

Confidence Intervals for Uncommon but Dramatic Responses to Treatment

Paul R. Rosenbaum

Department of Statistics, The Wharton School, University of Pennsylvania,
Philadelphia, Pennsylvania 19104-6340, U.S.A.
email: rosenbaum@stat.wharton.upenn.edu

SUMMARY. A small literature discusses locally most powerful rank tests when only a fraction of treated subjects respond to treatment. The ranks used in these tests are very different from conventional ranks, being relatively flat for low responses and then rising steeply, and the associated tests are much more powerful than conventional rank tests when, indeed, only a small fraction of treated subjects exhibit dramatic responses. Because the tests are derived from considerations of local power, they do not yield a plausible family of models for effect, and therefore they do not yield confidence intervals for the magnitude of effect formed by inverting the tests. There is a similarity between these tests and another family of tests, originally motivated by different considerations involving peak performance in small subsets. Exploiting this similarity, a method for obtaining confidence statements is proposed. In the case of observational studies, sensitivity to unobserved bias from nonrandom assignment of treatments is also examined. Two examples are used as illustrations: (i) a study of smoking during pregnancy and its effects on birth weight, in which smokers are matched to six controls, and (ii) a matched pair study of damage to DNA among workers at aluminum production plants.

KEY WORDS: Attributable effects; Causal effects; Observational study; Sensitivity analysis.

1. Introduction

1.1 Example: Smoking During Pregnancy and Birth Weight

Using data from the National Institute on Drug Abuse's "Washington, DC Metropolitan Area Drug Study (DC*MADS)," Figure 1 depicts the birth weights in grams of 48 babies born to mothers who smoked regularly during pregnancy and 288 = 6 × 48 matched control babies born to mothers who did not smoke during pregnancy. None of these mothers reported using illicit drugs or drinking alcohol regularly during pregnancy. The 48 smokers were each matched to six nonsmokers for two variables, whether or not the mother was on Medicaid, and whether or not the mother received regular prenatal care (i.e., at least once a month). Of the smokers, 46 of 48 reported smoking less than a pack a day and 2 of 48 reported smoking at least a pack a day; however, it is not clear how accurate these self-reports of amount smoked are. On the left in Figure 1a, there is a quantile–quantile plot of the two marginal distributions of birth weights, from 48 smokers and 288 controls, with the line of equality, $x = y$. On the right in Figure 1b are a pair of boxplots of birth weight. The 48 points in Figure 1a project horizontally into the smoker's boxplot in Figure 1b. (The 48 points for the 288 control babies are interpolated from their empirical distribution function.)

In the quantile–quantile plot, for heavier babies, the points approach the $x = y$ line, so that the upper quantiles of the two distributions are similar, but the lower quantiles are substantially lower for the babies whose mothers smoked. Figure 1 suggests an effect of smoking, but not a shift that could be

modeled as an additive effect. The pattern in Figure 1 is not uncommon: smoking seems to have had a harmful effect on the birth weights of some babies, and perhaps little or no effect on many others. The smallest 4 of 48 babies (8%) born to smoking mothers and the smallest 5 of 288 babies (2%) born to non-smoking mothers weighed less than 1 kg (2.2 pounds), and so are clearly babies at high risk. Perhaps there is heterogeneity in the biological effects of smoking, or perhaps the self-reports of moderate smoking by almost all the mothers contain some substantial understatement of the amounts smoked, or perhaps both are true.

1.2 Outline: Invertible Tests that Resemble the Locally Most Powerful Test

Several authors, including Lehmann (1953), Salsburg (1986, 1992), and Conover and Salsburg (1988), have proposed locally most powerful rank tests when only a fraction of subjects respond to treatment, and they have shown that some of these tests can have much greater power against this alternative than do conventional rank tests; see Section 1.3. Because the tests were built from considerations of local power, they do not generate confidence statements for the magnitude of effect by inverting the test. However, Section 1.3 shows that these locally optimal tests are very similar to a second family of tests, proposed by Stephenson (1981) and Stephenson and Ghosh (1985), who were originally motivated by different considerations. It is a familiar fact that the ranks in the Mann–Whitney–Wilcoxon test may be obtained by comparing all subjects two at a time, noting who had the higher response

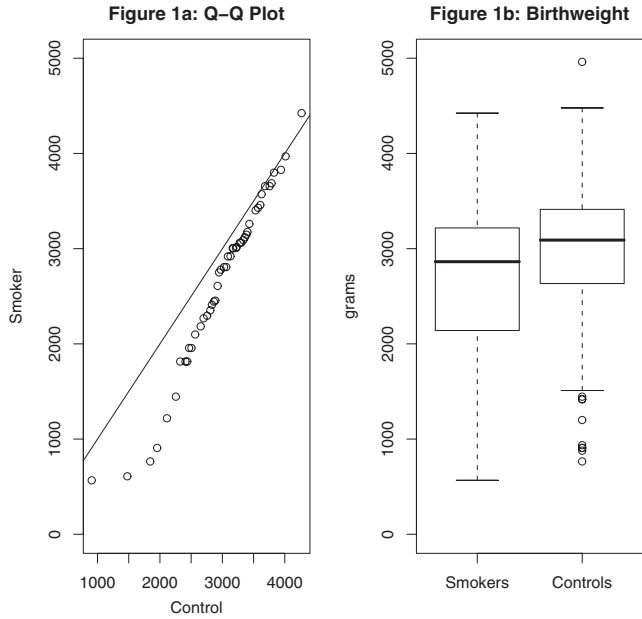


Figure 1. Birth weight of 48 babies born to mothers who smoked regularly during pregnancy and 288 = 6 × 48 matched control babies.

in each set of two subjects. Stephenson (1981) and Stephenson and Ghosh (1985) suggested comparing subjects not two at a time but m at a time, for fixed $m \geq 2$, and this yields ranks very similar to those proposed by Conover and Salsburg (1988). Notation is defined in Section 2.1, and a Web-based supplementary appendix Section 5 confirms that this second family of tests also has good power against alternatives in which only a fraction of subjects respond to treatment. The new confidence intervals are introduced in Section 2.2, with some suggestions for scaling to aid interpretation in Section 2.3. The example in Section 1.1 is examined again in Section 2.4. Section 2 concerns procedures that compete with the stratified Wilcoxon rank sum test, whereas in Section 3, attention shifts to paired data and competitors of Wilcoxon’s signed rank test. A second, paired example is discussed in Section 3.2. A brief summary is provided in Section 4.

