13
Analysis of Dose–Response Relationship Based on Categorical Outcomes

CHRISTY CHUANG-STEIN and ZHENGQING LI

13.1 Introduction

In an eloquent article prepared in defense of the dichotomy, Lewis (2004) wrote that one of the most important ways in which we learned to understand the world was to describe complicated phenomena using simple categories. Thus, it is hardly surprising that medical researchers often seek to categorize data in their attempt to make sense of unfamiliar measurement scales and treatment effects of uncertain implication. For this reason, threshold values based on continuous measurements are frequently used to help guide the decision to initiate medical interventions. Examples include a diastolic blood pressure greater than 90, a fasting cholesterol level higher than 200, and a CD4 count lower than 200. Normal ranges were constructed to screen subjects for possible lab abnormalities. Even though this black-and-white dichotomy appears to be crude in many situations, its simplicity helps human minds make decisions, decisions that are often binary in nature.

As we became more sophisticated in our views of the world, so did our descriptions of the surroundings. Being normal or abnormal alone is no longer enough. We want to know the extent of abnormality to decide if immediate actions are necessary. Experiencing pain alone is not enough to decide if pain relief medications are necessary. Similarly, recovery from a major trauma can mean recovery with major disability, with minor disability, or essentially with no noticeable disability. The human minds realized that creating a finer grid between the two extremes of black and white could help us make better decisions on many occasions.

Over the past 30 years, researchers have been busy developing scales to subdivide the space between the black and white extremes. The proliferation of scales is most prevalent in the area of outcome research where scales are used to record a patient's and the treating physician's global assessments of the clinical symptoms associated with the underlying disorders. Scales are also used to record the extent of relief patients receive from the medications. These activities have led to the collection of categorical data in many clinical trials.

In this chapter, we will focus on analyzing dose–response relationship when the primary endpoint is either ordinal or binary. We will treat the ordinal case first in Section 13.2 and regard the binary case as a special case of the former. The binary

case will be covered in Section 13.3. In Section 13.4, we will discuss multiple comparisons procedures that are applicable to categorical data when multiplicity adjustment is considered necessary because of the confirmatory nature of the trial. Readers are referred to Chapter 12 for a more general discussion on multiple comparisons. We will comment briefly in Section 13.5 the use of a titration design to explore the dose–response relationship with a binary outcome. In addition, we will discuss in that section issues related to sample size. Finally, we encourage our readers to use simulations to help evaluate the planned study at the design stage.

In this chapter, we provide numerical examples along with methodology. This is a deliberate effort to emphasize the applied nature of this chapter. To help implement the methodology, we include in the Appendix simple SAS codes that could be used to produce most of the results in Sections 13.2 and 13.3.

This chapter draws heavily from a review article by Chuang-Stein and Agresti (1997) on testing a monotone dose–response relationship with ordinal response data. Readers who wish to learn more about the technical details of the methodologies are encouraged to read the original publication.

13.2 When the Response is Ordinal

Consider the data in Table 13.1 where five ordered categories ranging from "death" to "good recovery" were used to describe the clinical outcome of patients who suffered from subarachnoid hemorrhage. The five outcome categories make up the Glasgow Outcome Scale (GOS). Three doses of an investigational drug (low, medium, and high) and a vehicle infusion (placebo) were included in the trial. For this type of data, one can either model the probability of an ordinal response as a function of the dose or conduct hypothesis testing to test for a dose–response relationship. In this section, we will briefly describe the modeling approach first followed by procedures that focus on hypothesis testing.

13.2.1 Modeling Dose–Response

Let p_{ij} be the probability that a subject in the ith dose group ($i = 1, 2, 3, 4$) will have a response in the jth ($j = 1, 2, \ldots, 5$) category. For each dose group, p_{ij}'s

Table 13.1. Responses measured on the Glasgow Outcome Scale from a trial comparing three doses of a new investigational treatment with a control (Chuang-Stein and Agresti, 1997)

	Glasgow Outcome Scale					
Treatment group	Death	Vegetative state	Major disability	Minor disability	Good recovery	Total
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

satisfy $\Sigma_j p_{ij} = 1$. We will use Y_{ij} to denote the number of subjects in the ith dose group whose responses are in the jth category. We assume that within each dose group $\{Y_{ij}, j = 1, \ldots, 5\}$ follows a multinomial distribution $(n_i; \{p_{ij}\})$ where $n_i = \Sigma_j Y_{ij}$. Here, we are treating $\{n_i\}$ as fixed constants since most trials have a target figure for $\{n_i\}$. Furthermore, we assume in this chapter that the response categories are arranged in such a way that higher response categories correspond to a more desirable outcome. The dose groups are arranged in an ascending order. If there is a placebo group, the placebo group will be the first dose group.

There are many ways to take advantage of the ordinal nature of the response when modeling the dose response relationship. The most popular one is probably the one using logits of the cumulative probabilities defined as (McCullagh, 1980)

$$\ln\left(\frac{\sum_{l=1}^{j} p_{il}}{\sum_{l=j+1}^{5} p_{il}}\right) = \alpha_j - \beta_i \tag{13.1}$$

In Eq. (13.1), "ln" represents the natural logarithm and $i = 1,2,3,4$ and $j = 1,2,3,4$. In Eq. (13.1), the parameters $\{\alpha_j\}$ associated with the response categories do not depend on the dose group. As a result, if one looks at the ratio of the cumulative odds between two dose groups, the ratio is constant across response categories. For this reason, model (13.1) is called the proportional odds model. The appropriateness of the proportional odds assumption can be checked by the likelihood ratio statistic obtained by comparing the proportional odds model to the saturated model.

