STATISTICS IN MEDICINE, VOL. 16, 2599–2618 (1997)

TUTORIAL IN BIOSTATISTICS

A REVIEW OF TESTS FOR DETECTING A MONOTONE DOSE–RESPONSE RELATIONSHIP WITH ORDINAL RESPONSE DATA

CHRISTY CHUANG-STEIN [1]* AND ALAN AGRESTI [2]

[1] *Clinical Development Biostatistics, Pharmacia & Upjohn Company, Kalamazoo, MI 49001, U.S.A.*
[2] *Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, U.S.A.*

SUMMARY

This tutorial reviews methods for testing independence between discrete levels of a dose and an ordered categorical response variable. The tests are designed to be powerful for cases in which the response improves monotonically as dosage level increases. First, we show how to apply some standard tests for doubly-ordered contingency tables. Then, we show how to construct tests as part of a model-building strategy. Other topics discussed include generalizations to stratified data, small-sample methods, and sample size and power considerations. © 1997 by John Wiley & Sons, Ltd.

Statist. Med., **16**, 2599–2618 (1997)
No. of Figures: 0 No. of Tables: 5 No. of References: 50

1. INTRODUCTION

The exploration of dose–response relationships is the focus of many studies in toxicology[1] and genetic toxicology.[2] This topic occupies an equally important place in animal growth promotion studies[3] and in the pre-marketing clinical testing of new drugs. In a typical pre-marketing dose–response study, a control and several doses of the drug are randomly assigned to the study subjects, with each subject receiving only one dose throughout the study (this is called the parallel-group design). The control is usually a placebo that provides necessary background information.

In dose–response studies, the response can either measure the efficacy of a treatment or the risk (side-effect) associated with an exposure. The exposure might be that to a new medication or to a risk factor such as cigarette smoking, with the dose quantifying the amount of exposure.

Statisticians have used the phrase *dose–response relationship* to represent a variety of things. Some refer to the shape of the exposure–outcome curve, no matter what that shape may be.[4] For

* Correspondence to: Christy Chuang-Stein, Clinical Development Biostatistics, Pharmacia & Upjohn Company, Kalamazoo, MI 49001, U.S.A. E-mail: jcchuang@au.pnu.com.

CCC 0277–6715/97/222599–20$17.50

Table I. Responses on the Glasgow Outcome Scale from a clinical trial with a placebo (control) and three treatment groups labelled as low dose, medium dose and high dose

Treatment group	Glasgow Outcome Scale					Total
	Death	Vegetative state	Major disability	Minor disability	Good recovery	
Placebo	59	25	46	48	32	210
Low dose	48	21	44	47	30	190
Medium dose	44	14	54	64	31	207
High dose	43	4	49	58	41	195

some, the objectives of a study assessing exposure-associated risk are to demonstrate a continuously increasing risk with increasing exposure.[5] Recently, other shapes have received attention, such as the umbrella pattern[6–10] and a plateaued drug effect beyond a certain level. Among the potential shapes, the monotone one is by far the most commonly discussed in the literature.

A rich literature exists on the exploration of dose–response relationships for the parallel-group design. The literature refers almost exclusively, however, to normal or binary response variables. The main purpose of this article is to summarize methods one can apply for ordinal responses – that is, responses measured with a set of ordered categories. Most of these methods were originally proposed for other applications, but are appropriate for dose–response relationships. The article assumes that the reader has familiarity with basic ideas of statistical inference, regression and ANOVA modelling, and chi-squared tests. Section 4 also assumes previous exposure to logit modelling, but otherwise the article does not require previous background in specialized methods for categorical response data.

Ruberg[11,12] noted that dose–response studies routinely ask four questions: (i) Is there any evidence of a drug effect? (ii) Which doses exhibit a response different from the control response? (iii) What is the nature of the dose–response relationship? (iv) Which is the optimal dose? One approaches the questions in this order, the later ones being more specific. In the drug development process, information obtained from dose–response studies is often used to select doses for subsequent confirmatory registration trials.

This article primarily focuses on the first question. Clearly, though, answers to the other questions are ultimately more informative. Though we occasionally refer to them as well, for lack of space we defer a detailed account to a follow-up paper. We summarize methods designed to detect an effect on an ordinal response when there is prior belief of a monotone dose–response relationship, expressed in the vague notion that a higher dose tends to produce a more desirable outcome. This prior belief relates to a monotone alternative to the 'no effect' hypothesis. We present the methods in the context of efficacy evaluation, though they also apply to risk assessment with a reversal in the direction of association. Strict monotonicity is not required, and we use 'monotone' interchangeably with 'non-decreasing'. This includes cases having only a high-dose effect or a constant drug effect at all the non-zero dose levels.[13]

Table I illustrates the type of data considered in this article. In Table I, five ordered categories ranging from 'death' to 'good recovery' describe the clinical outcome of patients who experienced trauma. In the literature on critical care, these five categories are often called the Glasgow Outcome Scale (GOS). Table I includes four treatment groups, with a vehicle infusion serving as the control. The three intravenous doses for the investigational medication are labelled as low,

medium and high. The original data have been modified somewhat to protect the identity of the trial. One study objective was to determine whether a more favourable GOS outcome tends to occur as the dose increases. This example and others in this article deal with fixed doses determined prior to the studies. As a result, levels of dose are treated as fixed rather than random. This is natural, since a pharmaceutical sponsor needs to justify its choice of the recommended dose in the new drug application for the compound. Furthermore, the availability of dosing strengths is often influenced and limited by manufacturing considerations.

Though Table I is simply a two-way contingency table, standard tests of independence for two-way tables such as the Pearson chi-squared test are inappropriate for testing against an ordered alternative. Those tests treat both classifications as nominal scale (unordered). When a monotone trend truly exists, methods designed to detect it are more powerful than such nominal-level procedures.

One possible approach dichotomizes the response and employs methods for binary responses. This approach is reasonable if the response categories are clearly divided into desirable and undesirable groups. Otherwise, this approach suffers from the lack of a clear choice for the collapsing and, as we see in Section 6, a loss of information and power.

The organization of this article is as follows: Sections 2 and 3, the heart of the article, present several tests of independence between dose and an ordinal response that are sensitive to the alternative of a monotone relationship. Section 2 discusses non-model-based tests while Section 3 focuses on model-based inference. Section 4 mentions generalizations to handle stratification. Section 5 discusses small-sample and sparse-data inference, and Section 6 comments on sample size and power. The final section summarizes and provides recommendations for conducting such an analysis.

2. SIGNIFICANCE TESTS FOR A MONOTONE DOSE–RESPONSE RELATIONSHIP

This section reviews significance tests for detecting monotone dose–response relationships. Section 3 discusses related tests for models for the relationship. Non-model-based inference, though less informative, is often considered simpler from a regulatory perspective because the tests do not need to validate any modelling assumptions;[14] however, we shall note in Section 3 that some tests from this section are equivalent to tests for certain models. This section presents four approaches: (i) tests based on association measures, including generalized Cochran–Mantel–Haenszel procedures; (ii) an adaptation of the Jonckheere–Terpstra test; (iii) adaptations of methods for continuous responses, including order-restricted inference; and (iv) treating the response distributions as survival distributions.

Let I denote the number of treatment groups, and let J denote the number of categories of the response variable, which is denoted by Y. Let x_{ij} denote the number of individuals in the ith treatment group whose response falls in the jth category, let $n_i = \sum_j x_{ij}$ denote the number of subjects in that group and let $N = \sum n_i$ denote the total sample size. We treat the counts in separate rows as independent multinomial samples. We arrange the I treatment groups from the lowest ($i = 1$) to the highest dose group ($i = I$), with d_i representing the dose level for the ith group, and the response categories from the least favourable ($j = 1$) to the most favourable ($j = J$).