1.3 Review: Rank Tests Using Polynomial Functions of Ranks

When J sorted responses have ranks $j = 1, 2, \dots, J$, several investigators with varied motivations have proposed using rank scores that are polynomial functions of j , including, for fixed m , the functions j^{m-1} , $\{J!(j+m-2)!\}/\{(J+m-1)!(j-1)!\}$, and $\binom{j-1}{m-1}$, each of which is a polynomial in j of the order j^{m-1} . Some investigators have been concerned with improved local power against constant treatment effects or shift alternatives, where fairly small increases in power were sometimes possible, but other investigators have been concerned with situations, such as Section 1.1, in which only a small fraction of treated subjects respond to treatment, where substantial increases in power are typical. None of these investigators proposed confidence intervals for the case of rare but dramatic responses to treatment, in part be-

cause their tests were built to be optimal against local alternatives, whereas a confidence interval is not a local statement. It turns out that these several tests behave similarly, and in later sections it is seen that one of the tests inverts gracefully to yield exact or approximate confidence intervals for measures of rare but dramatic responses to treatment.

In unmatched comparisons of a total of J independent observations divided between treated and control groups, with responses that have continuous distributions $G(\cdot)$ and $F(\cdot)$, respectively, one usually thinks of Wilcoxon’s rank sum test as well designed to detect a constant effect or shift alternative, $G(y) = F(y - \tau)$, and indeed it is the locally most powerful rank test against a shift alternative when $F(y) = 1/(1 + e^{-y})$ is the logistic distribution. What if there is not a constant effect, the same for everyone, but rather only a small fraction of subjects, p , with $0 < p < 1$, respond to the treatment? Perhaps surprisingly, Lehmann (1953, Section 6) showed that Wilcoxon’s rank sum test is also the locally most powerful rank test against the alternative $G(y) = (1 - p)F(y) + pF^2(y)$ as $p \rightarrow 0$. In this alternative, $1 - p$ of the treated subjects are unaffected by treatment with the same distribution $F(\cdot)$ as controls, but a tiny fraction p are affected, with a stochastically larger distribution $F^2(\cdot)$. After offering several biological mechanisms that might limit effects to a subset of treated subjects, Salsburg (1986) suggested the model $G(y) = (1 - p)F(y) + pF^m(y)$ where $m > 1$ need not be 2, and he suggested that the j th largest response might be assigned a rank proportional to j^{m-1} , which of course yields conventional ranks when $m = 2$. Extending Lehmann’s argument, Conover and Salsburg (1988) found the locally most powerful test against $G(y) = (1 - p)F(y) + pF^m(y)$ as $p \rightarrow 0$ to assign the j th largest response the rank

$$\begin{aligned} \rho_{mj} &= \frac{J!(j+m-2)!}{(J+m-1)!(j-1)!} \\ &= \left\{ \frac{J!}{(J+m-1)!} \right\} (j+m-2)(j+m-3) \cdots (j), \end{aligned}$$

which is a polynomial in j whose highest power of j is j^{m-1} , so for large J , Salsburg’s (1986) ranks j^{m-1} and Conover and Salsburg (1988)’s ranks ρ_{mj} behave similarly. Conover and Salsburg (1988) argued that $m = 5$ is often a good choice when only a small fraction of treated subjects respond strongly.

Motivated by different considerations to be discussed later, Stephenson (1981) and Stephenson and Ghosh (1985) replaced conventional ranks $j = 1, \dots, J$ by ranks $q_{mj} = \binom{j-1}{m-1}$, which are defined to be zero for $j < m$. A little arithmetic shows that these ranks are related to the Conover–Salsburg (1988) ranks by

$$\begin{aligned} \frac{q_{mj}}{\rho_{mj}} &= \binom{J+m-1}{J} \cdot \frac{(j-1)(j-2) \cdots (j-m+1)}{j(j+1) \cdots (j+m-2)} \\ &\text{for } j \geq m, \text{ and } \frac{q_{mj}}{\rho_{mj}} = 0 \text{ for } j < m, \end{aligned}$$

so that, as j and J increase, the ratio of these ranks tends to a constant. Figure 2 plots ρ_{mj}/ρ_{mJ} and q_{mj}/q_{mJ} against j , $j = 1, \dots, J$, for $J = 7, 25, 100$ and $m = 4, 5$.