One can further simplify model (13.1) by fitting β_i as a function of the dose as in Eq. (13.2) or the dose on the logarithmic scale if there is reason to believe that the treatment effect is a monotone function of the dose.

$$\ln\left(\frac{\sum_{l=1}^{j} p_{il}}{\sum_{l=j+1}^{5} p_{il}}\right) = \alpha_j - \beta\, d_i \tag{13.2}$$

Parameters in Eqs. (13.1) and (13.2) could be estimated by the maximum likelihood method. The procedure PROC LOGISTIC in SAS® can be employed to estimate the parameters. Testing the equality of the $\{\beta_i\}$ in Eq. (13.1) and $\beta = 0$ in Eq. (13.2) can be done using the likelihood ratio test. In either case, the likelihood ratio statistic has an asymptotic χ^2 distribution with degrees of freedom determined by the difference in the number of parameters included in the two models under comparison. For example, the likelihood ratio statistic for testing $\beta = 0$ in Eq. (13.2) has an asymptotic χ^2 distribution with 1 degree of freedom under the null hypothesis.

When employing model (13.2), one is typically interested in testing $\beta = 0$ against $\beta > 0$ so that rejecting $\beta = 0$ will infer that higher doses tend to produce more favorable response. Despite this, we will report two-sided p-values when testing the significance of the slope parameter to reflect the current regulatory requirement on reporting two-sided p-values even if the interest is clearly one-sided. Unless mentioned otherwise in this chapter, one-sided p-values can be obtained by halving the two-sided p-values.

By setting $\beta_1 = 0$ in model (13.1), we obtained the maximum likelihood estimates for β_2 to β_4 from PROC LOGISTIC as $\hat{\beta}_2 = 0.118$ (SE = 0.178), $\hat{\beta}_3 = 0.317$ (SE = 0.175), and $\hat{\beta}_4 = 0.521$ (SE = 0.178). Since the β's estimates increase with the dose, model (13.1) suggests that the cumulative odds for the lower response categories are a decreasing function of the dose. The likelihood ratio test for the goodness-of-fit of model (13.2) relative to model (13.1), obtained as the difference of -2 log-likelihood values between the two models, results in a likelihood ratio statistic of 0.13 with 2 degrees of freedom. The low value of the likelihood ratio statistic (therefore a high p-value) strongly suggests that the simpler model in (13.2) is appropriate for the data when compared to the model in (13.1).

The maximum likelihood estimate for β in (13.2) is $\hat{\beta} = 0.175$ (SE = 0.056). For Table 13.1, this means that as we move from one dose to the next higher dose, the odds of obtaining a more desirable outcome against a less desirable one is increased by 19% ($e^{0.175} = 1.19$). The Wald test for $\beta = 0$ produces a χ^2 statistic of 9.709 with 1 degree of freedom. This statistic is highly significant ($p = 0.002$ for a two-sided test), suggesting a monotone dose–response relationship on the cumulative odds scale.

Other choices to take advantage of the ordered categories include the adjacent-categories logit model that looks at the odds of being in two adjacent categories, i.e., $\ln(p_{ij}/p_{i,j+1})$, and the continuation-ratio logit model. The latter looks at $\ln(p_{ij}/\sum_{l=j+1}^{5} p_{il})$, the logarithmic odds of being in one category versus the categories above. While these other logit models are all reasonable models for ordinal response, the cumulative odds logit model is a natural extension of the binary response case because the former becomes the regular logit model when one chooses to collapse the ordinal response categories into two categories.

All the logit models can be further extended to include stratifying factors. Assuming that there are S strata defined by patient's characteristics at baseline, a straightforward extension of model (13.1) is model (13.3) in which the terms β_h^S, $h = 1, \ldots, H$, represent the stratum effect and β_i^D represent the dose effect. There is no treatment by stratum interaction in model (13.3). Furthermore, the proportional odds assumption now applies not only to dose groups but also to subgroups defined by the strata as well as those jointly defined by the dose and the stratum. PROC LOGISTIC can be used to estimate the model parameters and to test various hypotheses concerning β_i^D

$$\ln\left(\frac{\sum_{l=1}^{j} p_{ihl}}{\sum_{l=j+1}^{5} p_{ihl}}\right) = \alpha_j - \beta_h^S - \beta_i^D \tag{13.3}$$

13.2.2 Testing for a Monotone Dose–Response Relationship

A frequently asked question in dose–response studies is whether a monotone relationship exists between dose and the response. Section 13.2.1 discussed how the question on monotonicity could be addressed under a modeling approach. Following the discussion in Section 13.2.1, monotonicity can be interpreted as a more

favorable response with a higher dose. Since a more favorable outcome implies a smaller probability for the lower end of the response scale, the question on monotonicity can translate to a comparison on the cumulative probabilities. In other words, monotonicity can be evaluated by checking if $\sum_{l=1}^{j} p_{il}$ is a nonincreasing function of the dose $\{d_i, i = 1, 2, 3, 4\}$ for all $j = 1, \ldots, 4$. The latter implies testing a null hypothesis of equal distributions against a monotone stochastic ordering among the four dose groups as described below:

$$\mathrm{H}_0\colon\ p_{1j} = p_{2j} = p_{3j} = p_{4j}, \quad j = 1, \ldots, 4$$
$$\mathrm{H}_\mathrm{A}\colon \sum_{l=1}^{j} p_{1l} \geq \sum_{l=1}^{j} p_{2l} \geq \sum_{l=1}^{j} p_{3l} \geq \sum_{l=1}^{j} p_{4l}\ , \quad j = 1, \ldots, 4$$

Strict inequality holds for at least one j for one of the three inequalities included in the alternative hypothesis above. The subscript j above goes from 1 to 4 since $\sum_{l=1}^{5} p_{il} = 1$ for all i.

It should be pointed out that testing H_0 versus H_A as formulated above forces one to make a choice between a flat dose–response and a monotone one. Even though monotone dose–response is very common, other types of dose–response relationships are also possible. If there are reasons to anticipate beforehand that the dose–response relationship is substantially different from monotone, testing H_0 versus H_A as shown above will not be appropriate.

13.2.2.1 Tests Based on Association Measures

Since the response categories are ordinal, one can treat the response scale as quantitative and assign scores to the categories. One can also assign numerical values to the dose groups. With the assigned scores, one can use correlation-type association measures to tease out the linear component of the dose–response relationship and construct a χ^2 statistic with 1 degree of freedom to test for the significance of the correlation.

