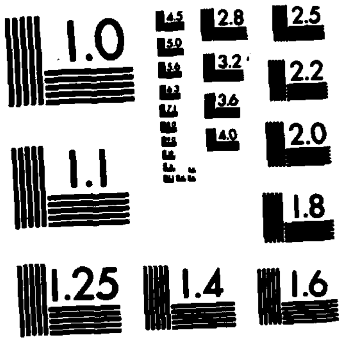
AD-A125 241 CONDITIONAL PERMUTATION TESTS AND THE PROPENSITY SCORE 1/1
IN OBSERVATIONAL STUDIES(U) WISCONSIN UNIV-MADISON
MATHEMATICS RESEARCH CENTER P R ROSENBAUM DEC 82
UNCLASSIFIED MRC-TSR-2463 DAAG29-80-C-0041 F/G 12/1 NL

END
FILMED
DTIC


1.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

MRC Technical Summary Report #2463

CONDITIONAL PERMUTATION TESTS AND THE PROPENSITY SCORE IN OBSERVATIONAL STUDIES

Paul R. Rosenbaum

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

December 1982

(Received September 2, 1982)

Approved for public releas
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Cancer Institute
9000 Rockville Pike
Bethesda, MD 20205

Wisconsin Alumni
Research Foundation
University of Wisconsin
Madison, WI 53705

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

CONDITIONAL PERMUTATION TESTS AND THE PROPENSITY SCORE IN OBSERVATIONAL STUDIES

Paul R. Rosenbaum*

Technical Summary Report #2463

December 1982

ABSTRACT

In observational studies, the distribution of treatment assignments is unknown, and therefore randomization tests are not generally applicable. However, permutation tests that condition on sample information about the treatment assignment mechanism can be applicable in observational studies, providing treatment assignment is strongly ignorable. These tests use the conditional distribution of the treatment assignments given a sufficient statistic for the unknown parameter of the propensity score. Several tests that are commonly used in observational studies are particular instances of this general procedure; moreover, conditional permutation tests and covariance adjustment are closely related. A backtrack algorithm is developed to permit efficient calculation of the exact conditional significance level, and two approximations are discussed. A clinical study of treatments for lung cancer is used to illustrate the technique. Conditional permutation tests extend previous large sample results on the propensity score by providing a general basis for exact inference in small observational studies when treatment assignment is strongly ignorable.

AMS (MOS) Subject Classifications: 62A20, 62A15, 62B99, 62F05, 62F04, 62G10, 62-04.

Key Words: Observational studies; randomization tests; ignorable treatment assignment; conditional inference; logistic models.

Work Unit Number 4 - Statistics and Probability

*Departments of Statistics and Human Oncology, University of Wisconsin - Madison.

The author acknowledges Donald B. Rubin for valuable conversations on the subject of this paper, and Paul P. Carbone for permission to use the data in §3.

Sponsored in part by the United States Army under Contract No. DAAG29-80-C-0041, in part by grant P30-CA-14520 from the National Cancer Institute to the Wisconsin Clinical Cancer Center, and in part by a grant from the Wisconsin Alumni Research Foundation.

SIGNIFICANCE AND EXPLANATION

An observational study is an attempt to draw inferences about the effects of treatments from nonexperimental data. In an experiment, treatments are assigned by the investigator, so it is possible to assure that the units which receive each treatment are comparable. In observational studies the units receiving the two treatments may differ markedly, since the treatment assignments were not under the control of the investigator. The current paper develops two extensions of a standard method in experiments -- Fisher's randomization test -- that are applicable in observational studies under explicit assumptions. An algorithm is developed for computing the required conditional permutation distribution.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

CONDITIONAL PERMUTATION TESTS AND THE PROPENSITY SCORE IN OBSERVATIONAL STUDIES

Paul R. Rosenbaum*

1. INTRODUCTION: Definitions; Fisher's Randomization Test

1.1. The Propensity Score in Observational Studies

The propensity score is the conditional probability of exposure to a particular treatment given a vector of observed covariates. Properties of the propensity score, its role in observational studies, and its relationship to various methods of bias reduction are described by Rosenbaum and Rubin (1983a). The methods they propose are applicable in large observational studies in which an estimate of the propensity score may substitute for the population propensity score. The current paper shows that, under conditions defined in §2, the propensity score can also provide a basis for exact inference in small observational studies if a sufficient statistic exists for the unknown parameter of the propensity score. Relevant notation and definitions from Rosenbaum and Rubin (1983a) are briefly reviewed in §1.2 and §1.3, and related to Fisher's randomization test in §1.4.

1.2. The Structure of Studies for Treatment Effects

In the case of two treatments numbered 1 and 0, the i^{th} of the N units under study has, in principle, both a response r_{1i} that would have resulted if it had received treatment 1, and a response r_{0i} that would have resulted if it had received treatment 0. Treatment effects are defined to be comparisons of r_{1i} and r_{0i}, such as $r_{1i} - r_{0i}$. Each unit receives only one treatment, so either r_{1i} or r_{0i} is observed, but not both. Therefore, inferences about the effects of treatments on single units, as distinct from collections or populations of units, are largely speculative: inferences about treatment effects are inherently statistical inferences. This structure is

*Departments of Statistics and Human Oncology, University of Wisconsin-Madison.
The author acknowledges Donald B. Rubin for valuable conversations on the subject of this paper, and Paul P. Carbone for permission to use the data in §3.

Sponsored in part by the United States Army under Contract No. DAAG29-80-C-0041, in part by grant P30-CA-14520 from the National Cancer Institute to the Wisconsin Clinical Cancer Center, and in part by a grant from the Wisconsin Alumni Research Foundation.

consistent with that traditionally used in the literature of experimental design, for example, in the books by Fisher (1935), Kempthorne (1952), and Cox (1958), and follows the development for observational studies in Rubin (1974, 1977, 1978), Hamilton (1979), Rosenbaum and Rubin (1983a,b) and Rosenbaum (1982). For further discussion of this structure and some of its limitations, see Cox (1958, chapter 2), Rubin (1978, §2.3; 1980) and Rosenbaum and Rubin (1983a, §1.1).

For the i^{th} unit of N units in the study $(i=1,\ldots,N)$, let z_i be the indicator for treatment assignment, with $z_i = 1$ if unit i is assigned treatment 1, and $z_i = 0$ if unit i is assigned to the treatment 0. Let $\underset{\sim}{x}_i$ be a vector of observed pretreatment measurements or covariates for the i^{th} unit; all of the measurements in $\underset{\sim}{x}$ were made prior to treatment assignment, but $\underset{\sim}{x}$ may not include all covariates used to make treatment assignments. The propensity score, $e(\underset{\sim}{x})$, is the conditional probability of assignment to treatment 1 given the observed covariates, that is,

$$e(\underset{\sim}{x}) = \text{pr}(z = 1 \mid \underset{\sim}{x})$$

where it is assumed that

$$\text{pr}(z_1,\ldots,z_N \mid \underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_N) = \prod_{i=1}^{N} e(\underset{\sim}{x}_i)^{z_i}\{1 - e(\underset{\sim}{x}_i)\}^{1-z_i} . \qquad (1.1)$$

Although this strict independence assumption is not essential, it simplifies notation and discussion.

