Koch, G. G. and Sollecito, W. A. (1984). Statistical Considerations in the Design, Analysis, and Interpretation of Comparative Clinical Studies: An Academic Perspective, *Drug Inf. J.*, 18: 131–151.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., and Lehnen, R. G. (1977). A General Methodology for the Analysis of Experiments with Repeated Measurement of Categorical Data, *Biometrics*, 33: 133–158.

Koch, G. G., Amara, I. A., Davis, G. W., and Gillings, D. B. (1982). A Review of Some Statistical Methods for Covariance Analysis of Categorical Data, *Biometrics*, 38: 563–595.

Koch, G. G., Gitomer, S. L., Skalland, L., and Stokes, M. E. (1983). Some Non-Parametric and Categorical Data Analyses for a Change-Over Design Study and Discussion of Apparent Carry-Over Effects, *Stat. Med.*, 2: 397–412.

Koch, G. G., Imrey, P. B., Singer, J. M., Atkinson, S. S., and Stokes, M. E. (1985a). *Analysis of Categorical Data*, Les Presses de l'Universite de Montreal, Montreal.

Koch, G. G., Sen, P. K., and Amara, I. A. (1985b). Logrank Scores, Statistics and Tests, *Encyclopedia of Statistical Sciences*, vol. 5 (N. L. Johnson and S. Kotz, eds.), Wiley, New York, pp. 136–142.

Koch, G. G., Singer, J. M., and Amara, I. A. (1985c). A Two-Stage Procedure for the Analysis of Ordinal Categorical Data, *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences, the Bernard G. Greenberg Volume* (P. K. Sen, ed.), North-Holland, New York, pp. 357–387.

Koch, G. G., Atkinson, S. S., and Stokes, M. E. (1986a). Poisson Regression, *Encyclopedia of Statistical Sciences*, vol. 7 (N. L. Johnson and S. Kotz, eds.), Wiley, New York, pp. 32–41.

Koch, G. G., Singer, J. M., Stokes, M. E., Carr, G. J., Cohen, S. B., and Forthofer, R. N. (1986b). "Some Aspects of Weighted Least Squares Analysis for Longitudinal Categorical Data," Proc. of the Workshop on Longitudinal Methods in Health Research, Berlin (forthcoming).

Koch, G. G., Elashoff, J. D., and Amara, I. A. (1987). Repeated Measurements Studies, Design and Analysis, *Encyclopedia of Statistical Sciences*, vol. 8 (N. L. Johnson and S. Kotz, eds.), Wiley, New York, forthcoming.

Lachin, J. M. (1981). Introduction to Sample Size Determination and Power Analysis for Clinical Trials, *Controlled Clin. Trials*, 2: 93–113.

Laird, N., and Olivier, D. (1981). Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques, *J. Am. Stat. Assoc.*, 76: 231–240.

Landis, J. R. and Koch, G. G. (1978). The Measurement of Observer Agreement for Categorical Data, *Biometrics*, 33: 159–174.

Landis, J. R. and Koch, G. G. (1979). The Analysis of Categorical Data in Longitudinal Studies of Behavioral Development, *Longitudinal Research in the Study of Behavior and Development* (J. R. Nesselroade and P. B. Baltes, eds.), Academic Press, New York, Chap. 9, pp. 233–261.

Landis, J. R., Stanish, W. M., Freeman, J. L., and Koch, G. G. (1976). A Computer Program for the Generalized Chi-Square Analysis of Categorical Data Using Weighted Least Squares (GENCAT), *Comput. Programs Biomed.*, 6: 196–231.

Landis, J. R., Cooper, M. M., Kennedy, T., and Koch, G. G. (1979). A Computer Program for Testing Average Partial Association in Three-Way Contingency Tables (PARCAT), *Comput. Programs Biomed.*, 9: 223–246.

Landis, J. R., Heyman, E. R., and Koch, G. G. (1978). Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests, *Int. Stat. Rev.*, 46: 237–254.

Landis, J. R., Miller, M. E., Davis, C. S., and Koch, G. G. (1987). Some General Methods for the Analysis of Categorical Data in Longitudinal Studies, *Stat. Med.*, 6, forthcoming.

Larntz, K. (1978). Small Sample Comparisons of Exact Levels for Chi-Squared Goodness of Fit Statistics, *J. Am. Stat. Assoc.*, 73: 253–263.

Lee, E. T. (1980). *Statistical Methods for Survival Data Analysis*, Lifetime Learning Publications, Belmont, Calif.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco.

Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other, *Ann. Math. Stat.*, 18: 50–60.

Mantel, N. (1963). Chi-Square Tests with One Degree of Freedom: Extensions of the Mantel-Haenszel Procedure, *J. Am. Stat. Assoc.*, 58: 690–700.

Mantel, N. (1966). Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration, *Cancer Chemother. Rep.*, 50: 163–170.

Mantel, N. and Fleiss, J. (1980). Minimum Expected Cell Size Requirements for the Mantel-Haenszel One-Degree of Freedom Chi-Square Test and a Related Rapid Procedure, *Am. J. Epidemiol.*, 112: 129–134.

Mantel, N. and Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease, *J. Nat. Cancer Inst.*, 22: 719–748.

McCullagh, P. (1980). Regression Models for Ordinal Data (with Discussion), *J. R. Stat. Soc. Ser. B*, 42: 109–142.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*, Chapman & Hall, New York.

Mehta, C. R. and Patel, N. R. (1983). A Network Algorithm for Performing Fisher's Exact Test in r × c Contingency Tables, *J. Am. Stat. Assoc.*, 78: 427–434.

Mehta, C. R., Patel, N. R., and Tsiatis, A. A. (1984). Exact Significance Testing to Establish Treatment Equivalence with Ordered Categorical Data, *Biometrics*, 40: 819–825.

Mehta, C. R., Patel, N. R., and Gray, R. (1985). Computing an Exact Confidence Interval for the Common Odds Ratio in Several 2 × 2 Contingency Tables, *J. Am. Stat. Assoc.*, 80: 969–973.

Overall, J. E. and Starbuck, R. R. (1983). F-Test Alternatives to Fisher's Exact Test and to the Chi-Square Test of Homogeneity in 2 by 2 Tables, *J. Educ. Stat.*, 8: 59–74.

Pagano, M. and Halvorsen, K. T. (1981). An Algorithm for Finding the Exact Significance Levels of r × c Tables, *J. Am. Stat. Assoc.*, 76: 931–934.

Peterson, B. L. (1986). Proportional Odds and Partial Proportional Odds Models for Ordinal Response Variables, Dissertation submitted to Department of Biostatistics, University of North Carolina, Chapel Hill.

Plackett, R. L. (1981). *The Analysis of Categorical Data*, Griffin, London.

Puri, M. L. and Sen, P. K. (1971). *Non-Parametric Methods in Multivariate Analysis*, Wiley, New York.

Salama, I. A., Quade, D., and Koch, G. G. (1984). Tables for Testing the Equality of Two Proportions when Prior Information on Their Common Value May Be Available, *Biometrical J.*, 25: 301–320.

SAS Institute, Inc. (1985). *SAS User's Guide: Statistics, Version 5 Edition*, SAS Institute, Cary, N.C.

Semenya, K. A., Koch, G. G., Stokes, M. E., and Forthofer, R. N. (1983). Linear Models Methods for Some Rank Function Analyses of Ordinal Categorical Data, *Commun. Stat.—Theory Methods*, 12: 1277–1298.

Stokes, M. E. and Koch, G. G. (1983). "A Macro for Maximum Likelihood Fitting of Log-Linear Models to Poisson and Multinomial Counts with Contrast Matrix Capability for Hypothesis Testing," Proc. of the 8th Annual SAS Users Group International Conference, SAS Institute, Cary, N.C., pp. 795–800.

Thomas, D. G. (1975). Exact and Asymptotic Methods for the Combination of 2 × 2 Tables, *Comput. Biomed. Res.*, 8: 423–446.