Let Y_i denote a response at dose i. Let $F_{ij} = P(Y_i \leqslant j)$. The null hypothesis of no difference among the I treatment groups is

$$H_0 : F_{1j} = F_{2j} = \cdots = F_{Ij} \text{ for all } j. \tag{1}$$

One way to operationalize the alternative hypothesis of 'monotone dose-response relationship' is in terms of a monotone *stochastic ordering* among the I cumulative distributions. This means that

$$H_1 : F_{1j} \geqslant F_{2j} \geqslant \cdots \geqslant F_{Ij} \text{ for all } j \tag{2}$$

with strict inequality for at least one j. Since higher response categories represent more favourable outcomes, this alternative implies a tendency for more favourable outcomes as the dose increases.

Since significance tests relate to particular hypotheses, they are confirmatory rather than exploratory in nature. The use of significance tests relates not to exploring the nature of the dose–response relationship, but rather to determining the probability of results at least as extreme as those observed in the direction of a monotone relationship, if the variables were truly independent. As a result, testing for a monotone relationship only makes sense when the pharmacology of the drug suggests that, within the safety limits, higher drug exposure results in efficacy that is at least as good as that at lower exposure. We suggest combining such formal analyses with informal checks of this prior belief, such as by plotting sample cumulative distributions. The modelling approaches in Section 3 have the advantage of a built-in goodness-of-fit check.

2.1. Tests based on association measures

Table I has ordered rows (the doses) and ordered columns (the ordinal response). The doses are quantitative, and one can treat the response scale as quantitative by assigning scores to the categories. Correlation-type association measures then summarize the linear component of the dose–response relationship. This strategy is reasonable if one expects roughly a linear trend on the chosen scales. Yates[15] presented a large-sample single-degree-of-freedom chi-squared test statistic based on this approach, essentially squaring the ratio of the sample correlation to its standard error. To form a P-value for the one-sided alternative of a positive trend, we use the signed square root of this statistic and refer to the right-hand tail probability from the standard normal curve. For binary responses, the closely related *Cochran–Armitage*[16] *test* is designed to detect a linear trend in a response proportion. Mantel[17] extended Yate's test to the stratified case.

A potential disadvantage of this strategy is the necessity of assigning scores. Normally, one would assign the actual dosage level or the log dose to the dose categories. Usually, the choice of response scores has little effect on the conclusion about whether an effect exists. It may have an effect, however, when the data are highly unbalanced, such as when some categories have many more observations than other categories.[18]

An alternative approach with correlation measures avoids the responsibility of selecting scores and uses the data to form them automatically. Specifically, one assigns ranks to the subjects and uses them as the category scores. For all subjects in a category, one assigns the average of the ranks that would apply for a complete ranking of the sample. These are called *midranks*. Let $x_{+j} = \sum_i x_{ij}$ denote the number of subjects in the sample who make response j. The midrank for category j equals

$$w_j = x_{+1} + \cdots + x_{+,j-1} + x_{+j}/2, \quad j = 1, \ldots, J.$$

The use of midrank scores for the responses and midrank scores for the drug doses yields a generalization of Spearman's rho for contingency tables with ordered categories.

The use of rank-based scores seems appealing, since one does not need to select arbitrary scores, but midrank scores do not necessarily provide distances between categories that correspond to a 'reasonable' metric.[18] In particular, for highly unbalanced response frequencies, adjacent categories having relatively few observations necessarily have similar midrank scores, even if they seem far apart in practical terms. For example, suppose few subjects fell in the first three categories on the scale (death, fair, good, very good, excellent); midranks then have similar scores for the categories 'death' and 'good'. It is usually better to use one's judgement by selecting scores that reflect perceived distances between categories. When uncertain about this choice, one should perform a sensitivity analysis, selecting two or three 'sensible' choices and checking that the conclusions are similar for each; for instance, for the scale just mentioned, one might compare results for scores (0, 1, 2, 3, 4), (0, 5, 7, 9, 10) and (0, 7, 8, 9, 10). Equally-spaced scores often provide a reasonable compromise when the category labels do not suggest any obvious choices, such as the response categories (worse, no change, better).

The test statistic for the fixed-score or rank-score correlation approach is a special case for a single table of a generalized Cochran–Mantel–Haenszel (CMH) statistic for testing conditional independence with several $I \times J$ contingency tables.[19] That chi-squared statistic, having d.f. $= 1$, summarizes correlation information between two ordinal variables, combined over several strata. For a single table such as Table I, it equals $(N-1)\,r^2$, where r denotes the sample correlation for the chosen scores. It is available in SAS (PROC FREQ) as the 'non-zero correlation' test, generated using option CMH1 in that procedure; one can use either fixed scores selected by the user or midrank scores. Table II shows SAS code for analysing Table I using (i) scores (1, 2, 3, 4) for dose and (1, 2, 3, 4, 5) for outcome and (ii) midrank scores. The signed square root of this generalized CMH statistic, $M = \sqrt{(N-1)}\,r$, is a standard normal test statistic that is sensitive to the direction of trend.

The correlation-based test applied to Table I has $M = 3{\cdot}10$ using any sets of equally-spaced scores for the rows and the columns and $M = 3{\cdot}07$ using midrank scores, both having one-sided P-values of 0·001. The response scores (0, 1, 6, 9, 10), which may reflect a more reasonable assessment of distances between outcome categories, yield $M = 3{\cdot}59 (P < 0{\cdot}001)$ when used in combination with equally-spaced row scores; the response scores 0, 0, 1, 3, 10, which give much more weight to the most favourable outcome, yield $M = 2{\cdot}35$ ($P = 0{\cdot}009$). Each statistic provides strong evidence against the hypothesis of identical response distributions at the various dose levels.

A similar association test strategy, but not requiring any scores, bases the test on a measure that strictly uses ordinal information. Examples include the generalizations of Kendall's tau for contingency tables that utilize the numbers C of concordant and D of discordant pairs in summarizing information about an ordinal trend (Agresti,[20] pp. 22, 34). The standard measures fall between -1 and $+1$, have expectations of zero under the null hypothesis, and have approximate large-sample normal distributions. One can form a z test statistic (that is, having a standard normal null distribution) by dividing any such measure by its large-sample standard error.

These measures describe the extent of monotonicity in the relationship, without focusing on a particular aspect of it, such as linearity. An example is *Goodman and Kruskal's gamma*, which is $(C-D)/(C+D)$. Gamma equals the difference between the proportion of concordant pairs and the proportion of discordant pairs, out of the untied pairs. *Somers' d*, which treats the variables asymmetrically, is the difference between these proportions out of those pairs of observations falling at different dose levels. For instance, Somers' d equals 1·0 if, for each dose level, every response at that dose level exceeds every response at every lower dose level.

Table II. Example of SAS code for performing various analyses with Table I

```
data cmh;
input dose outcome count @@;
group = 1;
cards;
1 1 59    1 2 25    1 3 46    1 4 48    1 5 32
2 1 48    2 2 21    2 3 44    2 4 47    2 5 30
3 1 44    3 2 14    3 3 54    3 4 64    3 5 31
4 1 43    4 2  4    4 3 49    4 4 58    4 5 41
;
proc freq; weight count; * CMH with scores entered in data;
  tables group * dose * outcome / cmh1;
proc freq; weight count; * CMH with mid-rank scores;
  tables group * dose * outcome / cmh1 scores = ridit;
proc freq; weight count; * association measures such as gamma;
  tables dose * outcome / measures;
proc catmod order = data; weight count; * mean response model;
  population dose;
  response 1 2 3 4 5; direct dose; * uses scores (1, 2, 3, 4, 5);
  model outcome = dose;
proc logistic; freq count; * proportional odds model (ML);
  model outcome = dose;
proc catmod; weight count; * proportional odds model (WLS);
  response clogits; direct dose;
  model outcome = _response_ dose;
proc catmod; weight count; * adjacent cat. logit model (WLS);
  response alogits; direct dose;
  model outcome = _response_ dose;
run;
```

Formulae for standard errors of the extensions of Kendall's tau are quite complex. The measures and their estimated standard errors are available in standard software, such as SAS (PROC FREQ), as illustrated in Table II. For Table I, gamma = 0·118 and has a standard error of 0·038, leading to test statistic $z = 3{\cdot}11$ and $P = 0{\cdot}001$; Somers' d provides similar results, its value of 0·092 having a standard error of 0·030.