1.3 A Critical Assumption: Strongly Ignorable Treatment Assignment

Randomized and nonrandomized trials differ in two ways. First, in a randomized trial, the propensity score is a known function, whereas, in an observational study the propensity score function is almost always unknown. Second, with properly collected data in a randomized trial, $\underset{\sim}{x}$ is known to contain all covariates that are both used to assign treatments and possibly related to the response (r_1, r_0). More formally, in a randomized trial, treatment assignment z and response (r_1, r_0), are known to be conditionally independent given $\underset{\sim}{x}$, or in Dawid's (1979) notation,

$$(r_1, r_0) \perp\!\!\!\perp z \mid \underset{\sim}{x} . \qquad (1.2)$$

This condition is usually not known to hold in a nonrandomized experiment. Moreover, in a randomized experiment, every unit in the experiment has a chance of receiving each treatment. Following Rosenbaum and Rubin (1983a,b), treatment assignment will be said to be <u>strongly ignorable</u> if (1.2) holds and $0 < \mathrm{pr}(z=1|\underset{\sim}{x}) < 1$ for all $\underset{\sim}{x}$. (As the term suggests, strong ignorability is a somewhat more restrictive condition than ignorability as defined by Rubin (1978).)

The assumption of strongly ignorable treatment assignment plays a critical role in inference from observational studies (e.g., Rosenbaum and Rubin 1983a, theorem 4, and §2 below), and therefore, the correctness and implications of this assumption will generally require investigation in each observational study. In the case of binary responses (r_1, r_0), Rosenbaum and Rubin (1983b) describe a method for assessing the sensitivity of conclusions to certain departures from strong ignorability. Rosenbaum (1982) reviews methods of testing the assumption of strong ignorability. Related discussion within a Bayesian framework is given by Rubin (1978, §4).

1.4 Fisher's (1935) Randomization Test in Randomized Experiments

Fisher's (1935) randomization test examines the sharp null hypothesis of zero difference in the effects of the treatments for each experimental unit, that is,

$$H_0: r_{1i} = r_{0i} \quad \text{for} \quad i = 1,2,\ldots,N \; . \tag{1.3}$$

Note that the sharp null hypothesis (1.3) states that the same response would have been observed from each unit had it received the alternative treatment.

Let r_{zi} be the observed response for unit i, that is, $r_{zi} = z_i r_{1i} + (1-z_i)\, r_{0i}$, and denote the vector of observed responses by $\underset{\sim}{r} = (r_{z1}, r_{z2}, \ldots, r_{zN})^T$, and the vector of treatment assignments by $\underset{\sim}{z} = (z_1, z_2, \ldots, z_N)^T$. Let $t(\underset{\sim}{z}, \underset{\sim}{r})$ be a statistic chosen to measure departures from the sharp null hypothesis (1.3); for example, $t(\underset{\sim}{z}, \underset{\sim}{r})$ might be the difference in sample mean responses to the two treatments, that is

$$t(\underset{\sim}{z}, \underset{\sim}{r}) = \{(\underset{\sim}{z}^T \underset{\sim}{r})/\underset{\sim}{z}^T \underset{\sim}{1}\} - \{(\underset{\sim}{1}-\underset{\sim}{z})^T \underset{\sim}{r}/(\underset{\sim}{1}-\underset{\sim}{z})^T \underset{\sim}{1}\} \tag{1.4}$$

where $\underset{\sim}{1}$ is an N dimensional vector of 1's. Alternatively, $t(\underset{\sim}{z}, \underset{\sim}{r})$ could be the difference between two robust measures of the typical response in the two treatment groups. Fisher proposed testing the sharp null hypothesis (1.3) using the tails of the permutation distribution of $t(\underset{\sim}{z}, \underset{\sim}{r})$ induced by the randomization, where $\underset{\sim}{r}$ is, in a sense, treated as a constant. Fixing $\underset{\sim}{r}$ at its observed value in this way is equivalent to conditioning on $\underset{\sim}{r}$ in a randomized experiment, but not generally in an observational study. As a result, randomization tests applied in randomized experiments have the correct size given the observed $\underset{\sim}{r}$, and therefore the correct unconditional size regardless of the distribution of $\underset{\sim}{r}$ (e.g. Lehmann 1959, chapter 5; Rubin 1980). In observational studies, a different type of conditioning may be required.

In order to motivate the general discussion in §2, it is useful to briefly review, with the current notation, the justification for conditioning on the observed response, $\underset{\sim}{r}$, in Fisher's randomization test. In a completely randomized experiment, the treatment assignment, z, has a known distribution that is independent of the response (r_1, r_0), and, moreover, $0 < pr(z=1) < 1$ so treatment assignment is strongly ignorable without any covariates, that is with $\underset{\sim}{x}$ equal to a null vector. The distribution of the observed response, r_z, generally depends on z; however, under the sharp null hypothesis (1.3), the observed response satisfies $r_z = r_1 = r_0$, so treatment assignment, z, is independent of the observed response, r_z. Therefore, under (1.3), the conditional distribution of the treatment assignments given the observed responses, $pr(\underset{\sim}{z}|\underset{\sim}{r})$, is equal to the marginal, randomization distribution of the treatment assignments, $pr(\underset{\sim}{z})$. Hence, under the sharp null hypothesis, (1.3), the conditional distribution of the test statistic, $t(\underset{\sim}{z},\underset{\sim}{r})$, given the value of the observed responses, $\underset{\sim}{r} = \underset{\sim}{c}$ say, equals the permutation distribution of $t(\underset{\sim}{z},\underset{\sim}{c})$ induced by the randomization. Formally, for each constant $\underset{\sim}{c}$,

$$pr\{t(\underset{\sim}{z},\underset{\sim}{r})|\underset{\sim}{r} = \underset{\sim}{c}\} = pr\{t(\underset{\sim}{z},\underset{\sim}{c})|\underset{\sim}{r} =\underset{\sim}{c}\} = pr\{t(\underset{\sim}{z},\underset{\sim}{c})\} \qquad (1.5)$$

from (1.3) and (1.2) with $\underset{\sim}{x}$ equal to a null vector. The conclusion that Fisher's test has the correct conditional size given the observed $\underset{\sim}{r}$, and therefore also the correct unconditional size, is an immediate consequence of (1.5).

Notice that the justification for Fisher's test rests on two conditions. First, the randomization or permutation distribution of $\underset{\sim}{z}$, and therefore also of $t(\underset{\sim}{z},\underset{\sim}{c})$ for each constant $\underset{\sim}{c}$, is known, since it is created by the experimenter. Second, treatment assignment is strongly ignorable without covariates, so under the sharp null hypothesis, the known permutation distribution of $t(\underset{\sim}{z},\underset{\sim}{c})$ equals the relevant conditional distribution of $t(\underset{\sim}{z},\underset{\sim}{r})$ given the observed responses, $\underset{\sim}{r} = \underset{\sim}{c}$; that is, condition (1.5) holds. In observational studies, even if treatment assignment is strongly ignorable, the distribution of treatment assignments is generally unknown, and therefore Fisher's randomization test is not generally applicable.