Upton, G. G. G. (1982). A Comparison of Alternative Tests for the 2 × 2 Comparative Trial, *J. R. Stat. Soc.*, 145: 86–105.

van Elteren, P. H. (1960). On the Combination of Independent Two-Sample Tests of Wilcoxon, *Bull. Int. Stat. Inst.*, 37: 351–361.

Walker, S. H. and Duncan, D. B. (1967). Estimation of the Probability of an Event as a Function of Several Independent Variables, *Biometrika*, 54: 167–179.

Wei, L. J., Stram, D., and Ware, J. H. (1985). Analysis of Repeated Ordered Categorical Outcomes with Possibly Missing Observations, Department of Biostatistics, Harvard School of Public Health, Boston, Technical Report No. 6.

Yates, F. (1934). Contingency Tables Involving Small Numbers and the χ^2 Test. *J. R. Stat. Soc. Suppl.*, 1: 217–235.

Analysis of Rank Measures of Association for Ordinal Data From Longitudinal Studies

GREGORY J. CARR, KERRY B. HAFNER, and GARY G. KOCH

Reprinted from the Journal of the American Statistical Association
Volume 84, Number 407, September 1989

Analysis of Rank Measures of Association for Ordinal Data From Longitudinal Studies

GREGORY J. CARR, KERRY B. HAFNER, and GARY G. KOCH*

Consideration is given to longitudinal data settings in which an ordinal response variable is observed at two or more visits for each subject in two or more ordered groups with at least moderately large sample sizes (e.g., overall $n \geq 40$). Rank measures of association between group and response are constructed at each time, and the covariance matrix of these measures is also estimated. Variation among the rank measures of association is then analyzed through weighted least squares methods with weights based on the estimated covariance matrix. Such methods permit evaluation of group × visit interaction as well as other effects of interest. Extensions of this type of analysis to situations with stratification, or concomitant variables, or missing data are also considered. Examples are provided for illustrative purposes.

KEY WORDS: Group × visit interaction; Multivisit clinical trials; Rank correlation coefficients; Repeated measures.

1. INTRODUCTION

In longitudinal data settings such as multivisit clinical trials, subjects are often randomized to two or more ordered groups, such as treatments with different doses of an active ingredient. The subjects are then observed for a response variable under each of a number of conditions, which are frequently occasions related to a progression of time. The statistical analysis involves estimation and tests of hypotheses for the group effects, condition effects, and group × condition interaction.

A number of analysis methods are available for longitudinal data. The choice of analysis method depends on the nature of the response measures, data structure, and sample size considerations. Two parametric methods, multivariate and univariate analysis of variance, are applicable to continuous responses for which an underlying multivariate normal distribution of the responses is tenable. See Gill (1978), Koch (1969), Koch, Elashoff, and Amara (1988), and Morrison (1976) for discussion and references concerning these methods.

Nonparametric procedures also exist for longitudinal data with continuous responses (Koch 1969, 1970; Koch, Amara, Stokes, and Gillings 1980). Group effects and linear model specifications of no group × condition interaction are assessed with multivariate extensions of the Kruskal–Wallis (1953) test. Tests of condition effects such as that of Friedman (1937) or related extensions may be conducted and resemble univariate or multivariate analysis of variance tests for ranks.

In categorical data settings, the previously described parametric and nonparametric methods are often not applicable, especially when response outcomes are purely ordinal (i.e., not meaningfully described with a set of numeric scores) and the extent of group × condition interaction is of particular interest. In addition, usage of ordinal data models such as the proportional odds model discussed in McCullagh (1980) and Stram, Wei, and Ware (1988) may not be appropriate because of uncertainty about the goodness of fit of the specific structure that it imposes. In this article, rank correlation methods, which are based on the relatively minimal assumptions of simple random sampling for longitudinal categorical data, are applied in such a way that group effects on response and group × condition interaction may be assessed in an effective and meaningful way. For a single ordinal response variable, a related Mann–Whitney (1947) type of statistic has been used in a similar manner to evaluate group effects and group × strata interaction for multistrata clinical trials [see Koch, Imrey, Singer, Atkinson, and Stokes (1985) and Koch and Edwards (1988) for examples]. The methods presented here provide extensions to analogous longitudinal data settings with the same nonparametric advantages. They also are based on a more convenient computational procedure.

2. METHODOLOGY

Let $i = 1, 2, \ldots, n$ index the subjects in a simple random sample from a population. For each subject, let X_i denote the value of an at least ordinal grouping variable such as amount of an active ingredient in a treatment, and let $Y_{i1}, Y_{i2}, \ldots, Y_{id}$ denote d at least ordinal response variables for the d conditions (or visits) in a longitudinal study.

One way of evaluating the association between the grouping variable and the response variables is based on the concepts of concordance and discordance that underlie Kendall (1962) rank correlation coefficients. In particular, relative to the jth response variable, the pair of points (X_i, Y_{ij}) and $(X_{i'}, Y_{i'j})$ are called concordant if $(X_i - X_{i'})(Y_{ij} - Y_{i'j}) > 0$, discordant if $(X_i - X_{i'})(Y_{ij} - Y_{i'j}) < 0$, and tied if $(X_i - X_{i'})(Y_{ij} - Y_{i'j}) = 0$. In other words, two points are concordant if the line connecting them has positive slope, discordant if it has negative slope, and tied if it has 0 or infinite slope. The Kendall (1963) tau, a rank correlation coefficient between the grouping variable and

* Gregory J. Carr is a graduate student, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7400. Kerry B. Hafner is Senior Statistician, Department of Statistics and Data Analysis, Marion Laboratories, Inc., Kansas City, MO 64134-0627. Gary G. Koch is Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7400. This research was partially supported by National Institute of Environmental and Health Sciences Training Grant 5T32-ES07018-11.

Reprinted from: © 1989 American Statistical Association
Journal of the American Statistical Association
September 1989, Vol. 84, No. 407, Biopharmaceutical Statistics

the jth response variable, is defined as

$$\hat{\tau}_j = \frac{\left\{\begin{matrix}\text{number pairs concordant} \\ \text{for } j\text{th response}\end{matrix}\right\} - \left\{\begin{matrix}\text{number pairs discordant} \\ \text{for } j\text{th response}\end{matrix}\right\}}{n(n-1)/2}. \tag{1}$$

It is indicative of the strength of association between two ordinal variables in the sense that its range is $[-1, 1]$; for this range, -1 corresponds to perfect negative correlations from discordance of all pairs and 1 corresponds to perfect positive correlation from concordance of all pairs. The value 0 for $\hat{\tau}_j$ is indicative of no association in the sense that statistical independence of the grouping variable and the jth response variable implies equal probabilities of concordance and discordance for randomly selected pairs of points (although it should be noted that 0 values are also compatible with certain patterns of dependence).

A limitation of $\hat{\tau}_j$ as a measure of association is that ties among the X_i or among the Y_{ij} cause its range to become narrower than $[-1, 1]$, and the extent of this becomes more substantial as the probability of ties increases. For the ordinal categorical data of interest in this article, many ties commonly occur. In the presence of such ties, a rank correlation coefficient with a definition that maintains the $[-1, 1]$ range is the Goodman–Kruskal (1963, 1972) gamma. It has the form

$$\hat{\gamma}_j = \frac{\left\{\begin{matrix}\text{number pairs concordant} \\ \text{for } j\text{th response}\end{matrix}\right\} - \left\{\begin{matrix}\text{number pairs discordant} \\ \text{for } j\text{th response}\end{matrix}\right\}}{\text{number pairs relevant for } j\text{th response}}, \tag{2}$$

where a pair of points is defined as relevant if it is either concordant or discordant (i.e., not tied). Other definitions of relevance can also be specified, and some that enable adjustment for stratification of one or more concomitant variables through matching specifications are discussed in Quade (1974) and Quade (1982).