As in other contexts, the non-null expected values of various score-based correlation measures or ordinal association measures depend on the distribution of subjects to the various dose levels; the measures tend to increase with greater dispersion in the dose values. The test statistics based on them provide a simple way of summarizing trend information, even though the sample measure may not be used to estimate a particular population parameter.

2.2. Jonckheere–Terpstra test

For any pair $a < b$ of doses, the midranks for response levels for the $2 \times J$ table formed from these two treatment groups equal

$$w_{(ab)j} = (x_{a1} + x_{b1}) + \cdots + (x_{a,j-1} + x_{b,j-1}) + (x_{aj} + x_{bj})/2, \ j = 1, \ldots, J.$$

The Jonckheere–Terpstra (JT) test[21] statistic sums the $I(I-1)/2$ one-sided Wilcoxon–Mann–Whitney statistics for comparing pairs of treatment groups, in the order given by the

doses. In other words, the test statistic is based on

$$\mathrm{JT} = \sum_{b=2}^{I} \sum_{a=1}^{b-1} \sum_{j} \left(w_{(ab)j}\, x_{bj} - \frac{n_b(n_b+1)}{2} \right).$$

For large samples, the standardized value $z = [\mathrm{JT} - E(\mathrm{JT})]/[\mathrm{var}(\mathrm{JT})]^{1/2}$ provides a test statistic. Again, the variance formula is complex (see StatXact,[22] p. 614). The StatXact software, which provides a great variety of small-sample and asymptotic analyses for categorical data, can conduct this test. The same comments apply to this strategy as to the rank-based association measure approach presented in the previous subsection. For Table I, $z = 3{\cdot}10$, having one-sided P-value of 0·001.

2.3. Tests treating the response as continuous

A common approach for analysing ordinal data is to assign scores to the response categories and use standard normal-theory methods, such as regression and analysis of variance. From our experience, treating ordinal data as continuous with constant variance can provide a useful approximation when the number of response categories is large, but may be inadequate when that number is less than five. At the highest dose or at the no-dose level, responses often fall mostly in one category, yet are more dispersed at other dose values. Though this can cause problems for model building, for instance with predicting means or cell probabilities, it is less problematic for significance testing. For testing with small samples in the two-sample case, Heeren and D'Agostino[23] showed that the actual level of the t-test may not exceed the nominal level by much, but it can be considerably less than the nominal level.

When predictors are categorical, one can account for non-constant response variance by basing regression parameter estimates and standard errors explicitly on multinomial rather than normal assumptions for the response distribution. See, for instance, the mean response model discussed by Grizzle *et al.*[24] and Agresti[20] (Section 9.6). A weighted least squares (WLS) solution is simple to implement for this method using SAS (PROC CATMOD), as illustrated in Table II. When the data do not display widely varying dispersion or when the model fits well, the two approaches (ordinary and weighted least squares) provide very similar results.

For Table I, using the dose scores, the regression t-test for a normal response has $t = 3{\cdot}12$ ($P = 0{\cdot}001$) for equally-spaced response and dose scores, and $t = 2{\cdot}35$ ($P = 0{\cdot}009$) for response scores (0, 0, 1, 3, 10). The corresponding results using the methodology of Grizzle *et al.* are $z = 3{\cdot}10$ ($P = 0{\cdot}001$) and $z = 2{\cdot}25$ ($P = 0{\cdot}012$). For response scores (1, 2, 3, 4, 5), the prediction equation for the mean response is $2{\cdot}699 + 0{\cdot}138$ (dose) both using ordinary and weighted least squares. For the scores (0, 0, 1, 3, 10), they are $2{\cdot}089 + 0{\cdot}256$ (dose) and $2{\cdot}099 + 0{\cdot}248$ (dose).

This approach has the advantages of fully utilizing the inherent quantitative nature of the variables and directing the focus toward model-building rather than significance testing. A disadvantage, compared to models discussed in Section 3, is that conclusions disregard the categorical nature of the response scale. For instance, models that treat the response as categorical provide predicted probabilities of response in each category.

2.4. Order-restricted tests treating the response as continuous

For the monotone stochastic ordering alternative, the approximate approach of treating the ordinal response as normal with constant variance can also utilize methods developed for testing

equality of normal means against order-restricted alternatives. We now review some methods in this class.

Bartholomew[25,26] proposed one of the earliest order-restricted methods. Denote the true mean and the sample mean of the ith dose group by μ_i and $\bar{y}_i$. Assuming normality, one obtains the maximum likelihood (ML) estimates $\{\hat{\mu}_i\}$ subject to the constraint $\mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_I$ by constructing the finest possible partition $\{R_l\}$ of treatment groups $\{1, \ldots, I\}$ so that

$$\frac{\sum_{i \in R_l} n_i \bar{y}_i}{\sum_{i \in R_l} n_i}$$

is strictly increasing in ℓ. For all i in R_l, $\hat{\mu}_i$ are identical and equal a weighted sample mean. The solution of order-restricted mean estimates is called the *isotonic regression* of $\bar{y}_i$ with respect to the simple order on the row means $\{\mu_i\}$, with $\{n_i\}$ as the weights.[27] The partition is easily determined with the *pooling adjacent violators algorithm*.[27,28]

Denote the jth observation in the ith group by y_{ij}. When the population variance is unknown, Bartholomew[26] proposed the test statistic

$$\bar{E}^2 = \frac{\sum_i n_i(\hat{\mu}_i - \bar{y})^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2} \tag{3}$$

where $\bar{y} = (\sum n_i \bar{y}_i)/N$ is the overall sample mean. The large-sample distribution of $\bar{E}^2$ is non-standard, being the same as that of a weighted average of beta random variables. Relatively large $\bar{E}^2$ values provide evidence against the null hypothesis. Robertson *et al.*[28] (Chapter 2) discussed this test, and Brunden[29] prepared an SAS program that computes the weights and supplies critical values.

For the equally-spaced response scores (1, 2, 3, 4, 5), the sample means are 2·852 (placebo), 2·947 (low dose), 3·116 (medium dose) and 3·256 (high dose). The sample means satisfy the order restriction, and $\bar{E}^2 = 0{\cdot}0122$. The sample means for the second choice of scores (0, 0, 1, 3, 10) are 2·429, 2·553, 2·686 and 3·246, again satisfying the order restriction, and $\bar{E}^2 = 0{\cdot}0081$. The upper 5th and 1st percentiles of the null $\bar{E}^2$ distribution (Brunden[29]) are 0·0056 and 0·0096, respectively, so the P-value is less than 0·01 for the equally-spaced scores and less than 0·05 (about 0·025) for the unequally-spaced scores.