For $J = 5, \dots, 100$, Figure 3 depicts the Pearson correlation between the Conover and Salsburg (1988) ranks ρ_{5j} with

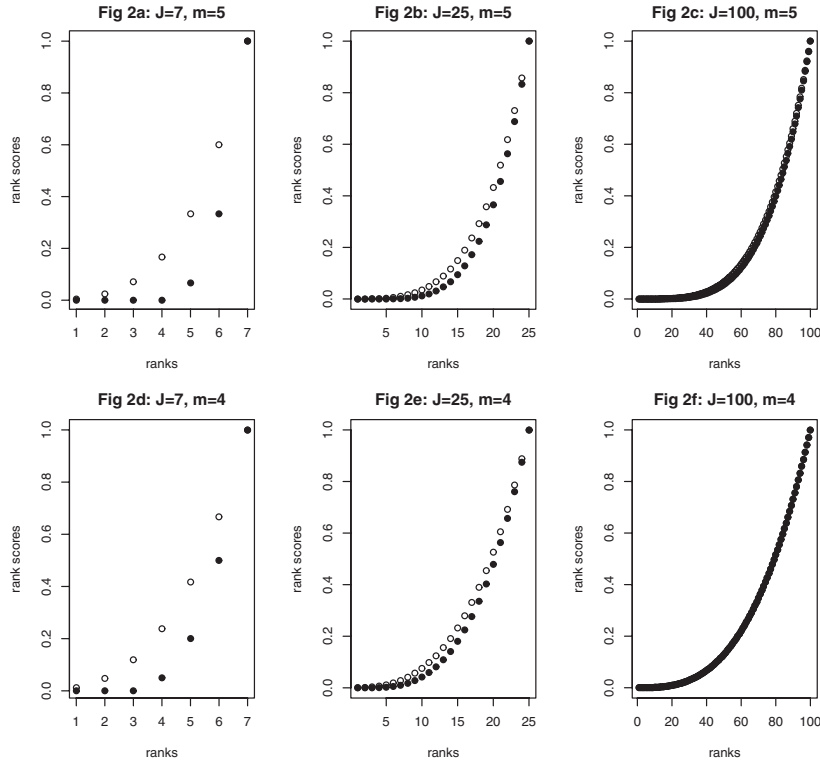


Figure 2. Conover and Salsburg locally most powerful ranks (○) and Stephenson’s subset ranks (●), scaled to have maximum rank of 1, for sample sizes $J = 7, 25, 100$, and $m = 4, 5$.

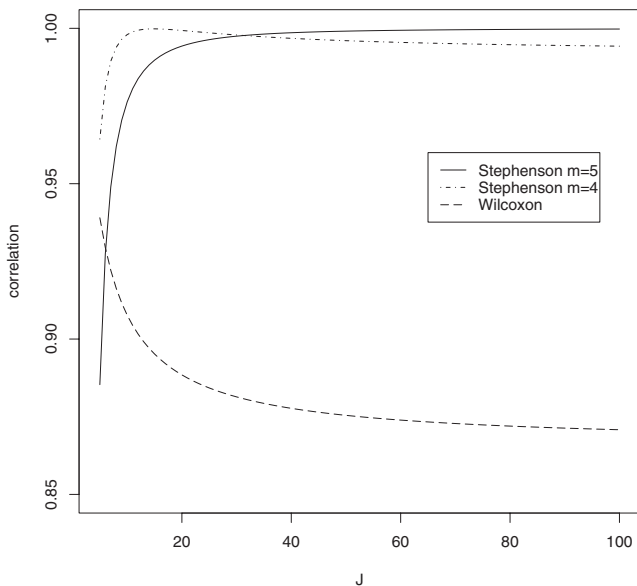


Figure 3. Pearson correlation between the Conover–Salsburg ranks, ρ_{5j} , and Wilcoxon ranks, j , and Stephenson ranks ϱ_{5j} and ϱ_{4j} , for sample sizes $J = 5, \dots, 100$.

$m = 5$ and three other ranks, namely the Wilcoxon ranks, j , the Stephenson ranks ϱ_{5j} , and ϱ_{4j} . Because ρ_{5j} , ϱ_{5j} , and ϱ_{4j} are all monotone increasing in j , the correlations cannot be extremely low. In Figure 3, the correlation between ρ_{5j} and j

is below 0.9 for $J > 12$, whereas the correlation between ρ_{5j} and ϱ_{5j} is above 0.98 for $J > 10$ and above 0.999 for $J > 47$. For small J , ρ_{5j} has a higher correlation with ϱ_{4j} than with ϱ_{5j} , with a crossover at $J = 32$. In Figure 3, the correlation between ρ_{5j} and ϱ_{4j} is 0.964 for $J = 5$, 0.981 for $J = 6$, 0.989 for $J = 7$, and above 0.99 for larger J . Not shown in Figure 3 is the correlation between ρ_{5j} and ϱ_{3j} , which is less than 0.98 for $J > 16$.

2. Peak Responses Caused by Treatment

2.1 The Highest Response Among m Subjects

There are I blocks, $i = 1, \dots, I$ and J subjects in each block, $j = 1, \dots, J$, of whom n are picked at random to receive treatment, denoted $Z_{ij} = 1$, the remaining $J - n$ to receive control, denoted $Z_{ij} = 0$, with $1 \leq n < J$, with independent assignments in different blocks. Two common cases are: (i) the completely randomized experiment with a single block, $I = 1$, and (ii) matching each individual treated subject, $n = 1$, to $J - 1$ controls. Section 1.1 describes an observational study, not an experiment, with $I = 48$, $J = 7$, $n = 1$. The j th subject in block i exhibits response r_{Tij} if assigned to treatment, $Z_{ij} = 1$, or response r_{Cij} if assigned to control, $Z_{ij} = 0$, and the effect of the treatment on this subject is a comparison of r_{Tij} and r_{Cij} (see Neyman, 1923; Rubin, 1974). The response actually observed from the j th subject in block i is $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$. Because (r_{Tij}, r_{Cij}) are not both observed for any subject ij , the effect of the treatment $r_{Tij} - r_{Cij}$ on subject ij cannot be calculated from observed data. Fisher’s (1935)

sharp null hypothesis of no treatment effect asserts $H_0: r_{Tij} = r_{Cij}$, $i = 1, \dots, I$, $j = 1, \dots, J$. Write \mathbf{Z} , \mathbf{R} , \mathbf{r}_T , and \mathbf{r}_C for the corresponding $I \times J$ matrices. It is convenient to assume that for each block i , the J potential responses to control r_{Cij} are not tied.

