Environmental Health Perspectives
Vol. 79, pp. 217-221, 1989

Trend Tests for Proportional Responses in Developmental Toxicity Experiments

by Vernon M. Chinchilli* and B. Christine Clark†

The data from developmental toxicity experiments usually are very difficult to analyze statistically because of the lack of independence among littermates and the random nature of the litter size. Only a few of the models that have been proposed in the literature have accounted for both of these features. One of the models proposed by Van Ryzin is invoked to construct a test of trend (dose response). The construction is achieved via a statistical technique called isotonic regression, which is applied to the moment estimators derived by Van Ryzin. The trend test based on isotonic regression is relatively straightforward to calculate, and when the number of dose groups (including control) is four or less, the significance of the observed result is easily determined. An example, in which fetolethality is the end point of interest, demonstrates the test.

Introduction

Developmental toxicity experiments with laboratory animals pose difficult problems in terms of statistical analysis. In such studies it is best to assume that the litter is the experimental unit and not the fetus or pup (*1,2*). This is because the parental animal is randomized to treatment, the parental animal receives the treatment directly, and fetuses or pups within a litter usually exhibit a litter effect in that they do not respond independently of one another. Ignoring the litter effect could lead to statistical tests that are too liberal in the sense that the Type I error rate is inflated (the null hypothesis is rejected more often than it should be).

Many of the responses that are measured in a typical experiment are proportional in nature, e.g., the proportion of affected fetuses in the litter. The naive approach to a statistical analysis is to regard the proportions as samples from a binomial distribution. However, this is not legitimate for developmental toxicity data because of the litter effect and because the litter size itself is random, i.e., not fixed.

Some of the statistical models and analyses that have been proposed for proportional data from developmental toxicity experiments have attempted to incorporate these aspects. Haseman and Kupper review various approaches to the analysis of a proportional response variable when comparing a treatment group to a control group in developmental toxicity studies (*3*). The approaches can be categorized into four groups: (a) generalized binomial models; (b) nonparametric analysis; (c) transformations of proportions; (d) resampling techniques.

The beta-binomial model (*4*) is the most popular generalized binomial model applied to developmental toxicity data. It is derived by assuming that X, the number of positive responses in a litter, follows a binomial distribution with probability of success θ, $0 < \theta < 1$, and that θ itself is a random variable following a beta distribution. Then the unconditional distribution of X is the beta-binomial. Whereas the binomial is a one-parameter model and assumes that the fetuses are independent (no litter effect), the beta-binomial is a two-parameter model and assumes a nonnegative correlation among fetuses. For the situation of a control group and more than one dose group, a trend test within the beta-binomial framework can be constructed as a logistic function of dose (*5,6*). One drawback with the beta-binomial model is that it does not allow for the random nature of the litter size.

Van Ryzin considered several different generalized binomial models in which the random litter size is taken into account (*7*). He derived a moment estimator for the probability of an effect within each group. An unfortunate aspect of his work is that it has not been applied very often, and his models have not been included in any of the simulation studies that have appeared in the statistical literature. One of his models is discussed in more detail in "Methods," and a trend

*Department of Biostatistics, Medical College of Virginia, Virginia Commonwealth University, Richmond, VA 23298.
†A. H. Robins Company, 1211 Sherwood Avenue, Richmond, VA 23220.

test based on isotonic regression and his moment estimators is constructed.

In an entirely different approach, Rai and Van Ryzin used a binomial framework and incorporated litter size as a covariate in modeling the conditional probability of response (*8*), then they arrived at the unconditional probability by assuming a Poisson distribution for the litter size. Their purpose was to provide a model for low-dose extrapolation and not necessarily a trend test. Williams has criticized this model because it does not account for extra-binomial variability (*6*).

