

Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot

BY PAUL R. ROSENBAUM

*Department of Statistics, The Wharton School of the University of Pennsylvania,
Philadelphia, Pennsylvania 19104-6302, U.S.A.*

rosenbaum@stat.wharton.upenn.edu

SUMMARY

In randomisation and permutation inference, pivotal arguments remove the hypothesised treatment effect, thereby basing inferences on the null distribution in which the treatment has no effect. This is common, for instance, with additive treatment effects. The current paper uses ‘attributable effects’ to expand substantially the scope of pivotal arguments. Attributable effects are defined for three cases, namely the 2×2 contingency table, displacement effects and the Mann–Whitney–Wilcoxon statistic, and in each case removing an appropriate attributable effect restores the familiar null randomisation distribution of the associated statistic, yielding exact inferences. The procedure extends immediately for use in sensitivity analysis in nonrandomised observational studies.

Some key words: Attributable risk; Control median test; Fisher’s exact test; Placement; Randomisation inference; Sensitivity analysis.

1. PIVOTS IN PERMUTATION INFERENCE

Fisher (1935) described randomisation as the reasoned basis for inference in experiments, for it created without assumptions the required distributions. This view encourages randomised experimentation and forces analyses of nonrandomised observational data to acknowledge explicitly, as part of the quantitative findings, greater uncertainty about the effects caused by treatments than would have been present had treatments been randomly assigned; see § 5.

A limitation is that many randomisation tests are not paired with confidence intervals and point estimates. If the treatment has an additive effect, τ , so that receiving the treatment increases response by τ , then a randomisation test can be inverted to yield a confidence interval and a point estimate for τ (Hodges & Lehmann, 1963; Moses, 1965; Lehmann, 1963, 1975), and similar considerations may be applied to multiplicative effects. This inversion uses a pivot in the following way: under the hypothesis $H_0: \tau = \tau_0$, subtracting τ_0 from this response of each treated subject restores the null hypothesis of no treatment effect, permitting the null randomisation distribution of the test statistic to be used. Though practical in this case, the model of an additive effect is limited, and is not applicable even for Fisher’s exact test for the 2×2 contingency table, where confidence intervals and point estimates usually invoke distributional assumptions not based on randomisation.

The current paper shows that this pivotal argument may be applied much more generally if inferences concern attributable effects. The attributable effect is not a parameter, but

rather an unobservable random variable indicating how the observed treated group would have responded differently under the control. Although not itself a parameter, inference about the attributable effect proceeds in an entirely standard fashion, because the attributable effect summarises many possible values of a high-dimensional parameter. Sections 2–4 discuss randomised experiments, while § 5 discusses sensitivity analysis in observational studies. In § 2, Fisher's exact test for a 2×2 table is inverted to yield exact inference about attributable effects. This exact inference uses no distributional assumption beyond the random assignment of treatments, and it is based on a hypergeometric distribution with altered marginal totals. In § 3, attributable effects are defined for ordered responses, such as continuous responses or discrete ordered scores, and again exact inferences about quantile displacements are obtained based solely on the random assignment of treatments with no assumption of additive effects. In § 4, attributable effects associated with the Mann–Whitney–Wilcoxon test are studied. The extension in § 5 to sensitivity analysis in observational studies is immediate.

Attributable effects have numerous links to other methods, including measures of attributable risk (MacMahon & Pugh, 1970, pp. 223–4; Cole & MacMahon, 1971; Walter, 1976; Hamilton, 1979; Gastwirth et al., 1999), the control-median procedure of Gart (1963) and Gastwirth (1968), the quantile-comparison function discussed by Li et al. (1996) and nonparametric methods based on placements (Orban & Wolfe, 1982; Kim & Wolfe, 1993), and I have selected terminology, such as attributable effects and displacements, to point to some of these links. Several of these papers make use of hypergeometric null distributions for the hypothesis of no treatment effect, while using asymptotic theory of infinite population sampling under the alternative. In contrast, the current paper obtains exact randomisation inferences when the null hypothesis of no effect does not hold, in some cases based on hypergeometric distributions, and extends the method to permit exact sensitivity analysis in nonrandomised studies.

2. RANDOMISED EXPERIMENT WITH BINARY RESPONSES

2.1. *Randomisation distribution under the alternative*

There are N subjects, $i = 1, \dots, N$, of whom m were exposed to treatment, signified by $Z_i = 1$, and $N - m$ were exposed to a control condition, signified by $Z_i = 0$, so that $m = \sum_{i=1}^N Z_i$. Write $Z = (Z_1, \dots, Z_N)^T$, and write B for the set containing the $|B| = N! / \{m!(N - m)!\}$ possible values of Z , that is, the set of N -dimensional vectors z with $z_i = 0$ or $z_i = 1$ such that $m = \sum_{i=1}^N z_i$. In a completely randomised experiment, Z is selected at random from B , so that $\text{pr}(Z = z) = 1/|B|$ for each $z \in B$, and in §§ 2–4 treatments are assigned at random.

Each subject i has two potential responses, r_{Ti} and r_{Ci} , which are the responses subject i would exhibit under treatment and control, respectively, so that r_{Ti} is observed if $Z_i = 1$ and r_{Ci} is observed if $Z_i = 0$ (Neyman, 1923; Rubin, 1974). In randomisation inference, probability enters only through the random assignment of treatments, so the potential responses ($r_{Ti}, r_{Ci}, i = 1, \dots, N$) are fixed features of the finite population of N subjects, whereas the observed response from subject i , namely $R_i = Z_i r_{Ti} + (1 - Z_i) r_{Ci}$, is a random variable depending on the random assignment of the treatment Z_i (Fisher, 1935). The vectors Z and $R = (R_1, \dots, R_N)^T$ are both observed and they record the treatment assignment and observed responses.