Write $|\mathcal{J}|$ for the number of elements of a finite set \mathcal{J} . Fix m with $2 \leq m \leq J$, and let \mathcal{S} be the set of all $\binom{J}{m}$ subsets $\mathcal{J} \subseteq \{1, 2, \dots, J\}$ such that $|\mathcal{J}| = m$. If \mathbf{A} is an $I \times J$ matrix, define for each i and j and for each $\mathcal{J} \in \mathcal{S}$

$$\begin{aligned} \lambda_{ij\mathcal{J}}(\mathbf{A}) &= 1 \text{ if } j \in \mathcal{J} \text{ and } A_{ij} = \max_{k \in \mathcal{J}} A_{ik} \\ &= 0 \text{ otherwise;} \end{aligned}$$

in other words, $\lambda_{ij\mathcal{J}}(\mathbf{A})$ indicates whether in block i , subject j had the highest A in \mathcal{J} . If the rows of \mathbf{A} have no ties, then (1) $1 = \sum_{j=1}^J \lambda_{ij\mathcal{J}}(\mathbf{A})$ for each $i = 1, \dots, I$ and each $\mathcal{J} \in \mathcal{S}$, and (2) the J values of $\sum_{\mathcal{J} \in \mathcal{S}} \lambda_{ij\mathcal{J}}(\mathbf{A})$, for $j = 1, \dots, J$, are some permutation of the J numbers $\binom{k-1}{m-1}$, $k = 1, \dots, J$, where $\binom{k-1}{m-1}$ is defined to equal 0 if $k < m$. With $m = 2$, the quantity $1 + \sum_{\mathcal{J} \in \mathcal{S}} \lambda_{ij\mathcal{J}}(\mathbf{A})$ is simply the rank of A_{ij} among A_{i1}, \dots, A_{iJ} . With $J = 7$, $m = 4$, the $J = 7$ rank scores $\sum_{\mathcal{J} \in \mathcal{S}} \lambda_{ij\mathcal{J}}(\mathbf{A})$ are 0, 0, 0, 1, 4, 10, 20, whereas with $J = 7$, $m = 5$, the $J = 7$ rank scores $\sum_{\mathcal{J} \in \mathcal{S}} \lambda_{ij\mathcal{J}}(\mathbf{A})$ are 0, 0, 0, 0, 1, 5, 15. As noted in Section 1.3, the use of $\sum_{\mathcal{J} \in \mathcal{S}} \lambda_{ij\mathcal{J}}(\mathbf{A})$ as a generalization of conventional ranks was proposed by Stephenson (1981) and Stephenson and Ghosh (1985), who also proposed $T = \sum_{i=1}^I \sum_{j=1}^J \sum_{\mathcal{J} \in \mathcal{S}} Z_{ij} \lambda_{ij\mathcal{J}}(\mathbf{R})$, with $I = 1$, as a generalization of Wilcoxon's rank sum test.

If $\lambda_{ij\mathcal{J}}(\mathbf{R}) = 1$, say that subject j in block i had a *peak response* among the m subjects k in block i with $k \in \mathcal{J}$. Conventional ranks measure peak responses among subjects taken $m = 2$ at a time, but if one is interested in dramatic responses among a small group of subjects, it may be preferable to consider peak responses among m subjects with $m > 2$. For instance, with $m = 5$, a peak response is higher than that of four other subjects.

2.2 Confidence Statements about Peak Responses Caused by Treatment

If a treated subject in a set \mathcal{J} has the highest response of subjects in set \mathcal{J} —that is, if the treated subject has the peak response in \mathcal{J} —then this might happen by the luck of the random assignment, or it might happen because the treatment caused an increase in the response of the treated subject. Under the null hypothesis of no effect; $H_0: r_{Tij} = r_{Cij}$, $i = 1, \dots, I$, $j = 1, \dots, J$, all such peak responses are just luck. How often is a peak response caused by an effect of the treatment?

The answer involves the quantity $Z_{ij}\{\lambda_{ij\mathcal{J}}(\mathbf{R}) - \lambda_{ij\mathcal{J}}(\mathbf{r}_C)\}$, which is zero for subjects assigned to control, $Z_{ij} = 0$, and is 1 or 0 or -1 for treated subjects, $Z_{ij} = 1$. If $Z_{ij}\lambda_{ij\mathcal{J}}(\mathbf{R}) = 1$, then the j th subject in block i received treatment, $Z_{ij} = 1$, and had the highest response R_{ij} among the m responses R_{ik} , $k \in \mathcal{J}$, so this treated subject had the peak response in \mathcal{J} ; otherwise, if $Z_{ij}\lambda_{ij\mathcal{J}}(\mathbf{R}) = 0$ with $Z_{ij} = 1$, this treated subject did not have the peak response in \mathcal{J} . If $Z_{ij}\lambda_{ij\mathcal{J}}(\mathbf{r}_C) = 1$, then the j th subject in block i received treatment, $Z_{ij} = 1$, and would have had the highest response had *all* m subjects in \mathcal{J} received control; otherwise, if $Z_{ij}\lambda_{ij\mathcal{J}}(\mathbf{r}_C) = 0$ with $Z_{ij} = 1$, then this treated subject would not have had the

