BIOMETRICS 40, 819–825
September 1984

Exact Significance Testing to Establish Treatment Equivalence with Ordered Categorical Data

Cyrus R. Mehta

Harvard School of Public Health, Dana-Farber Cancer Institute,
44 Binney Street, Boston, Massachusetts 02115, U.S.A.

Nitin R. Patel

Indian Institute of Management, Vastrapur,
Ahmedabad 380-015, India

and

Anastasios A. Tsiatis

Harvard School of Public Health, Dana-Farber Cancer Institute,
44 Binney Street, Boston, Massachusetts 02115, U.S.A.

SUMMARY

This communication concerns the problem of establishing the therapeutic equivalence of two treatments that are being compared on the basis of ordered categorical data. The problem is formulated as a significance test in which the null hypothesis specifies a treatment difference. An efficient numerical algorithm for computing the exact significance level is provided, along with a simple method for obtaining the asymptotic significance level. Both methods are applied to a clinical trial of a new agent versus an active control. Guidelines for when to use the exact procedure and when to rely on asymptotic theory are provided.

1. Introduction

In many clinical trials the purpose is to establish that a newly developed treatment is therapeutically equivalent to the current standard treatment for a given disease. For example, permission to market a new drug is often granted by the FDA, provided it can be demonstrated, in a clinical trial, that the drug is as effective as a control drug which is known from previous studies to be active. A serious misuse of significance tests can arise in such applications if the data are unable to reject the null hypothesis of no treatment difference. The investigator may then claim that the two treatments are equivalent. It is, however, entirely possible that a real difference does exist but, because of inadequate sample sizes, the significance test lacks the power to reject the null hypothesis.

Dunnett and Gent (1977) have investigated this problem for the special case in which two treatments with binomial outcomes are compared. They proposed that therapeutic equivalence be established by testing the hypothesis $H: \delta = \Delta$, where δ denotes the difference in response rates with the two treatments. Here Δ is not zero, as in the customary null hypothesis, but is determined from the practical aspects of the problem in such a way that the treatments can be considered equivalent for all practical purposes if their true difference does not exceed the specified Δ. In this communication the work of Dunnett and Gent is extended to ordered categorical data. An efficient numerical algorithm for computing the exact P-value of the Wilcoxon midranks test under any specified treatment difference is developed and applied to the analysis of a clinical trial that compares two drugs for the

Key words: Treatment equivalence; Exact Wilcoxon test with ties; Clinical trial.

treatment of active rheumatoid arthritis. The algorithm can also provide the exact confidence interval for any parameter which characterizes the treatment difference. Some guidelines for when to compute exact probabilities and when to rely on asymptotic results are provided.

2. Exact Permutation Distribution of the Wilcoxon Midranks Statistic

Suppose that the outcome of a clinical trial which compares m patients who receive a new therapy with n patients who receive a standard therapy is summarized by the $2 \times k$ contingency table:

Therapy	Ordered categorical outcome				Sample size
	1	2	...	k	
New	x_1	x_2	...	x_k	m
Standard	y_1	y_2	...	y_k	n
Ties	t_1	t_2	...	t_k	$m+n$

Let π_j denote the probability that a patient who receives the new therapy falls in Category j, and let π_j' denote the corresponding probability for a patient who receives the standard therapy. An appropriate test of the null hypothesis $H_0{:}\pi_j = \pi_j'$, $j = 1, 2, \ldots, k$, is the Wilcoxon midranks test (Lehmann, 1975; Klotz and Teng, 1977). This test is based on the permutation distribution of the statistic

$$W = \sum_{j=1}^{k} r_j X_j,$$

where $r_1 = \frac{1}{2}(t_1 + 1)$, and $r_j = t_1 + \cdots + t_{j-1} + \frac{1}{2}(t_j + 1)$, $j = 2, \ldots, k$, are the midranks of the k ordered categories.