Throughout the paper, the treatment is assumed to have a nonnegative effect, in that $r_{Ti} \geq r_{Ci}$, for $i = 1, \dots, N$. The model of a nonnegative effect cannot be verified or refuted

by inspecting the responses of individuals, because r_{Ti} and r_{Ci} are never jointly observed on the same person. In this section, the responses record the occurrence or absence of an event, such as death, with 1 or 0 signifying, respectively, that the event occurred or did not. For instance, if a hazardous treatment has a nonnegative effect, then an individual might die under both treatment and control, so that $(r_{Ti}, r_{Ci}) = (1, 1)$, or survive under both treatment and control, so that $(r_{Ti}, r_{Ci}) = (0, 0)$, or die if exposed to the hazardous treatment but not if exposed to control, so that $(r_{Ti}, r_{Ci}) = (1, 0)$, but exposure to the hazard does not cause someone to survive who would have died without the hazardous exposure, so that $(r_{Ti}, r_{Ci}) = (0, 1)$ never occurs. Hamilton (1979) reinterprets the standard epidemiological measures of effect on a binary response in terms of the model of nonnegative effects, and he notes that these measures are all compatible with the assumptions whenever treated subjects are at increased risk. The null hypothesis of no treatment effect asserts that $r_{Ti} = r_{Ci}$ for $i = 1, \dots, N$. Table 1 records the observed 2×2 contingency table.

Table 1. *The observed 2×2 contingency table*

Response	Treated	Control
1	$\sum Z_i r_{Ti}$	$\sum (1 - Z_i) r_{Ci}$
0	$\sum Z_i (1 - r_{Ti})$	$\sum (1 - Z_i) (1 - r_{Ci})$
Total	m	$N - m$

The observed counts in Table 1 do not, in general, follow a hypergeometric distribution. In contrast, the unobserved counts in Table 2 record the responses all N subjects would have exhibited had they all been assigned to the control, and they do follow a hypergeometric distribution in a randomised experiment. Tables 1 and 2 coincide under the null hypothesis of no treatment effect, $H_0: r_{Ti} = r_{Ci}$, for $i = 1, \dots, N$, and this fact forms the basis for Fisher's exact test of this null hypothesis, which uses the tail area of the corner cell, the number of events $\sum Z_i r_{Ti}$ in the treated group, as the significance level. The first goal is to find a pivot appropriate for binary responses which yields a hypergeometric distribution in a randomised experiment under the alternative hypothesis that the treatment causes an increase in some responses.

Table 2. *The responses that would have been observed had all subjects received the control*

Response	Treated	Control	Total
1	$\sum Z_i r_{Ci}$	$\sum (1 - Z_i) r_{Ci}$	$\sum r_{Ci}$
0	$\sum Z_i (1 - r_{Ci})$	$\sum (1 - Z_i) (1 - r_{Ci})$	$\sum (1 - r_{Ci})$
Total	m	$N - m$	N

2.2. *Testing hypotheses about treatment effects*

Write $\delta = (r_{T1} - r_{C1}, \dots, r_{TN} - r_{CN})^T$ for the vector of treatment effects. Now, δ is not observable, because only one of r_{Ti} and r_{Ci} is observed, but the observed data together with the nonnegative effects, $r_{Ti} \geq r_{Ci}$, do constrain the possible values of δ . Consider a specific hypothesis, $H_0: \delta = \delta_0$, for some specific vector δ_0 of 1's and 0's. Call δ_0 compatible with the observed data or, more briefly, 'compatible', if $\delta_{0i} = 0$ whenever $(R_i, Z_i) = (0, 1)$ or $(R_i, Z_i) = (1, 0)$, and 'incompatible' otherwise. This says that, because $r_{Ti} \geq r_{Ci}$, if a subject i receives the treatment $Z_i = 1$ and exhibits a response of $r_{Ti} = 0$, then r_{Ci} also

equals 0 and δ_i must equal 0. Similarly, if subject i receives the control $Z_i = 0$ and exhibits a response of $r_{Ci} = 1$, then r_{Ti} also equals 1 and δ_i must equal 0. An incompatible hypothesis, $H_0: \delta = \delta_0$, can be rejected with certainty, that is with type I error rate of zero. Obviously, the one true hypothesis, namely $H_0: \delta = \delta$, is compatible for every treatment assignment Z .

Write $A = \sum_{i=1}^N Z_i(r_{Ti} - r_{Ci}) = \sum_{i=1}^N Z_i\delta_i$. Then A is the number of events attributable to the treatment, that is the number of events which occurred in treated subjects who would not have had these events had they been exposed to the control instead. Whereas $\sum_{i=1}^N Z_i r_{Ti}$ is simply the observed number of events among treated subjects, the attributable effect, A , cannot be calculated from the data because r_{Ci} is never observed when $Z_i = 1$. Note that A is a random variable whose value cannot be observed, not a fixed parameter, since the value of A would change if Z changed, that is if different subjects were exposed to treatment.

The quantity $\sum Z_i r_{Ti} - \sum Z_i \delta_i = \sum Z_i r_{Ti} - A = \sum Z_i r_{Ci}$ is a pivotal quantity, in the sense that its distribution does not depend on the unknown parameter δ ; that is Table 3 equals Table 2.

Table 3. *The observed table adjusted for the attributable effect*

Response	Treated	Control
1	$\sum Z_i r_{Ti} - \sum Z_i \delta_i$	$\sum (1 - Z_i) r_{Ci}$
0	$\{\sum Z_i (1 - r_{Ti})\} + \sum Z_i \delta_i$	$\sum (1 - Z_i) (1 - r_{Ci})$
Total	m	$N - m$

Consider testing $H_0: \delta = \delta_0$. If δ_0 is incompatible, then reject the hypothesis with type I error rate of zero. Otherwise, if δ_0 is compatible, then test $H_0: \delta = \delta_0$ by calculating the hypothesised attributable effect $A_0 = \sum Z_i \delta_{0i}$ and use it in place of $\sum Z_i \delta_i$ in Table 3. Under the null hypothesis, the distribution of the corner cell $\sum Z_i r_{Ti} - \sum Z_i \delta_{0i}$ is given by the hypergeometric distribution for a table with the margins of Table 3. Note carefully, however, that Table 3 does not have the same row totals as the observed data in Table 1. The adjusted corner cell, $\sum Z_i r_{Ti} - \sum Z_i \delta_{0i}$, may be the basis for either a one-sided or a two-sided test.