peak response in \mathcal{J} had all m subjects received control. If $Z_{ij}\{\lambda_{ij\mathcal{J}}(\mathbf{R}) - \lambda_{ij\mathcal{J}}(\mathbf{r}_C)\} = 1$, then the j th subject in block i received treatment, $Z_{ij} = 1$, and had the highest response R_{ij} among the m responses R_{ik} , $k \in \mathcal{J}$, because of an effect of the treatment, as this subject would not have had the highest response had all m subjects received control. Similarly, if $Z_{ij}\{\lambda_{ij\mathcal{J}}(\mathbf{R}) - \lambda_{ij\mathcal{J}}(\mathbf{r}_C)\} = -1$, then the j th subject in block i received the treatment, and that subject's response R_{ij} was not the highest among R_{ik} , $k \in \mathcal{J}$, but r_{Cij} would have been the highest among the m responses, r_{Cik} , $k \in \mathcal{J}$ if all m subjects had received control. In other words, if $Z_{ij}\{\lambda_{ij\mathcal{J}}(\mathbf{R}) - \lambda_{ij\mathcal{J}}(\mathbf{r}_C)\} = -1$ then the treatment prevented the j th subject in block i from exhibiting a peak response that this subject would have exhibited in the complete absence of treatment. If $Z_{ij}\{\lambda_{ij\mathcal{J}}(\mathbf{R}) - \lambda_{ij\mathcal{J}}(\mathbf{r}_C)\} = 0$ with $Z_{ij} = 1$, then the application of the treatment to this subject did not change whether this subject had the peak response in \mathcal{J} .

How often is a peak response caused by an effect of the treatment? The answer would entail arithmetic using the $Z_{ij}\{\lambda_{ij\mathcal{J}}(\mathbf{R}) - \lambda_{ij\mathcal{J}}(\mathbf{r}_C)\}$'s if these quantities were observed. Of course, for $Z_{ij} = 1$, the value of $Z_{ij}\{\lambda_{ij\mathcal{J}}(\mathbf{R}) - \lambda_{ij\mathcal{J}}(\mathbf{r}_C)\}$ cannot be observed, because if $Z_{ij} = 1$ then $R_{ij} = r_{Tij}$, and (r_{Tij}, r_{Cij}) are never both observed. If the treatment had no effect, $r_{Tij} = r_{Cij}$, $i = 1, \dots, I$, $j = 1, \dots, J$, then $\lambda_{ij\mathcal{J}}(\mathbf{R}) - \lambda_{ij\mathcal{J}}(\mathbf{r}_C) = \lambda_{ij\mathcal{J}}(\mathbf{r}_C) - \lambda_{ij\mathcal{J}}(\mathbf{r}_C) = 0$ for all $i, j, \mathcal{J} \in \mathcal{S}$.

Consider the unobservable random variable,

$$\Lambda = \sum_{i=1}^I \sum_{j=1}^J \sum_{\mathcal{J} \in \mathcal{S}} Z_{ij}\{\lambda_{ij\mathcal{J}}(\mathbf{R}) - \lambda_{ij\mathcal{J}}(\mathbf{r}_C)\},$$

which is the number of times the treatment caused a treated subject to have the highest response in a group of m subjects minus the number of times the treatment prevented a treated subject from having the highest response in a group of m subjects. The value of Λ generally depends on the sample size, I and J , and alternative ways of scaling Λ to remove that dependence will be discussed. If the treatment has no effect, $r_{Tij} = r_{Cij}$, $i = 1, \dots, I$, $j = 1, \dots, J$, then $\Lambda = 0$.

In general, $\Lambda = T - \tilde{T}$ where T is the observed value of the statistic defined above, and $\tilde{T} = \sum_{i=1}^I \sum_{j=1}^J \sum_{\mathcal{J} \in \mathcal{S}} Z_{ij}\lambda_{ij\mathcal{J}}(\mathbf{r}_C)$ is not observed. Although \tilde{T} is not observed, its distribution in a randomized experiment is known: it is the sum of I independent random variables, where the i th random variable is formed by simple random sampling without the replacement of n of the J values $\binom{k-1}{m-1}$, $k = 1, \dots, J$. For instance, with $J = 7$, $m = 5$, $n = 1$, the ranks $\binom{k-1}{m-1}$ are 0, 0, 0, 0, 1, 5, 15, and the probability-generating function of \tilde{T} is $\{(4 + x + x^5 + x^{15})/7\}^I$, so that, for $I = 48$, $\Pr(\tilde{T} \leq 205) = 0.9501$. Similarly, with $J = 7$, $m = 4$, $n = 1$, the ranks $\binom{k-1}{m-1}$ are 0, 0, 0, 0, 1, 4, 10, 20, and the probability-generating function of \tilde{T} is $\{(3 + x + x^4 + x^{10} + x^{20})/7\}^I$, so that, for $I = 48$, $\Pr(\tilde{T} \leq 322) = 0.9506$. These exact calculations are derived by applying the fast Fourier transform to evaluate probability-generating functions, using the “convolve” function in R; see Pagano and Trichtler (1983) for related discussion.

Although Λ is an unobservable random variable, it is possible to create a confidence interval for Λ . See Weiss (1955) and Rosenbaum (2001) for a discussion of confidence intervals for a random variable.

PROPOSITION 1. *In a randomized experiment, if $\Pr\{\tilde{T} \leq \ell\} = 1 - \alpha$, then with $100(1 - \alpha)\%$ confidence, $\Lambda \geq T - \ell$.*

Proof. From $\Lambda = T - \tilde{T}$, it follows that $1 - \alpha = \Pr\{\tilde{T} \leq \ell\} = \Pr\{T - \Lambda \leq \ell\} = \Pr\{\Lambda \geq T - \ell\}$.

As noted above, the probability $\Pr\{\tilde{T} \leq \ell\}$ needed in Proposition 1 may be obtained exactly using probability-generating functions. Alternatively, an approximation may be based on the central limit theorem, $\Pr\{(\tilde{T} - \mu)/\sigma \geq v\} \rightarrow 1 - \Phi(v)$ as $I \rightarrow \infty$, where $\mu = E(\tilde{T})$ and $\sigma^2 = \text{var}(\tilde{T})$ are

$$\mu = nI\bar{q}, \quad \sigma^2 = \frac{nI(J-n)}{J(J-1)} \sum_{j=1}^J (q_j - \bar{q})^2, \quad \text{where}$$

$$q_j = \binom{j-1}{m-1} \quad \text{and} \quad \bar{q} = \frac{1}{J} \sum_{j=1}^J q_j. \tag{1}$$

With $I = 48, J = 7, n = 1, m = 5$, one has $\mu = 144, \sigma^2 = 1289.143$, so the exact $\Pr\{\tilde{T} \leq 205\} = 0.9501$ is approximated by $\Phi\{(205 - 144)/\sqrt{1289.143}\} = 0.955$.