A small P-value for an order-restricted test suggests strong evidence against the null hypothesis, but just as with previously mentioned tests, this does not imply that monotone ordering holds perfectly in the population of interest. Small P-values can occur when the expected order is violated somewhat in the sample, but the test statistic would be sufficiently unusual under the null. To illustrate, consider Table III, showing responses (worse, same, slightly better, much better) to three doses (placebo, low and high) for a hypothetical sample of 123 subjects. The sample means under the equally-spaced response scores (1, 2, 3, 4) are 2·72 (placebo), 2·62 (low dose), and 3·05 (high dose), violating the order restriction. The isotonic regression of these sample means with respect to the increasing order on the row means provides mean estimates of 2·67 (placebo and low dose) and 3·05 (high dose), and $\bar{E}^2 = 0{\cdot}038$. The sample means under another choice of scores ($-3, 0, 2, 5$) are 1·63, 1·36 and 2·39, again requiring pooling adjacent violators to obtain order-restricted mean estimates, which are 1·49, 1·49 and 2·39, and for which $\bar{E}^2 = 0{\cdot}033$. The upper 5th and 1st percentiles of the null $\bar{E}^2$ distribution with three treatment groups (Brunden[29]) are 0·031 and 0·055, respectively, so the P-value is a bit less than 0·05 for both choices of scores. For either choice of scores, the mean for the high dose group

Table III. Example that violates order restriction but yields small P-value

Treatment group	Response category			
	Worse	Same	Slightly better	Much better
Placebo	3	15	12	10
Low dose	4	17	12	9
High dose	2	8	17	14

is sufficiently large that, if there were truly no effect, it would be unusual to obtain such a large test statistic value.

When all $\{n_i\}$ equal some constant n and the objective is to compare several doses with a zero-dose control, Williams[30] proposed the test statistic

$$\bar{t} = \frac{\hat{\mu}_I - \bar{y}_1}{\sqrt{(2s^2/n)}} \tag{4}$$

where s^2 is an unbiased estimator of error variance. Williams tabulated the distribution of this statistic and generalized it[31] when each treatment sample size is a constant multiple of the control group sample size. Williams noted that $\bar{t}$ has higher power than (3) when a constant drug effect occurs ($\mu_2 = \cdots = \mu_I$), and showed its role in sequential testing to determine the lowest dose at which evidence exists of a drug effect. In other cases, (3) and other statistics that incorporate more of the sample information are likely to perform better. Capizzi *et al.*[32] reported a simulation study that further compared these two procedures with an adjustment of a trend test proposed by Tukey *et al.*[33] They found that the adjusted trend test tends to be more powerful than the other procedures, although circumstances exist where either Bartholomew's[34] or Williams'[31] test appears superior.

For Table I, we equated n to the arithmetic mean of n_i, since the study intended to assign the same number of patients to each treatment group. Williams' test yields $\bar{t} = 2{\cdot}877$ for equally-spaced scores and 2·367 for the unequally-spaced scores. From Williams,[30] the upper 5th and 1st percentiles of the null distribution for $\bar{t}$ are 1·739 and 2·377. Thus, the P-value is less than 0·01 for equally-spaced scores and barely exceeds 0·01 for the unequally-spaced scores. As for Table III, with n_i set to 41, Williams' test yields $\bar{t} = 1{\cdot}881$ for the equally-spaced scores and 1·748 for the unequally-spaced scores. The upper 5th and 1st percentiles of the null distribution for the case of three treatment groups are 1·731 and 2·400, so the P-values are similar to those obtained with the $\bar{E}^2$ statistic for this example.

Shirley[35] proposed a Wilcoxon-type version of Williams' test, with emphasis on comparing increasing doses of a substance with a zero-dose control. Hothorn[1] studied the robustness of Williams' and Shirley's[35] procedures as applied in toxicology studies. He concluded that Shirley's procedure tends to behave better when assumptions underlying the analysis of variance are violated. For additional discussion of order-restricted inference, see Robertson *et al.*,[28] Cohen and Sackrowitz,[36] Hayter,[37] Silvapulle and Silvapulle,[38] and the references therein.

For an ordinal response, the approaches just discussed are somewhat unsatisfactory, since they treat the response as normal with constant variance rather than multinomial. One might prefer an

order-restricted test derived specifically for the monotone stochastic ordering alternative, under the assumption of multinomial sampling. For $I = 2$, Grove[39] and Robertson and Wright[40] proposed a likelihood-ratio test for testing whether two multinomial distributions are identical against the alternative of a stochastic ordering. The large-sample distribution of the test statistic is chi-bar squared, the distribution of a weighted average of chi-squared variates with differing degrees of freedom. For an observed test statistic value t, the P-value has the form $\sum_{j=1}^{J} p_j \Pr(\chi^2_{j-1} > t)$, where $\{p_j\}$ are weights that are recursively calculated.

For $I > 2$, results for order-restricted comparisons of multinomial distributions are incomplete. Patefield[41] suggested using an alternative that is a special case of a stochastic ordering, but noted the computational complexity of maximizing the likelihood. Grove[42] proposed a large-sample chi-bar-squared test for a different type of ordered alternative. Tests for the ordinary stochastic ordering alternative for the several groups case do not seem to appear in the literature, but one such test is discussed in a recent technical report by Agresti and Coull (University of Florida, 1996).

2.5. Treating the response distributions as survival distributions

When response categories are ordered, such as in Table I, one can use methods that apply to life tables. To do this, one orders the response categories from the least to the most favourable ones and regards observations in column j like failures between times $j-1$ and j, the most favourable category (for example, good recovery in Table I) representing subjects with censored lifetimes beyond time $J-1$. In this formulation, $R_{ij} = \sum_{b \geqslant j} x_{ib}$ represents the number of subjects at risk prior to time j, for dose i. The approaches discussed in this subsection have the advantage of not requiring response scores.

For life-table analysis, Tarone[43] proposed a test for a trend in hazard functions as the dose level increases. The square of this trend statistic is a special case of the summary chi-squared statistic proposed by Mantel[17] for stratified tables with ordered levels. In this context, we express the data as $(J-1)$ separate $I \times 2$ contingency tables, where the jth one compares the I doses on a binary response in which the first level is category j of the original response and the second level combines responses in all categories higher than j. The $(J-1)$ component tables in this construction are independent, because the corresponding sets of 'continuation-ratio' binomial variates in the component tables are independent. The statistic uses the dose scores by summarizing the correlation between dose and this binary response across the $(J-1)$ ways of forming the binary response. One can compute this correlation-type statistic by applying the CMH1 option in PROC FREQ in SAS to the $(J-1)$ tables. For Table I, the chi-squared statistic equals 8·350 with d.f. = 1, with a one-sided P-value (for its positive square root) of 0·002.

For this statistic, reversing the order of the response categories yields a different value of the test statistic. For instance, applying the test in the reverse order to Table I, the chi-squared test statistic equals 7·08, giving a one-sided P-value of 0·004. This behaviour is not true for other tests discussed in this article, which have the same result for each of the two possible orders of categories for the ordinal scale.

3. MODEL-BASED INFERENCE ABOUT MONOTONE DOSE–RESPONSE RELATIONS

The tests in Section 2 are fine for detecting evidence against the null hypothesis in the direction of a positive trend. However, they do not lend much insight about the form of the relationship.

A model-based perspective is superior for this purpose. A good-fitting model describes the nature of the association, provides parameters for describing the strength of the relationship, provides predicted probabilities for the response categories at any dose, and helps us to determine the optimal dose. As a by-product, it also yields tests for the hypothesis of no effect. In fact, some tests presented in the previous section have natural connections with models. In this section, we again focus on the first question posed by Ruberg[11] – that is, whether an effect exists; however, the model-based approach is also well-suited for pursuing the other three questions.

The models we discuss are generalizations of logistic regression models that handle ordinal response categories. For further discussion of these and other models for ordinal responses, see Agresti[20] (Chapters 8 and 9) and McCullagh.[44]

3.1. Proportional odds models

Currently, the most popular model for ordinal responses uses logits of cumulative probabilities. A J-category response has $(J-1)$ non-redundant cumulative probabilities, $P(Y_i \leqslant j)$, $j = 1, \ldots, J-1$. For the dose–response problem, consider the model

$$\text{logit}[P(Y_i \leqslant j)] = \alpha_j - \beta_i, \quad i = 1, \ldots, I, \ j = 1, \ldots, J-1 \tag{5}$$

where $\text{logit}[P(Y_i \leqslant j)] = \log[P(Y_i \leqslant j)/P(Y_i > j)]$. This model adds effects $\{\beta_i\}$ of the drug dosages on the response to the null model that contains only parameters $\{\alpha_j\}$ pertaining to the logit of each cumulative probability. It treats the effects $\{\beta_i\}$ as identical for each cumulative probability; that is, the effect does not depend on j in the model formula.