2.3. Confidence sets and attributable effects

Conceptually, the test described above could be used to create a set estimate, say a 95% confidence set, for the N -dimensional parameter δ . As always, the 95% set estimate is the set of values of δ not rejected by a 5% level test (Lehmann, 1986, § 3.5). The 95% confidence set for δ is a subset of the set of all N -dimensional vectors with coordinates 1 or 0, and this random set contains the fixed, true δ in 95% of experiments. This confidence set is impractical because it is N -dimensional, but its existence is conceptually useful. In particular, all compatible hypotheses $H_0: \delta = \delta_0$ that give rise to the same attributable effect $\sum Z_i \delta_{0i}$ yield the same significance level, so they are all either included or excluded from the 95% confidence set for δ at the same time. As a result, it is tempting to say that we are testing the hypothesis that the attributable effect, namely $\sum Z_i \delta_i$, equals the hypothesised value, namely $\sum Z_i \delta_{0i}$. To say this is to speak informally, in that the effect attributable to treatment, $\sum Z_i \delta_i$, is a random variable, so one cannot test a hypothesis about its value. Nonetheless, there is little harm in using this suggestive terminology providing one keeps clearly in mind that one is actually summarising many hypothesis tests of the

form $H_0: \delta = \delta_0$ and noting that all hypotheses that give rise to a particular value of the attributable effect $\sum Z_i \delta_{0i}$ are simultaneously included or excluded from the confidence set.

Although the confidence set for δ is an awkward subset of the set of N -dimensional vectors with binary coordinates, there is a simple one-dimensional interval of values of the attributable effect $\sum Z_i \delta_{0i}$ which lead to δ_0 being included in the N -dimensional confidence set. This is proved in the Appendix. In other words, one can summarise the N -dimensional confidence set for δ by reporting the one-dimensional interval of values of the attributable effect $\sum Z_i \delta_{0i}$ which do not lead to rejection of δ_0 .

3. DISPLACEMENT EFFECTS ATTRIBUTABLE TO TREATMENT

3.1. Defining displacement effects with ordered responses

To extend the method of § 2 to ordered responses, let y_{Ti} be the ordered response subject i would exhibit under treatment and let y_{Ci} be the response this same subject would exhibit under control. Let $y_{C(N)} \geq \dots \geq y_{C(1)}$ be the N order statistics of responses to control that would have been observed had all N subjects been assigned to control, and let $y_{T(N)} \geq \dots \geq y_{T(1)}$ be the N order statistics that would have been observed had all N subjects been assigned to treatment. Since only m of the N subjects received the treatment, neither the $y_{C(j)}$ nor the $y_{T(j)}$ are observed. However, the observed data do permit inference about the $y_{C(j)}$ and the $y_{T(j)}$.

As before, it is assumed that the treatment has a nonnegative effect, $y_{Ti} \geq y_{Ci}$, for $i = 1, \dots, N$, and the data arise from a completely randomised experiment with m subjects assigned at random to treatment and the remaining $N - m$ assigned to control. In a randomised experiment, nonnegative effects imply that the treated responses are stochastically larger than the control responses, and, by a result of Strassen (1965), whenever observable distributions are stochastically ordered, there exists a joint distribution satisfying nonnegative effects.

Fix an integer k , so that $y_{C(k)}$ is the unobserved k/N quantile of responses to control. Let θ be a value strictly between $y_{C(k)}$ and $y_{C(k+1)}$, so that $y_{C(k+1)} > \theta > y_{C(k)}$, and say that subject i has a ‘displacement’ around θ if $y_{Ti} > \theta > y_{Ci}$. The observed data will reveal whether such a θ exists. Let $Y_i = Z_i y_{Ti} + (1 - Z_i) y_{Ci}$ be the observed response from subject i and let the order statistics of the Y_i 's be $Y_{(N)} \geq Y_{(N-1)} \geq \dots \geq Y_{(1)}$. Write $r_{Ci} = 1$ if $y_{Ci} > \theta$ and $r_{Ci} = 0$ otherwise, and write $r_{Ti} = 1$ if $y_{Ti} > \theta$ and $r_{Ti} = 0$ otherwise. Then there is a displacement if $\delta_i = r_{Ti} - r_{Ci} = 1$ and no displacement if $\delta_i = r_{Ti} - r_{Ci} = 0$. The number of displacements attributable to treatment is $A = \sum_i Z_i \delta_i$. Under the null hypothesis of no treatment effect, $H_0: y_{Ti} = y_{Ci}$, for $i = 1, \dots, N$, it follows that $\delta_i = 0$ for $i = 1, \dots, N$.

A traditional choice of k is $k/N = \frac{1}{2}$, somewhat analogous to the control median test of Gart (1963) and Gastwirth (1968). In some applications, it may be interesting to ask whether the treatment caused the response to be abnormally high, rather than higher than typical, in which case $k/N = 0.95$ might be appropriate.

The attributable effect,

$$A = \sum_{i=1}^N Z_i (r_{Ti} - r_{Ci}) = \sum_{i=1}^N Z_i \delta_i,$$

is the number of subjects who were caused to have responses above the k/N quantile of potential control responses because they were exposed to the treatment, and would have had responses below that quantile had they been exposed to the control instead. For instance, if $A/m = 50\%$ for $k/N = 95\%$, then half of the treated subjects were caused to

have responses above the 95% quantile of responses that would have been seen had all subjects received the control.

