Drug Information Journal, Vol. 31, pp. 463–469, 1997
Printed in the USA. All rights reserved.

0092-8615/97

TREND TESTS FOR DICHOTOMOUS ENDPOINTS WITH APPLICATION TO CARCINOGENICITY STUDIES

MARKUS NEUHÄUSER
Solvay Pharmaceuticals, Hannover, Germany
LUDWIG A. HOTHORN
Universität Hannover, LG Bioinformatik, Hannover, Germany

A trend test for dichotomous endpoints analogous to the nonparametric Jonckheere test is developed. The power of this and all other single trend tests for dichotomous endpoints strongly depends on the shape of the dose response curve. Combined tests which have a stable power over a wide range of the ordered alternative are suggested. One can combine several contrast tests to a so-called adjustive test which is more powerful than a Cochran-Armitage test with equally-spaced scores. The latter was recommended by Armitage (1) in case there is no a priori knowledge of the type of the trend.

Key Words: Carcinogenicity studies; Trend tests; Dichotomous endpoints; Order restricted alternative; Unknown shape

INTRODUCTION

DICHOTOMOUS ENDPOINTS frequently occur in toxicological studies, especially in carcinogenicity studies. One observes whether the animal has developed a tumor or not. In this paper only crude tumor rates are considered, with no mortality-adjustment. An extension for stratified trend tests will be presented in a further paper.

Tumors are rare events, therefore, the sample size per group ranges up to 100 rodents on authorities' recommendations. Due to mortality and sacrifice, the sample sizes are often unbalanced. Table 1 displays lung tumor data from a study on 1,2-dichloroethane (2, p. 82). One can see unbalanced sample sizes, which are common in carcinogenicity studies.

There are dichotomous endpoints in other toxicological studies as well, for example, nonneoplastic histopathological lesions. Moreover, it is possible to dichotomize a continuous endpoint according to a cut-off value of clinical relevance (3).

A dichotomous endpoint can be represented by a random variable, which has a binomial distribution. In carcinogenicity as well as in other toxicological studies the many-to-one design with a negative control group and k different dose groups is commonly used. The

Presented at the DIA Workshop "Statistical Methodology in Non-Clinical and Toxicological Studies," March 25–27, 1996, Bruges, Belgium.

Reprint address: Markus Neuhäuser, Solvay Pharmaceuticals, Hans-Böckler-Allee 20, D-30173 Hannover, Germany.

TABLE 1
Lung Tumor Data from a Study on 1,2-dichloroethane (2, p. 82)

	Dose		
	0	1	2
With tumor	2	7	15
Without tumor	35	41	21
Number of animals at risk	37	48	36

$k + 1$ independent binomial distributed random variables are denoted by $R_0, R_1, \ldots, R_k$, where R_0 belongs to the control group. Let n_i be the sample size of group i, $N = n_0 + \ldots + n_k$ the total sample size, p_i the binomial proportion, and $r = R_0 + \ldots + R_k$ the total number of tumors.

The null hypothesis is the equality of all binomial proportions, that is, the equality of all crude tumor rates. The common probability is marked with p:

$$H_0 : p_0 = p_1 = \ldots = p_k =: p.$$

The one-sided ordered alternative is considered:

$$H_A : p_0 \leq p_1 \leq \ldots \leq p_k \quad \text{with at least} \quad p_0 < p_k.$$

Under this alternative, different shapes are possible. In an extreme concave shape there are no differences between the treated groups. Only the control group has a smaller binomial proportion: $p_0 < p_1 = p_2 = \ldots = p_k$. More frequent in carcinogenicity studies are convex shapes. In an extreme convex shape only the highest dose group differs from the other groups: $p_o = p_1 = \ldots = p_{k-1} < p_k$. In most situations knowledge about the shape is lacking, that is, the shape is a priori unknown. The power of the trend tests, however, strongly depends on the shape. In the parametric case the likelihood ratio test according to Bartholomew (4) reveals a test with uniform power characteristics over the whole alternative H_A, because the likelihood ratio test statistic may be expressed as the maximum of an infinite number of contrast statistics (5, p. 189). Since one deals with exact tests, however, the asymptotic likelihood ratio test for the dichotomous case is not considered (5, p. 167; 6).