2.3 *Scaling Λ to Aid Interpretation*

The unobservable quantity $\Lambda = T - \tilde{T}$ counts the increase in peak performance by treated subjects caused by the effects of the treatment, and because it is a count, it generally depends upon the sample size, I and J . It aids in interpretation to divide Λ by a known constant, so that the magnitude of the scaled version has a meaning that is not linked to the sample size. Two possible divisors are considered.

With the constant $E(\tilde{T}) = \mu = nI\bar{q}$ defined in (1), the quantity Λ/μ is the multiplicative increase above chance expectation in the number of times treated subjects had a peak performance in subsets of m subjects because of the effects caused by the treatment. If the treatment has no effect, then $\Lambda = 0$, so $\Lambda/\mu = 0$ also. From Proposition 1, with 95% confidence, $\Lambda/\mu \geq (T - \ell)/\mu$. Also, $(T - \mu)/\mu$ is a consistent point estimate of $\Lambda/\mu = (T - \tilde{T})/\mu$ in the sense that the difference between these two random variables converges in probability to 0 as $I \rightarrow \infty$.

An alternative divisor creates a correlation-like measure scaled to be 0 for a chance agreement and 1 for perfect agreement. The maximum possible value of T is $t_{\max} = I \sum_{j=J-n+1}^J \binom{j-1}{m-1}$. The quantity $\Lambda/(t_{\max} - \mu) = (T - \tilde{T})/(t_{\max} - \mu)$ takes the value 0 under the null hypothesis of no treatment effect, and it has expectation 1 if the treatment always raises the responses of all treated subjects above the responses of all controls in the same matched set. From Proposition 1, with 95% confidence, $\Lambda/(t_{\max} - \mu) \geq (T - \ell)/(t_{\max} - \mu)$, and $(T - \mu)/(t_{\max} - \mu)$ is a consistent point estimate of $\Lambda/(t_{\max} - \mu) = (T - \tilde{T})/(t_{\max} - \mu)$ as $I \rightarrow \infty$.

2.4 *Example, Continued: Smoking During Pregnancy and Birth Weight*

In Section 1.1, the concern is with the smallest babies, so ‘‘peak’’ refers to smallest rather than largest in a group of m babies. Recall that there were $I = 48$ matched sets consisting of $n = 1$ baby whose mother smoked during pregnancy and $J - n = 6$ babies whose mothers did not smoke during pregnancy. The current section will analyze the data in Section 1.1 as if one mother had been picked at random to smoke in each

matched set of $J = 7$. Nonrandom treatment assignment is discussed in Section 2.5.

Consider, first, $m = 5$, along the lines suggested by Salsburg (1986) and Conover and Salsburg (1988). In a matched set of $J = 7$ babies, there are $\binom{j-1}{m-1} = \binom{7-1}{5-1} = 15$ subsets consisting of the $n = 1$ baby whose mother smoked during pregnancy and four other babies. If the one exposed baby had the smallest birth weight of all seven babies, then that baby is smallest in all 15 subsets and receives a rank score of 15, but if the exposed baby had the second smallest birth weight, then that baby is smallest in the $\binom{7-2}{5-1} = 5$ of the 15 subsets that exclude the smallest baby, so the rank is 5, etc., for possible ranks 15, 5, 1, 0, 0, 0, 0. If smoking had no effect on birth weight and smoking behavior had been randomly assigned to one mother in each matched set, then by chance alone, we expect the exposed baby to be the smallest in $3 = (15 + 5 + 1 + 0 + 0 + 0 + 0)/7$ of the 15 subsets. For the $I = 48$ matched sets together, by chance alone we expect the baby whose mother smoked to be the smallest in $\mu = nI\bar{q} = 48 \times 3 = 144$ of the $t_{\max} = I \sum_{j=J-n+1}^J \binom{j-1}{m-1} = 48 \binom{7-1}{5-1} = 48 \times 15 = 720$ subsets of $m = 5$ babies. In fact, the baby whose mother smoked was the smallest not in $\mu = 144$ sets of $m = 5$ babies, but in $T = 264$ sets of $m = 5$ babies. As calculated in Section 2.3, $\Pr\{\tilde{T} \leq 205\} = 0.9501$, so from Proposition 1, we are 95% confident that in at least $\Lambda \geq T - \ell = 264 - 205 = 59$ of these subsets, smoking caused the exposed baby to have the smallest birth weight of $m = 5$ babies. The point estimate, $(T - \mu)/\mu = (264 - 144)/144 = 0.83$, suggests that smoking caused an increase of 83% in the number of peakedly small babies over what was expected by chance, but we are 95% confident of only an increase of $(264 - 205)/144 = 59/144 = 0.41$ or 41%.

Consider subsets of size $m = 2$ rather than $m = 5$. As in Section 2.1, the ranks for $m = 2$ are one less than Wilcoxon’s ranks, yielding the same significance levels and power. The stratified Wilcoxon statistic is $T + I = 186 + 48 = 234$, where $T = 186, \mu = nI\bar{q} = 48 \times (0 + 1 + \dots + 6)/7 = 144$, so that the exposed baby was the smaller baby in 186 pairwise comparisons of two babies, with 144 expected by chance. As $0.9551 = \Pr\{\tilde{T} \leq 167\}$, we are 95% confident that at least $\Lambda \geq T - \ell = 186 - 167 = 19$ of the 186 comparisons were caused by smoking. The point estimate is an increase of $(T - \mu)/\mu = (186 - 144)/144 = 0.29$ or 29% above chance, but we are 95% confident of only an increase of $(186 - 167)/144 = 19/144 = 0.13$ or 13%.