This form of model is called *proportional odds*.[44] Independence of dose and response is equivalent to $\beta_1 = \cdots = \beta_I$, each cumulative probability then being identical for all doses. Using a minus sign before the effect of dose in equation (5) implies that the higher the value of β_i relative to other $\{\beta_a\}$, the *lower* the cumulative probability tends to be at dose i, and hence the higher the response tends to be at dose i compared to other doses. The response distributions are stochastically ordered according to $\{\beta_i\}$. The case of a monotone relationship with direction (2) corresponds to $\beta_1 \leqslant \cdots \leqslant \beta_I$.

A monotone relation in which the trend is linear in dose scores $\{d_i\}$ has the simpler model form

$$\text{logit}[P(Y_i \leqslant j)] = \alpha_j - \beta d_i, \quad i = 1, \ldots, I, \ j = 1, \ldots, J-1 \tag{6}$$

with $\beta > 0$ implying (2). The ordinary logistic regression model with a linear dose effect is the special case $J = 2$. For this ordinal model, the odds that the response falls above any given category are multiplied by $\exp(\beta)$ for each unit increase in dose.

The ML fit of any model of this type yields estimated cumulative probabilities at each dose, and hence predicted numbers of observations (fitted values) in the cells of the table. One can test the fit using Pearson or likelihood-ratio chi-squared statistics that compare the observed cell counts to the model's fitted values. The adequacy of these goodness-of-fit tests improves as the cell counts increase in size, the Pearson test being preferred if the cell counts are relatively small.

Model (6) treats the doses as ordinal, whereas model (5) treats them as nominal. To increase power for testing independence when one expects a monotone trend, it is better to use model (6) as the alternative rather than (5). This leads to single degree-of-freedom chi-squared tests for testing independence ($\beta = 0$ in this model).

The likelihood-ratio test has chi-squared statistic given by double the difference in maximized log-likelihoods between the fit of model (6) and the simpler independence model having $\beta = 0$.

Table IV. Example of part of SAS output (using PROC LOGISTIC) for fitting proportional odds model (6) to Table I

Model Fitting Information and Testing Global Null Hypothesis
BETA = 0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
− 2 LOG L	2470·961	2461·349	9·612 with 1 DF (p = 0·0019)
Score			9·429 with 1 DF (p = 0·0021)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCP1	1	− 0·7192	0·1588	20·5072	0·0001
INTERCP2	1	− 0·3186	0·1564	4·1486	0·0417
INTERCP3	1	0·6917	0·1579	19·1809	0·0001
INTERCP4	1	2·0570	0·1737	140·2550	0·0001
DOSE	1	− 0·1755	0·0563	9·7087	0·0018

The Wald test is based on the square of the ratio of the ML estimate of β to its standard error. A third test, the 'efficient score' test, is based on the derivative of the log-likelihood function at $\beta = 0$. This score test is equivalent to a generalized CMH correlation test using the dose scores for X and midranks for categories of Y. One can perform all three tests using SAS software for this model, PROC LOGISTIC, as illustrated in Table II. For any of these statistics, one can refer the signed square root (that is, having the same sign as $\hat{\beta}$) to the standard normal distribution to construct a one-sided P-value. PROC LOGISTIC provides the ML fit of the model; PROC CATMOD can also fit the model, but only using WLS. The two approaches give similar results for large samples, but ML is preferred for small samples.

For Table I, the more general model (5) fitted with the constraint $\hat{\beta}_1 = 0$ has estimates $\hat{\beta}_2 = 0{\cdot}118$ (ASE = 0·178), $\hat{\beta}_3 = 0{\cdot}317$ (ASE = 0·175), and $\hat{\beta}_4 = 0{\cdot}521$ (ASE = 0·178). The estimates suggest a monotone increase in response as a function of dose. The likelihood-ratio test that $\beta_2 = \beta_3 = \beta_4 = 0$ has test statistic equal to 9·75 with d.f. = 3 ($P = 0{\cdot}021$). There is evidence of a drug effect, though only the estimate for the high dose level shows substantial evidence of differing from the placebo.

Table IV shows sample SAS output for the simpler model (6), using dose scores (1, 2, 3, 4). One can compare the fit of this model to that of the model (5) with separate dose effects using the difference of −2 log-likelihood values for the two models. Under the hypothesis that the simpler model is adequate, this difference is an approximate chi-squared statistic with d.f. equal to the number of dose levels −2. In this case, the test statistic comparing the models is $2461{\cdot}35 - 2461{\cdot}22 = 0{\cdot}13$ with d.f. = 2, indicating that the simpler model is adequate. It has $\hat{\beta} = 0{\cdot}176$ (ASE = 0·056) (SAS actually reports the negative of this value, since it parameterizes the model with + rather than − as the coefficient of the effect). The z test statistics equal $3{\cdot}10 = \sqrt{9{\cdot}612}$ for the likelihood-ratio test, $3{\cdot}12 = \sqrt{9{\cdot}709}$ for the Wald test, and $3{\cdot}07 = \sqrt{9{\cdot}429}$ for the score test, all having a one-sided P-value of 0·001.

These three tests tend to show similar results for large samples. They are valid for smaller samples than one needs for performing goodness-of-fit tests for the model. In fact, even if model (6) does not fit well (as is, in fact, the case for these data), the test statistics provide relatively

powerful tests, compared to tests that ignore the ordering of doses or responses, as long as the linear term in the model represents a major component of the departure from independence. That is, one does not need to test the goodness-of-fit of the model before conducting the association test. In this regard, the remark of Mantel[17] in a similar context is instructive, 'that a linear regression is being tested does not mean that an assumption of linearity is being made. Rather it is that that test of a linear component of regression provides power for detecting any progressive association which may exist.'

Proportional odds models have several appealing properties. If the model holds for a particular set of response categories, it holds with the same parameter effects when the response scale is collapsed in any way. This behaviour is true, approximately, for the sample data. For instance, if we combine the major and minor disability categories in Table I and again fit model (6), we get $\hat{\beta} = 0{\cdot}185$ (ASE = 0·060), compared to $\hat{\beta} = 0{\cdot}176$ (ASE = 0·056) for the complete table. When this model fits well, different studies using different definitions of response categories should reach similar conclusions. In addition, it is unnecessary to assign scores to the response categories.

Once an effect is established, the more complex model (5) is useful for comparing response distributions at different dosage levels. For instance, the difference $\hat{\beta}_i - \hat{\beta}_j$ divided by its standard error is a standard normal test statistic for judging whether the pair of doses i and j is significantly different. One can use Bonferroni methods for simultaneous comparisons; for instance, to simultaneously compare all $(I - 1)$ pairs of adjacent dose levels with an overall type I error probability of no greater than 0·05, one uses nominal size $0{\cdot}05/(I - 1)$ for each pairwise test.

3.2. Other ordinal models

Though the proportional odds model is currently a popular one for modelling ordinal response data, one could alternatively use other ordinal models to detect a monotone dose effect. For instance, McCullagh[44] discussed transforms other than the logit for the cumulative probability, such as the probit and ones (log-log and complementary log-log) for which the cumulative probability approaches 0 at a different rate than it approaches 1. The probit usually provides similar results as the logit, in terms of testing for an effect. McCullagh showed that the probit and logit are most appropriate when an underlying continuous response is roughly bell-shaped, and when a similar form of model holds for that continuum. For instance, if an underlying normal response has approximately a linear relationship with dose, then the logit or the probit of the cumulative probabilities with a linear dose effect tends to fit well. Log-log links, on the other hand, are appropriate when an underlying response is highly skewed. All these options are available with PROC LOGISTIC in SAS.