SINGLE TESTS

The United States *Federal Register* (7) recommends that the analysis of tumor incidence data is carried out with a Cochran-Armitage (1,8) trend test. The test statistic of the Cochran-Armitage test (hereafter, C-A test) is defined as this term:

$$T_{CA} = \sqrt{\frac{N}{(N-r)r}} \cdot \frac{\sum_{i=0}^{k} \left(R_i - \frac{n_i}{N} r\right) d_i}{\sqrt{\sum_{i=0}^{k} \frac{n_i}{N} d_i^2 - \left(\sum_{i=0}^{k} \frac{n_i}{N} d_i\right)^2}},$$

with dose scores d_i. Armitage's (1) test statistic is the square of this term (T_{CA}^2). As one-sided tests are carried out for an increase of tumor rates, the square is not considered. Instead, the above-mentioned test statistic which is presented by Portier and Hoel (9) is used. This test statistic is asymptotically standard normal distributed. The C-A test is asymptotically efficient for all monotone alternatives (10), but this result only holds asymptotically. And tumors are rare events, so the binomial proportions are small. In this situation approximations may become unreliable (11).

Therefore, exact tests which can be performed using two different approaches: conditional and unconditional are considered. In the first case, the total number of tumors r is regarded as fixed. As a result the null distribution of the test statistic is independent of the common probability p. The exact conditional null distribution is a multivariate hypergeometric distribution.

The unconditional model treats the sum of all tumors as a random variable. Then the exact unconditional null distribution is a multivariate binomial distribution. The distribution depends on the unknown probability. This nuisance parameter can be eliminated by considering the worst-case scenario according to Suissa and Shuster (12). Storer and Kim (13) utilize the maximum-likelihood estimate for p and obtain an approximate unconditional test which is numerically simpler than the test of Suissa and Shuster (12).

Mehta and Hilton (14) wrote: "Whether the conditional or unconditional distribution . . . should be used for inference has been the subject of intense controversy over a period of 50 years. It involves deep questions about the logic of inductive inference and the foundations of hypothesis testing. It is still unresolved because ultimately the choice is philosophical rather than statistical."

Thus, no one model is preferred, instead both are studied. In the unconditional model the suggestion by Storer and Kim (13) is used.

Because the C-A test is only asymptotically optimal, other trend tests are also considered. Contrast tests can be applied for the dichotomous case as well, the test statistic is in general:

$$T_{CT} = \sum_{i=0}^{k} a_i \frac{R_i}{n_i} \quad \text{with} \quad \sum_{i=0}^{k} a_i = 0.$$

This contrast test statistic does not take the value of the dose scores into account. Therefore, a lower power could be expected than for the regression test from the C-A type. It will be demonstrated, however, that the power of contrast tests can also be higher. Robertson et al. (5, p. 168) mentioned contrast tests for dichotomous data, but they only refer to the parametric results which asymptotically hold in the dichotomous case.

For dichotomous endpoints the same contrasts can be used which are used for parametric tests: the Helmert (15,16) and reverse-Helmert (16) contrasts, the linear contrast (16), the maximin contrast (17), the pairwise contrast (16,18), and so forth. The different contrasts are displayed for four groups ($k = 3$) in Table 2. A trend test (19) analogous to the nonparametric Jonckheere (20) test was also developed, with the test statistic as the following sum:

TABLE 2
Contrasts for $k = 3$

Contrast	a_0	a_1	a_2	a_3
Helmert	−1	−1	−1	3
Reverse-Helmert	−3	1	1	1
Maximin	−0.87	−0.13	0.13	0.87
Pairwise	−1	0	0	1
Linear	−3	−1	1	3
Analogon	−1	−0.67	−0.17	1.83
Reverse-analogon	−1.83	0.17	0.67	1