2. Conditional Permutation Tests Under a Logistic Model for the Propensity Score

2.1 A Basic Theorem

This section shows that Fisher's randomization test may be extended so that it is applicable in observational studies providing (a) treatment assignment is strongly ignorable, and (b) the propensity score follows a logistic model (Cox 1970), that is,

$$\log \frac{e(\underset{\sim}{x})}{1-e(\underset{\sim}{x})} = \underset{\sim}{\beta}^{T}\underset{\sim}{f}(\underset{\sim}{x}) \qquad (2.1)$$

where $\underset{\sim}{\beta}$ is an unknown vector parameter, and $\underset{\sim}{f}(\cdot)$ is a known, vector-valued function of $\underset{\sim}{x}$, such as $\underset{\sim}{f}(\underset{\sim}{x}) = (1,\underset{\sim}{x})^{T}$. Since $\underset{\sim}{f}(\underset{\sim}{x})$ may include polynomial terms in $\underset{\sim}{x}$, condition (2.1) is not particularly restrictive. Let $\underset{\sim}{F}$ and $\underset{\sim}{X}$ be the matrices whose N rows are, respectively, the values of $\underset{\sim}{f}(\underset{\sim}{x}_i)^{T}$ and $\underset{\sim}{x}_i^{T}$, $i = 1,2,\ldots,N$. By a familiar argument (Cox 1970, §4.2), $\underset{\sim}{z}^{T}\underset{\sim}{F}$ is sufficient for $\underset{\sim}{\beta}$ in (2.1).

The proposed test is similar to a randomization test, but with a nuisance parameter, $\underset{\sim}{\beta}$, describing the treatment assignment mechanism. To eliminate the nuisance parameter, we use the conditional distribution of the treatment assignments $\underset{\sim}{z}$ given the sufficient statistic, $\underset{\sim}{z}^{T}\underset{\sim}{F}$, for $\underset{\sim}{\beta}$. Unlike a randomization test, this conditional test compares the observed test statistic, $t(\underset{\sim}{z},\underset{\sim}{r})$, to the value, $t(\underset{\sim}{b},\underset{\sim}{r})$, that would have been obtained under the null hypothesis with a different treatment assignment, indicated by the binary vectcr $\underset{\sim}{b}$, only if $\underset{\sim}{b}$ is similar to the observed treatment assignment in the sense that $\underset{\sim}{b}^{T}\underset{\sim}{F} = \underset{\sim}{z}^{T}\underset{\sim}{F}$. Clearly, alternative treatment assignments satisfying $\underset{\sim}{b}^{T}\underset{\sim}{F} = \underset{\sim}{z}^{T}\underset{\sim}{F}$ will typically exist only when the values in $\underset{\sim}{F}$ are fairly coarse; see for example §3.

Theorem 1. Suppose the propensity score follows the logistic model (2.1).

(A) Then the conditional distribution of treatment assignments $\underset{\sim}{z}$ given $(\underset{\sim}{z}^{T}F,\underset{\sim}{X})$ is free of unknown parameters and assigns the same probability to each binary vector $\underset{\sim}{b}$ satisfying $\underset{\sim}{b}^{T}\underset{\sim}{F} = \underset{\sim}{z}^{T}\underset{\sim}{F}$.

(B) Under the sharp null hypothesis (1.3), if treatment assignment is strongly ignorable, then the conditional distribution of the test statistic, $t(\underset{\sim}{z},\underset{\sim}{r})$ given $\underset{\sim}{r} = \underset{\sim}{c}$ and

$(\mathbf{z}^T\mathbf{F},\mathbf{X})$ equals the known conditional permutation distribution of $t(\mathbf{z},\mathbf{c})$ given $(\mathbf{z}^T\mathbf{F},\mathbf{X})$ that is determined from part (A); i.e.,

$$pr\{t(\mathbf{z},\mathbf{r})|\mathbf{r}=\mathbf{c},\ \mathbf{z}^T\mathbf{F},\ \mathbf{X}\} = pr\{t(\mathbf{z},\mathbf{c})|\mathbf{z}^T\mathbf{F},\ \mathbf{X}\}$$

for each constant $\mathbf{c}$.

(C) If treatment assignment is strongly ignorable, and if, for each fixed $\mathbf{c}$ and $\mathbf{a}$, the set $W(\mathbf{c},\mathbf{a})$ satisfies $pr\{t(\mathbf{z},\mathbf{c}) \in W(\mathbf{c},\mathbf{a})|\mathbf{z}^T\mathbf{F} = \mathbf{a},\mathbf{X}\} = \alpha$ (respectively $\leq \alpha$) under the sharp null hypothesis (1.3), then a test which rejects whenever $t(\mathbf{z},\mathbf{r}) \in W(\mathbf{r},\mathbf{z}^T\mathbf{F})$ has level α (respectively $\leq \alpha$) for all values of the unknown parameter $\boldsymbol{\beta}$.

<u>Proof</u>: Part A is straightforward since $\mathbf{z}^T\mathbf{F}$ is sufficient for $\boldsymbol{\beta}$ in (2.1). To prove part B, note that strong ignorability and the sharp null hypothesis (1.3) imply

$$\mathbf{z} \perp\!\!\!\perp \mathbf{r} \mid \mathbf{X}$$

and hence, essentially following Lemma 4.2(ii) of Dawid (1979),

$$\mathbf{z} \perp\!\!\!\perp \mathbf{r} \mid \mathbf{X},\ \mathbf{z}^T\mathbf{F}$$

since $\mathbf{F}$ is a function of $\mathbf{X}$. Therefore, strong ignorability and the null hypothesis imply

$$\begin{aligned} pr\{t(\mathbf{z},\mathbf{r}) \mid \mathbf{r} = \mathbf{c},\ \mathbf{z}^T\mathbf{F},\ \mathbf{X}\} &= pr\{t(\mathbf{z},\mathbf{c}) \mid \mathbf{r} = \mathbf{c},\ \mathbf{z}^T\mathbf{F},\ \mathbf{X}\} \\ &= pr\{t(\mathbf{z},\mathbf{c}) \mid \mathbf{z}^T\mathbf{F},\ \mathbf{X}\} \qquad (2.2)\end{aligned}$$

as required for part B. Part C follows immediately from (2.2). //

If the treatment effect is constant in the sense that $r_1 = r_0 + \Delta$, or that $r_1 = \Delta\, r_0$, for some scalar Δ, then a confidence interval for Δ may be constructed by inverting the test (Lehmann 1959, §5.4). The test described by Theorem 1 will, however, have the nominal level whether or not the treatment effect is constant.

2.2 <u>An Artificial Example</u>

The following artificial example is intended to clarify the procedure described in §2.1 and to simplify discussion of the backtrack algorithm in §2.3; a practical example appears in §3. The data in Table 1 were generated by setting

$$r_{0i} = 5\ x_i$$

and

$$r_{1i} = 5\ x_i + 1$$

TABLE 1. DATA FOR THE ARTIFICIAL EXAMPLE.

			Observed
	Covariate	Treatment	Response
Unit	x_i	z_i	r_{zi}
A	1	1	6
B	0	1	1
C	1	0	5
D	1	0	5
E	1	0	5
F	1	0	5
G	0	0	0
H	0	0	0
I	0	0	0
J	0	0	0

$$r_{zi} = 5\ x_i + z_i$$

where the treatment effect $r_{1i} - r_{0i}$ equals 1 for each unit, and the observable response is

$$r_{zi} = 5\ x_i + z_i\ .$$

Two units received treatment 1, and eight units received treatment 0.

The test statistic used here is the sample total in the treatment 1 group, that is, $t(\underset{\sim}{z},\underset{\sim}{r}) = \underset{\sim}{z}^T\underset{\sim}{r}$. It is straightforward to show that the critical region induced by this statistic is the same as the critical region induced by the difference in sample means (1.4). See Kempthorne (1952) for details.