In a nonparametric approach, the proportional responses for the litters are ranked and a Kruskal-Wallis test applied, which reduces to the Wilcoxon rank sum test or Mann-Whitney test in the case of only one dose group. Gaylor discussed in detail nonparametric analyses for developmental toxicity studies (*9*). The nonparametric version of the trend test is known as Jonckheere's test. See Lin and Haseman (*10*) on how to modify Jonckheere's test to allow for ties, a situation usually encountered with the proportional responses observed in developmental toxicity studies. These rank tests are computationally easy and avoid some of the distributional difficulties of the generalized binomial models. Although random litter sizes do not invalidate this approach, the rank tests may not handle this problem in an optimal manner. For instance, a proportion of 1/3 receives the same rank as 4/12, even though the latter proportion provides more information.

In the transformation approach, the observed proportions for the litters are transformed to approximate normal random variables so that one-way ANOVA techniques can be applied. Two useful transformations are the arc-sine and the Freeman-Tukey binomial arc-sine transformations (*3*). This is justified if the proportions behave in a binomial manner, which usually is not the case with developmental toxicity data.

Gladen invoked jackknife methodology to estimate the probability of response within each treatment group weighted according to litter size (*11*). After using this method, a one-way ANOVA of the jackknifed estimators, weighted by their estimated variances, can be justified asymptotically. Analogous approaches using other resampling plans, such as the bootstrap (*12*), have been suggested but their application to developmental toxicity studies has not appeared in the literature.

A number of researchers have conducted simulation studies to compare the performance of these procedures. Haseman and Kupper reviewed these studies and concluded that none of the above procedures seems to be superior in terms of statistical power and attaining the desired significance level (*3*). After their review article, other simulation studies appeared in the literature (*13,14*). However, all of these simulation studies have not seriously examined the situation with random litter sizes. Out of all the procedures discussed, only the generalized binomial models of Van Ryzin (*6*) and the nonparametric analyses with rank tests legitimately account for random litter size.

Methods

Model

One of the models considered by Van Ryzin (*7*) is used to illustrate the construction of a trend test via isotonic regression. At this point the model is described just for one litter. Let X denote the number of affected fetuses or pups out of a litter of size N, and let P denote the probability that a fetus or pup is affected. Out of the triplet (X, N, P) of random variables, X and N are observable, while P is not. It is assumed that the conditional distribution of X given that N = n and P = p is binomial, i.e.,

$$\Pr[X = x \mid N = n,\ P = p] = \binom{n}{x} p^x (1-p)^{n-x},\quad x = 0,1,\ldots,n \text{ and } 0 < p < 1. \qquad [1]$$

Next, it is assumed that the expected value of P, denoted by $E(P)$, is the unknown parameter μ, and that the conditional expectation of P given N = n is $\alpha\beta^n$, where α and β are unknown parameters with $0 < \alpha \leq 1$ and $0 < \beta \leq 1$. Van Ryzin (*7*) thought this regression model of the conditional expectation of P given N = n to be realistic because (a) the probability of response P is not independent of the litter size, and (b) for values of $\beta < 1$, the probability of any individual fetus being affected is a decreasing function of the litter size, i.e., the proportion affected in a large litter is likely to be smaller than the proportion affected in a small litter.

The final assumption is that N has an unconditional Poisson distribution:

$$\Pr[N = n] = e^{-\lambda}\lambda^n/n!,\quad n = 0,1,2,\ldots \qquad [2]$$

Van Ryzin (*7*) defined the following two variables

$$Y = \begin{cases} X/N, & \text{if } N > 0 \\ 0, & \text{if } N = 0 \end{cases} \qquad [3]$$

and

$$V = \begin{cases} 1, & \text{if } X = N = 1 \\ 0, & \text{otherwise} \end{cases},$$

and noted that

$$\mu = E(Y) + \frac{E(V)}{\log\{E(X)\} - \log\{E(V)\}}. \qquad [4]$$

Now suppose that there are k groups (usually a control and k–1 dose groups) with m_i litters in the ith group, i = 1,2, . . . , k. Let X_{ij} denote the number of affected fetuses out of the N_{ij} fetuses in the jth litter of the ith group, i = 1,2, . . . , k and j = 1,2, . . . , m_i. It is assumed that within the ith group the m_i triplets of random variables (X_{ij}, N_{ij}, P_{ij}) are identically distributed with $\mu_i = E(P_{ij})$ being the expected value of the probability of