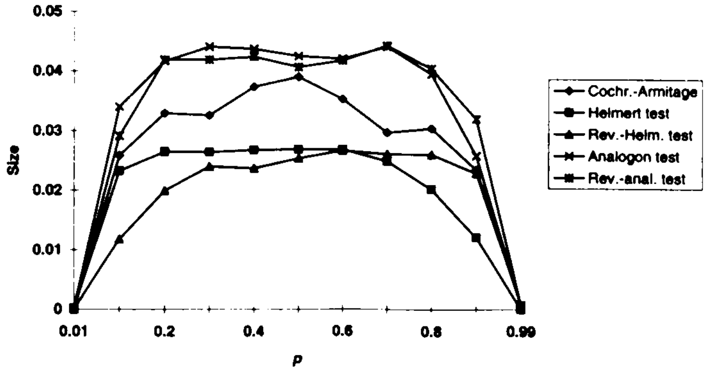

FIGURE 1. Results of the simulation study: Size of conditional tests ($k = 2$, $n_i = 20$, $\alpha = 0.05$).

$$T_{AT} = \sum_{i=1}^{k} \left(\frac{R_i + \ldots + R_k}{n_i + \ldots. + n_k} - \frac{R_{i-1}}{n_{i-1}} \right).$$

This test was called the analogon test; it is a contrast test, too. A reverse-analogon test can also be considered.

These tests are compared in a simulation study. The size and the power for a variety of shapes were estimated by simulating 10,000 replications. For the C-A test equally-spaced scores are used. They are suggested by Armitage (1) and Graubard and Korn (21) in case there is no a priori knowledge of the type of the trend.

Figure 1 shows the size of some exact conditional tests. The C-A test is not of the largest size. For small probabilities, which are common in cancerogenicity studies, all tests are very conservative.

In convex shapes the analogon test and the Helmert test are more powerful than the C-A test (Figure 2). For concave shapes the reverse tests have the highest power. For linear shapes the maximin, linear, reverse-analogon, and C-A test have a similar power. Consequently, the C-A test is not the best test. Which test is more powerful depends on the shape (Table 3).

If the total sum r is not fixed, the sample space is larger, so that there is a less discrete

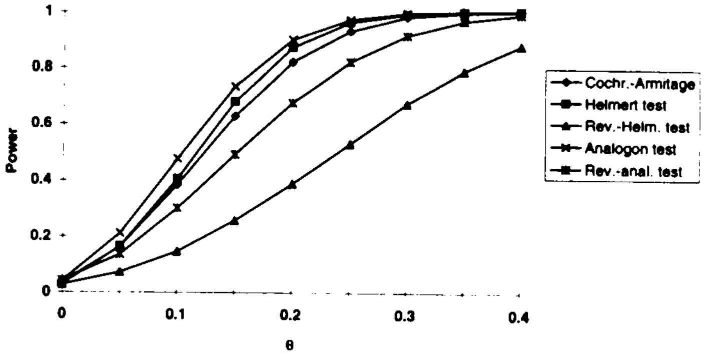

FIGURE 2. Results of the simulation study: Power of conditional tests for a convex shape ($p_0 = p_1 = 0.1$, $p_2 = 0.1 + \theta$, $k = 2$, $n_i = 50$, $\alpha = 0.05$).

TABLE 3
Results of the Simulation Study: Power of Conditional Tests for Various Shapes ($k = 2$, $n_i = 50$, $\alpha = 0.05$)

	Power				
	Single tests			Adjustive permutation tests	
Shape p_0, p_1, p_2	Analogon test	Reverse analogon test	Equally spaced C-A test	Analogon + reverse-analogon	Helmert + reverse-Helmert
Linear					
0.1, 0.15, 0.2	0.35	0.37	0.35	0.38	0.31
0.1, 0.2, 0.3	0.75	0.79	0.78	0.79	0.72
0.1, 0.25, 0.4	0.95	0.97	0.97	0.97	0.95
Convex					
0.1, 0.1, 0.2	0.48	0.30	0.39	0.44	0.38
0.1, 0.1, 0.3	0.90	0.67	0.82	0.87	0.85
0.1, 0.1, 0.4	0.99	0.91	0.98	0.99	0.99
Concave					
0.1, 0.2, 0.2	0.25	0.45	0.33	0.39	0.33
0.1, 0.3, 0.3	0.55	0.89	0.75	0.84	0.81
0.1, 0.4, 0.4	0.80	0.99	0.96	0.99	0.99
Umbrella					
0.1, 0.2, 0.15	0.07	0.26	0.12	0.20	0.18
0.1, 0.3, 0.2	0.11	0.62	0.27	0.53	0.56
0.1, 0.4, 0.3	0.34	0.95	0.72	0.92	0.94