Together, the two parts of Table 2 list the elements of the sample space associated with Fisher's randomization test. There are $\binom{10}{2} = 45$ elements in the sample space corresponding to the $\binom{10}{2}$ ways of selecting the two units that will receive the treatment 1. Eleven of the 45 treatment reassignments produce response totals greater than or equal to the observed response total of $\underset{\sim}{z}^T\underset{\sim}{r} = 7$, so Fisher's one sided significance level is 11/45 = .24.

In an observational study, a logistic model, (2.1), for the propensity score with $\underset{\sim}{f}(x_i) = (1,\ x_i)^T$ would lead us to restrict attention to treatment assignments, $\underset{\sim}{b}$, that are similar to the observed treatment assignment in the sense that $\underset{\sim}{b}^T\underset{\sim}{F} = \underset{\sim}{z}^T\underset{\sim}{F} = (2,1)$, that is, treatment assignments in which the treatment group includes one unit with $x_i = 1$ and one unit with $x_i = 0$. This conditional sample space contains the 25 elements in the top half of Table 2. The observed treatment total, $\underset{\sim}{z}^T\underset{\sim}{r} = 7$, is the largest of the 25 treatment totals from the conditional sample space, so the conditional one-sided significance level is 1/25 = .04. In this instance, the one-sided .05 level conditional critical region, $W(\underset{\sim}{r},\underset{\sim}{z}^T\underset{\sim}{F})$, contains only the observed treatment total. By Theorem 1, this test would have the nominal level if treatment assignment is strongly ignorable and model (2.1) holds.

In a completely randomized experiment, both the unconditional and the conditional tests have the nominal level, although the conditional test performs a kind of covariance adjustment; see §2.5. However, in an observational study, only the conditional test can be used because the distribution of treatment assignments generally depends on unknown parameters.

TABLE 2. The Conditional and Unconditional Permutational Sample Space.

Elements of the Conditional Sample Space

	$\underset{\sim}{z}^T\underset{\sim}{r}$		$\underset{\sim}{z}^T\underset{\sim}{r}$		$\underset{\sim}{z}^T\underset{\sim}{r}$
AB*	7	CG	5	EH	5
AG	6	CH	5	EI	5
AH	6	CI	5	EJ	5
AI	6	CJ	5	FG	5
AJ	6	DG	5	FH	5
BC	1	DH	5	FI	5
BD	1	DI	5	FJ	5
BE	1	DJ	5		
BF	1	EG	5		

Additional Elements of the Unconditional Sample Space

	$\underset{\sim}{z}^T\underset{\sim}{r}$		$\underset{\sim}{z}^T\underset{\sim}{r}$
AC	11	DF	10
AD	11	EF	10
AE	11	GH	0
AF	11	GI	0
BG	1	GJ	0
BH	1	HI	0
BI	1	HJ	0
BJ	1	IJ	0
CD	10		
CE	10		
CF	10		
DE	10		

* Letter pairs indicate the units receiving treatment 1.

2.3 A Backtrack Algorithm

In general, calculation of the conditional significance level requires identification of all binary vectors $\underset{\sim}{b}$ such that $\underset{\sim}{b}^T\underset{\sim}{F} = \underset{\sim}{z}^T\underset{\sim}{F}$. For small N, this task is not as difficult as one might suppose, providing we avoid checking most of the 2^N possible binary vectors. This section describes an efficient but easily implemented backtrack algorithm. For a general discussion of backtrack algorithms, see Whitehead (1973, §2.3) or Horowitz and Sahni (1978, chapter 7).

Each binary vector, $\underset{\sim}{b}$, is a path through $N + 1$ nodes of a binary tree; see Figure 1. We begin at the root of the tree, exploring each branch until it becomes apparent that no $\underset{\sim}{b}$ in that branch will satisfy $\underset{\sim}{b}^T\underset{\sim}{F} = \underset{\sim}{z}^T\underset{\sim}{F}$. If we abandon a branch at a node at level k, then we have eliminated 2^{N-k+1} of the 2^N possible binary vectors.

Without loss of generality, we may assume that $\underset{\sim}{F}$ is strictly nonnegative, that is, $f_{im} > 0$ for $i = 1,2,\ldots,N$ and $m = 1,2,\ldots,M$, where f_{im} is the element in the ith row and mth column of $\underset{\sim}{F}$. Suppose we are at a node at level $k + 1$ defined by $(b_1, b_2,\ldots, b_k)$ where $k < N$. A simple rule is to abandon the branch beginning at this node if for some m, $1 < m < M$, either

$$\sum_{i=1}^{k} b_i f_{im} + \sum_{i=k+1}^{N} f_{im} < \sum_{i=1}^{N} z_i f_{im} \tag{2.3}$$

or

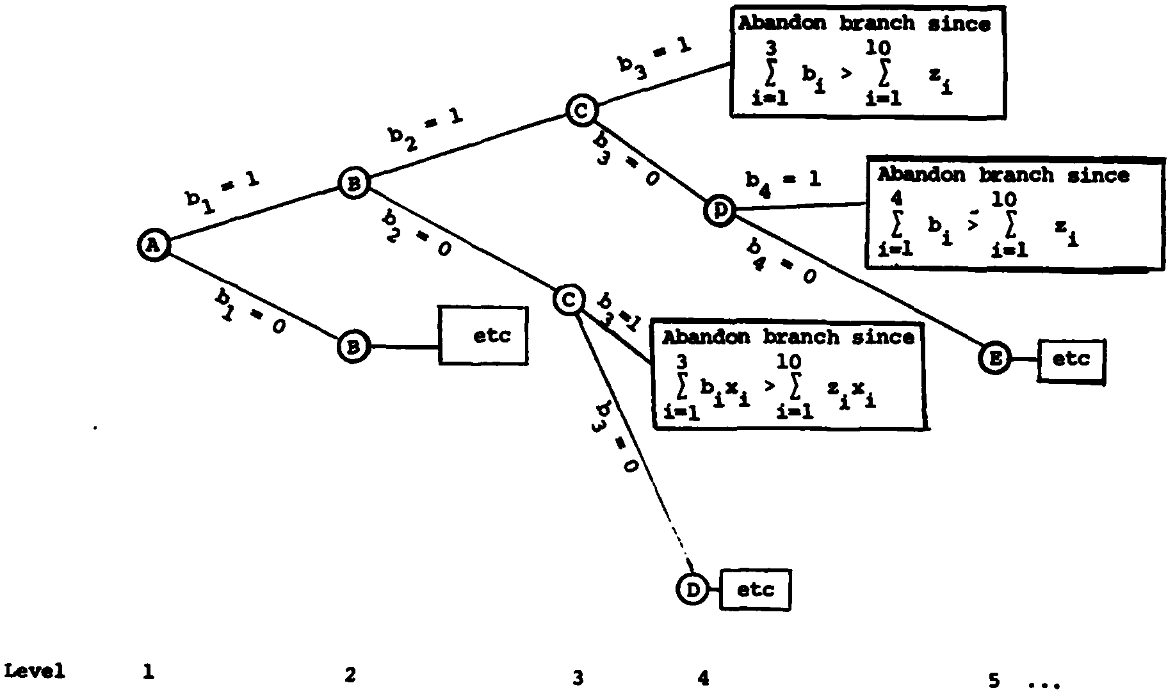
$$\sum_{i=1}^{k} b_i\, f_{im} > \sum_{i=1}^{N} z_i\, f_{im} \tag{2.4}$$

If condition (2.3) holds at a node at level $k + 1$, then $\sum_{i=1}^{k} b_i\, f_{im}$ is already too small, in the sense that every binary vector $\underset{\sim}{b}$ whose first k coordinates correspond to the given node of the tree will satisfy

$$\sum_{i=1}^{N} b_i\, f_{im} < \sum_{i=1}^{N} z_i\, f_{im} .$$

Similarly, if (2.4) holds, then $\sum_{i=1}^{k} b_i\, f_{im}$ is already too large. The procedure is illustrated in Figure 1 using the artificial data from §2.2.