3.2. Inference about displacement effects in experiments

The structure here is similar to § 2, but $y_{C(k)}$ is not observed, so r_{Ti} and r_{Ci} cannot be calculated for any subject i . This uncertainty about $y_{C(k)}$ is not a problem, however.

PROPOSITION 1. *If $a = \sum_i Z_i \delta_i$, then $Y_{(k+1-a)} > \theta > Y_{(k-a)}$.*

Proof. There are exactly $N - k$ subjects with $y_{Ci} > \theta$, and since $y_{Ti} \geq y_{Ci}$ it follows that these $N - k$ subjects all have $Y_i > \theta$. Since $a = \sum_i Z_i \delta_i$ there are exactly a other subjects, not included among the $N - k$ subjects, with $Y_i = y_{Ti} > \theta > y_{Ci}$. For the remaining $k - a$ subjects, $\theta > y_{Ti} \geq y_{Ci}$, so $\theta > Y_i$. Thus there are exactly $N - k + a$ subjects with $Y_i > \theta$ and exactly $k - a$ subjects with $\theta > Y_i$. This means that

$$Y_{(N)} \geq Y_{(N-1)} \geq \dots \geq Y_{(k+1-a)} > \theta > Y_{(k-a)} \geq \dots \geq Y_{(1)}. \quad \square$$

Proposition 1 is used in the following way. To test the hypothesis $H_0: \delta = \delta_0$, calculate the observed order statistics, $Y_{(N)} \geq \dots \geq Y_{(1)}$, and the hypothesised attributable effect, $A_0 = \sum_i Z_i \delta_i$. If $Y_{(k+1-A_0)} = Y_{(k-A_0)}$, then there cannot be A_0 displacements around a θ such that $y_{C(k+1)} > \theta > y_{C(k)}$, so reject H_0 with type one error rate of zero. Assume therefore from now on that $Y_{(k+1-A_0)} > Y_{(k-A_0)}$, in which case, if the hypothesis is true, then $Y_{(k+1-A_0)} > \theta > Y_{(k-A_0)}$. Call the hypothesis compatible with the observed data if $\delta_{0i} = 0$ for all i such that either $Z_i = 0$ and $Y_i > Y_{(k-A_0)}$ or $Z_i = 1$ and $Y_{(k-A_0)} \geq Y_i$; otherwise, call the hypothesis incompatible. If the hypothesis is incompatible, reject it with certainty. If it is compatible, then a treated subject, $Z_i = 1$, has $r_{Ti} = 1$ if $Y_i > Y_{(k-A_0)}$ and $r_{Ti} = 0$ if $Y_{(k-A_0)} \geq Y_i$, while a control subject, $Z_i = 0$, has $r_{Ci} = 1$ if $Y_i > Y_{(k-A_0)}$ and $r_{Ci} = 0$ if $Y_{(k-A_0)} \geq Y_i$. Write $I(E) = 1$ if event E occurs, $I(E) = 0$ otherwise, and compute Table 4. Under the null hypothesis, Table 4 equals Tables 2 and 3, and the same test may be performed. Note that Table 4 always has row totals $N - k$ and k .

Table 4. *The observed table for testing an attributable displacement effect $H_0: \delta = \delta_0$*

Response	Treated	Control
1	$\sum Z_i I(Y_i > Y_{(k-A_0)}) - A_0$	$\sum (1 - Z_i) I(Y_i > Y_{(k-A_0)})$
0	—	—
Total	m	$N - m$

‘—’ signifies ‘by subtraction’.

3.3. An example: Cytogenetic effects of benzene

In an observational study, Tunca & Egeli (1996) attempted to estimate the cytogenetic changes caused by long-term occupational exposure to benzene. They compared $m = 58$ shoe workers near Bursa, Turkey, to $N - m = 20$ controls who were also residents of Bursa, Turkey, but who were believed not to have exposure to benzene. The shoe workers used glues containing substantial quantities of benzene, for periods ranging from 5 to 50 years. Tunca & Egeli (1996, p. 1316) describe working conditions with insufficient ventilation and benzene exposure levels 10 to 30 times the maximum allowable concen-

tration. In the current section, the data will be analysed as if they came from an experiment, while sensitivity to bias from nonrandom assignment of treatments is examined in § 5.

For each of the $N = 78$ subjects, roughly 20 metaphases were analysed for chromosomal aberrations, but the number of metaphases analysed varied from person to person. Although several types of aberration were examined, for illustration here attention will focus on the percent of gaps found for each subject. Although recorded to several decimal places, the data are fairly coarse, exhibiting ties. For instance, for several subjects, 17 metaphases were analysed and gaps were found in two, so the percentage of gaps is $\frac{2}{17} = 11.76\%$, and this number appears several times. Sorted into order, the percentages of gaps for the 58 shoe workers are 0.00, 0.00, 5.55, 6.66, 7.69, 8.33, 8.33, 8.69, 9.09, 9.09, 9.52, 10.00, 10.00, 10.00, 10.00, 10.52, 11.11, 11.11, 11.76, 11.76, 11.76, 11.76, 11.76, 12.00, 12.50, 13.04, 13.33, 13.33, 13.33, 13.33, 13.51, 14.28, 15.38, 15.38, 15.38, 16.66, 17.64, 17.64, 18.18, 18.60, 19.04, 19.23, 19.23, 20.00, 20.00, 20.00, 20.00, 20.00, 20.00, 21.05, 23.07, 23.80, 24.13, 25.00, 27.27, 29.41, 30.00 and 33.33. For the 20 controls, the sorted percentages of gaps are 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 5, 5, 5, 5, 5, 5, 8 and 10. Many of the shoe workers had many more gaps than did most of the controls.