The Goodman–Kruskal rank correlation coefficient $\hat{\gamma}_j$ is a ratio estimator of the corresponding population parameter $\gamma_j = (\tau_j/\phi_j)$, where τ_j represents the difference between the probabilities of concordance and discordance for randomly selected pairs of points involving the grouping variable and the jth response variable and ϕ_j represents the probability of relevance (i.e., the probability of no tie for the pair or, equivalently, the sum of the probabilities of concordance and discordance). When the grouping variable and the jth response variable are independent, $\tau_j = \gamma_j = 0$. Thus the differences of the $\hat{\gamma}_j$ from 0 express the strength of association between the grouping variable and the jth response variable, and the differences among the $\hat{\gamma}_j$ across conditions are measures of the extent of group × condition interaction. The evaluation of condition effects for the responses is not feasible with the $\hat{\gamma}_j$, since they are determined within the respective conditions separately. Therefore, methods outside the scope of this article are needed; see Koch (1970) and Landis, Miller, Davis, and Koch (1988) for relevant discussion.

2.1 Properties of the $\hat{\gamma}_j$

A basic consideration for statistical inference concerning the γ_j is the recognition that their estimators $\hat{\gamma}_j$ are ratios of two U statistics. Thus consistent estimators for their variances are available and normal approximations are applicable to their distributions in moderately large samples. For a relatively detailed discussion concerning U statistics and their desirable asymptotic properties, see Davis and Quade (1968), Hoeffding (1948), Quade (1974), Sen (1960), and Serfling (1980, chap. 5). The subsequent discussion summarizes the noteworthy considerations for situations with longitudinal data.

The U statistic expression for $\hat{\gamma}_j$ is the ratio $\hat{\gamma}_j = (\hat{\tau}_j/\hat{\phi}_j)$, where $\hat{\tau}_j$ and $\hat{\phi}_j$ are means of estimators for τ_j and ϕ_j from all possible subsamples of two subjects. For τ_j, the estimator corresponding to the ith and i'th subjects is

$$\begin{aligned} U_{ii',j} &= 1 \quad \text{if } (X_i - X_{i'})(Y_{ij} - Y_{i'j}) > 0 \\ &= 0 \quad \text{if } (X_i - X_{i'})(Y_{ij} - Y_{i'j}) = 0 \\ &= -1 \quad \text{if } (X_i - X_{i'})(Y_{ij} - Y_{i'j}) < 0; \end{aligned} \tag{3}$$

for ϕ_j, it is $W_{ii',j} = |U_{ii',j}|$. Since $U_{ii',j}$ is 1, 0, or -1 as two randomly selected subjects are concordant, tied, or discordant for the association of the grouping variable and the jth response variable, it is an unbiased estimator for τ_j. In addition, subsamples of size 2 are the smallest samples that can yield an unbiased estimator for τ_j; for this reason, the $U_{ii'j}$ are usually called kernels of degree 2 for the τ_j in references dealing with U statistics. Similarly, $W_{ii',j}$ is 1 or 0 as relevance (i.e., no ties) for a pair of subjects applies or not, so it is an unbiased estimator for ϕ_j; it is also a kernel of degree 2 for ϕ_j.

Relative to the $U_{ii',j}$ and the $W_{ii',j}$,

$$\begin{aligned} \hat{\tau}_j &= \left[\sum_{i=1}^{n}\left\{\sum_{i'\neq i}^{n} U_{ii',j}/(n-1)\right\}\right]\Big/ n \\ &= \left[\sum_{i=1}^{n} U_{ij}\right]\Big/ n \end{aligned} \tag{4}$$

and

$$\begin{aligned} \hat{\phi}_j &= \left[\sum_{i=1}^{n}\left\{\sum_{i'\neq i}^{n} W_{ii',j}/(n-1)\right\}\right]\Big/ n \\ &= \left[\sum_{i=1}^{n} W_{ij}\right]\Big/ n \end{aligned} \tag{5}$$

are the U statistics for estimating τ_j and ϕ_j. By construction, they are unbiased estimators for τ_j and ϕ_j. The quantities U_{ij} and W_{ij}, for which $\hat{\tau}_j$ and $\hat{\phi}_j$ are sample means, are usually called components of their corresponding U statistics. From them, the variances of the $\hat{\tau}_j$ and the $\hat{\phi}_j$ can be consistently estimated with

$$\begin{aligned} v_{\tau,j} &= (4/n^2)\sum_{i=1}^{n}(U_{ij} - \hat{\tau}_j)^2, \\ v_{\phi,j} &= (4/n^2)\sum_{i=1}^{n}(W_{ij} - \hat{\phi}_j)^2, \end{aligned} \tag{6}$$

respectively, and the covariance of $\hat{\tau}_j$ and $\hat{\phi}_j$ can be consistently estimated with

$$v_{\tau\phi,j} = (4/n^2) \sum_{i=1}^{n} (U_{ij} - \hat{\tau}_j)(W_{ij} - \hat{\phi}_j). \quad (7)$$

It is of interest to note, however, that references such as Davis and Quade (1968) equivalently make use of somewhat more conservative estimators, which are $n/(n-1)$ times larger than (6) and (7). For large sample sizes n, the structure of $(\hat{\tau}_j, \hat{\phi}_j)$ as a pair of U statistics enables them to have an approximately bivariate normal distribution with mean vector (τ_j, ϕ_j). The estimator $\hat{\gamma}_j = (\hat{\tau}_j/\hat{\phi}_j)$ is a ratio of two U statistics, so from the results of Quade (1974), which involve the approximation of $\hat{\gamma}_j$ with the linear Taylor series expansion

$$\gamma_j + (\hat{\tau}_j - \tau_j)/\phi_j - \tau_j(\hat{\phi}_j - \phi_j)/\phi_j^2 \quad (8)$$

about (τ_j, ϕ_j), it follows that $\hat{\gamma}_j$ has a normal distribution asymptotically with mean γ_j and a variance that can be consistently estimated with

$$v_{\gamma,j} = \hat{\gamma}_j^2\{(v_{\tau,j}/\hat{\tau}_j^2) - 2(v_{\tau\phi,j}/\hat{\tau}_j\hat{\phi}_j) + (v_{\phi,j}/\hat{\phi}_j^2)\}. \quad (9)$$

The expression (9) is identical to the variance estimator for $\hat{\gamma}_j$ given in Goodman and Kruskal (1963, 1972) and Brown and Benedetti (1977) for contingency tables with ordered categories (see Hardison 1981, p. 33). It also has been implemented in widely used computer procedures such as P4F in BMDP (see Brown 1985) or FREQ in SAS Institute Inc. (1987).

2.2 Multivariate Properties of the $\hat{\gamma}_j$

Since the $\hat{\tau}_j$ and the $\hat{\phi}_j$ for all d conditions in a longitudinal study are jointly U statistics, the elements of their covariance matrix can be consistently estimated in a manner similar in spirit to (6)–(7). More specifically, let

$$\mathbf{F}_i = (U_{i1}, U_{i2}, \ldots, U_{id}, W_{i1}, W_{i2}, \ldots, W_{id})' \quad (10)$$

denote the vector of the components U_{ij} and W_{ij} from (4) and (5) for the ith subject. Then, the sample mean vector

$$\bar{\mathbf{F}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{F}_i = (\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_d, \hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_d)' = (\hat{\boldsymbol{\tau}}', \hat{\boldsymbol{\phi}}')' = \hat{\boldsymbol{\xi}} \quad (11)$$

is a vector of U statistics, which is an unbiased estimator for $\mathbf{E}\{\bar{\mathbf{F}}\} = (\boldsymbol{\tau}', \boldsymbol{\phi}')' = \boldsymbol{\xi}$. In addition, for large n, $\bar{F} = (\hat{\boldsymbol{\tau}}', \hat{\boldsymbol{\phi}}')'$ has, approximately, a multivariate normal distribution with mean vector $\boldsymbol{\xi} = (\boldsymbol{\tau}', \boldsymbol{\phi}')'$ and a covariance matrix that can be consistently estimated with

$$\mathbf{V}_{\bar{\mathbf{F}}} = (4/n^2) \sum_{i=1}^{n} (\mathbf{F}_i - \bar{\mathbf{F}})(\mathbf{F}_i - \bar{\mathbf{F}})'. \quad (12)$$