The most commonly used scores for the response categories are the equally spaced ones. When the desirability of moving from one category to the next depends strongly on the categories involved, other scores might be more appropriate. For example, it might be more appropriate to assign scores $\{0, 1, 2, 4, 8\}$ than $\{1, 2, 3, 4, 5\}$ to the response categories in Table 13.1. From our experience, conclusions are generally robust to the scores assigned to the response categories unless the data are highly imbalanced with many more observations falling in some categories than others. Because of this potential issue, it is often prudent to check the robustness of the conclusion by using several sets of scores.

For the dose group, one can use equally spaced scores, the actual doses, or the logarithmic doses to represent the treatment groups. Since trials typically randomize patients to the treatment groups, treatment groups are represented either similarly or according to a prespecified ratio. As a result, the above three choices of the numerical scores for the dose groups usually lead to similar conclusions on the existence of a linear relationship between the dose and the response.

One approach that does not require assigning scores is to use ranks of the observations. All observations in the same response category will have the same

rank r_j defined in Eq. (13.4). These ranks $\{r_j\}$ are called the midranks. Midranks are nothing but the averages of all ranks that would have been assigned to the observations in the same response category if we rank observations from the entire trial

$$r_j = y_{+1} + y_{+2} + \cdots + \frac{y_{+j}}{2} \tag{13.4}$$

In Eq. (13.4), $y_{+l} = \sum_{i=1}^{4} y_{il}$ is the total number of subjects with a response in the lth category.

The use of midranks seems appealing because one does not need to assign any scores. On the other hand, midranks cannot address the unequal spacing of the response categories from the clinical perspective. In addition, when one particular response category has very few observations, this response category will have a midrank similar to the preceding one. The latter might not be desirable if the two categories represent very different outcomes with drastically different medical implications.

Using PROC FREQ (CMH1 option) with scores $\{1, 2, 3, 4\}$ for the four dose groups and $\{1, 2, 3, 4, 5\}$ for the five response categories, we obtained a χ^2 test statistic of 9.61 with 1 degree of freedom. Under the null hypothesis of a zero correlation, this statistic produced a two-sided p-value of 0.002. Using $\{1, 2, 3, 4\}$ for doses and midranks for the response categories yielded a χ^2 statistic of 9.42 with a two-sided p-value of 0.002. Finally, using $\{1, 2, 3, 4\}$ for the dose groups and $\{0, 1, 2, 4, 8\}$ for the response categories produced a χ^2 statistic of 7.39 with a two-sided p-value of 0.007. In this case, the three sets of response scores produced similar results, all confirming a higher chance for a more favorable outcome with higher doses.

13.2.2.2 Tests Treating the Response as Continuous

Another application of assigning scores to the ordered categories is to treat the data as if they come from continuous distributions and apply standard normal theory methods to the data. The latter include approaches such as the analysis of variance and regression-type of analysis. There is evidence that treating ordinal data as continuous can provide a useful approximation as long as the number of categories is at least five (Heeren and D'Agostino, 1987). Under this approach, all methods developed for continuous data can be applied here. Interested readers are referred to other chapters in this book for a detailed account of the methods for continuous data.

One can, however, take into account the nonconstant response variance by explicitly incorporating the multinomial distribution structure when estimating the mean response for each dose group. Assuming that scores $\{s_j\}$ are assigned to the five response categories, the mean response score for the ith dose group, denoted by m_i, is defined by

$$m_i = \frac{\sum_{j=1}^{5} s_j \times p_{ij}}{\sum_{j=1}^{5} s_j}$$

Under the above definition, m_i can be thought of as a weighted average of $\{p_{ij}\}$ within each dose.

Grizzle et al. (1969) proposed to model $\{m_i\}$ as a function of the dose as in Eq. (13.5). The $\{m_i\}$ can be estimated by replacing p_{ij} with the observed proportions. Using multinomial distributions, one can calculate the variance of the estimated mean response for each group. The inverse of the variances can then be used as the weights when estimating the regression coefficients in Eq. (13.5) using the least-squares methods,

$$m_i = a + b \times (\text{dose}_i) \tag{13.5}$$

SAS® procedure PROC CATMOD with weight option can be used to estimate the parameters a and b in Eq. (13.5). Testing a monotone dose–response relationship (i.e., $b = 0$) can be accomplished using a χ^2 test statistic. Using $\{1, 2, 3, 4\}$ for the dose groups and $\{1, 2, 3, 4, 5\}$ for the response categories, PROC CATMOD produced a χ^2 statistic of 9.58 with a two-sided p-value of 0.002. Using the same dose scores but $\{0, 1, 2, 4, 8\}$ as the scores for the response categories produced a χ^2 statistic of 7.08. The latter has a two-sided p-value of 0.008.

Applying weighted least squares method to the mean response model in Eq. (13.5), we conclude that there is a monotone relationship between the mean response and the dose. As the dose increases, the mean response increases accordingly.

13.2.2.3 Jonckheere–Terpstra Test

Assume d_i and $d_{i'}$ are two doses such that $d_{i'} > d_i$. Consider the Wilcoxon-Mann-Whitney (WMW) statistic for testing equal distributions in response to these two doses against a stochastic ordering with response to dose $d_{i'}$ being stochastically greater than that to dose d_i. Let $\{r_{(i,i)j}\}$ represent the midranks constructed from dose groups d_i and $d_{i'}$ only, i.e.,

$$r_{(i,i')j} = (y_{i1} + y_{i'1}) + (y_{i2} + y_{i'2}) + \cdots + \frac{(y_{ij} + y_{i'j})}{2}$$

The WMW statistic for comparing groups i and i' can be constructed as

$$WMW_{i,i'} = \sum_{j=1}^{5} r_{(i,\ i')j}\, y_{i'j} - \frac{n_{i'}(n_{i'} + 1)}{2}$$

If there is a stochastic ordering between the response distributions to doses d_i and $d_{i'}$, we would expect the observed rank sum for dose $d_{i'}$ to be greater than the rank sum expected for that group if there is no difference between the two groups. In other words, we would expect $\text{WMW}_{\text{i,i'}}$ to be generally positive under the alternative hypothesis of a stochastic ordering between d_i and $d_{i'}$.