The $N = 78$ order statistics of responses to treatment and control, namely

$$\{y_{Ti}, y_{Ci}, i = 1, \dots, 78\},$$

are not observed. The median of $N = 78$ observations is the average of the 39th and 40th order statistics. Therefore, let $k = 39$, so that subject i is defined to have a displacement above the control median if there is a θ such that $y_{C(40)} > \theta > y_{C(39)}$ and $y_{Ti} > \theta > y_{Ci}$. Roughly speaking, subject i has a displacement above the median if he would have had fewer than the median number of gaps that would have been seen had all $N = 78$ subjects escaped exposure to benzene, but instead would have had strictly more than that median had the subject been exposed to benzene.

Table 5 tests three groups of hypotheses, namely each compatible hypothesis $H_0: \delta = \delta_0$ with $A_0 = 1$, with $A_0 = 19$ and with $A_0 = 25$. Since $Y_{(39)} = 11.76 > 11.11 = Y_{(38)}$, if $A = 1$, then $y_{C(40)} > \theta > y_{C(39)}$ for any θ between 11.11 and 11.76. Since none of the observed responses for controls is above 11.11, the first 2×2 table in Table 5 is obtained, with significance level $P < 0.00001$, so it is not plausible that only $A = 1$ treated subject experienced a displacement from 11.11 or below to 11.76 or above. Similarly, it is not plausible that only $A = 19$ treated subjects had displacements from $Y_{(20)} = 5$ or below to $Y_{(21)} = 5.55$

Table 5. Testing three hypotheses in the benzene data

$A = 1, Y_{(39-1)} = 11.11, P < 0.00001$				$A = 19, Y_{(39-19)} = 5, P = 0.000025$			
	Treated	Control	Total		Treated	Control	Total
$y_{Ci} > 11.11$	39	0	39	$y_{Ci} > 5$	37	2	39
$11.11 \geq y_{Ci}$	19	20	39	$5 \geq y_{Ci}$	21	18	39
Total	58	20	78	Total	58	20	78

$A = 25, Y_{(39-25)} = 4, P = 0.219$			
	Treated	Control	Total
$y_{Ci} > 4$	31	8	39
$4 \geq y_{Ci}$	27	12	39
Total	58	20	78

or above. Now, $Y_{(15)} = \dots = Y_{(20)} = 5$. It is plausible that $A = 25$ or fewer treated subjects had displacements from $Y_{(14)} = 4$ or below to $Y_{(15)} = 5$ or above.

4. RANK SUM STATISTICS AND ATTRIBUTABLE EFFECTS

4.1. *Attributable effects in randomised experiments*

This section considers pivotal inference for attributable effects based on the Wilcoxon rank sum test and the Mann–Whitney U -test. It is convenient to assume, at first, that the N potential responses to control are not tied, $y_{C(N)} > \dots > y_{C(1)}$, and to discuss adjustments for ties separately. Ties do not present technical problems, but it is easier to interpret the attributable effect when ties are either absent or rare. Write

$$\tilde{W}_{ij} = \begin{cases} 1 & \text{if } y_{Ti} > y_{Cj}, \\ 0 & \text{otherwise,} \end{cases} \quad W_{ij} = \begin{cases} 1 & \text{if } y_{Ci} > y_{Cj}, \\ 0 & \text{otherwise,} \end{cases}$$

so that $W_{ii} = 0$ and $q_i = 1 + \sum_{j=1}^N W_{ij}$ is the rank of y_{Ci} among the N potential responses to control, y_{Cj} , for $j = 1, \dots, N$. Write $\delta_{ij} = \tilde{W}_{ij} - W_{ij}$. Since the treatment is assumed to have a nonnegative effect, $y_{Ti} \geq y_{Ci}$, it follows that $\delta_{ij} \geq 0$. Under the null hypothesis of no treatment effect, $H_0: y_{Ti} = y_{Ci}$, for $i = 1, \dots, N$, it follows that $\delta_{ij} = \tilde{W}_{ij} - W_{ij} = 0$ for all i, j . If $\delta_{ij} = 1$, then exposure of subject i to the treatment would cause subject i to have a higher response than subject j would have under the control, whereas i would have had a response no higher than subject j had they both received the control. Although there are N^2 different δ_{ij} , the maximum value of $\sum_{i=1}^N \sum_{j=1}^N \delta_{ij}$ is not N^2 but rather $N + N(N - 1)/2$ because all N of the δ_{ii} could equal 1, but $1 = W_{ij} + W_{ji}$ so that $1 \geq \delta_{ij} + \delta_{ji}$ for $i \neq j$.

Now \tilde{W}_{ij} is observed only if subject i received the treatment and subject j received the control, that is only if $Z_i = 1$ and $Z_j = 0$, so that many \tilde{W}_{ij} are not observed, but the Mann–Whitney U -statistic $V = \sum_{i=1}^N \sum_{j=1}^N Z_i(1 - Z_j)\tilde{W}_{ij}$ is observed. As is well known, the Mann–Whitney statistic differs by a constant from Wilcoxon's rank sum statistic. Under the null hypothesis of no treatment effect in a randomised experiment, $\tilde{W}_{ij} = W_{ij}$, so that then V would equal $\sum_{i=1}^N \sum_{j=1}^N Z_i(1 - Z_j)W_{ij}$ which in turn equals $-m(m + 1)/2 + \sum Z_i q_i$ which has the conventional null randomisation distribution of the Mann–Whitney statistic. If we write $A = \sum_{i=1}^N \sum_{j=1}^N Z_i(1 - Z_j)\delta_{ij}$, it follows that $V - A = \sum_{i=1}^N \sum_{j=1}^N Z_i(1 - Z_j)W_{ij}$ is a pivot which, under the alternative hypothesis, has the null distribution of the Mann–Whitney statistic, with expectation $m(N - m)/2$ and variance $m(N - m)(N + 1)/12$. The quantity $A/\{m(N - m)\}$ is the proportion of the $m(N - m)$ possible comparisons of one of the m treated subjects with one of the $N - m$ controls in which exposure to the treatment caused the treated subject to have a higher response than the control, where this treated subject would not have had a higher response than this control had they both been exposed to the control. Under the null hypothesis of no treatment effect, $H_0: y_{Ti} = y_{Ci}$, for $i = 1, \dots, N$, the attributable proportion is $A/\{m(N - m)\} = 0$.