Other ordinal models utilize single-category probabilities rather than cumulative probabilities. For instance, the *adjacent-categories logit* model with a linear dose effect has form

$$\log[P(Y_i = j)/P(Y_i = j + 1)] = \alpha_j - \beta d_i, \quad i = 1, \ldots, I, \ j = 1, \ldots J - 1. \tag{7}$$

The same dose effect β occurs for logits for each pair of adjacent response categories. Independence is the special case $\beta = 0$, and one can test this with a likelihood-ratio, Wald, or efficient score test. The score test is equivalent to the generalized CMH correlation test using the dose scores for X and equally-spaced scores for the response categories. One can fit models of this form using PROC CATMOD in SAS.[45] With SAS, an ML fit is possible, but it is much simpler to prepare code for the WLS fit; see Table II. Model (7) is also equivalent to an ordinal log-linear

model that uses these scores for the two classifications, called the *linear-by-linear association* model (Agresti,[20] Section 8.1).

The parameters for the effect in models (7) and (6) refer to different types of odds ratios. For instance, $\exp(\beta)$ in model (7) refers to the multiplicative effect of a one-unit increase in dose on the odds of response in the higher instead of the lower of any two adjacent categories. The two models usually fit well in similar situations and provide similar results in the tests.

For instance, with Table I, the estimated effect in model (7) is $\hat{\beta} = 0{\cdot}070$, with standard error 0·023. The z test statistic versions of the Wald and likelihood-ratio statistics equal 3·09 and 3·11, again giving one-tailed P-values of 0·001. The model fits fairly well (Pearson goodness-of-fit statistic = 15·8, d.f. = 11). Both it and the proportional odds model (6) show some lack of fit in the second column for the last row, the response count of 4 in this cell being significantly smaller than the value of nearly 14 that the model predicts.

Finally, the *continuation-ratio* logit model with common effect for each logit has form

$$\log[P(Y_i = j)/P(Y_i \geqslant j+1)] = \alpha_j - \beta d_i, \quad i = 1, \ldots, I, \ j = 1, \ldots, J-1. \tag{8}$$

A score test for this model is equivalent to the test of Tarone[43] for survival data discussed in Section 2.5; that is, it is a generalized CMH test based on the sets of probabilities used in these logits. The statistic value and resulting P-value differs from the ones using continuation-ratio logits of form $\log[P(Y_i = j+1)/P(Y_i \leqslant j)]$.

4. GENERALIZATIONS FOR STRATIFIED DATA

Typically, one studies dose–response relations while controlling for factors that could influence the relationship. For instance, one might display the relationship separately for men and for women, for different age groups, for different centres from which the data are obtained, or for different stages or levels of severity of the medical condition being treated. To illustrate, Table V is a stratified version of Table I that classifies subjects according to the trauma severity at the time of study entry. The study was designed to enroll about the same number of mild versus moderate/severe patients, and the randomization was carried out with severity grade as a stratifying factor. In general, the stratification can be part of the study design or represent post-study control to form more homogeneous subgroups.

For a stratified table, interest focuses not only on the effect of the dose on the response within each stratum, but also on whether there is interaction. Does the dose effect vary according to the stratum?

4.1. Models for the stratified case

The model-building approach can easily accommodate stratified data. We illustrate this with the proportional odds model. For level h of S strata, let Y_{hi} denote a response for a subject at dose level i. The proportional odds model

$$\text{logit}[P(Y_{hi} \leqslant j)] = \alpha_j - \beta_h^{\mathrm{S}} - \beta_i^{\mathrm{D}}, h = 1, \ldots, S, \ i = 1, \ldots, I, \quad j = 1, \ldots, J-1 \tag{9}$$

has dose effects $\{\beta_i^{\mathrm{D}}\}$ and stratum effects $\{\beta_h^{\mathrm{S}}\}$, but assumes a lack of dose-by-stratum interaction. That is, the effects of the doses on the response are assumed to be the same in each stratum. The special case of this model replacing β_i^{D} by $\beta^{\mathrm{D}} d_i$ for the dose scores $\{d_i\}$ is relevant for detecting a particular type of monotone trend. One can then construct single-degree-of-freedom

Table V. Data from Table I stratified by trauma severity

Trauma severity	Treatment group	Glasgow Outcome Scale				
		Death	Vegetative state	Major disability	Minor disability	Good recovery
Mild	Placebo	2	4	29	43	26
	Low dose	2	4	25	39	23
	Medium dose	1	3	23	49	24
	High dose	0	1	21	47	26
Moderate/ severe	Placebo	57	21	17	5	6
	Low dose	46	17	19	8	7
	Medium dose	43	11	31	15	7
	High dose	43	3	28	11	15

chi-squared statistics (or, taking square roots, z statistics) testing whether $\beta^{D} = 0$ using the likelihood-ratio, Wald, or score approaches, in the same way as just discussed for two-way tables.

To illustrate, applying the simpler model with a linear dose effect and dose scores (1, 2, 3, 4) to Table V, we get $\hat{\beta} = 0{\cdot}205$ (ASE $= 0{\cdot}058$). The Wald chi-squared statistic equals 12·5, and the likelihood-ratio statistic comparing this model to the simpler one without the dose effect equals the difference in -2-log-likelihood values for the two models, which is also 12·5 ($z = 3{\cdot}53$). The P-value is less than 0·001.

More generally, one could extend model (9) or the simpler one with the linear effect by permitting dose-by-stratum interaction. The model simply then adds cross-product terms of the dose and strata variables (or dummy variables). One can test the hypothesis of no interaction by comparing the -2 log-likelihood values for this model and the corresponding model without interaction. When the degree of interaction seems substantively important, one can estimate and test the effect separately in each stratum using the dose effect estimates pertaining to that stratum, or one could simply fit the original model (for example, 6) separately to each stratum to obtain the separate effects. (This approach is not equivalent, because it estimates intercept parameters separately with each fit.) On the other hand, when the dose effects do not vary much among the strata, the overall test based on a lack of interaction tends to be much more powerful, and the overall estimate tends to be more efficient, since they summarize information across the strata.

In fact, there is some evidence of interaction in Table V. For the models with linear dose effect, the likelihood-ratio statistic comparing the model with separate slopes to the model with a single slope equals 3·85 with d.f. $= 1$. The model with separate slopes has estimates $\hat{\beta} = 0{\cdot}099$ (ASE $= 0{\cdot}082$) for the mild trauma group and $\hat{\beta} = 0{\cdot}327$ (ASE $= 0{\cdot}082$) for the moderate/severe trauma group. Hence, there is a strong evidence of a dose effect only for the latter group. There are other approaches one could use both to check for interaction and to describe the separate effects, but we do not discuss them here because of space limitations.

4.2. Non-model-based approaches for the stratified case

The CMH approach generalizes naturally to combining information from several strata; in fact, the original statistic presented by Mantel and Haenszel[46] was designed specifically for the stratified case with two groups and a binary response. For the case of several doses and an ordinal

response, the correlation statistic (Mantel[17]) provides a large-sample chi-squared statistic with d.f. = 1 for detecting a linear trend in the effect. One can, as usual, treat the signed square root as a standard normal statistic. The CMH approach, like model (9), works well when the dose effects are similar in each stratum. It is available with the CMH1 option in PROC FREQ in SAS. For Table V, this approach used with equally-spaced scores for doses and response outcomes yields a chi-squared statistic of 16·2 and normal statistic of 4·0, for which the P-value is less than 0·001.

Similarly, one could consider stratified versions of tests discussed in Section 2 that are special cases of a generalized CMH test, such as Tarone's test.[42] In principle, this type of construction could also be used with other sorts of statistics, such as the Jonckheere-Terpstra statistic.