The distribution of W is derived by keeping t_j fixed for each j, and permuting the allocation of the ties to the new and standard therapies in all possible ways. This conditional distribution of W depends on the π_j and π_j' values only through $k-1$ odds ratio parameters $\phi_j = (\pi_j' \pi_{j+1})/(\pi_j \pi_{j+1}')$, $j = 1, 2, \ldots, k-1$. Let

$$G_\Phi(w) = \mathrm{pr}(W \geqslant w \mid \phi_j, t_j, j = 1, 2, \ldots, k),$$

where $\phi_k \equiv 1$ for notational convenience. One can then show that

$$G_\Phi(w) = \sum_{\mathbf{x} \in S_w} \left\{ \prod_{j=1}^{k} \binom{t_j}{x_j} \phi_j^{-s_j} \right\} \Big/ \sum_{\mathbf{x} \in S} \left\{ \prod_{j=1}^{k} \binom{t_j}{x_j} \phi_j^{-s_j} \right\}, \tag{2.1}$$

where

$$s_j = \sum_{1}^{j} x_i,$$

$S = \{\mathbf{x}{:}\mathbf{x} \text{ is } 2 \times k \text{ with row sums } (m, n) \text{ and column sums } (t_1, t_2, \ldots, t_k)\}$ and

$$S_w = \left\{ \mathbf{x}{:}\mathbf{x} \in S \text{ and } \sum_{j=1}^{k} r_j x_j \geqslant w \right\}.$$

Under H_0, $\phi_j = 1$ for all j, and (2.1) reduces to

$$G_0(w) = \sum_{\mathbf{x} \in S_w} \prod_{j=1}^{k} \binom{t_j}{x_j} \Big/ \binom{m+n}{m}. \tag{2.2}$$

Observe that (2.2) contains no unknown parameters and can, in principle, furnish an exact P-value under H_0. If the purpose of the clinical trial is to establish that the two treatments are equivalent, then, following Dunnett and Gent, one must hypothesize a clinically important treatment difference $H_1: \phi_j = \phi_j^*$, $j = 1, 2, \ldots, k-1$. The P-value associated with this hypothesis can be computed from (2.1). A small P-value would imply treatment equivalence. In practice, since (2.1) and (2.2) are very difficult to compute, most investigators would rely on asymptotic results. In §3, we present an efficient numerical algorithm for the exact computation of (2.1). The computation of (2.2) then follows as a special case.

3. Network Algorithm

The usual way to compute (2.2) would be to enumerate explicitly each table in S and record its binomial product and midrank sum, i.e.

$$\prod_{j=1}^{k} \binom{t_j}{x_j} \phi^{-s_j} \quad \text{and} \quad \sum_{j=1}^{k} r_j x_j.$$

The denominator of (2.1) would then be the sum of the binomial products of all the tables in S. The numerator would be the sum of the binomial products of the tables with midrank sums equalling or exceeding a fixed quantity, w, i.e. all tables in S_w. The efficiency of our algorithm is based on an innovation which circumvents the need to enumerate explicitly all the tables in S. By converting the original problem into one of identifying paths through a network, our algorithm can identify simultaneously, large sets of tables that belong to S_w and eliminate from the reference set large sets of tables that do not belong to S_w.

It is convenient to represent the reference set as a network of nodes and arcs. The network is constructed in $k+1$ stages labelled $0, 1, \ldots, k$. At any Stage j, there exists a set of nodes each labelled by a unique ordered pair of integers (j, l_j). Arcs emanate from each node at Stage j and every arc is directed to exactly one node at Stage $j+1$. The network is constructed recursively by specifying all nodes of the form $(j+1, l_{j+1})$, which succeed the node (j, l_j) and are connected to it by arcs. The range of l_{j+1} for these successor nodes is given by

$$\max\left(l_j,\ l_j + \sum_{i=1}^{j+1} t_i - n\right) \leqslant l_{j+1} \leqslant \min(l_j + t_{j+1},\ l_j + m), \quad j = 0, 1, \ldots, k-1. \tag{3.1}$$