distribution. Hence, in unconditional tests, higher values in size and power are possible. The power is practically identical for sample sizes of 40 and up (14). In the case of smaller sample sizes, the relative power comparisons of the conditional tests hold in the unconditional model. Detailed simulation results are referred to by Neuhäuser (22).

COMBINED TESTS

The shape is usually a priori unknown. But the power of the different single tests depends on the shape. Which test should be applied?

One can use the C-A (single) test according to guidance. The authors' recommendation is a test principle, which is called an *adjustive test* (22). Two tests are carried out and the null hypothesis is rejected if at least one of the two P-values is smaller than $\alpha/2$. One test should be powerful for convex shapes, the other for concave shapes. An adjustive test is also possible with more than two tests.

A "Bonferronization," however, is not the optimal approach. Instead, the maximum of some test statistics is used as a new test statistic. It is similar to the multiple contrast method (5, p. 188ff), but in this general situation the distribution of this new statistic is unknown. Therefore, a permutation test is carried out. Analogous to the maximin efficiency robust test (23) one can put two extreme contrasts (optimal for convex and concave shapes) together to an adjustive test, for example, the analogon and the reverse-analogon test or the Helmert and the reverse-Helmert test.

The adjustive permutation test is almost as powerful as the best test and more powerful than a C-A test with equally-spaced scores. Figure 3 shows this result for a convex shape. Table 3 contains representative results for various shapes. In reality the shape is unknown,

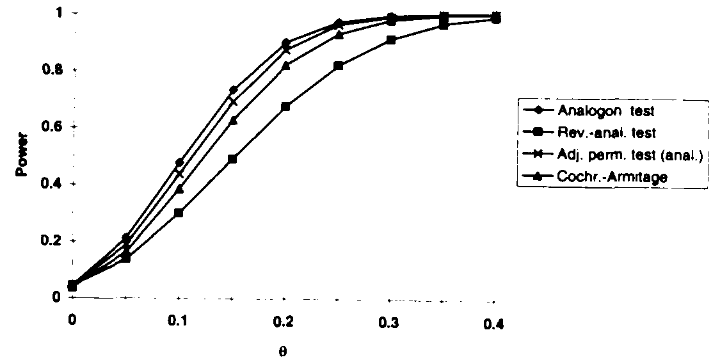

FIGURE 3. Results of the simulation study: Power of conditional tests for a convex shape ($p_0 = p_1 = 0.1$, $p_2 = 0.1 + \theta$, $k = 2$, $n_i = 50$, $\alpha = 0.05$).

so one does not know which test is best. The adjustive test, however, is always relatively good. The adjustive permutation test with analogon contrasts is more powerful than the one with Helmert contrasts (Table 3).

The evaluation of the example data is shown in Table 4. The *P*-values of the adjustive tests are close to each of the smallest *P*-value in their groups. One can compute a single exact C-A test with StatXact or SAS. Conditional contrast tests and adjustive permutation tests can be carried out with an algorithm according to Williams (11). An easy modification makes the unconditional test possible, but these tests need a long computation time. An alternative is a simulation-based permutation test according to Berry (24). An algorithm for the adjustive versions will be published.