5. SMALL-SAMPLE AND SPARSE-DATA INFERENCE

The test statistics presented in this article are large-sample statistics. For chi-squared statistics, the convergence to chi-squared distributions tends to be faster for statistics having smaller values of d.f., such as the single-degree-of-freedom statistics.

For any particular statistic referring to the two-way contingency table of dose by ordinal response, one can construct a small-sample 'exact' test using the generalized hypergeometric distribution that results from conditioning on the row and column totals. This approach generalizes Fisher's exact test for 2-by-2 tables, with the conditioning argument yielding a distribution not depending on unknown nuisance parameters. Exact tests are available in StatXact[22] for several statistics, including the Jonchkeere–Terpstra statistic and correlation-type statistics with fixed or rank scores.[47] (The correlation-type statistics use the 'linear-by-linear' option in StatXact.) Currently these tests are restricted to the single-stratum case.

For stratified data, only the case of two dose groups is currently addressed by standard software, using the CMH correlation type approach for a set of fixed or midrank response scores (StatXact). In principle, though, the methodology of small-sample exact tests extends directly to the more general case of several dose groups.[48] Specialized FORTRAN programs exist for these analyses.

6. SAMPLE SIZE AND POWER

Whitehead[49] discussed sample size formulae for an ordered categorical response with the proportional odds model, though only for the case of two groups (for example, two doses). Suppose we want power $1 - \beta$ in an α-level test for detecting an effect of size β_0 in that model. The sample is to be allocated to the two groups in the ratio A to 1, and $\bar{p}_j$ denotes the anticipated marginal proportion in response category j. Whitehead[49] stated that the required sample size for a two-sided test is then approximately

$$N = 3(A + 1)^2(z_{\alpha/2} + z_\beta)^2/[A\beta_0^2(1 - \sum \bar{p}_j^3)]$$

where z_a is the $100(1 - \alpha)$ percentile of the standard normal distribution.

This requires anticipating the marginal proportions as well as the size of the effect. Setting $\bar{p}_j = 1/J$ provides a lower bound for N. Whitehead[49] showed that the sample size does not depart much from this bound unless a single dominant response category occurs. Hilton and Mehta[50] provided a somewhat different approach to sample size determination, based on evaluating the exact conditional distribution with a network algorithm, or simulating that distribution.

With equal marginal probabilities, Whitehead's[49] formula is useful for showing the effect of the choice of number of response categories. The ratio of the sample size $N(J)$ needed for J categories relative to the sample size $N(2)$ needed for two categories is

$$N(J)/N(2) = 0{\cdot}75/[1 - 1/J^2].$$

Relative to a continuous response ($J = \infty$), using J categories provides efficiency $(1 - 1/J^2)$. The loss of information from collapsing to a binary response is substantial, but there is little gain from using more than about five categories. For fixed J, equal allocation ($A = 1$) produces the smallest sample size.

The case of $I > 2$ groups does not seem to have been considered. However, various rather *ad hoc* ways exist of approaching the problem. For instance, many tests discussed in this article are based on asymptotically normal statistics, such as a measure of association (for example, correlation, gamma) or an estimate of a model parameter ($\hat{\beta}$ for the proportional odds model). Let $\hat{\theta}$ denote a generic asymptotically normal estimator of a parameter θ, with variance of the form V/N. Then, for a fixed non-null value θ_0 of θ, standard arguments show that the required sample size for a one-sided test is

$$N = (z_\alpha + z_\beta)^2 V/\theta_0^2.$$

To use this formula, the steps are to: (i) choose an anticipated set of non-null cell probabilities; (ii) find the value of θ_0 corresponding to those probabilities; (iii) find V for those probabilities, and (iv) substitute V into this formula using the required size and power. In some cases V has closed form, based on the delta method, and in some cases it requires iterative methods. Even when it has closed form, though, the formula is typically messy computationally. A simple approach to determining V (and θ_0) enters the anticipated probabilities as data into standard software, in which case V equals the square of the reported asymptotic standard error.

For illustrative purposes, suppose we had anticipated probabilities proportional to the counts in Table I. For the proportional odds model, we observed $\hat{\beta} = 0{\cdot}1755$ and a standard error of 0·0563 for these data having a sample size of 802. Setting $V/802 = (0{\cdot}0563)^2$ yields $V = 2{\cdot}542$. To have power 0·90 in an $\alpha = 0{\cdot}05$ level one-sided test of $\beta = 0$ when the true relationship has $\beta_0 = 0{\cdot}1755$ requires a sample size of about $N = (1{\cdot}645 + 1{\cdot}282)^2(2{\cdot}542)/(0{\cdot}1755)^2 = 707$.

For stratified data, Whitehead[49] noted that logistic regression and an ordinal extension such as the proportional odds model may require a somewhat increased sample size to preserve the desired power. However, the variation among strata in the category probabilities has to be quite extreme before sample size is greatly affected.

SUMMARY AND RECOMMENDATIONS

We have presented a variety of tests for detecting a monotone relation between dose and an ordinal response. Of the non-model-based methods, the tests based on the correlation seem the most flexible. These connect closely with methods used for continuous reponses, which can be regarded as a limiting case as the number of response categories and doses increases indefinitely. Though directed toward a narrow alternative, namely linearity for the choice of scores, this provides the advantage of good power if a strong linear component exists for the true association.

The order-restricted approach has the advantage of specifying the alternative in a broader and more realistic manner. Disadvantages include a rather awkward limiting distribution, a lack of a full theoretical and methodological development for a categorical response when the number of

doses exceeds two or the data are stratified, and potential power loss compared to a linear trend statistic when the true relation has a strong linear component. Some preliminary power studies by one of the authors for a separate project suggest that the order-restricted approach has better power than the linear trend test if the response is essentially identical for all positive dose groups but those groups have better response than the control group. On the other hand, for other patterns of monotone increase that do not depart so drastically from linearity, the linear trend statistic is more powerful.

Section 2 addressed the dose–response relationship within the significance-testing framework. Our overall preference, however, is for a model-based approach, since it provides a fuller description of the dose–response relationship. For instance, estimated odds ratios describe the strength of the effect, and fitted values provide estimates of response probabilities that are smoother and tend to have smaller mean squared errors than the sample proportions. Goodness-of-fit tests check the model adequacy, and residuals can indicate potential departures from the trend predicted by the model. Moreover, the fit of a model such as (5) enables us to consider the more important follow-up questions, such as determining which doses have significantly different responses and which dose is optimal.

Some statisticians avoid the model-building approach for fear of increasing the number of assumptions, with the resulting test being less robust. However, many of the standard tests have connections with models, being efficient score tests. For large samples, one obtains similar results from a likelihood-ratio test for a model parameter as one does from a score test. For Table I, for instance, model-based and non-model-based tests gave similar results.

Focusing on models makes one recognize the structure under which a particular test is natural. Moreover, a model-based test can provide a powerful approach even if the model does not fit well. For instance, for testing conditional independence in stratified tables, generalizations of the Cochran–Mantel–Haenszel test are popular. These tests are score tests for models that assume homogeneity of odds ratios across strata. However, one does not need to assume such homogeneity to use the tests, and they perform well whenever the true degree of heterogeneity is not severe.

Finally, emphasizing models has the advantage of decreasing reliance on significance tests as the primary mode of analysis. Though this paper has surveyed a variety of such tests for detecting monotone dose-response relationships, ultimately estimation of parameters yields more informative conclusions.

ACKNOWLEDGEMENTS

The work of Agresti was partially supported by an NIH grant.