an affected fetus, i = 1,2, . . . , k. The following moment estimators

$$\bar{Y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} Y_{ij},\ \bar{V}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} V_{ij},$$

and

$$\bar{X}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} X_{ij} \qquad [5]$$

lead to a moment estimator of μ_i from Equation 4 as

$$\hat{\mu}_i = \begin{cases} \bar{Y}_i + \bar{V}_i/\{\log(\bar{X}_i)-\log(\bar{V}_i)\}, & \text{if } \bar{X}_i > \bar{V}_i > 0 \\ \bar{Y}_i, & \text{if } \bar{X}_i > \bar{V}_i = 0 \\ 1, & \text{if } \bar{X}_i = \bar{V}_i \end{cases} \qquad [6]$$

for the ith group, i = 1,2, . . . , k. Van Ryzin (*7*) proved that as $m_i \to \infty$, i.e., the number of litters in the ith group gets very large, then (a) $\hat{\mu}_i \to \mu_i$ (strong consistency), and (b) $(m_i)^{1/2}(\hat{\mu}_i - \mu_i)/\hat{v}_i$ has a standard normal distribution, where

$$\hat{v}_i^2 = \sum_{r=1}^{3}\sum_{s=1}^{3} \hat{\sigma}_{irs}\hat{\theta}_{ir}\hat{\theta}_{is}, \qquad [7]$$

$$\hat{\sigma}_{i11} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}(Y_{ij}-\bar{Y}_i)^2,$$

$$\hat{\sigma}_{i22} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}(V_{ij}-\bar{V}_i)^2, \qquad [8a]$$

$$\hat{\sigma}_{i33} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}(X_{ij}-\bar{X}_i)^2,$$

$$\hat{\sigma}_{i12} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}(Y_{ij}-\bar{Y}_i)(V_{ij}-\bar{V}), \qquad [8b]$$

$$\hat{\sigma}_{i13} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}(Y_{ij}-\bar{Y}_i)(X_{ij}-\bar{X}),$$

$$\hat{\sigma}_{i23} = \frac{1}{m_i-1}\sum_{j=1}^{m_i}(V_{ij}-\bar{V}_i)(X_{ij}-\bar{X}), \qquad [8c]$$

$$\hat{\sigma}_{i21} = \hat{\sigma}_{i12},\ \hat{\sigma}_{i31} = \hat{\sigma}_{i13},\ \hat{\sigma}_{i32} = \hat{\sigma}_{i23}, \qquad [8d]$$

$$\hat{\theta}_{i1} = 1, \qquad [9a]$$

$$\hat{\theta}_{i2} = \begin{cases} \{\log(\bar{X}_i)-\log(\bar{V}_i)+1\}/\{\log(\bar{X}_i)-\log(\bar{V}_i)\}^2 & \text{if } \bar{X}_i > \bar{V}_i > 0 \\ 0, & \text{otherwise} \end{cases} \text{ and}$$

$$\hat{\theta}_{i3} = \begin{cases} -\bar{V}_i/[\bar{X}_i\{\log(\bar{X}_i)-\log(\bar{V}_i)\}^2], & \text{if } \bar{X}_i > \bar{V}_i > 0 \\ 0, & \text{otherwise} \end{cases} \text{ and} \qquad [9b]$$

$$[9c]$$

Trend Test

The moment estimators $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k$ are useful in the construction of a trend test, i.e., of testing

$$H_0: \{\mu_1 = \mu_2 = \ldots = \mu_k\}$$
$$\text{vs.} \qquad [10]$$
$$H_1: \{\mu_1 \le \mu_2 \le \ldots \le \mu_k\},$$

where at least one inequality in H_1 is strict. The construction is accomplished via the method of isotonic regression (*15*).