For four dose groups, there are six pairs of dose groups and six WMW statistics to compare the response distributions within each pair. Constructing WMW in such a way that the WMW statistic always looks at the difference between the expected and observed rank sums of the higher dose groups, we can express the

Jonckheere–Terpstra (Jonckheere, 1954; Terpstra, 1952) statistic (*JT*-statistic for short) as below

$$JT = \sum_{i'=2}^{4} \sum_{i=1}^{i'-1} WMW_{i,i'}$$

For large samples, the standardized value

$$z = \frac{JT - E(JT)}{\sqrt{var(JT)}}$$

provides a test statistic that has a standard normal distribution under the null hypothesis of equal response distributions across dose groups. Both SAS (PROC FREQ) and the StatXact (1995) software could conduct this test. For the data in Table 13.1, the standardized *JT*-statistic is 3.10, producing a two-sided p-value of 0.002. The conclusion from the *JT*-test is similar to that obtained from other approaches.

13.2.2.4 Summary

If the number of categories is at least five and the sample size is reasonable, the simplest approach is to treat the response as if it is continuous. This approach is particularly relevant if one intends to look at the change in the response at a follow-up visit from that at the baseline. In this case, change from baseline can be constructed using the scores assigned to the ordinal categories. This approach could be extended to include baseline covariates using an analysis of covariance model.

On the other hand, if there is much uncertainty in assigning scores to the categories or if the primary interest is in estimating the probability of a response in a particular category, modeling approach becomes a natural choice. Modeling approach is especially useful for dose-finding studies at the early stage of drug development when there is very little information on the dose response relationship. In this case, modeling allows us to borrow information from adjacent doses to study the effect of any particular dose.

13.3 When the Response is Binary

Binary endpoints are very popular in clinical trials. Frequently, "success" and "failure" are used to describe the outcome of a treatment. Even if the endpoint is continuous, there is an increasing tendency to define criteria and classify subjects as a "responder" or a "nonresponder". For example, patients in antidepressants trials are frequently referred to as a responder if they experience a 50% reduction in the HAM-D score from their baseline values. The American College of Rheumatology (ACR) proposed to use ACR20 as the basis to determine if the treatment is a success or not for an individual. ACR20 is defined as

- $\geq$ 20% improvement in tender joint count
- $\geq$ 20% improvement in swollen joint count
- $\geq$ 20% improvement in at least three of the following five assessment

- Patient pain assessment
- Patient global assessment
- Physician global assessment
- Patient self-assessed disability
- Acute phase reactant

The above definition combines multiple endpoints into a single dichotomous endpoint. By setting a criterion to classify treatment outcome, the medical community implicitly provides a target for treatment success. The popularity of responder analysis arises from the above desire even though dichotomization can lead to the loss of information (Senn, 2003).

There are situations when binary response makes sense. Examples include "alive" or "dead" for patients in salvage trials with end stage cancer. In anti-infective trials, it is natural to consider if an individual is cured of the underlying infection, both clinically and microbiologically. There are many situations where dichotomizing subject's response in a manner that makes clinical sense is not a trivial matter. This is especially so when the response is measured using an instrument. Does a 50% improvement in the HAM-D scale from the baseline in depressed patients translate to clinically meaningful improvement? Is the rule we use to dichotomize patients sensitive to drug effect? All these are important questions when determining the responder definition.

We will assume that a binary endpoint is appropriately defined and the objective is to explore the relationship between the likelihood of the desirable outcome and the dose. Using Table 13.1 as an example, we will assume that it is reasonable to collapse the three categories of death, vegetative state, and major disability into one category and combining minor disability and good recovery into another. The first (combined) category is deemed undesirable while the second (combined) category is the desirable one. Following our previous notations, we will label the two response categories as $j = 1$ and 2. The four dose groups will be labeled as $i = 1, 2, 3, 4$, respectively. Response categories after combination are given in Table 13.2.

The binary case can be considered as a special case of the ordinal response discussed in Section 13.2. For example, the logit model in (13.6) is subsumed in the proportional odds logit model described in (13.1). Similarly, the logit model

Table 13.2. Collapsing the first three response categories and the last two response categories in Table 13.1 to form a binary response consisting of "undesirable" and "desirable" categories

	Outcome		
Treatment group	Undesirable	Desirable	Total
Placebo	130	80	210
Low dose	113	77	190
Medium dose	112	95	207
High dose	96	99	195

in (13.7) is subsumed in the proportional odds logit model in (13.2).

$$\ln\left(\frac{p_{i2}}{p_{i1}}\right) = \alpha - \beta_i \tag{13.6}$$

$$\ln\left(\frac{p_{i2}}{p_{i1}}\right) = \alpha - \beta\, d_i \tag{13.7}$$

Because of the relationship between the above and their counterparts for the ordinal response, estimation and testing related to models (13.6) and (13.7) can be conducted similar to those for models (13.1) and (13.2). Setting $\beta_1 = 0$, the maximum likelihood estimates for $\{\beta_i\}$ in (13.6) are $\hat{\beta}_2 = 0.102$ (SE = 0.205), $\hat{\beta}_3 = 0.321$ (SE = 0.199), and $\hat{\beta}_4 = 0.516$ (SE = 0.202). Maximum likelihood estimate for β in Eq. (13.7) is $\hat{\beta} = 0.176$ (SE = 0.064). The Wald test for $\beta = 0$ produced a two-sided p-value of 0.006.