The vector of Goodman–Kruskal rank correlation coefficients $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_d)'$ consists of the ratios $\hat{\gamma}_j = (\hat{\tau}_j/\hat{\phi}_j)$ of U statistics for the d conditions. Thus, by the same arguments given for the separate $\hat{\gamma}_j$ in Section 2.1, the vector $\hat{\boldsymbol{\gamma}}$ has a multivariate normal distribution asymptotically with mean vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_d)'$ and a covariance matrix for which the linear Taylor series expansions in (8) can be used to obtain a consistent estimator. The matrix expression for this consistent estimator is

$$\mathbf{V}_{\hat{\boldsymbol{\gamma}}} = \mathbf{D}_{\hat{\boldsymbol{\gamma}}}[\mathbf{D}_{\hat{\boldsymbol{\tau}}}^{-1}, -\mathbf{D}_{\hat{\boldsymbol{\phi}}}^{-1}]\mathbf{V}_{\bar{\mathbf{F}}}[\mathbf{D}_{\hat{\boldsymbol{\tau}}}^{-1}, -\mathbf{D}_{\hat{\boldsymbol{\phi}}}^{-1}]'\mathbf{D}_{\hat{\boldsymbol{\gamma}}}, \quad (13)$$

where $\mathbf{D}_{\mathbf{g}}$ denotes the diagonal matrix formed from the vector $\mathbf{g}$. The diagonal elements of $\mathbf{V}_{\hat{\boldsymbol{\gamma}}}$ are the estimators $v_{\gamma,j}$ in (9) for the variances of the $\hat{\gamma}_j$, and the off-diagonal elements are analogous estimators for the covariances between $\hat{\gamma}_j$ and $\hat{\gamma}_{j'}$ for $j \neq j'$.

2.3 Statistical Tests and Model Fitting for $\hat{\gamma}_j$

When the estimators $\hat{\boldsymbol{\gamma}}$ are obtained from a large enough sample size (e.g., $n \geq 40$) for multivariate normal approximations to apply, it is possible to test linear hypotheses concerning them or to fit linear models that describe their variation. For the hypothesis $H_0 : \mathbf{C}\boldsymbol{\gamma} = \mathbf{0}_c$, where $\mathbf{C}$ is a $(c \times d)$ matrix with full rank $c \leq d$ and $\mathbf{0}_c$ is a $(c \times 1)$ vector of 0s, the Wald statistic

$$Q_W(\mathbf{C}\hat{\boldsymbol{\gamma}}) = \hat{\boldsymbol{\gamma}}'\mathbf{C}'[\mathbf{C}\mathbf{V}_{\hat{\boldsymbol{\gamma}}}\mathbf{C}']^{-1}\mathbf{C}\hat{\boldsymbol{\gamma}} \quad (14)$$

has, approximately, a chi-squared distribution with rank$(\mathbf{C}) = c$ df. In this regard, when $\mathbf{C}$ is a $(d-1) \times d$ matrix of contrasts such as $\mathbf{C} = [\mathbf{I}_{(d-1)} - \mathbf{1}_{(d-1)}]$, where $\mathbf{I}_g$ is the $(g \times g)$ identity matrix and $\mathbf{1}_g$ is a $(g \times 1)$ vector of 1s, $Q_W(\mathbf{C}\hat{\boldsymbol{\gamma}})$ in (14) provides a test of group × condition interaction with df $= (d-1)$. In addition, for $\mathbf{C} = \mathbf{1}_d'$, a test statistic with df $= 1$ is obtained for group effects averaged over conditions.

As noted in Koch, Landis, Freeman, Freeman, and Lehnen (1977) and Koch et al. (1985), the Wald statistic $Q_W(\mathbf{C}\hat{\boldsymbol{\gamma}})$ in (14) is identical to the weighted least squares goodness-of-fit statistic

$$Q_W(\mathbf{Z}) = (\hat{\boldsymbol{\gamma}} - \mathbf{Z}\hat{\boldsymbol{\beta}})'\mathbf{V}_{\hat{\boldsymbol{\gamma}}}^{-1}(\hat{\boldsymbol{\gamma}} - \mathbf{Z}\hat{\boldsymbol{\beta}}) \quad (15)$$

for the linear model $\boldsymbol{\gamma} = \mathbf{Z}\boldsymbol{\beta}$, where $\mathbf{Z}$ is a full rank, $d \times (d-c)$ matrix orthogonal to $\mathbf{C}$ and $\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{V}_{\hat{\boldsymbol{\gamma}}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}_{\hat{\boldsymbol{\gamma}}}^{-1}\hat{\boldsymbol{\gamma}}$ is the weighted least squares estimator for the model parameters $\boldsymbol{\beta}$. For models with satisfactory fit, the estimates $\hat{\boldsymbol{\beta}}$ have, approximately, a multivariate normal distribution in large samples with mean vector $\boldsymbol{\beta}$ and a covariance matrix that can be consistently estimated with $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = (\mathbf{Z}'\mathbf{V}_{\hat{\boldsymbol{\gamma}}}^{-1}\mathbf{Z})^{-1}$. In addition, statistical tests for linear hypotheses $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{C}$ has full rank, can be applied with the Wald statistic

$$Q_W(\mathbf{C}\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}'\mathbf{C}'(C\mathbf{V}_{\hat{\boldsymbol{\beta}}}\mathbf{C}')^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}, \quad (16)$$

which has, approximately, a chi-squared distribution with df $=$ rank$(\mathbf{C})$.

For situations where the parameters γ_j are not near the center of the $[-1, 1]$ range, very large sample sizes may be necessary for the Goodman–Kruskal rank correlation coefficients $\hat{\gamma}_j$ to have approximately normal distributions. One way to address this issue is to direct analyses involving Wald statistics and weighted least squares methods at the

Fisher (1925, chap. 6) transformations

$$\hat{\theta}_j = .5 \log_e \{(1 + \hat{\gamma}_j)/(1 - \hat{\gamma}_j)\} \quad (17)$$

of the $\hat{\gamma}_j$. A consistent estimator for the covariance matrix of $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d)$ from methods involving linear Taylor series expansions is $\mathbf{V}_{\hat{\boldsymbol{\theta}}} = D_{\hat{\boldsymbol{\eta}}}^{-1}\mathbf{V}_{\hat{\boldsymbol{\gamma}}}\mathbf{D}_{\hat{\boldsymbol{\eta}}}^{-1}$, where $\mathbf{D}_{\hat{\boldsymbol{\eta}}}$ is a diagonal matrix with the quantities $\hat{\eta}_j = (1 - \hat{\gamma}_j^2)$ on the diagonal.

From the definition of the $\hat{\gamma}_j$ in terms of concordant and discordant pairs of observations, it follows that the $\hat{\theta}_j$ have the form

$$\hat{\theta}_j = .5 \log_e \left\{\frac{\text{number pairs concordant}}{\text{number pairs discordant}}\right\}, \quad (18)$$

so they have an interpretation analogous to log-odds ratios for categorical data. In addition, the $\hat{\theta}_j$ have a symmetric nature relative to concordances and discordances that can enhance applicability of normal approximations to their distributions. The nature of Wald statistics for testing hypotheses concerning $\hat{\boldsymbol{\theta}}$ or weighted least squares procedures for fitting models to them is similar to that previously summarized for $\hat{\boldsymbol{\gamma}}$. For additional discussion of such methods in reference to rank correlation coefficients obtained from ordinal categorical data or related functions of such indexes, see Forthofer and Koch (1973) and Koch et al. (1985).

2.4 Stratification

For many clinical trials, stratified research designs are used; within each stratum, the set of subjects under study is considered to be a simple random sample. A familiar example is the multicenter clinical trial with stratification for the centers at which the study is undertaken. In addition, strata are often based on factors with potentially strong association with response variables (e.g., medical history), and so need to be taken into account in the evaluation of the association between an ordinal set of groups and the response variables.