For convenience, let

$$\hat{w}_i = m_i/\hat{v}_i^2,\ i = 1,2,\ldots,k, \qquad [11]$$

denote a set of weights for the k groups ($\hat{w}_i$ represents the inverse of the asymptotic variance of $\hat{\mu}_1$). The pooled estimator of the common expected probability of response under H_0 is the weighted average (*14*)

$$\hat{\mu}^* = \left(\sum_{i=1}^{k}\hat{w}_i\hat{\mu}_i\right)/\left(\sum_{i=1}^{k}\hat{w}_i\right). \qquad [12]$$

Next, the isotonic regression estimators of $\mu_1, \mu_2, \ldots, \mu_k$ under H_1 are needed. These can be found using the pool adjacent violators algorithm (*15*). This proceeds as follows. If the estimators already satisfy the relationship $\hat{\mu}_1 \le \hat{\mu}_2 \le \ldots \le \hat{\mu}_k$, then the isotonic regression estimators are taken as $\hat{\mu}_i^* = i = 1,2, \ldots, k$. On the other hand, suppose that $\hat{\mu}_i > \hat{\mu}_{i+1}$ for some i. Then these two groups are pooled and a weighted average calculated as

$$\hat{\mu}_i^* = \hat{\mu}_{i+1}^* = (\hat{w}_i\hat{\mu}_i + \hat{w}_{i+1}\hat{\mu}_{i+1})/(\hat{w}_i + \hat{w}_{i+1}). \qquad [13]$$

The algorithm of pooling adjacent groups continues until $\hat{\mu}^*, \hat{\mu}_2^*, \ldots, \hat{\mu}_k^*$ satisfy the order imposed by H_i.

The log-likelihood ratio statistic of H_0 versus H_1

based on $\hat{\mu}^*$ in Equation 12 and the isotonic regression estimators $\hat{\mu}^*, \hat{\mu}^*_2, \ldots, \hat{\mu}^*_k$ is

$$L = \sum_{i=1}^{k} \hat{w}_i(\hat{\mu}^*_i - \hat{\mu}^*)^2. \qquad [14]$$

The test constructed from the statistic in Equation 14 is called the $\bar{\chi}^2$ test (*15*). Note that it is possible for L to be zero if each $\hat{\mu}^*_i = \hat{\mu}^*$ i = 1,2, . . . , k, which occurs when all k groups must be pooled to attain the isotonic regression estimates under H_1. Although L in Equation 14 is relatively easy to calculate, even its null probability distribution is somewhat involved if k > 4. The large-sample distribution of L under H_0 is

$$\Pr[L \geq c] = \begin{Bmatrix} \sum_{i=2}^{k} P(i,k)\Pr[\chi^2_i \geq c], & c > 0 \\ P(1,k) & , c = 0 \end{Bmatrix}, \qquad [15]$$

where χ^2_i denotes a chi-square random variable with i degrees of freedom and P(1,k), . . . , P(k,k) denote a set of probabilities which sum to 1. The derivation of the P(i,k) is beyond the scope of this article, so the interested reader is referred to Barlow et al. (*15*). In the case of k = 2, 3, or 4 (one, two, or three dose groups), closed form solutions for the P(i,k) exist. First, define the quantities

$$\hat{\rho}_{12} = -\left[\frac{\hat{w}_1\hat{w}_3}{(\hat{w}_1+\hat{w}_2)(\hat{w}_2+\hat{w}_3)}\right]^{1/2},$$

$$\hat{\rho}_{23} = -\left[\frac{\hat{w}_2\hat{w}_4}{(\hat{w}_2+\hat{w}_3)(\hat{w}_3+\hat{w}_4)}\right]^{1/2}, \qquad [16a]$$

$$\hat{\rho}_{12\cdot3} = -\left[\frac{\hat{w}_1(\hat{w}_3+\hat{w}_4)}{(\hat{w}_1+\hat{w}_2)(\hat{w}_2+\hat{w}_3+\hat{w}_4)}\right]^{1/2},$$

$$\hat{\rho}_{13\cdot2} = -\left[\frac{\hat{w}_1\hat{w}_4}{(\hat{w}_1+\hat{w}_2+\hat{w}_3)(\hat{w}_2+\hat{w}_3+\hat{w}_4)}\right]^{1/2}, \qquad [16b]$$