Write Δ for the $N \times N$ matrix containing the δ_{ij} , and consider testing the hypothesis $H_0: \Delta = \Delta_0$, where Δ_0 is an $N \times N$ matrix of 1's and 0's. The observed pattern of responses can rule out some Δ_0 's with certainty. This can happen in one of three ways:

- (i) if treated subject i is observed to have a higher response than treated subject k , then it must be the case that $\delta_{ij} \geq \delta_{kj}$ for every j ;
- (ii) if control subject k is observed to have a higher response than control subject j , then it must be the case that $\delta_{ij} \geq \delta_{ik}$ for every i ; and
- (iii) if treated subject i is observed to have a lower response than control subject j ,

then $\delta_{ij} = 0$.

Call Δ_0 incompatible if it violates one of these conditions; otherwise, call Δ_0 compatible. Test $H_0: \Delta = \Delta_0$ by rejecting Δ_0 with certainty if Δ_0 is incompatible, and otherwise, if Δ_0 is compatible, calculate the attributable effect under this hypothesis, $A_0 = \sum_{i=1}^N \sum_{j=1}^N Z_i(1 - Z_j)\delta_{0ij}$, rejecting the null hypothesis if $V - A_0$ falls in the extreme tail of the conventional null distribution of the Mann–Whitney statistic. Again, the set of Δ_0 not rejected in this way by a 0.05 level test is a 95% confidence set for the $(N \times N)$ -dimensional parameter Δ , and this confidence set can be summarised by reporting a one-dimensional interval of values of the attributable proportion $A/\{m(N - m)\}$.

4.2. Adjustments for ties

The procedure for adjusting for ties is simple to use, but if ties are extremely numerous then the significance level and confidence intervals may be slightly conservative, and the scoring of ties begins to affect the interpretation of the attributable proportion. If there are ties among the potential responses to control, define

$$W_{ij} = \begin{cases} 1 & \text{if } y_{Ci} > y_{Cj}, \\ 0 & \text{if } y_{Ci} < y_{Cj}, \\ \frac{1}{2} & \text{if } y_{Ci} = y_{Cj}, \end{cases}$$

so that $W_{ii} = \frac{1}{2}$ and $q_i = \frac{1}{2} + \sum_{j=1}^N W_{ij}$ is the rank of y_{Ci} among the N potential responses to control, y_{Cj} , for $j = 1, \dots, N$, with average ranks used for ties. If there is no tie, these q_i are the same as in § 4.1. Similarly, define

$$\tilde{W}_{ij} = \begin{cases} 1 & \text{if } y_{Ti} > y_{Cj}, \\ 0 & \text{if } y_{Ti} < y_{Cj}, \\ \frac{1}{2} & \text{if } y_{Ti} = y_{Cj}. \end{cases}$$

Now, $\delta_{ij} = \tilde{W}_{ij} - W_{ij}$ may take values 0, $\frac{1}{2}$ or 1, where an unchanged inequality scores $\delta_{ij} = 0$, a reversed inequality scores $\delta_{ij} = 1$, and a switch between an inequality and a tie scores $\delta_{ij} = \frac{1}{2}$. This scoring method is not unreasonable for a small number of ties, and it allows inference to be based on the usual permutation distribution of the Mann–Whitney–Wilcoxon statistic with ties. When ties are extremely numerous, the mid-scoring of ties will affect the interpretation of the attributable effect A , so alternative formulations, such as the displacements in § 3, may be more natural.

The permutation distribution of the Mann–Whitney–Wilcoxon statistic with ties exists, but cannot be usefully tabulated, because different patterns of ties give rise to slightly different distributions. The common practice is to approximate the permutation distribution of the statistic by a Normal distribution using the expectation and variance of $-\frac{1}{2}m(m + 1) + \sum_{i=1}^N Z_i q_i$, which, in a randomised experiment, is simply a constant plus the total of m fixed scores drawn at random without replacement from a population of size N . This works for testing a specific hypothesis, $H_0: \Delta = \Delta_0$. In all previous cases in §§ 2–4, two hypotheses that gave rise to the same attributable effect A also gave rise to the same inference, but ties can change this slightly. Different Δ_0 's may yield the same A but may yield different ties in forming the q_i 's leading to slightly different variances for the Mann–Whitney–Wilcoxon statistic and hence slightly different inferences. Wherever ties fall, the null expectation of the statistic remains $m(N - m)/2$, but ties reduce the variance

below the untied variance, $m(N - m)(N + 1)/12$, although the reduction is often very small when ties are not extremely numerous. Consider the following procedure: compute the Mann–Whitney–Wilcoxon statistic V with the adjustment for ties, and accept into a one-sided 95% confidence set all hypotheses Δ_0 giving rise to A_0 such that

$$1.65 \geq \frac{V - A_0 - m(N - m)/2}{\{m(N - m)(N + 1)/12\}^{\frac{1}{2}}}.$$

Then in large samples this confidence set will have coverage of approximately 95% or greater, and the set would be summarised by reporting the corresponding range of attributable proportions, $A_0/\{m(N - m)\}$.