Starting with the unique initial node $(0, 0)$ at Stage 0, we apply (3.1) successively to Stages $1, 2, \ldots, k$, and we end with a unique terminal node (k, m). A path through the network is defined as a succession of connected arcs directed from the initial node to the terminal node. Each path is of the form $(0, 0) \to (1, l_1) \to \cdots \to (k, m)$ and corresponds to a table in S with $x_j = l_j - l_{j-1}$, $j = 1, 2, \ldots, k$. For example, suppose S is the set of all 2×3 contingency tables with $m = n = 3$ and $t_1 = t_2 = t_3 = 2$. The network representation of S is shown in Fig. 1. Each path in this network corresponds to one and only one table in S. For instance, the dotted path corresponds to the table

$$\begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix}.$$

Each arc $(j, l_j) \to (j+1, l_{j+1})$ has associated with it a *rank length* and a *binomial length*, given by

$$r_{j+1}(l_{j+1} - l_j) \quad \text{and} \quad \binom{t_{j+1}}{l_{j+1} - l_j} \phi^{-L_j},$$

respectively, where $L_j = \sum_{i=0}^{j} (l_{i+1} - l_i)$. The rank length of an entire path through the

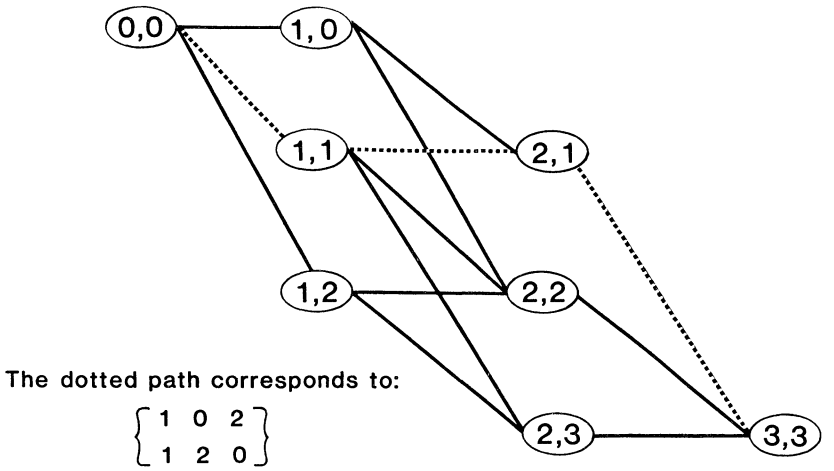

Figure 1. Network representation of $S = (x{:}m = n = 3;\ t_1 = t_2 = t_3 = 2)$.

network is the sum of rank lengths of the individual arcs that constitute the path. The binomial length of a path is the product of the binomial lengths of the individual arcs constituting the path. For instance, the rank length of the dotted path in Fig. 1 is 12.5 and its binomial length (under $\phi_j = 1$ for all j) is 2.

For each node, we now compute

$SP(j, l_j)$, the shortest rank length among all the paths from Node (j, l_j) to Node (k, m),

$LP(j, l_j)$, the longest rank length among all the paths from Node (j, l_j) to Node (k, m),

$TP(j, l_j)$, the sum of the binomial lengths of all the paths from Node (j, l_j) to Node (k, m).

These computations can be made rapidly by means of a single backwards induction pass through the network. See, for example, Wagner (1969, p. 256). Notice that $TP(0, 0)$ is precisely the denominator of (2.1) so that this otherwise formidable computation has been done with very little effort.

The network also provides a very efficient way to compute the numerator of (2.1). Suppose that in the process of enumerating paths through the network we have reached the node (j, l_j^*) via the subpath $(0, 0) \to (1, l_1^*) \to \cdots \to (j-1, l_{j-1}^*)$. Denote the rank length of this subpath by

$$b = \sum_{i=1}^{j} r_i(l_i^* - l_{i-1}^*).$$

Let Q denote the set of all complete paths from $(0, 0)$ to (k, m) that share the above subpath. Under certain conditions, we can know immediately the contribution of Q to the numerator of (2.1). If

$$b\{LP(j, l_j^*)\} < w, \tag{3.2}$$

we know that every path in Q has a rank length of less than w. Hence, none of these paths contributes to the numerator of (2.1) and the entire set can be dropped from further consideration. Next, if

$$b\{SP(j, l_j^*)\} \geqslant w, \tag{3.3}$$

every path in Q has a rank length equal to or greater than w and the entire set of paths contributes to the numerator of (2.1). The contribution of these paths is, of course, $TP(j, l_j^*)$. If neither (3.2) nor (3.3) hold, we extend the common subpath to a node $(j + 1, l_{j+1}^*)$ in accordance with (3.1) and proceed to verify (3.2) and (3.3) in the same manner as before.