And

$$\hat{\rho}_{23\cdot1} = -\left[\frac{(\hat{w}_1+\hat{w}_2)\hat{w}_4}{(\hat{w}_1+\hat{w}_2+\hat{w}_3)(\hat{w}_3+\hat{w}_4)}\right]^{1/2}. \qquad [16c]$$

Then

$$P(2,2) = P(1,2) = 1/2; \qquad [17]$$

$$P(3,3) = (1/4) + (2\pi)^{-1}\arcsin(\hat{\rho}_{12}),$$

$$P(2,3) = 1/2, \qquad [18a]$$

and

$$P(1,3) = (1/4) - (2\pi)^{-1}\arcsin(\hat{\rho}_{12}); \qquad [18b]$$

$$P(4,4) = (1/8) + (4\pi)^{-1}\{\arcsin(\hat{\rho}_{12}) + \arcsin(\hat{\rho}_{23})\}, \qquad [19a]$$

$$P(3,4) = (3/8) + (4\pi)^{-1}\{\arcsin(\hat{\rho}_{12\cdot3}) + \arcsin(\hat{\rho}_{13\cdot2}) + \arcsin(\hat{\rho}_{23\cdot1})\}, \qquad [19b]$$

$$P(2,4) = (1/2) - P(4,4),$$

and

$$P(1,4) = (1/2) - P(3,4). \qquad [19c]$$

For cases with k > 4, numerical integration methods are needed to calculate the P(i,k). For a discussion of this and the noncentral distribution of the test statistic in Equation 15 under H_1, see Barlow et al. (*15*).

Results

Rai and Van Ryzin (*8*) listed the data from a dominant lethal assay (*16*). Male mice were assigned to one of three radiation groups (0, 300, or 600 rads) and then mated to females. The response of interest is the proportion of dead fetuses out of the number of implantations. The sample sizes for the three groups are $m_1 = 683$, $m_2 = 604$, $m_3 = 486$, respectively. Unfortunately, the investigators had excluded the results of those females with ≤ 4 or ≥ 11 implants (*16*). Although this is somewhat of an unusual data set in terms of the number of litters, it is used here to illustrate the calculations.

Applying Equations 5 through 9 results in the following moment estimates and estimated standard deviations: $\hat{\mu}_1 = 0.1030$, $\hat{v}_1 = 0.1284$, $\hat{\mu}_2 = 0.2566$, $\hat{v}_2 = 0.1829$, $\hat{\mu}_3 = 0.3998$, and $v_3 = 0.2119$. This and Equation 11 lead to the set of estimated weights as $\hat{w}_1 = 41449.2$, $\hat{w}_2 = 18046.0$, and $\hat{w}_3 = 10821.3$. The test of trend, i.e., H_0: $\{\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}_3\}$ versus H_1: $\{\hat{\mu}_1, \hat{\mu}_2$ and $\hat{\mu}_3\}$ with at least one strict inequality in H_1, is easy to calculate for this example because $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\mu}_3$ already satisfy the order specified in H_1, so that the isotonic regression estimates are given by $\hat{\mu}^*_i = \hat{\mu}_i$, i = 1, 2, or 3. According to Equation 12, the pooled estimate under H_0 is $\hat{\mu}^* =$

0.1881. Then the observed value of the test statistic from Equation 14 is 869.83. In order to determine the significance of this statistic, from Equations 16a and 18, it is calculated that $\rho_{12} = -0.5110$, $P(3,3) = 0.1646$, $P(2,3) = 0.5$, and $P(1,3) = 0.3354$. From the large-sample distribution of L in Equation 15, the p-value is $0.1646 \cdot \Pr[\chi^2_3 > 869.83] + 0.5 \cdot \Pr[\chi^2_3 > 869.83]$ is less than 10^{-8}. The strength of this result is due to the very large sample size ($m_1 + m_2 + m_3 = 1773$) and the clear-cut dose-response effect.