Figure 1. Tree Structure and Algorithm for the Artificial Data


$b_1 = 1$
$b_1 = 0$
$b_2 = 1$
$b_2 = 0$
$b_3 = 1$
$b_3 = 0$
$b_3 = 1$
$b_3 = 0$
$b_4 = 1$
$b_4 = 0$
A
B
B
C
C
D
D
E
etc
etc
etc
Abandon branch since $\sum_{i=1}^{3} b_i > \sum_{i=1}^{10} z_i$
Abandon branch since $\sum_{i=1}^{4} b_i > \sum_{i=1}^{10} z_i$
Abandon branch since $\sum_{i=1}^{3} b_i x_i > \sum_{i=1}^{10} z_i x_i$
Level
1
2
3
4
5 ...

If several units have identical values of the vector $\underset{\sim}{f}(\underset{\sim}{x})$, then a more efficient algorithm may be constructed. Suppose $\underset{\sim}{f}(\underset{\sim}{x}_u) = \underset{\sim}{f}(\underset{\sim}{x}_v)$ for some $u < v$. If $\underset{\sim}{b} = (b_1,b_2,\ldots,b_u,\ldots,b_v,\ldots,b_N)$ solves $\underset{\sim}{b}^T\underset{\sim}{F} = \underset{\sim}{z}^T\underset{\sim}{F}$, then so does $\underset{\sim}{b}^* = (b_1,b_2,\ldots,b_v,\ldots,b_u,\ldots,b_N)$, where $\underset{\sim}{b}^*$ is obtained from $\underset{\sim}{b}$ by interchanging b_u and b_v. Therefore, the solutions of $\underset{\sim}{b}^T\underset{\sim}{F} = \underset{\sim}{z}^T\underset{\sim}{F}$ may be partitioned into equivalence classes, where $\underset{\sim}{b}$ and $\underset{\sim}{b}^*$ are in the same equivalence class if $\underset{\sim}{b}^*$ may be obtained from $\underset{\sim}{b}$ by permuting coordinates of $\underset{\sim}{b}$ associated with identical values of $\underset{\sim}{f}(\underset{\sim}{x})$. To obtain all solutions of $\underset{\sim}{b}^T\underset{\sim}{F} = \underset{\sim}{z}^T\underset{\sim}{F}$, it is sufficient to obtain one $\underset{\sim}{b}$ from each equivalence class of solutions using a backtrack algorithm, and then to obtain the other members of the same equivalence class by appropriate permutations of the coordinates of $\underset{\sim}{b}$. This procedure is a version of isomorph rejection; see Whitehead (1973, §2.4). To obtain one $\underset{\sim}{b}$ from each equivalence class, we may use a backtrack algorithm that abandons a branch at a node at level k if (2.3) or (2.4) holds, or if

$$b_k > b_u \quad \text{for some} \quad u < k \quad \text{such that} \quad \underset{\sim}{f}(\underset{\sim}{x}_u) = \underset{\sim}{f}(\underset{\sim}{x}_k) \quad . \qquad (2.5)$$

In a backtrack algorithm, additional conditions such as (2.5) generally reduce the number of branches that require investigation, thereby generally increasing efficiency. In special cases, more efficient methods are available; see §2.4. Approximations to the null distribution of the test statistic are given in §4, and other related large sample procedures are described by Rosenbaum and Rubin (1983a, §3).

2.4 Standard Tests for Observational Studies Derived as Conditional Tests Given a Sufficient Statistic for the Propensity Score

This section shows that several commonly used tests can be viewed as conditional permutation tests given a sufficient statistic for the propensity score. In these tests, the response r_{ti} is a discrete random variable taking one of R possible values.

In the Mantel-Haenszel(1959) and Mantel (1963) approximate procedures and the corresponding exact procedures given by Birch(1964, 1965) and Cox(1966), there are M subclasses, resulting in an $R$x2x$M$ contingency table (i.e., observed response r_z by treatment z by subclass $\underset{\sim}{x}$). Define $\underset{\sim}{F}$ so that $f_{im} = 1$ if unit i falls in subclass m, and $f_{im} = 0$ otherwise, so that under the logit model (2.1), the unknown conditional probability of assignment to treatment 1 given the covariates $\underset{\sim}{x}$ is constant within each subclass. If treatment assignment is strongly ignorable, then it follows from Theoren 1.B that, under the null hypothesis (1.3), conditioning on $(\underset{\sim}{r},\underset{\sim}{z}^T\underset{\sim}{F},\underset{\sim}{X})$ restricts the sample space to those Rx2xM tables in which each of the J subtables has the same margins as the observed table. This leads to the Birch-Cox exact distributions and the Mantel-Haenszel and Mantel approximations. If treatment assignment is strongly ignorable, subclassification on the propensity score can produce subclasses with the properties required by Theorem 1. For discussion of subclassification on the propensity score, see Rosenbaum and Rubin (1983a, §3.3).

McNemar's(1947) test for paired binary responses is the special case in which each subclass has just two units, with one receiving each treatment. The pairs are typically constructed by matched sampling (e.g., Rubin, 1973). Model (2.1) implies that units have been selected by matched sampling from a population of treated and control units in such a way that the conditional probability of assignment to treatment 1 given covariates $\underset{\sim}{x}$ is constant within each pair. If treatment assignment is strongly ignorable, then matched sampling of treated and control units with the same value of the propensity score can produce matched pairs with the properties required by Theorem 1. For discussion of propensity matching, see Rosenbaum and Rubin (1983a, §3.2).

2.5 Conditional Permutation Tests and Covariance Adjustment

This section examines the relationship between conditional permutation tests and covariance adjustment. In §1.3 and §2.2, the difference in sample means (1.4), or equivalently the treatment 1 total, $\underset{\sim}{z}^T\underset{\sim}{r}$, was used as a test statistic, $t(\underset{\sim}{z},\underset{\sim}{r})$. An alternative test statistic is the difference in means after covariance adjustment for $\underset{\sim}{F}$, that is, the first coordinate of the estimated coefficient vector in the least squares regression of $\underset{\sim}{r}$ on $(\underset{\sim}{z},\underset{\sim}{F})$. The randomization distributions (§1.4) of these two test statistics can lead to markedly different conclusions. We now show that the conditional permutation tests (§2.1) based on these two statistics lead to identical critical regions, and therefore to identical tests and confidence intervals, providing the model (2.1) includes a constant term (or, formally, providing the column rank of $(\underset{\sim}{1},\underset{\sim}{F})$ equals the column rank of $\mathbf{F}$). In a sense, the conditional test performs a covariance adjustment; however, the test has the nominal level even if the linear regression model is incorrect.