4. Establishing the Therapeutic Equivalence of a New Agent and an Active Control

A six-month double-blind randomized multicenter study was used to compare a new agent with an active control in patients with acute rheumatoid arthritis. At the end of the trial each patient was evaluated on the basis of a five-point global assessment scale; the results are given in Table 1.

These data are clearly unable to reject H_0: the exact two-sided Wilcoxon midrank test gave $P = .24$. It is important to see what evidence the data provide for the therapeutic equivalence of the two drugs. To this end, we conveniently specify a clinically important therapeutic advantage for the active control in terms of a constant odds ratio $\phi^* > 1$. Specifically, we will test the null hypothesis H_1: $\phi_j = \phi^*$, $j = 1, 2, \ldots, k - 1$, against the one-sided alternative hypothesis H_2: $\phi_j < \phi^*$, $j = 1, 2, \ldots, k - 1$. Letting w_0 denote the observed value of the W statistic, we define the exact one-sided P-value to be $\mathrm{pr}(W \leqslant w_0)$. A one-sided test in the left tail is quite appropriate here since it will prevent us declaring erroneously that the two treatments are equivalent when, in fact, the new agent is worse than the active control. Table 2 displays exact and asymptotic one-sided P-values for various choices of ϕ^*.

The exact P-value was computed by the network algorithm. Observe that direct application of the algorithm to (2.1) yields $\mathrm{pr}(W \geqslant w_0)$, whereas we want $\mathrm{pr}(W \leqslant w_0)$. To overcome this difficulty we used the algorithm to compute $G_\eta(v_0)$, as defined by (2.1) with $\eta = 1/\phi$ and $v_0 = \sum \mathrm{r}_j y_j$. One can easily show that $G_\eta(v_0) = \mathrm{pr}(W \leqslant w_0)$.

To compute the asymptotic P-value, observe that the distribution of the normalized Wilcoxon statistic for $2 \times k$ contingency tables,

$$T = \frac{W - m(m + n + 1)/2}{[\{mn(m + n + 1)/12\} - \{(mn) \sum_{i=1}^{k} (t_i^3 - t_i)\}/\{12(m + n)(m + n - 1)\}]^{1/2}},$$

will be approximately a standard normal under H_0, provided m and n are sufficiently large and $\max_{1<i<k}\{t_i/(m + n)\}$ is bounded away from unity (see Lehmann, 1975, pp. 20–33). As is generally the case for normal statistics derived from independent multinomial distributions, under local alternatives (i.e. ϕ^* values close to 1), the distribution of T can be approximated as a normal with mean θ and variance 1, where θ is equal to the value of the normalized Wilcoxon statistic applied to data corresponding to the expected counts of a $2 \times k$ table under H_1. In our particular case, the expected counts preserve the marginal totals of the observed counts and are compatible with the $k - 1$ odds ratios of H_1.

Table 2 provides us with the information needed to decide whether or not the two drugs are therapeutically equivalent. For example, suppose that an odds ratio of 1.1 in favor of the active control is considered to be just clinically significant. This means that it is

Table 1
Clinical trial of a new agent and an active control

Drug	Global assessment				
	Much improved	Improved	No change	Worse	Much worse
New agent	24	37	21	19	6
Active control	11	51	22	21	7

Table 2
Exact and asymptotic P-values for various choices of ϕ^*

	ϕ^*				
	1.0	1.05	1.1	1.15	1.2
Exact *P*-value:	.119	.056	.024	.0096	.0035
Asymptotic *P*-value:	.119	.057	.025	.0098	.0037

acceptable for patients treated with the new agent to have up to 10% excess risk of falling into Category $j + 1$ rather than into Category j, as compared to patients treated with the active control. Since the *P*-value associated with H_1: $\phi^* = 1.1$ is only .024, we reject this hypothesis and conclude that the two treatments are therapeutically equivalent.