As for approaches that assign scores to response categories, one can easily show that with only two response categories, one will reach identical conclusions regardless of the scores assigned. Since there are only two response categories for a binary outcome, the approach of treating the data as if they are from continuous distributions (Section 13.2.2.2) is generally not encouraged. On the other hand, the mean response model with parameters estimated using the weighted least-squares method is still appropriate. In the latter case, one can choose (0, 1) scores so that the mean response is actually the probability of the desirable response. The mean response model in (13.5) now reduces to

$$p_{i2} = a + b \times (\text{dose})_i \tag{13.8}$$

In general, fitting model (13.8) could be a challenge if one wants to incorporate the constraint that $\{p_{i2}, i = 1, 2, 3, 4\}$ are between 0 and 1. For Table 13.2, with numerical scores $\{1, 2, 3, 4\}$ for the doses, the weighted least-squares estimates for a and b are 0.331 and 0.043 with standard errors of 0.042 and 0.016, respectively. These estimates produced weighted least-squares estimates of 0.374, 0.417, 0.460, and 0.503 for $\{p_{i2}, i = 1, 2, 3, 4\}$.

For the binary case, the association-based approach is closely related to the Cochran-Armitage (1955) test that is designed to detect a linear trend in the response probabilities with dose. Mancuso et al. (2001) proposed to use isotonic regression to increase the power of common trend tests in situations where a monotone dose response relationship is imposed. They developed the isotonic versions of the Cochran-Armitage type trend tests and used bootstrap method to find the empirical distributions of the test statistics. Using simulations, they demonstrated that the order-restricted Cochran-Armitage type trend tests could increase the power of the regular Cochran-Armitage trend test. When using $\{1, 2, 3, 4\}$ as the scores for doses, the Cochran-Armitage test for detecting a linear trend in $\{p_{i2}\}$ produced a χ^2 statistic of 7.65 with a two-sided p-value of 0.006.

For the data in Table 13.1 and the collapsed data in Table 13.2, all approaches confirm a monotone dose–response relationship. As the dose increases, so is the

probability for a more favorable outcome. Since the binary case is a simplified case of the ordinal data, we will not devote more attention to this special case.

13.4 Multiple Comparisons

Dose–response studies can be conducted at different stages of a drug development program. They can be the studies to establish proof of concept or to establish a dose to bring into the confirmatory phase. Because the objectives of dose–response studies at various development phases are different, the analytic approaches to handling the data should vary accordingly. For dose response studies to establish proof of concept, the focus is to see if the response varies with the dose to suggest any drug activity. Therefore, the analysis will focus on estimation. This also applies to many Phase IIb studies that are designed to correctly identify the dose(s) with adequate treatment benefit. In this case, studying the trend and identifying doses by the observed mean responses will be key since the studies might not be sufficiently powered to detect a clinically meaningful difference between doses. There are also situations where studies are powered to differentiate between pairs of groups, but not powered to do so with adjustment for multiple comparisons. In any cases, the analysis should be conducted to specifically address their objectives.

When multiple doses are included in a confirmatory trial and the goal is to test the efficacy of each dose against the control (often a placebo), statistical analyses should be adjusted for multiple comparisons. The latter will be the focus of this section. Unlike the previous two sections where hypothesis testing, when employed, is to check for a monotone dose–response relationship, a monotone dose response is not necessarily the basis for hypothesis testing in this section. Instead, definitively differentiating between treatment groups (especially doses of an investigational medication from the control) will be the primary objective.

The primary objective of a multiple testing procedure is to control the overall probability of erroneously rejecting at least one null hypothesis irrespective of which and how many of the null hypotheses of interest are in fact true. Many multiple testing procedures have been proposed in the literature. In general, they fall in two classes. The first class includes procedures that are developed specifically for continuous data such as the Dunnett's method (1965) and the procedure by William (1971). The second class includes procedures that are "distributional free" in the sense that their implementation does not depend on any particular distributional assumption. Most of the procedures in this class are derived from the closed testing procedure proposed by Marcus et al. (1976) and work directly with the p-values produced from individual tests. As such, procedures in the second class are readily applicable to categorical data.

We will assume in this section that the objective of pairwise comparisons is to unequivocally identify doses that have significantly different effect from the control. For data in Tables 13.1 and 13.2, this means comparing low, medium, and high doses against the placebo. Except for the Dunnett's procedure described in Section 13.4.5, we will focus on approaches that compare p-values to adjusted

significance levels with adjustments determined by the multiple testing procedures. We will look at four most commonly used procedures.

When the proportional odds model (13.1) is employed to analyze the ordinal response data such as in Table 13.1 with the convention of $\beta_1 = 0$, comparing each dose group to the placebo is equivalent to testing if $\beta_i = 0$ for $i = 2, 3, 4$. Dividing β_i estimate by its asymptotic standard error, we obtained z-statistics of 0.663, 1.811, and 2.937 for testing $\beta_2 = 0$, $\beta_3 = 0$, and $\beta_4 = 0$, respectively. The two-sided p-values associated with these three z-statistics under their respective null hypotheses are 0.507, 0.070, and 0.003.

We can also obtain p-values for comparing each dose to the placebo by assigning scores to the ordinal categories and treating the data as if they are continuous. When doing this, multiple comparison procedures developed for normal distribution could be applied. Alternatively, one can apply the Wilcoxon-Mann-Whitney test to compare each dose group to the placebo. Similarly, one can either use the modeling approach or compare two proportions directly for the binary case. For Sections 13.4.1 through 13.4.4, we will assume that p-values corresponding to the hypotheses of interest have been produced. We will assume throughout Section 13.4 that the overall Type I error rate is to be controlled at the 5% level.

We would like to point out that multiplicity adjustment for model-based approaches with large sample sizes may be done using parametric re-sampling techniques. Macros for doing these are provided in Westfall et al. (1999) and an example for binary outcome is given in Chapter 12 of that book. For the rest of this chapter, large sample asymptotic normal approximations are used to derive the significance levels.

13.4.1 Bonferroni Adjustment

Under this approach, we will compare each p-value to 0.0167 (=0.05/3) since we will make three comparisons. Because of its simplicity, Bonferroni adjustment is often used despite its conservativeness. Taking the three p-values cited above, i.e., 0.507, 0.070 and 0.003, only 0.003 is smaller than 0.0167. Thus, applying the Bonferroni procedure, one could only conclude that the high dose produced a significantly better result than the placebo.