4.3. Example: Cytogenetic effects of benzene

In the benzene example, there were $m = 58$ shoe workers exposed to benzene and $N - m = 20$ unexposed controls, giving $m(N - m) = 58 \times 20 = 1160$ comparisons of a benzene worker and a control. In $V = 1117$ or $V/\{m(N - m)\} = 96\%$ of these comparisons, the benzene worker had a higher percent of chromosome gaps, whereas, under the null hypothesis of no effect in a randomised experiment, this was expected in $m(N - m)/2 = 1160/2 = 580$ or 50% of the comparisons, so that the observed proportion is 46% higher than expected in the absence of a treatment effect. A hypothesis $H_0: \Delta = \Delta_0$ which has an attributable proportion $A_0/\{m(N - m)\}$ below

$$33.9\% = V/\{m(N - m)\} - \frac{1}{2} - 1.65[(N + 1)/\{12m(N - m)\}]^{\frac{1}{2}}$$

is rejected as implausible at the 0.05 level in a large-sample one-sided test. In the absence of a treatment effect the attributable proportion would have been zero, but in a randomised experiment we would have been 95% confident that the attributable proportion is at least 33.9%. In this calculation, ties slightly affected the computation of V , but were not used in computing the variance, so that the large-sample confidence level is actually somewhat greater than 95%.

5. SENSITIVITY ANALYSIS IN OBSERVATIONAL STUDIES

In an observational study, treatments are not randomly assigned, so the scientist makes a diligent and determined effort to record important pretreatment characteristics and to compare, perhaps by matching or stratification, treated and control subjects who appeared comparable in terms of these observed covariates. Even with best efforts, scientists are typically concerned that some important covariate escaped measurement, creating an unobserved or hidden bias, that is a systematic departure from random assignment for ostensibly similar subjects. A sensitivity analysis asks about the degree to which the conclusions of an observational study might be altered by varied assumptions about the magnitude of the departure from random assignment that the unobserved covariate produces. One model for sensitivity analysis (Rosenbaum, 1987, 1995a, § 4) begins with treatment assignments Z_i that are independent with unknown probabilities such that two subjects may differ in their odds of exposure to treatment by at most a factor of $\Gamma \geq 1$, that is

$$\Gamma \geq \{\text{pr}(Z_i = 1) \text{pr}(Z_j = 0)\} / \{\text{pr}(Z_i = 0) \text{pr}(Z_j = 1)\} \geq \Gamma^{-1}$$

for all i, j . The distribution of Z is then returned to the set B by conditioning on

$m = \sum Z_i$. When $\Gamma = 1$, this yields the randomisation distribution and randomisation inferences, whereas for fixed $\Gamma > 1$ the distribution of treatment assignments Z is unknown but departs from randomisation to a bounded extent. For several fixed values of Γ , a sensitivity analysis computes bounds on inference quantities, such as significance levels or the endpoints of confidence intervals, thereby indicating the magnitude of bias that would be needed to alter the conclusions of the study. An earlier sensitivity analysis by Cornfield et al. (1959) had an important influence on the debate about the effects of smoking on the risk of lung cancer. Other methods of sensitivity analysis are discussed by Rosenbaum & Rubin (1983), Rosenbaum (1986), Angrist et al. (1996), Copas & Li (1997) and Lin et al. (1998).

The methods for attributable effects in §§ 2–4 extend immediately for use in sensitivity analysis. Specifically, for testing the null hypothesis of no treatment effect, $H_0: r_{Ti} = r_{Ci}$, for $i = 1, \dots, N$, methods exist for displaying the sensitivity of significance levels for the 2×2 table using the extended hypergeometric distribution as bounds (Rosenbaum, 1995b) and for the Wilcoxon rank sum test (Rosenbaum & Krieger, 1990). In both cases, and also in the case of displacement effects, subtraction of the attributable effect restores the null distribution, so that inferences about attributable effects follow immediately from sensitivity analyses for the null hypothesis of no effect, in parallel with the case of randomised experiments in §§ 2–4.

The sensitivity analysis will be illustrated for the benzene example in § 3.2, the procedure being unchanged except that the resulting 2×2 tables in Table 5 are compared to the extended hypergeometric distributions to obtain an upper bound on the significance level. For instance, for testing hypotheses $H_0: \delta = \delta_0$ yielding $A_0 = 19$ displacements attributable to treatment, the upper bounds on the significance level for $\Gamma = 1, 2, 3$ and 4 are 0.000025, 0.0027, 0.020 and 0.058. For $A_0 = 19$ to be even marginally plausible, failure to control for an unobserved covariate would need to alter the odds of treatment by at least a factor of $\Gamma = 4$.

ACKNOWLEDGEMENT

This research was supported by a grant from the Methodology, Measurement, and Statistics Program and the Statistics and Probability Program of the U.S. National Science Foundation and by the Research Foundation of the University of Pennsylvania.

APPENDIX

Adjustment of extended hypergeometric distributions

Table A1 is a 2×2 contingency table in which the upper left-hand corner cell has been reduced by a count of a and the lower left-hand corner cell has been increased by a , with consequent changes in the row totals. The observed table is obtained by letting $a = 0$. If $D - a$ in Table A1 has the extended hypergeometric distribution, then $\text{pr}(D - a = b)$ is $\pi(a, b)$ given by

$$\pi(a, b) = \left\{ \binom{L-a}{b} \binom{N-L+a}{m-b} \Gamma^b \right\} / \left\{ \sum_{c=\max(0, m+L-a-N)}^{\min(m, L-a)} \binom{L-a}{c} \binom{N-L+a}{m-c} \Gamma^c \right\},$$

if $\min(m, L - a) \geq b \geq \max(0, m + L - a - N)$, and is given by $\pi(a, b) = 0$ otherwise; and $\text{pr}(D - a \geq k)$ is $\lambda(a, k)$, given by $\lambda(a, k) = \sum_{b=\max(k, m+L-a-N)}^{\min(m, L-a)} \pi(a, b)$.