CONCLUSIONS

A single C-A test is not the optimal test. The asymptotic version should be avoided for the sample sizes and binomial proportions which are common in carcinogenicity studies. In these studies the shape is a priori unknown, but for rare tumors convex shapes are more likely than other shapes. Considering different studies, both sexes and all tumor sites, however, the shape has to be considered as a priori unknown. Armitage (1) wrote: "In the absence of any a priori knowledge of the type of the trend to be expected, it seems reasonable to choose . . . equally-spaced [scores]." This recommendation was confirmed 22 years later by

TABLE 4
***P*-Values of the Lung Tumor Data**

	P-Value	
Test	Conditional	Unconditional
Analogon test	0.000050	0.000053
Reverse-analogon test	0.000122	0.000177
Adjustive permutation test	0.000052	0.000058
Helmert test	0.000198	0.000124
Reverse-Helmert test	0.005338	0.003997
Adjustive permutation test	0.000926	0.000815
C-A test with equally-spaced scores	0.000066	0.000037

Graubard and Korn (21). The authors' recommendation, an adjustive permutation test, is a more powerful strategy. There are further applications of the adjustive approach, for example, with nonparametric trend tests (22). These further applications are not presented here.

REFERENCES

1. Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics.* 1955;11:375–386.
2. Gart JJ, Krewski D, Lee PN, Tarone RE, Wahrendorf J. *Statistical methods in cancer research, Vol. III. The design and analysis of long term animal experiments.* Lyon: International Agency for Research on Cancer; 1986.
3. Suissa S, Blais L. Binary regression with continuous outcomes. *Stat Med.* 1995;14:247–255.
4. Bartholomew DJ. Ordered tests in the analysis of variance. *Biometrika.* 1961;48:325–332.
5. Robertson T, Wright FT, Dykstra RL. *Order restricted statistical inference.* New York: Wiley; 1988.
6. Oluyede BO. Tests for equality of several binomial populations against an order restricted alternative and model selection for one-dimensional multinomials. *Biom J.* 1994;36:17–32.
7. *Federal Register.* No. 50, Vol. 50. Washington; 1985.
8. Cochran WG. Some methods for strengthening the common χ^2 tests. *Biometrics.* 1954;10:417–451.
9. Portier C, Hoel D. Type 1 error of trend tests in proportions and the design of cancer screens. *Comm Stat-Theory Meth.* 1984;A13:1–14.
10. Tarone RE, Gart JJ. On the robustness of combined tests for trends in proportions. *J Am Stat Assoc.* 1980;75: 110–116.
11. Williams DA. Tests for differences between several small proportions. *App Stat.* 1988;37:421–434.
12. Suissa S, Shuster JJ. Exact unconditional sample sizes for the 2 × 2 binomial trial. *J R Statist Soc.* 1985;A148: 317–327.
13. Storer BE, Kim C. Exact properties of some exact test statistics for comparing two binomial proportions. *J Am Stat Assoc.* 1990;85:146–155.
14. Mehta CR, Hilton JF. Exact power of conditional and unconditional tests: Going beyond the 2 × 2 contingency table. *Am Stat.* 1993;47:91–98.
15. Ruberg SJ. Contrasts for identifying the minimum effective dose. *J Am Stat Assoc.* 1989;84:816–822.
16. Tamhane AC, Hochberg Y, Dunnett CW. Multiple test procedures for dose finding. *Biometrics.* 1996;52:21–37.
17. Abelson RP, Tukey JW. Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Ann Math Statist.* 1963;34:1347–1369.
18. Hothorn LA, Lehmacher W. A simple testing procedure 'control versus *k* treatments' for one-sided ordered alternatives, with application in toxicology. *Biom J.* 1991;33:179–189.
19. Neuhäuser M, Hothorn LA. Trendtests für dichotome Endpunkte. Presented on the 41st Biometric Conference of the German Region of the International Biometric Society; 1995.
20. Jonckheere AR. A distribution-free *k*-sample test against ordered alternatives. *Biometrika.* 1954;41:133–145.
21. Graubard BI, Korn EL. Choice of column scores for testing independence in ordered 2 × *K* contingency tables. *Biometrics.* 1987;43:471–476.
22. Neuhäuser M. Trendtests bei a priori unbekanntem Erwartungswertprofil. (Trend tests for a priori unknown shapes.) Doctoral Thesis, Dortmund; 1996.
23. Gastwirth JL. The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *J Am Stat Assoc.* 1985;80:380–384.
24. Berry JJ. A simulation-based approach to some nonparametric statistics problems. *Observations.* 1995;5:19–26.