2.5 *Sensitivity Analysis for Nonrandom Treatment Assignment*

The analysis in Section 2.4 viewed smoking during pregnancy as a treatment that was randomly assigned to one mother in each matched set, but of course this is not true. A sensitivity analysis in an observational study asks how the conclusions might be altered by departures from random assignment of various magnitudes. The first such sensitivity analysis was performed by Cornfield et al. (1959), and it concerned the association between heavy smoking and lung cancer. Cornfield et al. showed that to explain away that association as not caused by smoking, and instead created by failure to match on an unobserved covariate u , one would have to postulate an enormous departure from random assignment, as measured by a parameter describing the relationship between treatment

assignment and u . A related but slightly more general approach assumes that two subjects with the same observed covariates might differ in their odds of treatment by at most a factor of $\Gamma \geq 1$ because of the failure to also match for an unobserved covariate u , and then for several values of Γ calculates the range of possible values of inference quantities, such as significance levels, point estimates, or confidence intervals (see Rosenbaum, 1987, 2002, Section 4). Because T is simply a linear rank statistic with somewhat unusual rank scores, existing methods of sensitivity analysis for linear rank statistics apply immediately. For matching with multiple controls, the large sample procedure is described in Gastwirth, Krieger, and Rosenbaum (2000) or Rosenbaum (2002, Section 4.5), where all technical details may be found. Here, the large sample sensitivity analysis will be briefly illustrated for inference about Λ in Section 2.4 for smoking and birth weight. Other methods of sensitivity analysis are discussed by Lin, Psaty, and Kronmal (1998); Robins, Rotnitzky, and Scharfstein (1999); Copas and Eguchi (2001); and Imbens (2003).

For $\Gamma = 1$, one obtains the randomization inference, as described in Section 2.4, with $m = 5$, where the null hypothesis of no effect is rejected with an approximate one-sided significance level of 0.00035. If Γ were 1.5, if the odds ratio linking smoking with an unobserved binary covariate were 1.5, then two mothers matched for observed covariates might differ by a factor of 1.5 in their odds of smoking. For $\Gamma = 1.5$, $m = 5$, the maximum possible significance level is 0.021, whereas at about $\Gamma = 5/3$, the maximum possible significance level is 0.045 and so is close to crossing the conventional 0.05 level. If the stratified Wilcoxon test with $m = 2$ is used in place of Stephenson's ranks with $m = 5$, then for $\Gamma = 1, 1.5$, and $5/3$, the upper bounds on the significance level are, respectively, 0.0012, 0.035, and 0.066; therefore, in this one example, there is slightly less sensitivity to bias from unobserved covariates using $m = 5$ than using $m = 2$.

With $I = 48$, $J = 7$, $n = 1$, $m = 5$, as in Section 2.4, and with $\Gamma = 1.5$, the largest upper tail probabilities for \tilde{T} come from an approximate normal distribution with expectation 186 and variance 1535.25, using the method in Gastwirth et al. (2000, Section 3.1) or Rosenbaum (2002, Section 4.5.2), so this approximation yields $\Pr(\tilde{T} > \ell) \leq 0.05$ with $\ell \doteq 186 + 1.65\sqrt{1535.25} = 250.65 \doteq 251$, where $0.05 \doteq 1 - \Phi(1.65)$, and $\Phi(\cdot)$ is the standard normal cumulative distribution. In parallel with Proposition 1, if Γ were 1.5, with at least 95% confidence, $\Lambda \geq T - \ell = 264 - 251 = 13$ rather than $\Lambda \geq 59$ for $\Gamma = 1$ in Section 2.4, so a bias of magnitude $\Gamma = 1.5$ could explain some of the extremely low birth weights of babies born to smokers, but not all of them. At $\Gamma = 5/3$, the continuous normal approximation only permits one to say with 95% confidence that Λ is greater than zero. The comparison in Figure 1 is insensitive to small biases, but could be produced by a moderate bias of $\Gamma = 2$, so it is much more sensitive to unobserved bias than, say, Hammond's (1964) study of smoking as a cause of lung cancer, which is insensitive to a very large bias of $\Gamma = 5$ (see Rosenbaum, 2002, Table 4.1). Moderate biases are not inconceivable in Figure 1, because a mother who smokes during pregnancy may also be less cautious in the use of other substances, such as alcohol, medications, or narcotics.

3. Paired Data and Signed Ranks

3.1 Confidence Intervals for Peak Responses

If the method in Section 2 were applied to paired data, $J = 2$, $n = 1$, the statistic T would become the sign test statistic. With paired data, Wilcoxon's signed rank statistic ranks the *absolute* values of the matched pair differences and sums the ranks of the positive differences. The signed rank test, which looks at relative magnitudes across different pairs, is more efficient than the sign test for normal data. For paired data, Stephenson (1981) used the same ranks as in Section 2, namely $\binom{i-1}{m-1}$, but applied to the i th largest *absolute* difference, rather than to the responses themselves, and summed the ranks of the positive differences. As Stephenson notes, for $m = 2$, this is nearly the same as Wilcoxon's signed rank test.