13.4.2 Bonferroni–Holm Procedure

This procedure calls for ordering the p-values from the smallest to the largest. In our case, this lead to the order of 0.003 (high dose), 0.070 (medium dose), and 0.507 (low dose). If the smallest p-value is smaller than 0.0167 (=0.05/3), we will move to the next smallest p-value; otherwise we will stop and conclude that no dose group is significantly different from the placebo. In our case, 0.003 is less than 0.0167, so we continue to 0.070, the next smallest p-value. We will compare 0.070 to 0.025 (0.05/2). Since 0.070 is greater than 0.025, we will stop the comparison and conclude that only the high dose produced results that are significantly different from the placebo. Should the second smallest p-value be

smaller than 0.025, we would proceed to the next p-value in the ordered sequence. In other words, we continue the process with a significance level that is 0.05 divided by the number of hypotheses remaining to be tested at each stage unless the p-value under comparison exceeds the current significance level. When this occurs, we will conclude significance for all comparisons before the present one.

13.4.3 Hochberg Procedure

This procedure is among the most popular multiple comparison procedures by pharmaceutical statisticians. Instead of ordering the p-values from the smallest to the largest, this procedure orders the p-values from the largest to the smallest. In our example, the ordered p-values are 0.507 (low dose), 0.070 (medium dose), and 0.003 (high dose). The largest p-value will be compared to 0.05. If it is smaller than 0.05, we will stop the testing and conclude significance for all comparisons; otherwise we will move to the second highest p-value. In our case, the largest p-value, i.e., 0.507 is greater than 0.05, so we will continue. The second largest p-value will be compared to 0.025 (=0.05/2). If it is smaller than 0.025, we will stop and conclude significance for this comparison and all subsequent ones that produced p-values smaller than the current one. Since 0.07 is greater than 0.025, we will continue. The smallest p-value in our example will be compared to 0.0167 (0.05/3). Since 0.003 is smaller than 0.0167, we will conclude a significant difference between the high dose and the placebo. Under the Hochberg procedure, the process starts with the largest p-value and the significance level decreases as we proceed. The significance level for the kth step is given by $0.05/k$. Unlike the Holm procedure, the Hochberg procedure continues the testing until we reach a statistical significance, otherwise it will conclude that none of the doses is statistically different from the placebo.

13.4.4 Gate-Keeping Procedure

This procedure is also known as predetermined step-down or the hierarchy procedure (Bauer and Budde, 1994; Bauer et al., 1998). In short, this procedure follows a prespecified sequence. Testing will be conducted at the 0.05 level at each stage and it will continue as long as the p-value is significant at the 0.05 level. Testing will stop at the first instance when a p-value is above 0.05.

This procedure is used very frequently when there is a prior belief of a monotone dose–response relationship and therefore it is logical to start with the highest dose first. This procedure is especially helpful when looking for the minimum effective dose (Tamhane et al., 1996). To look for a minimum effective dose under the strong belief of a monotone dose–response relationship, one can start by comparing the highest dose with the control and working our way down the doses. The minimum effective dose is often defined as the smallest dose for which the null hypothesis of no effect is rejected. Another appealing feature is that all comparisons are conducted at the level of 0.05. Despite its appeal and ease to implement, if the prior belief turns out to be false and the dose–response relationship turns out to be

umbrella-shaped, the predetermined step-down procedure can miss the opportunity to identify effective doses.

In our example, if we choose (high, medium, low) as the testing sequence based on biologic considerations, we will reach the same conclusion as the previous procedures. That is, the high dose is the only one demonstrating a statistically different effect from the placebo.

13.4.5 A Special Application of Dunnett's Procedure for Binary Response

Chuang-Stein and Tong (1995) examined three approaches for comparing several treatments with a control using a binary outcome. The first approach relies on the asymptotic theory applied to the Freeman and Tukey (1950) transformation of the observed proportions. The second finds an acceptance region based on the binomial distributions estimated under the joint null hypotheses. The third approach applies Dunnett's procedure to the binary data. The authors found that for sample sizes typical of the confirmatory trials, applying Dunnett's critical values to the z-statistics obtained from comparing proportions results in an actual overall Type I error rate generally at the desirable level.

For the data in Table 13.2, the z-statistics for comparing each dose against the placebo are 0.497 for the low dose, 1.618 for the medium dose, and 2.585 for the high dose. Dunnett's critical value for three comparisons and a sample sizes greater than 160 per group is 2.212 (Hsu, 1996, Table E.3). Compared to 2.212, only the comparison between the high dose and the placebo reached statistical significance at the 0.05 level.

Occasionally, one might want to compare among doses that have been established to be efficacious. Our recommendation is to make these comparisons without worrying about multiplicity adjustment. This is because the latter are secondary to the primary objective of identifying efficacious doses.

Ruberg (1995a,b) noted that dose–response studies routinely ask four questions. They are (1) Is there any evidence of a drug effect? (2) Which doses exhibit a response different from the control group? (3) What is the nature of the dose–response relationship? (4) Which is the optimal dose? One can discuss the first three questions either in the context of safety (Hothorn and Hauschke, 2000) or efficacy data. The prevailing practice is to focus on safety and efficacy data separately without making a conscious effort to integrate them. Compared to the first three, the last question can only be answered when safety and efficacy are considered jointly. The latter is outside the scope of this chapter.

13.5 Discussion

Ordinal data occur frequently in real life. Likert scale is frequently used to record a subject's response to a question or to an external intervention. Because of the

way the scale is constructed, it is intuitive to use scores such as {–2, –1, 0, 1, 2} for the five-category scale and {–3, –2, –1, 0, 1, 2, 3} for the seven-category scale. Other examples of ordinal response come from using instruments to record outcome reported by both patients and their treating physicians. Many instruments contain questions that are ordinal in nature. Even though the score summed over the various questions is often the primary point of interest, analysis of specific questions leads to the analysis of ordinal data.