In drawing inferences, one entertains various values of a . The proposition compares the tail probability $\lambda(a, k)$ for various values of a . Note that this is not a conventional comparison because distributions on 2×2 tables with different marginal totals are compared. Proposition A1 is the

Table A1. A 2×2 table adjusted for attributable effects

Response	Treated	Control	Total
1	$D - a$	$L - D$	$L - a$
0	$m - D + a$	$N - m - L + D$	$N - L + a$
Total	m	$N - m$	N

basis for the claim that the set of values of δ_0 that are not rejected corresponds to an interval of values of the attributable effect $A_0 = \sum Z_i \delta_{0i}$. Proposition A1 looks at the chance that $D - a$ is greater than or equal to k for two values of a ; however, note carefully that, when a changes, the margins in Table A1 and the relevant extended hypergeometric distribution of D also change.

PROPOSITION A1. *If $a \geq a^*$, then $\lambda(a^*, k) \geq \lambda(a, k)$.*

Proof. It suffices to prove this for $a^* = a - 1$, as the general result follows by induction. Moreover, it suffices to prove that if $b < b^*$ then

$$\pi(a, b)\pi(a - 1, b^*) \geq \pi(a, b^*)\pi(a - 1, b),$$

by the familiar fact that TP_2 ordering implies stochastic ordering (Barlow & Proschan, 1975, § 5.4; Eaton, 1987, p. 4; Rosenbaum, 1999). There are two cases to be considered. First, it is straightforward to verify that if $\pi(a, b)\pi(a - 1, b^*) = 0$ then $\pi(a, b^*)\pi(a - 1, b) = 0$ also. Secondly, if $\pi(a, b^*)\pi(a - 1, b) > 0$ then $\pi(a, b)\pi(a - 1, b^*)$ and $\pi(a, b^*)\pi(a - 1, b)$ have the same denominator, and simple algebra applied to the numerators suffices to complete the proof. \square

REFERENCES

- ANGRIST, J., IMBENS, G. & RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* **91**, 444–69.
- BARLOW, R. & PROSCHAN, F. (1975). *Statistical Theory of Reliability and Life Testing*. New York: Holt, Rinehart and Winston.
- COLE, P. & MACMAHON, B. (1971). Attributable risk percent in case-control studies. *Br. J. Prev. Social Med.* **25**, 242–4.
- COPAS, J. B. & LI, H. G. (1997). Inference for nonrandom samples (with Discussion). *J. R. Statist. Soc. B* **59**, 55–96.
- CORNFIELD, J., HAENSZEL, W., HAMMOND, E., LILIENFELD, A., SHIMKIN, M. & WYNDER, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Nat. Cancer Inst.* **11**, 1269–75.
- EATON, M. L. (1987). *Lectures on Topics in Probability Inequalities*. Amsterdam: Centrum voor Wiskunde en Informatica.
- FISHER, R. A. (1935). *Design of Experiments*. Edinburgh: Oliver and Boyd.
- GART, J. (1963). A median test with sequential application. *Biometrika* **50**, 55–62.
- GASTWIRTH, J. L. (1968). The first-median test: a two-sided version of the control median test. *J. Am. Statist. Assoc.* **63**, 692–706.
- GASTWIRTH, J. L., KRIEGER, A. M. & ROSENBAUM, P. R. (1999). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* **85**, 907–20.
- HAMILTON, M. A. (1979). Choosing the parameter for 2×2 and $2 \times 2 \times 2$ table analysis. *Am. J. Epidemiol.* **109**, 362–75.
- HODGES, J. L. & LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34**, 598–611.
- KIM, D. & WOLFE, D. A. (1993). Properties of distribution-free two-sample procedures based on placements. *Far East J. Math. Sci.* **1**, 179–90.
- LEHMANN, E. L. (1963). Nonparametric confidence intervals for a shift parameter. *Ann. Math. Statist.* **34**, 1507–12.
- LEHMANN, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden Day.
- LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*. New York: Wiley.
- LI, G., TIWARI, R. C. & WELLS, M. T. (1996). Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *J. Am. Statist. Assoc.* **91**, 689–98.

- LIN, D., PSATY, B. & KRONMAL, R. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54**, 948–63.
- MACMAHON, B. & PUGH, T. F. (1970). *Epidemiology*. Boston, MA: Little, Brown.
- MOSES, L. (1965). Confidence limits from rank tests. *Technometrics* **7**, 257–60.
- NEYMAN, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, § 9 (in Polish). *Roczniki Nauk Rolniczych Tom X*, pp. 1–51. Reprinted in English (1990), *Statist. Sci.* **5**, 465–80, with Discussion by T. Speed and D. B. Rubin.
- ORBAN, J. & WOLFE, D. A. (1982). A class of distribution-free two-sample tests based on placements. *J. Am. Statist. Assoc.* **77**, 666–72.
- ROSENBAUM, P. R. (1986). Dropping out of high school in the United States: An observational study. *J. Educ. Statist.* **11**, 207–24.
- ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74**, 13–26. Correction (1988), **75**, 396.
- ROSENBAUM, P. R. (1995a). *Observational Studies*. New York: Springer-Verlag.
- ROSENBAUM, P. R. (1995b). Quantiles in nonrandom samples and observational studies. *J. Am. Statist. Assoc.* **90**, 1424–31.
- ROSENBAUM, P. R. (1999). Holley's inequality. In *Encyclopedia of Statistical Sciences, Update Volume 3*, Ed. S. Kotz, C. B. Read and D. Banks, pp. 328–30. New York: Wiley.
- ROSENBAUM, P. R. & KRIEGER, A. M. (1990). Sensitivity analysis for two-sample permutation inferences in observational studies. *J. Am. Statist. Assoc.* **85**, 493–8.
- ROSENBAUM, P. R. & RUBIN, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc. B* **45**, 212–8.
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.
- STRASSEN, V. (1965). The existence of probability distributions with given marginals. *Ann. Math. Statist.* **36**, 423–39.
- TUNCA, B. T. & EGELI, U. (1996). Cytogenetic findings on shoe workers exposed long-term to benzene. *Envir. Health Perspect.* **104**, Suppl. 6, 1313–7.
- WALTER, S. D. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics* **32**, 829–49.

[Received September 1998. Revised November 1999]