Discussion

The technique of isotonic regression, as discussed in "Methods," can be applied to any model that provides an estimate of the expected probability of response for each dose group. In fact, Van Ryzin developed a number of models to which isotonic regression could be applied (*7*). However, one model assumes independence of the random variables P_{ij} and N_{ij}, and another assumes that $\Pr[N_{ij} = O]$ is known, both of which are not likely situations.

Another model that Van Ryzin (*7*) proposed is analogous to the one discussed in "Methods," except that the geometric distribution replaces the Poisson distribution. Clark has extended this class of models by using the negative binomial distribution (*17*), which is a generalization of both the one-parameter geometric and Poisson distributions. However, Williams (*6*) has indicated that the Poisson (and correspondingly the geometric and negative binomial) distribution predicts too large a variance for the litter size data. With respect to the moment estimator $\hat{\mu}_i$ in Equation 6, $i = 1,2,\ldots,k$, the assumed litter size distribution has only a minimal effect because only a few V_{ij} are positive.

REFERENCES

1. Palmer, A. K. Statistical analysis and choice of sampling units. Teratology 10: 301–302 (1974).
2. Haseman, J. K., and Hogan, M. D. Selection of the experimental unit in teratology studies. Teratology 12: 165–172 (1975).
3. Haseman, J. K., and Kupper, L. L. Analysis of dichotomous response data from certain toxicological experiments. Biometrics 35: 281–293 (1979).
4. Williams, D. A. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics 31: 949–952 (1975).
5. Kupper, L. L., Portier, C., Hogan, M. D., and Yamamoto, E. The impact of litter effects on dose-response modeling in teratology. Biometrics 42: 85–98 (1986).
6. Williams, D. A. Dose-response models for teratological experiments. Biometrics 43: 1013–1016 (1987).
7. Van Ryzin, J. Estimating the mean of a random binomial parameter with trial size random. Sankhya, Ser. B 37: 10–27 (1975).
8. Rai, K., and Van Ryzin, J. A dose-response model for teratological experiments involving quantal responses. Biometrics 41: 1–9 (1985).
9. Gaylor, D. W. Methods and concepts of biometrics applied to teratology. In: Handbook of Teratology, Vol. 4, Research Procedures and Data Analysis (J. G. Wilson and F. C. Fraser, Eds.), Plenum Press, New York, 1978, pp. 429–444.
10. Lin, F. O., and Haseman, J. K. A modified Jonckheere test against ordered alternatives when ties are present at a single extreme value. Biometr. Zeit. 18: 623–631 (1976).
11. Gladen, B. The use of the jackknife to estimate proportions from toxicological data in the presence of litter effects. J. Am. Stat. Assoc. 74: 278–283 (1979).
12. Efron, B. The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia, 1982.
13. Shirley, E. A. C., and Hickling, R. An evaluation of some statistical methods for analysing numbers of abnormalities found amongst litters in teratology studies. Biometrics 37: 819–829 (1981).
14. Paul, S. R. Analysis of proportions of affected fetuses in teratological experiments. Biometrics 38: 361–370 (1982).
15. Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. Statistical Inference under Order Restrictions. Wiley, New York, 1972.
16. Lüning, K. G., Sheridan, W., Ytterborn, K. H., and Gullberg, U. The relationship between the number of implantations and the rate of intrauterine death in mice. Mutat. Res. 3: 444–451 (1966).
17. Clark, B. C. Developmental Toxicity Data: Trend Tests for Mean Proportional Responses from Litters of Random Size. Unpublished Ph.D. dissertation, Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, 1987.