To prove the equivalence of conditional tests based on the two test statistics, it is sufficient to show that the covariance adjusted difference, $t^*(\mathbf{b},\mathbf{r})$ say, is a strictly monotone function of $\mathbf{b}^T\mathbf{r}$, for each treatment assignment, $\mathbf{b}$, in the conditional sample space, that is, for $\mathbf{b}$ such that $\mathbf{b}^T\mathbf{F} = \mathbf{z}^T\mathbf{F}$. Without loss of generality, assume $\mathbf{F}$ is of full column rank. Familiar arguments (e.g. Seber, 1977, p. 65) show that the covariance adjusted difference with treatment assignment $\mathbf{b}$ is

$$t^*(\mathbf{b},\mathbf{r}) = \frac{\mathbf{r}^T(\mathbf{I}-\mathbf{P})\mathbf{b}}{\mathbf{b}^T(\mathbf{I}-\mathbf{P})\mathbf{b}}$$

where $\mathbf{P} = \mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T$ and $\mathbf{I}$ is the $N \times N$ identity matrix. Over the conditional sample space, $\mathbf{b}^T\mathbf{F}$ is constant, and therefore $\mathbf{P}\mathbf{b}$ and $\mathbf{b}^T\mathbf{P}\mathbf{b}$ are constant. Moreover, by assumption $\mathbf{1} = \mathbf{F}\mathbf{d}$ for some $\mathbf{d}$, so $\mathbf{b}^T\mathbf{b} = \mathbf{b}^T\mathbf{1} = \mathbf{b}^T\mathbf{F}\mathbf{d}$ is constant. Therefore,

$$t^*(\mathbf{b},\mathbf{r}) = \frac{\mathbf{r}^T\mathbf{b} - k_1}{k_2}$$

for constants k_1 and k_2, so $t^*(\mathbf{b},\mathbf{r})$ is a strictly monotone function of $\mathbf{r}^T\mathbf{b}$, as required to complete the proof.

3. A Clinical Example: Tumor Response In Lung Cancer Patients

3.1. The Conditional Permutation Test

The example in this section illustrates the use of the exact conditional test with adjustments for several covariates in a small observational comparison. The data are adapted from a clinical study of lung cancer in which two slight variants of the same treatment appeared to produce differing tumor response rates. Given the expectation that this minor variation in the treatment would not alter the response rate, it is natural to ask to what extent the observed difference in response rates is surprising, given the characteristics of the patients involved. The data appear in Table 3.

Table 3. Data on 14 Lung Cancer Patients

Patient (i)	Tumor Response* (r_{zi})	Treatment (z_i)	Cell type	Previous Treatment	Performance Status	Subclass** (j)
1	0	0	Squamous	None	0	1
2	0	0	Large cell	None	1	2
3	0	0	Squamous	Radiation	1	3
4	0	0	Squamous	Radiation	1	3
5	0	0	Squamous	Radiation	2	4
6	1	1	Squamous	Radiation	1	3
7	0	1	Squamous	Radiation	1	3
8	0	1	Adenocarcinoma	Radiation	1	5
9	1	1	Squamous	None	1	6
10	0	1	Large cell	None	2	7
11	0	1	Squamous	Radiation & Chemotherapy	2	8
12	0	1	Squamous	Chemotherapy	1	9
13	0	1	Squamous	None	0	1
14	2	1	Squamous	None	1	6

*0 = no response; 1 = partial response; 2 = complete response

**Used in §3.2 and §4.1.

There are three covariates: previous treatment (none, radiation only, chemotherapy only, radiation plus chemotherapy), cell type (squamous, large cell, adenocarcinoma), and performance status (grades 0, 1, 2). In the logit model (2.1), previous treatment was coded as three binary variables, and cell type as two binary variables. The conditional permutation test considers all reassignments of treatments to patients such that (a) 9 patients receive treatment one, and of those 9 patients, (b) one has adenocarcinoma, (c) one has large cell carcinoma, (d) seven have squamous cell carcinoma, (e) three have had only previous radiation therapy, (f) one has had only previous chemotherapy, (g) one has had previous radiation and chemotherapy, and (h) the average performance status is $10/9 \doteq 1.1$. There are 28 such treatment reassignments, as compared to $\binom{14}{5} = 2002$ reassignments in the unconditional, randomization sample space.

Tumor response is defined in terms of a reduction in the size of the tumor. In Table 3, no tumor response has been scored as 0; a partial response as 1; a complete response as 2. All of the tumor responses were observed in patients receiving treatment 1, yielding a total score among patients receiving treatment 1 of $t(\underset{\sim}{z},\underset{\sim}{r}) = \underset{\sim}{z}^T\underset{\sim}{r} = 4$. The conditional permutation distribution of this total under the sharp null hypothesis (1.3) assigns probability $8/28 = .29$ to a total of 4, probability $13/28 = .46$ to 3, probability $4/28 = .14$ to 2, and $3/28 = .11$ to 1. The expected total score under the null hypothesis is 2.93. If the two variations of the treatment were in fact identical, and if treatment assignment is strongly ignorable, there would be little reason to be surprised by the observed total response score in treatment group 1, since 29% of all treatment assignments that are similar to the observed treatment assignment would have resulted in a total of 4.

3.2. Comparison with Other Tests: The Randomization Test; A Test Based on Subclassification

We now compare the results obtained in §3.1 with the results of two other tests: Fisher's randomization test, and an exact test of zero partial association between the response and the treatment within each of the $4 \times 3 \times 3 = 36$ subclasses defined by the covariates. Fisher's test corresponds to a sample space containing $\binom{14}{5} = 2002$ treatment assignments, with an expected total response score under the null hypothesis of 2.57, and a one-sided significance level of $\binom{11}{5}/\binom{14}{5} = .23$. Of course, Fisher's test may not be applicable since treatments were not randomly assigned.

An alternative procedure, described in §2.4, is to form 36 subclasses from the three covariates, and to test for partial association within subclasses. There is a $3 \times 2 \times 36$ dimensional contingency table for the 3 response scores, 2 treatments, and 36 subclasses; each of the 36 three-by-two subtables has fixed margins. In this instance, only 9 of the 36 subclasses contain at least one patient; see the last column in Table 3. Six patients fall in subclasses with no other patient, and two patients who fall in the same subclass had both received treatment 1. In effect, none of these 8 patients contribute to the permutation distribution, since their treatments cannot be reassigned subject to the marginal constraints. It is disturbing to note that the one patient with a complete response and one of the two patients with a partial response are among the 8 patients who do not contribute to the test. The conditional sample space contains $\binom{4}{2}\binom{2}{1} = 12$ treatment reassignments, with an expected total response score of 3.5 under the null hypothesis. The one-sided significance level is .5.

The conditional test described in §3.1 has the advantage of permitting adjustment for covariates with fewer restrictions on the conditional sample space than result from the subclassification procedure. Both tests require the assumption of strongly ignorable treatment assignment with covariates $\underset{\sim}{x}$; however, the tests assume different logistic models for the propensity score.

4. Approximations to the Null Distribution

4.1. An Approximation Based on Exact Conditional Moments

This section develops an approximation to the null distribution of $t(\underset{\sim}{z},\underset{\sim}{r})$ using its exact conditional moments. The approximation generalizes the procedures of Mantel and Haenszel (1959) and Mantel (1963).