In this chapter, we discussed both the modeling and the testing approaches. In our opinion, modeling approach, for all its advantages, is underutilized. Modeling approach can handle covariates and predict the chance for achieving certain response for a given dose as well as the uncertainty associated with the prediction. In addition, a fitted model can be used to estimate the dose within the dosing range that has a desirable probability to produce certain response. By plotting the observed logit against the dose (or ln dose), one can get some indication whether the assumption of a monotone dose–response relationship is likely to be supported by the data or not. For example, if there is a downward trend in response when the dose moves toward the high end, one might want to consider including a quadratic term in dose (or ln dose) to describe the umbrella-like relationship. In addition to modeling, distribution-free tests for umbrella alternatives were studied by Chen and Wolfe (1990).

To be useful, models typically come with accompanying assumptions to aid interpretation. The proportional odds logit models require constant odds ratios among dose groups on the cumulative probability scale. Models such as those in (13.2) and (13.7) describe a linear dose effect. Some researchers (e.g., Mantel, 1963) considered such requirements appropriate as long as the required conditions constitute a major component of the phenomenon under examination. For example, the linear models as in (13.2) and (13.7) are reasonable as long as the linearity assumption holds for the underlying dose–response relationship. In using a linear model, we are able to construct a powerful test for a hypothesis that suggests a monotone dose–response relationship. In most cases, the linearity assumption and the proportional odds assumption can be checked via the goodness-of-fit based on the likelihood ratio tests.

Calculating sample size for binary outcome when comparing each dose group against the control is straightforward. If multiplicity adjustment is needed, a conservative approach is to use the Bonferroni significance level as the Type I error in the calculation. Sample size calculated in this way will be adequate when other more efficient multiple comparison procedures are used in the analysis. If the number of categories associated with an ordinal response is at least five and the analysis calls for treating the data as continuous, the calculation of the sample size can proceed as for the continuous case. A detailed discussion on sample sizes needed for dose response studies is provided in Chapter 14 of this book.

Sample size calculation for the modeling approach is more complicated. Whitehead (1993) discussed the case of comparing two groups based on the proportional odds logit model. Suppose we want an 80% power in a two-sided 5% test for detecting a size of β_0 in a model like (13.2). Assuming a randomization ratio of A

to 1 to the two groups and $\{\overline{p}_j,\ j = 1, \ldots, J\}$ the anticipated marginal proportions of the response categories, Whitehead showed that the total required sample size is

$$N = \frac{3(A+1)^2(z_{0.975} + z_{0.80})^2}{A\beta_0^2\left(1 - \sum_j \overline{p}_j^3\right)}$$

where z_c above represents the $100 \times c\%$ percentile of the standard normal distribution. A lower bound for N can be obtained by substituting $1/J$ for $\overline{p}_j$ in the above formula. Whitehead showed that the required sample size did not differ much from this lower bound unless a single dominant response category occurred. It can be easily seen that equal allocation, i.e., $A = 1$, produces the smallest sample size.

Using Whitehead formula, Chuang-Stein and Agresti (1997) discussed the effect of the choice of the number of response categories on the sample size. In particular, they discussed the sample size required for J categories $N(J)$ and that required for two categories $N(2)$ for the case of equal marginal response probabilities. The ratio of $N(J)$ to $N(2)$ is

$$\frac{N(J)}{N(2)} = \frac{0.75}{1 - J^{-2}}$$

For $J = 5$, the above ratio is about 78%, suggesting a substantial loss of information when collapsing five response categories into two. This observation is consistent with our earlier comments on the loss of information when dichotomizing a nonbinary response. Even though Whitehead's original discussion was applied to two-arm trials, the discussion is relevant to dose–response studies when the primary focus is to compare each dose group against, for example, the placebo.

For dose–response studies conducted at the earlier development phase, the objective might not be to statistically differentiate between doses, but to correctly identify doses that have better efficacy. For example, when studying a new antibiotic at the Phase II stage, the primary objective of a dose–response trial is often to pick a dose to bring to the confirmatory phase. Such a trial might contain only doses of the new antibiotic. The major consideration for sample size decision is to make sure that we have enough patients at each dose so that the probability of correctly identifying the dose with the best efficacy using the observed success rate is at a desirably high level. For example, we might want to have an 80% chance that the observed success rates will correctly reflect the ordering in the true response rates when the underlying true rates are 60% and 50%, respectively. If this is the objective, then we will need approximately 35 patients per dose group. On the other hand, if one wants to differentiate between these two doses with an 80% power using a hypothesis testing procedure at a two-sided 5% significance level, one will need 408 patients per dose group. The latter is excessive for a Phase II antibiotic trial, especially when in vitro testing and animal model have already confirmed the antibacterial activity of the compound under investigation. Some design considerations for dose–response studies can be found in Wong and Lachenbruch (1998).

The analysis approaches discussed in this chapter are applicable to a parallel design under which subjects are randomized to receive one of the treatments (doses) under comparison. In some therapeutic areas, early phase dose–response studies are done using a titration design. Under a titration design, subjects typically start with the lowest dose and have their doses titrated upwards until treatment intolerance or the obtainment of a response. The exploration of a dose–response relationship with a binary outcome in a titration study requires special care because of the selective nature of the titration scheme. For more details on the analysis of such studies, the readers are referred to Chuang (1987).

There is a great flexibility in analyzing dose–response studies when the endpoint is measured on an ordinal scale. Since most of the discussion in this chapter is for the average situation, one might want to consider evaluating the use of a general-purpose approach for a particular situation more thoroughly. To this end, we would like to encourage our readers to diligently use simulations to evaluate various design options. The latter includes the sample size. For example, one can simulate studies under various conditions to see if the planned size provides adequate power to address the research objectives. In addition to sample size, how missing data are handled (e.g., the baseline response category carried forward, the worse response category experienced by the individual, or the worst response category) could have a significant impact on power. Furthermore, most multiple comparison procedures, when applied to categorical data, rely on the asymptotic behaviors of the underlying test statistics. Whether the asymptotic approximation is adequate for a particular application needs to be assessed for the planned sample size, the response profiles, the dropout patterns, as well as the choices of the analytical approaches (modeling vs. comparing proportions directly). With the convenience of modern computing power, it is highly desirable to take advantage of these tools so we have a good understanding of the operating characteristics of the procedures chosen before we initiate a clinical trial.