When the sample sizes n_h for the respective strata $h = 1, 2, \ldots, q$ are large, the methods summarized in Sections 2.1 and 2.2 can be used to construct the within-stratum vectors $\hat{\boldsymbol{\gamma}}_h$ of Goodman–Kruskal rank correlation coefficients and the consistent estimators $\mathbf{V}_{\hat{\boldsymbol{\gamma}},h}$ for their covariance matrices. In addition, $\hat{\boldsymbol{\gamma}}_h$ have, approximately, multivariate normal distributions and are mutually independent. On the basis of these considerations, the methods discussed in Section 2.3 can be used to analyze the compound vector $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_1', \hat{\boldsymbol{\gamma}}_2', \ldots, \hat{\boldsymbol{\gamma}}_q')'$ for variation across both strata and conditions relative to its estimated covariance matrix $\mathbf{V}_{\hat{\boldsymbol{\gamma}}}$ (which has the $\mathbf{V}_{\hat{\boldsymbol{\gamma}},h}$ as diagonal blocks and 0s elsewhere). Results from such analysis can include tests for group × strata interaction and group × strata × condition interaction as well as group effects and group × condition interaction.

Another situation of interest involves several or many strata (e.g., $q \geq 4$) with intermediate sample sizes (e.g., $15 \leq n_h \leq 30$). One way to analyze the association between the grouping variable and the response variables for the large overall sample size $n = \sum_{h=1}^{q} n_h$, with adjustment for the strata, is based on $\bar{\mathbf{F}}_\cdot = \sum_{h=1}^{q} (n_h/n)\bar{F}_h = (\hat{\boldsymbol{\tau}}_\cdot', \hat{\boldsymbol{\phi}}_\cdot')'$, where the $\bar{\mathbf{F}}_h$ are vectors of U statistics with definitions like (11). As a consequence of the linear structure of $\bar{\mathbf{F}}_\cdot$, the vector $\hat{\boldsymbol{\gamma}}_\cdot$ of ratio estimates $\hat{\gamma}_{\cdot j} = (\hat{\tau}_{\cdot j}/\hat{\phi}_{\cdot j})$ has, approximately, a multivariate normal distribution. The estimated covariance matrix $\mathbf{V}_{\hat{\boldsymbol{\gamma}},\cdot}$ for $\hat{\boldsymbol{\gamma}}_\cdot$ is obtained by applying (13) with $\hat{\boldsymbol{\tau}}$ and $\hat{\boldsymbol{\phi}}$ replaced by $\hat{\boldsymbol{\tau}}_\cdot$ and $\hat{\boldsymbol{\phi}}_\cdot$ and $\mathbf{V}_{\bar{\mathbf{F}}}$ replaced by $\mathbf{V}_{\bar{\mathbf{F}},\cdot} = \sum_{h=1}^{q} (n_h/n)^2\mathbf{V}_{\bar{\mathbf{F}},h}$, where $\mathbf{V}_{\bar{\mathbf{F}},h}$ is the counterpart of (12) for the hth stratum. Through $\hat{\boldsymbol{\gamma}}_\cdot$ and $\hat{\mathbf{V}}_{\hat{\boldsymbol{\gamma}},\cdot}$, the methods discussed in Section 2.3 can be used to assess group effects and group × condition interaction in a way that is adjusted for the strata.

The estimators $\hat{\gamma}_{\cdot j}$ are similar in spirit to the matched correlation coefficients discussed in Quade (1974); the latter involve weighted averages of the $\bar{\mathbf{F}}_h$ with weights proportional to the numbers of relevant pairs from the strata rather than the numbers of subjects.

2.5 Missing Data

In most longitudinal studies, some subjects will have missing data for the responses for one or more conditions. When such missing data have low prevalence and are not a noteworthy source of bias, it is reasonable to view them as noninformative. This perspective provides a basis for managing them as causing ties in the pairs where they apply in the definition (1) for the $\hat{\tau}_j$ and the definition (2) for the $\hat{\gamma}_j$; that is, $U_{ii',j}$ in (3) is modified to be 0 when X_i, $X_{i'}$, Y_{ij}; or $Y_{i'j}$ is missing as well as when $(X_i - X_{i'})(Y_{ij} - Y_{i'j}) = 0$. All other aspects of the analysis of the $\hat{\gamma}_j$ are the same as discussed in Sections 2.1–2.4.

2.6 Simulation Study

A small simulation study was undertaken to evaluate the properties of statistical tests based on the Goodman–Kruskal rank correlation coefficients $\hat{\gamma}_j$ or their Fisher transformed values $\hat{\theta}_j$ from (17). Its design was focused on two groups of patients, for whom bivariate normally distributed response variables were observed at two visits, and so pertains to an ideal setting. The per group sample size was 15 or 30 (so the overall sample size was 30 or 60). The bivariate normal distributions had specified correlations $\rho = .3, .5$, and .7, unit variances, and specified means corresponding to the following three situations:

1. Situation 1 had 0 means for both groups at both visits (a null hypothesis situation).
2. Situation 2 had 0 means for Group 1 at both visits and means = .5 for Group 2 at both visits (a group effect situation).
3. Situation 3 had 0 means for Group 1 at both visits and at Visit 1 for Group 2 and mean = 1 at Visit 2 for Group 2 (an interaction situation).

The bivariate normal random variables were generated with the normal function in the SAS/IML Software of SAS Institute Inc. (1985). In each of the 18 combinations of sample size, situation for means, and correlation, the $\hat{\gamma}_j$, the $\hat{\theta}_j$, and their corresponding estimated covariance

matrices were constructed from each of 1,000 sets of simulated bivariate data with matrix computations in SAS/IML. For the tests of group × visit interaction, p values were based on Wald statistics like (14) for $(\hat{\gamma}_1 - \hat{\gamma}_2)$ and $(\hat{\theta}_1 - \hat{\theta}_2)$; for the tests of average group effects, they were based on Wald statistics for $(\hat{\gamma}_1 + \hat{\gamma}_2)$ and $(\hat{\theta}_1 + \hat{\theta}_2)$. The prevalences of $p \leq .050$ for these tests for the respective specifications in the simulation are shown in Table 1. For Situation 1, where there are no group effects and no group × visit interaction, the prevalences of $p \leq .050$ for the tests based on the $\hat{\gamma}_j$ tend to exceed the nominal .050 value to a small extent, but those for the tests based on the $\hat{\theta}_j$ are reasonably near .050. Similar comments apply to the tests of no group × visit interaction in Situation 2. The other specifications in Table 1 correspond to tests in the presence of group effects or group × visit interaction, and the prevalences of $p \leq .050$ represent powers. It can be verified that these powers are similar to those of the analogous t tests for group effects and group × visit interaction; in addition, such agreement is somewhat better for the tests based on the $\hat{\theta}_j$. In summary, the results of the study suggest that statistical tests based on the $\hat{\gamma}_j$ and the $\hat{\theta}_j$ have reasonable behavior, although the properties of the tests based on the $\hat{\theta}_j$ seem somewhat better.

3. EXAMPLES

Two examples are used to illustrate aspects of the application of the methodology in Section 2. The first is based on a multicenter study for a respiratory illness and has complete data for the response variable at four visits. The second is based on a study for a skin disorder and has three visits for which some patients have missing data for the response variable.

3.1 A Multicenter Study

This example is based on data given in Koch, Carr, Amara, Stokes, and Uryniak (1989) for a multicenter study of $n = 111$ patients who had a respiratory illness. Its research design had stratification for two centers (Clinic 1 and Clinic 2) within which patients were randomly assigned to one of two treatments (placebo = 0, active = 1). The response variables were ordinal classifications (0 = terrible, 1 = poor, 2 = fair, 3 = good, 4 = excellent) of the health status of each patient at each of four successive visits after the initiation of treatment. The baseline status (i.e., prior to treatment) of patients was also evaluated according to the same ordinal classification.