As noted in §2.3, we need not generate all solutions, $\underset{\sim}{b}$, of $\underset{\sim}{b}^T\underset{\sim}{r} = \underset{\sim}{z}^T\underset{\sim}{r}$ using the backtrack algorithm; rather, we may generate one solution from each equivalence class of solutions using the backtrack algorithm, and then obtain the other solutions in the same equivalence class by permuting units with identical values of $\mathbf{f}(\mathbf{x})$. Often, the number of equivalence classes will be quite small, while the number of individual solutions will be quite large: in the example in §3, there are 28 solutions, but only 3 equivalence classes of solutions. An approximate procedure is to: (a) identify the equivalence classes of solutions using the backtrack algorithm, (b) obtain by standard methods the conditional expectations and variances of the test statistic, $t(\mathbf{z},\mathbf{r})$, within each equivalence class, (c) combine these expectations and variances in an appropriate way to obtain $\bar{E} = E\{t(\mathbf{z},\mathbf{r})|\mathbf{r}, \mathbf{z}^T\mathbf{f}, \mathbf{x}\}$ and $\bar{V} = \mathrm{var}\{t(\mathbf{z},\mathbf{r})|\mathbf{r}, \mathbf{z}^T\mathbf{f}, \mathbf{x}\}$ under the null hypothesis, and (d) test the hypothesis (1.3) by referring a suitable standardized deviate, such as $\{t(\mathbf{z},\mathbf{r}) - \bar{E}\}/\sqrt{\bar{V}}$, to tables of the normal distribution.

Divide the N units into J subclasses or strata based on $\mathbf{f}(\mathbf{x})$, where there are J distinct values of $\mathbf{f}(\mathbf{x})$. (See for example the last column in Table 3.) Let N_j be the number of units in the jth subclass, and let $\bar{r}_j$ and s_j^2 be the mean and variance of the observed responses of all units in subclass j, where s_j^2 is set to zero if N_j equals one, and s_j^2 is the sum of squared deviations around $\bar{r}_j$ divided by $N_j - 1$ if $N_j > 2$. The kth equivalence class of solutions may be characterized by a vector $(a_{1k},a_{2k},\ldots,a_{Jk})^T$ where a_{jk} is the number of units in subclass j assigned to treatment 1; see Table 4.

The following theorem provides expressions for the null expectation $\bar{E}$ and variance $\bar{V}$ when $t(\mathbf{z},\mathbf{r}) = \mathbf{z}^T\mathbf{r}$.

<u>Theorem</u>: Suppose that treatment assignment is strongly ignorable, and that (2.1) holds. Then under the null hypothesis (1.3), the expectation and variance of the test statistic $t(\mathbf{z},\mathbf{r}) = \mathbf{z}^T\mathbf{r}$ are

$$\bar{E} = \sum_k E_k p_k \tag{4.1}$$

and

$$\bar{V} = \sum_k V_k p_k + \sum_k (E_k - \bar{E})^2 p_k \tag{4.2}$$

where

$$E_k = \sum_j a_{jk}\bar{r}_j \; , \tag{4.3}$$

$$V_k = \sum_j \frac{a_{jk}(N_j - a_{jk})}{N_j} s_j^2 \; , \tag{4.4}$$

Table 4. Calculations for the Approximate Test Based on Exact Moments

Subclass (j)	Cell Type	Previous Treatment	Performance Status	Equivalence Class of Solutions (k)			$\bar{r}_j$	s_j^2	N_j
				1	2	3			
1	Squamous	None	0	1*	1	2	0	0	2
2	Large cell	None	1	0	1	0	0	0	1
3	Squamous	Radiation	1	2	1	1	.25	.25	4
4	Squamous	Radiation	2	0	1	1	0	0	1
5	Adenocarcinoma	Radiation	1	1	1	1	0	0	1
6	Squamous	None	1	2	2	1	1.5	.5	2
7	Large cell	None	2	1	0	1	0	0	1
8	Squamous	Radiation & Chemotherapy	2	1	1	1	0	0	1
9	Squamous	Chemotherapy	1	1	1	1	0	0	1

$\Pi_j \binom{N_j}{a_{jk}}$ = # solutions in the kth equivalence class	12	8	8
P_k	.429	.285	.285
E_k	3.503	.251	.75
V_k	.250	.188	.438

$\underline{g}^T\underline{y} = 4$

$\bar{E} = 2.93$

$\bar{V} = .851$

$(|\underline{g}^T\underline{y} - \bar{E}| - \frac{1}{2})/\sqrt{\bar{V}} = .62$

$1 - \Phi(.62) = .27$

* a_{jk} = # units in subclass j assigned to treatment 1 in the kth equivalence class of solutions.

and

$$p_k = \frac{\prod_j \binom{N_j}{a_{jk}}}{\sum_k \prod_j \binom{N_j}{a_{jk}}} \quad . \qquad (4.5)$$

Remarks: The probability p_k is the proportion of all solutions of $\underset{\sim}{b}^T\underset{\sim}{r} = \underset{\sim}{z}^T\underset{\sim}{r}$ that fall in the kth equivalence class. The expectation, E_k, and variance, V_k, corresponding to the kth equivalence class of solutions, are the expectation and variance of the treatment 1 total, $\underset{\sim}{z}^T\underset{\sim}{r}$, in a stratified randomized experiment in which a_{jk} of the N_j units in subclass j are randomly assigned to treatment 1. In the variance, (4.4), the factor $(N_j - a_{jk})/N_j$ is a finite population correction. If there is only one equivalence class, then $\bar{E}$ and $\bar{V}$ are the expectations and variances appearing in the Mantel-Haenszel (1959) approximation for binary responses (see also Birch (1964, §4) and Cox (1966, §3)), and in the Mantel (1963) approximation for scored responses (see also Birch (1965, §5)).

Proof: Let C_k be the kth equivalence class of solutions. Clearly,

$$\begin{aligned}\bar{E} &= \sum_k E(\underset{\sim}{z}^T\underset{\sim}{r}|\underset{\sim}{r}=\underset{\sim}{c},\ \underset{\sim}{z}^T\underset{\sim}{r},\ \underset{\sim}{X},\ \underset{\sim}{z} \in C_k)\mathrm{pr}(\underset{\sim}{z} \in C_k|\underset{\sim}{r},\ \underset{\sim}{z}^T\underset{\sim}{r},\ \underset{\sim}{X}) \\ &= \sum_k E(\underset{\sim}{z}^T\underset{\sim}{c}|\underset{\sim}{X},\ \underset{\sim}{z} \in C_k)\mathrm{pr}(\underset{\sim}{z} \in C_k|\underset{\sim}{X},\underset{\sim}{z}^T\underset{\sim}{r})\end{aligned}$$

for each constant $\underset{\sim}{c}$, by (1.2) and (1.3), and the fact that $\underset{\sim}{z}^T\underset{\sim}{r}$ is constant for all solutions, and in particular is constant for all solutions in C_k. By Theorem 1.A and simple combinatorial arguments, it follows that $\mathrm{pr}(\underset{\sim}{z} \in C_k|\underset{\sim}{X},\underset{\sim}{z}^T\underset{\sim}{r}) = p_k$. Now, all solutions in C_k assign a_{jk} units from subclass j to treatment 1, and moreover, all solutions in C_k are equally probable by Theorem 1.A, so with $\underset{\sim}{c}$ equal to the observed response $\underset{\sim}{r}$, the permutational expectation $E(\underset{\sim}{z}^T\underset{\sim}{c}|\underset{\sim}{X},\ \underset{\sim}{z} \in C_k)$ equals E_k. We have proved (4.1). Similarly,