References

Agresti, A. 2000. *Categorical Data Analysis*, 2nd ed. New York: Wiley.

Armitage, P. 1955. Tests for linear trends in proportions. *Biometrics* 11:375–386.

Bauer, P., and Budde, M. 1994. Multiple testing for detecting efficient dose steps. *Biometrical Journal* 36:3–15.

Bauer, P., Rohmel, J., Maurer, W., and Hothorn, L. 1998. Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* 17:2133–2146.

Chen, Y.I., and Wolfe, D.A. 1990. A study of distribution-free tests for umbrella alternatives. *Biometrical Journal* 32:47–57.

Chuang, C. 1987. The analysis of a titration study. *Statistics in Medicine* 6:583–590.

Chuang-Stein, C., and Tong, D.M. 1995. Multiple comparison procedures for comparing several treatments with a control based on binary data. *Statistics in Medicine* 14:2509–2522.

Chuang-Stein, C., and Agresti, A. 1997. A review of tests for detecting a monotone dose–response relationship with ordinal response data. *Statistics in Medicine* 16:2599–2618.

Dunnett, C.W. 1965. A multiple comparisons procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 60:573–583.

Freeman, M.F., and Tukey, J.W. 1950. Transformations related to the angular and the square root. *Annals of Mathematical Statistics* 21:607–611.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969) Analysis of categorical data by linear models. *Biometrics* 25, 489–504.

Heeren, T., and D'Agostino, R. 1987. Robustness of the two independent sample t-test when applied to ordinal scale data. *Statistics in Medicine* 6:79–90.

Hothorn, L.A., and Hauschke, D. 2000. Identifying the maximum safe dose: A multiple testing approach. *Journal of Biopharmaceutical Statistics* 10:15–30.

Hsu, J.C. 1996. *Multiple Comparisons—Theory and Methods*. London, UK: Chapman & Hall.

Jonckheere, A.R. 1954. A distribution-free K sample test against ordered alternatives. *Biometrika* 41:133–145.

Lewis, J.A. 2004. In defence of the dichotomy. *Pharmaceutical Statistics* 3:77–79.

Mancuso, J.Y., Ahn, H. and Chen, J.J. (2001) Order-restricted dose-related trend tests. Statistics in Medicine 20, 2305–2318.

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

Marcus, R., Peritz, E., and Gabriel, K.R. 1976. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63:655–660.

McCullagh, P. 1980. Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society Series B* 42:109–142.

Ruberg, S.J. 1995. Dose-response studies I: Some design considerations. *Journal of Biopharmaceutical Statistics* 5:1–14.

Ruberg, S.J. 1995. Dose-response studies II: Analysis and interpretation. *Journal of Biopharmaceutical Statistics* 5:15–42.

SAS, version 9.0. 2003. Cary, NC: SAS Institute Inc.

Senn, S. 2003. Disappointing dichotomies. *Pharmaceutical Statistics* 2:239–240.

StatXact 1995. *StatXact3 for Windows: Statistical Software for Exact Nonparametric Inference, User Manual*. Cytel Software.

Tamhane, A.C., Hochberg, Y., and Dunnett, C.W. 1996. Multiple test procedures for dose finding. *Biometrics* 52:21–37.

Terpstra, T.J. 1952. The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indigationes Mathematicae* 14:327–333.

Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D., and Hochberg, Y. 1999. *Multiple Comparisons and Multiple Tests Using the SAS System*. Cary, North Carolina: SAS Institute Inc.

Whitehead, J. 1993. Sample size calculations for ordered categorical data. Statistics in Medicine 12, 2257–2271.

William, D.A. 1971. A test for difference between treatment means when several dose levels are compared with a zero dose control. *Biometrics* 27:103–117.

Wong, W.K., and Lachenbruch, P.A. 1998. Designing studies for dose response. *Statistics in Medicine* 15:343–359.

Appendix: SAS Code for Performing Various Analyses

SAS code for performing various analyses with data in Table 13.1

```
data one;
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 data=one; *CMH test with scores entered in the data;
weight count;
tables group*dose*outcome/cmh1;
run;

proc freq data=one; *CMH test with mid-rank scores;
weight count;
tables group*dose*outcome/cmh1 scores=ridit;
run;

proc catmod data=one order=data; *mean response model;
weight count;
population dose;
response 1 2 3 4 5; direct dose; *use scores (1,2,3,4,5);
model outcome=dose;
run;

proc logistic data=one; *proportional odds model (ML);
freq count;
model outcome=dose;
run;

proc catmod data=one; *proportional odds model (WLS);
weight count;
response clogits;
direct dose;
model outcome=_response_ dose;
run;

proc catmod data= one; *adjacent cat. Logit model (WLS);
weight count;
response alogits;
direct dose;
```

```
model outcome=_response_ dose;
run;
```

SAS code for performing various analyses after classifying data in Table 13.2

```
data two;
input dose outcome count;
cards;
1 0 130
1 1 80
2 0 113
2 1 77
3 0 112
3 1 95
4 0 96
4 1 99
proc logistic data=two; *treating dose levels as a continuous variable;
freq count;
model outcome=dose;
run;

data three;
set two; *create dummy variables for dose levels;
if dose=2 then idose2=1; *placebo group is treated as a reference level;
else idose2=0;
if dose=3 then idose3=1;
else idose3=0;
if dose=4 then idose4=1;
else idose4=0;
run;

proc logistic data=three; *treating dose levels as nominal categories;
freq count;
model outcome=idose2 idose3 idose4;
run;

proc freq data=two; *Cochran-Armitage trend test;
weight count;
tables dose*outcome/trend;
run;
```