5. Discussion

We have provided an algorithm for computing the exact *P*-value of the Wilcoxon midrank test under arbitrarily specified treatment differences; this makes it possible to establish the equivalence of two treatments. The procedure avoids explicit enumeration of all tables in the conditional reference set, S. This is the main reason why our algorithm remains computationally feasible even when the sample size for the clinical trial is fairly large. On the DEC 2060 computer system, it took us only 21 cpu seconds to execute each exact computation in Table 2. On the other hand, the same computation took 396 cpu seconds when the conventional approach of enumerating explicitly all the tables in S was followed. For larger dimensional tables, the savings would be even greater as some of our other network algorithms show (Mehta and Patel, 1983).

Our parameterization of treatment differences in §4 is, of course, arbitrary. However, this parameterization has the advantage that it is independent of nuisance parameters such as π_j and π_j', $j = 1, 2, \ldots, k$. It is easy to show moreover that $\phi > 1$ implies that the π and π' values are stochastically ordered, i.e. $\pi_1 + \pi_2 + \cdots + \pi_j \geqslant \pi_1' + \pi_2' + \cdots + \pi_j'$ for all j. The algorithm remains valid with any alternative parameterization. Similarly, the algorithm can be applied, with very little modification, to two-sided significance tests rather than the one-sided tests considered here.

Although the exact and asymptotic *P*-values reported in Table 2 are almost identical, the asymptotic results are less accurate for small clinical trials with large treatment differences. The results of a hypothetical clinical trial involving only 10 patients per arm are shown in Table 3. The corresponding exact and asymptotic *P*-values are given in Table 4. The accuracy of the asymptotic *P*-value diminishes as ϕ^* increases.

In practice, clinical trials involving only a few patients would not be justified unless large treatment differences were expected or clinically important. Table 4 shows that this is precisely the situation in which exact hypothesis tests are preferred to asymptotic ones. For the hypothetical example above, suppose that a clinically important treatment difference is defined by $\phi^* \geqslant 2$. Then exact computation yields a *P*-value $< .05$, which implies that the

Table 3
Results of a small hypothetical clinical trial

Drug	Ordered category				
	Much improved	Improved	No change	Worse	Much worse
Treatment	2	2	2	2	2
Control	1	3	3	3	0

Table 4
*Exact and asymptotic P-values for the hypothetical clinical trial, for selected values of ϕ**

	ϕ^*					
	1	1.2	1.4	1.6	1.8	2.0
Exact *P*-value:	.572	.362	.279	.120	.066	.037
Asymptotic *P*-value:	.561	.363	.224	.138	.089	.058

treatment and control arms are equivalent. The asymptotic computation does not furnish evidence of treatment equivalence at the .05 level.

An alternative but equally valid way to obtain evidence of treatment equivalence is to determine a confidence interval for ϕ. This too is easily achieved by our algorithm since $G_\phi(w)$ can be readily computed for several values of ϕ at the fixed observed value of w. Note that since S and S_w remain the same at every value of ϕ, only one iteration of the network algorithm is needed to evaluate $G_\phi(w)$ for several different values of ϕ. A 95% confidence interval can then be computed by determining ϕ_1 and ϕ_2 so that $G_{\phi_1}(w) = .975$ and $G_{\phi_2}(w) = .025$.

Acknowledgements

This investigation was supported by Grants, Numbers CA-23415 and CA-33019, awarded by the National Cancer Institute, DHHS, and by a Mellon Foundation Career Development Grant.

Résumé

Cet article discute le problème suivant: comment établir l'équivalence thérapeutique de deux traitements comparés à l'aide de données qualitatives ordonnées. Ce problème est formulé comme un test de signification dans lequel l'hypothèse nulle spécifie une différence entre les traitements. Un algorithme numérique efficace pour calculer le niveau de signification exact est fourni ainsi qu'une méthode simple pour obtenir un niveau de signification asymptotique. Les deux méthodes sont appliquées à un essai clinique. Des indications sont données quant au choix entre la procédure exacte et la théorie asymptotique.

References

Dunnett, C. W. and Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 × 2 tables. *Biometrics* **33,** 593–602.

Klotz, J. and Teng, J. (1977). One-way layout for counts and the exact enumeration of the Kruskal–Wallis *H* distribution with ties. *Journal of the American Statistical Association* **72,** 165–169.

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

Mehta, C. R. and Patel, N. R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association* **78,** 427–434.

Wagner, H. M. (1969). *Principles of Operations Research.* New Jersey: Prentice Hall.

Received September 1982; *revised August* 1983