For Clinic 1 and Clinic 2, the samples sizes $n_1 = 56$ and $n_2 = 55$ are moderately large. Within each of them, the vectors $\hat{\boldsymbol{\gamma}}_h$ of Goodman–Kruskal rank correlation coefficients between treatment and the response variables and their corresponding estimated covariance matrices $\mathbf{V}_{\hat{\gamma},h}$ are constructed by the methods in Section 2. These results are shown in Table 2. Since treatment is a dichotomous grouping variable, the $\hat{\boldsymbol{\gamma}}_h$ are similar in spirit to Mann–Whitney (1947) statistics by being indicative of the extent to which a nontied pair of subjects from the active and placebo groups is more likely to have the more favorable response occur for the active subject than for the placebo subject. Thus they can be interpreted as measure of treatment effect.

Statistical tests concerning variation of the elements of $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_1', \hat{\boldsymbol{\gamma}}_2')'$ across both strata and conditions can be applied with the Wald statistics in (14). In Table 3, the $\mathbf{C}$ matrices, Wald statistics, and p values are given for tests concerning treatment × visit interaction and treatment × center interaction. The results of these tests indicate that treatment × visit interaction is significant ($p = .044$) for the combined centers. The nature of this interaction corresponds to stronger associations between treatment and response at Visits 2 and 3 than Visits 1 and 4. The test statistics for treatment × center interaction and treatment × center × visit interaction were both nonsignificant ($p \geq .100$), so the $\hat{\boldsymbol{\gamma}}_j$ can be interpreted as homogeneous across the strata (i.e., $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2$ applies for the parameters that they estimate). Since all of the elements of $\hat{\boldsymbol{\gamma}}$ are

Table 1. Prevalences of $p \leq .050$ From Simulation Study of Rank Measures of Association to Evaluate Statistical Tests for Average Group Effects and Group × Visit Interaction

			$\rho = .3$		$\rho = .5$		$\rho = .7$	
Measure[a]	Situation[b]	Sample size	Average	Interaction	Average	Interaction	Average	Interaction
I	1	30	.077	.068	.057	.055	.066	.065
I	1	60	.061	.048	.049	.054	.072	.060
I	2	30	.445	.045	.394	.051	.355	.041
I	2	60	.668	.050	.615	.071	.555	.051
I	3	30	.439	.584	.369	.717	.337	.867
I	3	60	.641	.880	.607	.945	.535	.997
II	1	30	.059	.060	.037	.052	.038	.061
II	1	60	.052	.041	.039	.053	.059	.060
II	2	30	.392	.040	.336	.046	.288	.039
II	2	60	.645	.048	.581	.064	.506	.049
II	3	30	.365	.560	.316	.707	.271	.863
II	3	60	.612	.875	.579	.942	.503	.996

[a] Measure I is the Goodman–Kruskal rank correlation coefficient and Measure II is its Fisher transformed value.

[b] Situation 1 corresponds to 0 means for two groups at two visits; Situation 2 corresponds to 0 means for Group 1 at two visits and means = .5 for Group 2 at two visits; and Situation 3 corresponds to 0 means for Group 1 at two visits and Group 2 at Visit 1 and mean = 1 for Group 2 at Visit 2. In all cases, variances = 1 and between-visit correlations = ρ as indicated. For each of the 18 combinations of Situations 1–3, $\rho = .3, .5, .7$, and sample size = 30, 60, there were 1,000 replicates.

Table 2. Goodman–Kruskal Rank Correlation Coefficients Between Treatment and Response, Standard Errors, and Estimated Covariance Matrix for Respiratory Illness Study

Clinic	Visit	Rank correlation estimates	Standard error	Estimated covariance matrix			
1	1	.148	.194	.0375	.0201	.0225	.0206
1	2	.467	.163		.0265	.0162	.0172
1	3	.292	.182			.0332	.0244
1	4	.242	.188				.0352
2	1	.421	.185	.0340	.0136	.0127	.0144
2	2	.647	.137		.0189	.0164	.0166
2	3	.559	.153			.0234	.0189
2	4	.433	.182				.0330

positive, the overall evaluation of the association between group and response can be reasonably based on the average index $\hat{\bar{\gamma}} = (\mathbf{1}_8'\hat{\boldsymbol{\gamma}}/8) = .401$ and its standard error, $SE(\hat{\bar{\gamma}}) = (\mathbf{1}_8'\mathbf{V}_{\hat{\gamma}}\mathbf{1}_8/64)^{1/2} = .102$. The extent to which $\hat{\bar{\gamma}}$ is greater than 0 is strongly significant ($p < .001$), so one can interpret active treatment as providing more favorable response than placebo. In view of the treatment × visit interaction identified, such tendencies are stronger at Visits 2 and 3 than Visits 1 and 4.

An alternative approach for this example is based on the use of the stratification adjusted estimators $\hat{\boldsymbol{\gamma}}_{\cdot}$ discussed in Section 2.4 and their estimated covariance matrix $\mathbf{V}_{\hat{\gamma}_{\cdot}}$. Considerations of these results, which are shown in Table 4, is reasonable because of the previously noted homogeneity of the $\hat{\boldsymbol{\gamma}}_h$ across the strata. Application of the Wald statistic in (14) to test group × visit interaction for $\hat{\boldsymbol{\gamma}}_{\cdot}$ via $\mathbf{C} = [\mathbf{I}_3 \; -\mathbf{1}_3]$ yields $Q_W(\mathbf{C}\hat{\gamma}_{\cdot}) = 8.20$, which is significant ($p = .042$) relative to the chi-squared distribution with df = 3. In addition, the average index $\hat{\bar{\gamma}}_{\cdot} = (\mathbf{1}_4'\hat{\boldsymbol{\gamma}}_{\cdot}/4) = .397$, relative to its standard error, $SE(\hat{\bar{\gamma}}_{\cdot}) = (\mathbf{1}_4'\mathbf{V}_{\hat{\gamma}_{\cdot}}\mathbf{1}_4/16)^{1/2} = .103$, is significantly ($p < .001$) greater than 0. Thus the conclusions for the stratification adjusted estimators $\hat{\boldsymbol{\gamma}}_{\cdot}$ are essentially the same as those previously given for the within-stratum estimators $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_1', \hat{\boldsymbol{\gamma}}_2')'$.

A straightforward refinement of the previously described methods of analysis enable the role of one or more concomitant variables, such as baseline status, to be taken into account. As a consequence of randomization, no association is expected between treatment and baseline status. In other words, randomization implies compatibility of the data from this example with the hypotheses $H_0 : \gamma_{h0} = 0$, where the γ_{h0} denote the parameters corresponding to the Goodman–Kruskal rank correlation coefficients $\hat{\gamma}_{h0}$ between treatment and baseline status. The $\hat{\gamma}_{h0}$ and their standard errors are shown in Table 5, where their support for H_0 is evident in the nonsignificance ($p \geq .100$) of their differences from 0. Also in Table 5 are the covariances between the $\hat{\gamma}_{h0}$ and the $\hat{\gamma}_{hj}$ in Table 2 for the rank correlations between treatment and the responses at the visits $j = 1, 2, 3, 4$. These covariances reflect the extent to which the association between treatment and the responses at Visits 1–4 is related to the association between baseline status and the responses at Visits 1–4. Their nature can be taken into account through covariance adjustments along the lines discussed in Koch, Amara, Davis, and Gillings (1982). These methods involve the use of weighted least squares to fit a linear model that incorporates $H_0 : \gamma_{h0} = 0$ for the augmented compound vector

$$\boldsymbol{\gamma}_A = (\gamma_{10}, \gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}, \gamma_{20}, \gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{24})' \quad (19)$$

of rank correlation coefficients pertaining to both baseline and the responses at Visits 1–4 for both strata. One such model, which also incorporates the conditions of no treatment × center interaction and no treatment × center × visit interaction on the basis of the results in Table 3, has the specification

$$\boldsymbol{\gamma}_A = [\mathbf{0}_4\mathbf{I}_4\mathbf{0}_4\mathbf{I}_4]'\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\beta}, \quad (20)$$

where $\mathbf{0}_4$ denotes the (4 × 1) vector of 0s and $\mathbf{I}_4$ denotes the (4 × 4) identity matrix. In (20), $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)'$ is the vector of parameters corresponding to the covariance adjusted rank correlations for the association between treatment and the responses at Visits 1–4. An important aspect of models like (20) is that they enable adjustment for an ordinal concomitant variable under minimal assumptions (i.e., simple random sampling) without further stratification and its corresponding thinning of the sample size to degrees that would weaken the applicability of normal approximations. In addition, adjustment for more than one concomitant variable is feasible as long as there is no association between them and treatment.