$$\begin{aligned}\bar{V} &= \sum_k \mathrm{var}(\underset{\sim}{z}^T\underset{\sim}{r}|\underset{\sim}{r},\ \underset{\sim}{z}^T\underset{\sim}{r},\ \underset{\sim}{X},\ \underset{\sim}{z} \in C_k)\mathrm{pr}(\underset{\sim}{z} \in C_k|\underset{\sim}{r},\ \underset{\sim}{z}^T\underset{\sim}{r},\ \underset{\sim}{X}) \\ &\quad + \mathrm{var}\{E(\underset{\sim}{z}^T\underset{\sim}{r}|\underset{\sim}{r},\ \underset{\sim}{z}^T\underset{\sim}{r},\ \underset{\sim}{X},\ \underset{\sim}{z} \in C_k)|\underset{\sim}{r},\ \underset{\sim}{z}^T\underset{\sim}{r},\ \underset{\sim}{X}\} \\ &= \sum_k V_k p_k + \sum_k (E_k - \bar{E})^2 p_k \quad ,\end{aligned}$$

as required. //

Table 4 illustrates the procedure for the example in §3. The approximate significance level is .27, compared to the exact significance level of .29.

4.2. A Large Sample Approximation

This section describes a large sample approximation to the test defined in Theorem 1 when the test statistic is the total response in treatment group 1, that is, when $t(\underset{\sim}{z},\underset{\sim}{r}) = \underset{\sim}{z}^T\underset{\sim}{r}$. Let $v = \log[\mathrm{pr}(z=1|\underset{\sim}{x},r_z)/\{1 - \mathrm{pr}(z=1|\underset{\sim}{x}, r_z)\}]$, and let $\underset{\sim}{v}$ be the corresponding vector for the N units under study. Consider the following logistic model for $\underset{\sim}{z}$:

$$\underset{\sim}{v} = \underset{\sim}{X}\underset{\sim}{\gamma} + \underset{\sim}{r}\theta \qquad (4.6)$$

where $\underset{\sim}{\gamma}$ and θ are, respectively, unknown vector and scalar parameters. Note that (1.2), (1.3) and (2.1) imply $\theta = 0$ in (4.6). Indeed, the exact test defined in Theorem 1 is, under the null hypothesis (1.3), formally identical to the exact, uniformly most powerful similar region test, described by Cox (1970, §4.2), of the hypothesis that $\theta = 0$. To demonstrate the equivalence, it is sufficient to note that the test statistics, the null distributions, and hence the critical regions are the same. Since the tests are equivalent for every finite sample, their asymptotic properties under the null hypothesis are also identical, so a test of (1.3) may be based on the familiar large sample properties of tests of $\theta = 0$ in (4.6); see Cox (1970, §6.4) for discussion of these tests.

REFERENCES

Birch, M.W. (1964). The detection of partial association, I: The 2x2 case. Journal of the Royal Statistical Society, Series B, 26, 313-324.

Birch, M.W. (1965). The detection of partial association, II: The general case. Journal of the Royal Statistical Society, Series B, 27, 111-124.

Cox, D.R. (1958). The Planning of Experiments. New York: John Wiley and Sons.

Cox, D.R. (1966). A simple example of a comparison involving quantal data. Biometrika 53, 215-220.

Cox, D.R. (1970). The Analysis of Binary Data. London: Methuen.

Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion). Journal of the Royal Statistical Society, Series B, 41, 1-31.

Fisher, R.A. (1935). The Design of Experiments. (1st Edition) Edinburgh: Oliver and Boyd.

Hamilton, M.A. (1979). Choosing a parameter for 2 x 2 table or 2 x 2 x 2 table analysis. American Journal of Epidemiology, 109,362-375.

Horowitz, E. and Sahni, S. (1978). Fundamentals of Computer Algorithms. Potomac, Maryland: Computer Science Press.

Kempthorne, O. (1952). The Design and Analysis of Experiments. New York: John Wiley and Sons.

Lehmann, E. (1959). Testing Statistical Hypotheses. New York: John Wiley.

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.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute 22, 719-748.

McNemar, Q. (1947). Note on the sampling error of differences between correlated proportions or percentages. Psychometrika 12, 153-157.

Rosenbaum, P.R. (1982). Testing the assumption of strongly ignorable treatment assignment in observational studies: A review within a general framework. Submitted to the Journal of the American Statistical Association.

Rosenbaum, P.R. and Rubin, D.B. (1983a). The central role of the propensity score in observational studies for causal effects. To appear in Biometrika, 70, #1.

Rosenbaum, P.R. and Rubin, D.B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary response. To appear in the Journal of the Royal Statistical Society, Series B, 45, #2.

Rubin, D. B. (1973). Matching to remove bias in observational studies. Biometrics 29, 159-183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66, 688-701.

Rubin, D.B. (1977). Assignment to treatment group on the basis of a covariate. Journal of Educational Statistics 2, 1-26.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. Annals of Statistics 6, 34-58.

Rubin, D.B. (1980). Discussion of "Randomization Analysis of Experimental Data: The Fisher Randomization Test". Journal of the American Statistical Association 75, 591-593.

Seber, G. A. F. (1977). Linear Regression Analysis. New York: John Wiley.

Whitehead, E. G. (1973). Combinatorial Algorithms. New York: Courant Institute of Mathematical Sciences, New York University.

PRR/jb/jvs

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
# 2463

2. GOVT ACCESSION NO.
A125241

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Conditional Permutation Tests and the Propensity Score in Observational Studies

5. TYPE OF REPORT & PERIOD COVERED
Summary Report - no specific reporting period

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Paul R. Rosenbaum

8. CONTRACT OR GRANT NUMBER(s)
P30-CA-14520
DAAG29-80-C-0041

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Mathematics Research Center, University of Wisconsin
610 Walnut Street
Madison, Wisconsin 53706

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Work Unit Number 4 - Statistics & Probability

11. CONTROLLING OFFICE NAME AND ADDRESS
See Item 18 below

12. REPORT DATE
December 1982

13. NUMBER OF PAGES
23

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office	National Cancer Institute	Wisconsin Alumni
P. O. Box 12211	9000 Rockville Pike	Research Foundation
Research Triangle Park	Bethesda, MD 20205	Univeristy of Wisconsi
North Carolina 27709		Madison, WI 53705

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Observational studies; randomization tests; ignorable treatment assignment; conditional inference; logistic models.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
In observational studies, the distribution of treatment assignments is unknown, and therefore randomization tests are not generally applicable. However, permutation tests that condition on sample information about the treatment assignment mechanism can be applicable in observational studies providing treatment assignment is strongly ignorable. These tests use the conditional distribution of the treatment assignments given a sufficient statistic for the unknown parameter of the propensity score. Several tests that are commonly used in observational studies are particular instances of this general procedure;

ABSTRACT (continued)

moreover, conditional permutation tests and covariance adjustment are closely related. A backtrack algorithm is developed to permit efficient calculation of the exact conditional significance level, and two approximations are discussed. A clinical study of treatments for lung cancer is used to illustrate the technique. Conditional permutation tests extend previous large sample results on the propensity score by providing a general basis for exact inference in small observational studies when treatment assignment is strongly ignorable.