REFERENCES

1. Hothorn, L. 'Robustness study on Williams- and Shirley- procedure, with application in toxicology', *Biometrical Journal*, **31**, 891–903 (1989).
2. Piegorsch, W. W. 'Nonparametric methods to assess non- monotone dose response: Applications to genetic toxicity', in Sen, P. K. and Salama, I. A. (eds.), *Order Statistics and Nonparametrics*: *Theory and Applications*, Elsevier Science Publishers, B. V., 1992.
3. Dalal, S. N. and Lawson, J. S. 'Methods for the dose response analysis in animal growth promotion studies', *ASA Proceedings of the Biopharmaceutical Section*, 171–176 (1989).
4. Maclure, M. and Greenland, S. 'Tests for trend and dose response: Misinterpretations and alternatives', *American Journal of Epidemiology*, **135**, 96–104 (1992).
5. Breslow, N. E. and Day, N. E. 'Statistical methods in cancer research', in *The Design and Analysis of Cohort Studies*, Vol. 2, International Agency for Research on Cancer, 1987, p. 97.

6. Mack, G. A. and Wolfe, D. A. 'K-sample rank tests for umbrella alternatives', *Journal of the American Statistical Association*, **76**, 175–181 (1981).
7. Hettmansperger, T. P. and Norton, R. M. 'Tests for patterned alternatives in k-sample problems', *Journal of the American Statisical Association*, **82**, 292–299 (1987).
8. Chen, Y. I. and Wolfe, D. A. 'Modifications of the Mack–Wolfe umbrella tests for a generalized Behrens–Fisher problem', *Canadian Journal of Statistics*, **18**, 245–253 (1990).
9. Chen, Y. I. and Wolfe, D. A. 'A study of distribution-free tests for umbrella alternatives', *Biometrical Journal*, **32**, 47–57 (1990).
10. Chen, Y. I. 'On the comparison of umbrella pattern treatment means with a control mean', *Biometrical Journal*, **35**, 689–700 (1993).
11. Ruberg, S. J. 'Dose-response studies. I. Some design considerations', *Journal of Biopharmaceutical Statistics*, **5**, 1–14 (1995).
12. Ruberg, S. J. 'Dose-response studies. II. Analysis and Interpretation', *Journal of Biopharmaceutical Statistics*, **5**, 15–42 (1995).
13. Kodell, R. L. and Chen, J. J. 'Characterization of dose-response relationships inferred by statistically significant trend tests', *Biometrics*, **47**, 139–146 (1991).
14. Zeger, S. L. and Liang, K. Y. 'Dose-response estimands. Comment on "Compliance as an explanatory variable in clinical trials"', *Journal of the American Statistical Association*, **86**, 18–19 (1991).
15. Yates, F. 'The analysis of contingency tables with grouping based on quantitative characters', *Biometrika*, **35**, 176–181 (1948).
16. Armitage, P. 'Tests for linear trends in proportions', *Biometrics*, **11**, 375–386 (1955).
17. Mantel, N. 'Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure', *Journal of the American Statistical Association*, **58**, 690–700 (1963).
18. Graubard, B. I. and Korn, E. L. 'Choice of column scores for testing independence in ordered 2XK contingency tables', *Biometrics*, **43**, 471–476 (1987).
19. Landis, J. R., Heyman, E. R. and Koch, G. G. 'Average partial association in three-way contingency tables: A review and discussion of alternative tests', *International Statistical Review*, **46**, 237–254 (1978).
20. Agresti, A. *Categorical Data Analysis*, Wiley, 1990.
21. Jonckheere, A. R. 'A distribution-free K-sample test against ordered alternatives', *Biometrika*, **41**, 133–145 (1954).
22. StatXact. *StatXact3 for Windows*: *Statistical Software for Exact Nonparametric Inference*, *User Manual*, Cytel Software, 1995.
23. Heeren, T. and D'Agostino, R. 'Robustness of the two independent samples *t*-test when applied to ordinal scaled data', *Statistics in Medicine*, **6**, 79–90 (1987).
24. Grizzle, J. E., Starmer, C. F. and Koch, G. G. 'Analysis of categorical data by linear models', *Biometrics*, **25**, 489–504 (1969).
25. Bartholomew, D. J. 'Ordered tests in the analysis of variance', *Biometrika*, **48**, 325–332 (1961).
26. Bartholomew, D. J. 'A test of homogeneity of means under restricted alternatives', *Journal of the Royal Statistical Society*, *Series B*, 239–281 (1961).
27. Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. *Statistical Inference Under Order Restrictions*, Wiley, 1972.
28. Robertson, T., Wright, F. T. and Dykstra, R. L. *Order- Restricted Statistical Inference*, Wiley, 1988.
29. Brunden, M. N. 'A review of isotonic regression theory – means from normal distribution', TR 9164-95-001, The Upjohn Co., Kalamazoo, MI, 1995.
30. Williams, D. A. 'A test for differences between treatment means when several dose levels are compared with a zero dose control', *Biometrics*, **27**, 103–117 (1971).
31. Williams, D. A. 'The comparison of several dose levels with a zero dose control', *Biometrics*, **28**, 519–531 (1972).
32. Capizzi, T., Survill, T. T., Heyse, J. F. and Malani, H. 'An empirical and simulated comparison of some tests for detecting progressiveness of response with increasing doses of a compound', *Biometrical Journal*, **34**, 275–289 (1992).
33. Tukey, J. W., Ciminera, J. L. and Heyse, J. F. 'Testing the statistical certainty of a response to increasing doses of a drug', *Biometrica*, **41**, 295–301 (1985).
34. Bartholomew, D. J. 'A test of homogenity for ordered alternatives', *Biometrika*, **46**, 36–48 (1959).

35. Shirley, E. 'A non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment', *Biometrics*, **33**, 386–389 (1977).
36. Cohen, A. and Sackrowitz, H. B. 'Improved tests for comparing treatments against a control and other one-sided problems', *Journal of the American Statistical Association*, **87**, 1137–1144 (1992).
37. Hayter, A. J. 'A one-sided Studentized range test for testing against a simple ordered alternative', *Journal of the American Statistical Association*, **85**, 778–785 (1990).
38. Silvapulle, M. J. and Silvapulle, P. 'A score test against a one-sided alternative', *Journal of the American Statistical Association*, **90**, 342–349 (1995).
39. Grove, D. M. 'A test of independence against a class of ordered alternatives in a $2 \times c$ contingency table', *Journal of the American Statistical Association*, **75**, 454–459 (1980).
40. Robertson, T. and Wright, F. T. 'Likelihood-ratio tests for and against a stochastic ordering between multinomial populations', *Annals of Statistics*, **9**, 1248–1257 (1981).
41. Patefield, W. M. 'Exact tests for trends in ordered contingency tables', *Journal of the Royal Statistical Society, Series C*, **31**, 32–43 (1982).
42. Grove, D. M. 'Positive association in a two-way contingency table: Likelihood ratio tests', *Communications in Statistics, A: Theory and Methods*, **13**, 931–945 (1984).
43. Tarone, R. E. 'Tests for trend in life table analysis', *Biometrika*, **62**, 679–682 (1975).
44. McCullagh, P. 'Regression models for ordinal data (with discussion)', *Journal of the Royal Statistical Society, Series B*, **42**, 109–142 (1980).
45. Stokes, M. E., Davis, C. S. and Koch, G. G. *Categorical Data Analysis Using the SAS System*, SAS Institute Inc, 1995.
46. Mantel, N. and Haenszel, W. 'Statistical aspects of the analysis of data from retrospective studies of disease', *Journal of the National Cancer Institute*, **22**, 719–748 (1959).
47. Agresti, A., Mehta, C. R. and Patel, N. R. 'Exact inference for contingency tables with ordered categories', *Journal of the American Statistical Association*, **85**, 453–458 (1990).
48. Kim, D. and Agresti, A. 'Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables', *Computational Statistics and Data Analysis*, **24**, 89–104 (1997).
49. Whitehead, J. 'Sample size calculations for ordered categorical data', *Statistics in Medicine*, **12**, 2257–2271 (1993).
50. Hilton, J. F. and Mehta, C. R. 'Power and sample size calculations for exact conditional tests with ordered categorical data', *Biometrics*, **49**, 609–616 (1993).