The goodness of fit of the model in (20) is evaluated with the Wald statistic in (15). Since $Q_W(\mathbf{Z}) = 1.68$ is nonsignificant ($p = .947$) relative to the chi-squared distribution with df = 6, usage of the model (20) is considered reasonable. The weighted least squares estimates for the

Table 3. Wald Statistics for Tests Concerning Treatment × Visit Interaction and Treatment × Center Interaction for Respiratory Illness Study

Source of variation	C matrix*	Degrees of freedom	Wald statistic	Approximate p value
Treatment × center × visit	$[\mathbf{I}_3 \; -\mathbf{1}_3 \; -\mathbf{I}_3 \; \mathbf{1}_3]$	3	.36	.949
Treatment × visit (average over centers)	$[\mathbf{I}_3 \; -\mathbf{1}_3 \; \mathbf{I}_3 \; -\mathbf{1}_3]$	3	8.12	.044
Treatment × center (average over visits)	$[\mathbf{1}_4' \; -\mathbf{1}_4']$	1	1.24	.265
Treatment × center (all visits jointly)	$[\mathbf{I}_4 \; -\mathbf{I}_4]$	4	1.52	.823

* $\mathbf{I}_g$ denotes the (g × g) identity matrix and $\mathbf{1}_g$ denotes the (g × 1) vector of 1s.

Table 4. Stratification Adjusted Rank Correlation Coefficients, Standard Errors, and Estimated Covariance Matrix for Respiratory Illness Study

Visit	Rank correlation estimates	Standard errors	Estimated covariance matrix			
1	.277	.135	.0183	.0085	.0090	.0090
2	.557	.107		.0114	.0082	.0085
3	.423	.120			.0144	.0111
4	.333	.131				.0173

parameters $\boldsymbol{\beta}$ and their corresponding standard errors are as follows:

Visit	1	2	3	4
$\hat{\beta}_j$	.290	.557	.443	.342
$SE(\hat{\beta}_j)$	.107	.089	.101	.117

(21)

Since the $\hat{\boldsymbol{\beta}}_j$ are covariance adjusted rank correlation coefficients for the respective visits, the hypothesis of no treatment × visit interaction can be tested through the Wald statistic in (16) with $\mathbf{C} = [\mathbf{I}_3, -\mathbf{1}_3]$. The result of this test, $Q_W(\mathbf{C}\hat{\boldsymbol{\beta}}) = 9.44$, is significant ($p = .024$) relative to the chi-squared distribution with df = 3. In addition, the overall association between group and response is significantly ($p < .001$) indicated through the extent to which the average index $\bar{\hat{\boldsymbol{\beta}}} = (\mathbf{1}_4'\hat{\boldsymbol{\beta}}/4) = .408$ is greater than 0 relative to its estimated standard error, $SE(\bar{\hat{\boldsymbol{\beta}}}) = (\mathbf{1}_4' V_{\hat{\boldsymbol{\beta}}}\mathbf{1}_4/16)^{1/2} = .082$, where $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$ is the estimate for the covariance matrix of $\hat{\boldsymbol{\beta}}$ given in Section 2.3.

A noteworthy feature of the covariance adjusted rank correlation coefficients $\hat{\beta}_j$ in (21) is their tendency to have smaller standard errors than their stratification adjusted counterparts $\hat{\gamma}_{\cdot j}$ in Table 4. A reason for this gain in precision is their determination by a method that takes the association between baseline and the responses at Visits 1–4 into account. The strong nature of such association is reflected by a range of .400 to .800 for the Goodman–Kruskal rank correlation coefficients between baseline and the responses at Visits 1–4 for the two treatment groups (placebo and active) relative to standard errors in the range of .075 to .150. In addition, for these rank correlation coefficients, which pertain to pairs of multicategory ordinal variables, application of the methods used to obtain Table 3 yielded results that indicated no baseline × treatment interaction and significant baseline × visit interaction.

In summary, analyses of rank correlation coefficients between treatment and responses at four visits were applied to this example in a direct way, with stratification adjustment and with covariance adjustment. Their results indicated significant ($p < .050$) treatment × visit interaction. In addition, the association between treatment and response was interpreted as significant ($p < .010$) on the basis of the average rank correlation over visits. Such association corresponded to more favorable response for patients receiving active treatment; the tendency for this type of association to be stronger at Visits 2 and 3 than Visits 1 and 4 was the source of the treatment × visit interaction. These conclusions agree with those given in Koch et al. (1989) for other methods of analysis for this example.

Table 5. Goodman–Kruskal Rank Correlation Coefficients Between Treatment and Baseline, Standard Errors, and Covariance Matrix Estimates Relative to Visits 1–4 for Respiratory Illness Study

Clinic	Rank correlation estimates	Standard errors	Components of estimated covariance matrix relative to Visits 1–4				
1	−.115	.201	.0404	.0275	.0205	.0207	.0190
2	.122	.201	.0404	.0126	.0049	.0063	.0096

3.2 A Study With Missing Data

Data from a study to compare two treatments (test drug = 1, placebo = 2) for $n = 172$ patients with a skin disorder are given in Stanish, Gillings, and Koch (1978), where they are used to illustrate analyses based on multivariate sets of ratios of means. The responses of patients to their assigned treatment were evaluated at three follow-up visits in terms of the ordinal categories of rapidly improving (1), slowly improving (2), stable (3), slowly worsening (4), and rapidly worsening (5). As often occurs in longitudinal studies, however, a few patients had missing data for at least one of the visits. In particular, the number of patients with missing data at the respective visits were as follows for the two treatments:

	Visit 1	Visit 2	Visit 3	Sample size
Test drug	3	8	9	88
Placebo	0	8	21	84

(22)

Although there were six investigators in this study, to simplify the discussion, the analyses illustrated here presume that the patients are representative of a single population; that is, no stratification adjustment (see Sec. 2.4) is applied, but it could be.

Through the strategies outlined in Section 2.5 for the management of missing data, the vector $\hat{\boldsymbol{\gamma}}$ of Goodman–Kruskal rank correlation coefficients between treatment and the response variables and their corresponding estimated covariance matrix $\mathbf{V}_{\hat{\boldsymbol{\gamma}}}$ are constructed. These results are shown in Table 6. Application of the Wald statistic in (14) to test group × visit interaction for $\hat{\boldsymbol{\gamma}}$ via $\mathbf{C} = [\mathbf{I}_3, -\mathbf{1}_3]$ yields $Q_W(\mathbf{C}\hat{\boldsymbol{\gamma}}) = 3.24$, which is nonsignificant ($p = .198$) relative to the chi-squared distribution with df = 2. Since the $\hat{\gamma}_j$ are similar for the three visits, their average $\bar{\hat{\gamma}} = (\mathbf{1}_3'\hat{\boldsymbol{\gamma}}/3) = .778$ can be reasonably used to evaluate the association between group and response. Relative to $SE(\bar{\hat{\gamma}}) = (\mathbf{1}_3'\mathbf{V}_{\hat{\boldsymbol{\gamma}}}\mathbf{1}_3/9)^{1/2} = .052$, such association is significant

Table 6. Goodman–Kruskal Rank Correlation Coefficients Between Treatment and Response, Standard Errors, and Estimated Covariance Matrix for Skin Disorder Study

Visit	Rank correlation estimates	Standard errors	Estimated covariance matrix		
1	.717	.068	.004578	.002465	.002051
2	.810	.055		.003067	.002399
3	.807	.057			.003280

($p < .001$) and is indicative of the strong tendency for the responses to test drug to be more favorable than the responses to placebo.

Since the $\hat{\gamma}_j$ in Table 6 are far from the center of the $[-1, 1]$ range, analyses of the Fisher transformed value $\hat{\theta}_j$ from (17) are also of interest. For the respective visits, the $\hat{\theta}_j$ and their estimated standard errors are as follows:

	Visit 1	Visit 2	Visit 3	
$\hat{\theta}_j$	.90	1.13	1.12	(23)
$SE(\hat{\theta}_j)$	.14	.16	.16	

For $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)'$, the Wald statistic to test group × visit interaction is $Q_W(\mathbf{C}\hat{\boldsymbol{\theta}}) = 3.37$; it is nonsignificant ($p = .186$) and is nearly the same as its previously discussed counterpart for $\hat{\boldsymbol{\gamma}}$. The across-visit average of the $\hat{\theta}_j$ is $\bar{\hat{\theta}} = (\mathbf{1}_3'\hat{\boldsymbol{\theta}}/3) = 1.05$, and $SE(\bar{\hat{\theta}}) = (\mathbf{1}_3'\mathbf{V}_{\hat{\boldsymbol{\theta}}}\mathbf{1}_3/9)^{1/2} = .14$. Thus $\{\bar{\hat{\theta}}/SE(\bar{\hat{\theta}})\} = 7.50$ is substantially smaller than $\{\bar{\hat{\gamma}}/SE(\bar{\hat{\gamma}})\} = 14.96$; it is also more compatible with the results from other tests indicating the significant association between group and response. Thus the $\hat{\theta}_j$ are interpreted as providing a more appropriate basis of analysis for this example than the $\hat{\gamma}_j$. This perspective is also in agreement with the results of the simulation work reported in Section 2.6.

[*Received April 1988. Revised October 1988.*]

REFERENCES

Brown, M. B. (1985), "Frequency Tables," in *BMDP Statistical Software Manual*, ed. W. J. Dixon, Los Angeles: University of California Press, pp. 143–206.

Brown, M. B., and Benedetti, J. K. (1977), "Sampling Behavior of Tests for Correlation in Two-Way Contingency Tables," *Journal of the American Statistical Association*, 72, 309–315.

Davis, C. E., and Quade, D. (1968), "On Comparing the Correlations Within Two Pairs of Variables," *Biometrics*, 24, 987–995.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, London: Oliver & Boyd.

Forthofer, R. N., and Koch, G. G. (1973), "An Analysis for Compounded Functions of Categorical Data," *Biometrics*, 29, 143–157.

Friedman, M. (1937), "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *Journal of the American Statistical Association*, 32, 675–699.

Gill, J. L. (1978), *Design and Analysis of Experiments in the Animal and Medical Sciences*, Ames: Iowa State University Press.

Goodman, L. A., and Kruskal, W. H. (1963), "Measures of Association for Cross Classifications III: Approximate Sampling Theory," *Journal of the American Statistical Association*, 58, 310–364.

——— (1972), "Measures of Association for Cross Classifications IV: Simplification of Asymptotic Variances," *Journal of the American Statistical Association*, 67, 415–421.

Hardison, C. D. (1981), "Small-Sample Properties of a Family of Nonparametric Partial Correlation Measures," Mimeo 1372, University of North Carolina at Chapel Hill, Institute of Statistics.

Hoeffding, W. (1948), "A Class of Statistics With Asymptotically Normal Distribution," *The Annals of Mathematical Statistics*, 19, 293–325.

Kendall, M. G. (1962), *Rank Correlation Methods* (3rd ed.), New York: Hafner.

Koch, G. G. (1969), "Some Aspects of the Statistical Analysis of 'Split Plot' Experiments in Completely Randomized Layouts," *Journal of the American Statistical Association*, 64, 485–505.

——— (1970), "The Use of Non-parametric Methods in the Statistical Analysis of a Complex Split Plot Experiment," *Biometrics*, 26, 105–128.

Koch, G. G., Amara, I. A., Davis, G. W., and Gillings, D. B. (1982), "A Review of Some Statistical Methods for Covariance Analysis of Categorical Data," *Biometrics*, 38, 563–595.

Koch, G. G., Amara, I. A., Stokes, M. E., and Gillings, D. B. (1980), "Some Views on Parametric and Non-parametric Analysis for Repeated Measurements and Selected Bibliography," *International Statistical Review*, 48, 249–265.

Koch, G. G., Carr, G. J., Amara, I. A., Stokes, M. E., and Uryniak, T. J. (1989), "Categorical Data Analysis," in *Statistical Methodology in the Pharmaceutical Sciences*, ed. D. A. Berry, New York: Marcel Dekker, pp. 391–475.

Koch, G. G., and Edwards, S. (1988), "Clinical Efficacy Trials With Categorical Data," in *Biopharmaceutical Statistics for Drug Development*, ed. K. E. Peace, New York: Marcel Dekker, pp. 403–457.

Koch, G. G., Elashoff, J. D., and Amara, I. A. (1988), "Repeated Measurements—Design and Analysis," in *Encyclopedia of Statistical Sciences* (Vol. 8), eds. S. Kotz and N. L. Johnson, New York: John Wiley, pp. 46–73.

Koch, G. G., Imrey, P. B., Singer, J. M., Atkinson, S. S., and Stokes, M. E. (1985), *Analysis of Categorical Data*, Montreal: Les Presses de l'Universite de Montreal.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H., Jr., and Lehnen, R. G. (1977), "A General Methodology for the Analysis of Experiments With Repeated Measurement of Categorical Data," *Biometrics*, 33, 133–158.

Kruskal, W. H., and Wallis, W. A. (1953), "Use of Ranks in One Criterion Variance Analysis," *Journal of the American Statistical Association*, 46, 583–621.

Landis, J. R., Miller, M. E., Davis, C. S., and Koch, G. G. (1988), "Some General Methods for the Analysis of Categorical Data in Longitudinal Studies," *Statistics in Medicine*, 7, 109–137.

Mann, H. B., and Whitney, D. R. (1947), "On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other," *The Annals of Mathematical Statistics*, 18, 50–60.

McCullagh, P. (1980), "Regression Models for Ordinal Data" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 42, 109–142.

Morrison, D. F. (1976), *Multivariate Statistical Methods* (2nd ed.), New York: McGraw-Hill.

Quade, D. (1974), "Nonparametric Partial Correlation," in *Measurement in the Social Sciences: Theories and Strategies*, ed. H. M. Blalock, Jr., Chicago: Aldine, pp. 369–398.

——— (1982), "Nonparametric Analysis of Covariance by Matching," *Biometrics*, 38, 597–611.

SAS Institute Inc. (1985), *SAS/IML User's Guide* (Version 5 ed.), Cary, NC: Author.

——— (1987), "The FREQ Procedure," in *SAS/STAT Guide for Personal Computers* (Version 6 ed.), Cary, NC: Author, pp. 519–548.

Sen, P. K. (1960), "On Some Convergence Properties of *U*-Statistics," *Calcutta Statistical Association Bulletin*, 10, 1–18.

Serfling, R. (1980), *Approximation Theorems in Mathematical Statistics*, New York: John Wiley.

Stanish, W. M., Gillings, D. B., and Koch, G. G. (1978), "An Application of Multivariate Ratio Methods for the Analysis of a Longitudinal Clinical Trial With Missing Data," *Biometrics*, 34, 305–317.

Stram, D. O., Wei, L. J., and Ware, J. H. (1988), "Analysis of Repeated Ordered Categorical Outcomes With Possibly Missing Observations and Time-Dependent Covariates," *Journal of the American Statistical Association*, 83, 631–637.