Fix an integer m , $2 \leq m \leq I$, and let \mathcal{P} be the set of all $\binom{I}{m}$ subsets \mathcal{I} of m pairs, $\mathcal{I} \subseteq \{1, 2, \dots, I\}$, $|\mathcal{I}| = m$. If \mathbf{A} is an $I \times 2$ matrix, then define for $i = 1, \dots, I$, $j = 1, 2$, and for each $\mathcal{I} \in \mathcal{P}$

$$\begin{aligned} \phi_{ij\mathcal{I}}(\mathbf{A}) &= 1 \text{ if } i \in \mathcal{I} \text{ and } A_{ij} = \max(A_{i1}, A_{i2}) \quad \text{and} \\ |A_{i1} - A_{i2}| &= \max_{k \in \mathcal{I}} |A_{k1} - A_{k2}| \\ &= 0 \text{ otherwise;} \end{aligned}$$

in other words, $\phi_{ij\mathcal{I}}(\mathbf{A})$ indicates whether in pair i , subject j had the higher response and whether the absolute difference in pair i was the largest absolute difference among the m pairs $k \in \mathcal{I}$. If \mathbf{A} has no ties of any kind, then (1) $1 = \sum_{i \in \mathcal{I}} \{\phi_{i1\mathcal{I}}(\mathbf{A}) + \phi_{i2\mathcal{I}}(\mathbf{A})\}$ for each $\mathcal{I} \in \mathcal{P}$, and (2) the I values of $\sum_{\mathcal{I} \in \mathcal{P}} \{\phi_{i1\mathcal{I}}(\mathbf{A}) + \phi_{i2\mathcal{I}}(\mathbf{A})\}$, for $i = 1, \dots, I$, are some permutation of the I numbers $\binom{k-1}{m-1}$, $k = 1, \dots, I$, where $\binom{k-1}{m-1}$ is defined to equal 0 if $k < m$. With $m = 2$, the quantity $1 + \sum_{\mathcal{I} \in \mathcal{P}} \{\phi_{i1\mathcal{I}}(\mathbf{A}) + \phi_{i2\mathcal{I}}(\mathbf{A})\}$ is simply the rank of $|A_{i1} - A_{i2}|$ among $|A_{11} - A_{12}|, \dots, |A_{I1} - A_{I2}|$. Stephenson (1981) proposed $H = \sum_{i=1}^I \sum_{j=1}^2 \sum_{\mathcal{I} \in \mathcal{P}} Z_{ij} \phi_{ij\mathcal{I}}(\mathbf{R})$, as a generalization of Wilcoxon's signed rank test, which it closely resembles for $m = 2$. Write $\tilde{H} = \sum_{i=1}^I \sum_{j=1}^2 \sum_{\mathcal{I} \in \mathcal{P}} Z_{ij} \phi_{ij\mathcal{I}}(\mathbf{r}_C)$.

The quantity $Z_{ij}\{\phi_{ij\mathcal{I}}(\mathbf{R}) - \phi_{ij\mathcal{I}}(\mathbf{r}_C)\}$ can take values 1, 0, or -1 . If it is 1, then subject ij was exposed to treatment, $Z_{ij} = 1$, and because of this, pair i had the largest absolute treated-minus-control difference in responses among the m pairs $k \in \mathcal{I}$, and the difference was positive, $Z_{ij}\phi_{ij\mathcal{I}}(\mathbf{R}) = 1$, but this would not have happened had all subjects received control, $Z_{ij}\phi_{ij\mathcal{I}}(\mathbf{r}_C) = 0$. Using the same English word from Section 2 for this slightly altered situation, $Z_{ij}\{\phi_{ij\mathcal{I}}(\mathbf{R}) - \phi_{ij\mathcal{I}}(\mathbf{r}_C)\} = 1$ means that the treatment caused ij to have a peak response. Conversely, $Z_{ij}\{\phi_{ij\mathcal{I}}(\mathbf{R}) - \phi_{ij\mathcal{I}}(\mathbf{r}_C)\} = -1$, means that the treatment prevented ij from having a peak response, and 0 means that the treatment did not change whether ij had a peak response. If the treatment has no effect, then $\phi_{ij\mathcal{I}}(\mathbf{R}) - \phi_{ij\mathcal{I}}(\mathbf{r}_C) = 0$ for all ij .

Consider the unobservable random variable

$$\Delta = \sum_{i=1}^I \sum_{j=1}^2 \sum_{\mathcal{I} \in \mathcal{P}} Z_{ij} \{\phi_{ij\mathcal{I}}(\mathbf{R}) - \phi_{ij\mathcal{I}}(\mathbf{r}_C)\} = H - \tilde{H}.$$

If the treatment has no effect, $r_{Tij} = r_{Cij}$, $i = 1, \dots, I$, $j = 1, \dots, J$, then $\Delta = 0$. In general, Δ is the net increase in peak responses among treated subjects caused by the

treatment. The proof of Proposition 2 parallels the proof of Proposition 1.

PROPOSITION 2. *In a randomized experiment, if $\Pr(\tilde{H} \leq h) = 1 - \alpha$, then with $100(1 - \alpha)\%$ confidence, $\Delta \geq H - h$.*

In a randomized experiment without ties, writing $q_i = \binom{i-1}{m-1}$, with q_i defined to be 0 for $i < m$, the probability generating function of \tilde{H} is $2^{-I} \prod_{i=1}^I \{1 + x^{q_i}\}$, the expectation and variance of \tilde{H} are $E(\tilde{H}) = (1/2) \sum_{i=1}^I q_i$ and $\text{var}(\tilde{H}) = (1/4) \sum_{i=1}^I q_i^2$, so by the central limit theorem, as $I \rightarrow \infty$, $\Pr(\tilde{H} \leq h)$ may be approximated by $\Phi[\{h - E(\tilde{H})\} / \sqrt{\text{var}(\tilde{H})}]$. In parallel with Section 2.5, the large sample sensitivity bounds in Rosenbaum (1987) for $\Pr(\tilde{H} \leq h)$ are $\Phi[\{h - \bar{\zeta}\} / \sqrt{\nu}]$ and $\Phi[\{h - \bar{\zeta}\} / \sqrt{\nu}]$ where $\bar{\zeta} = \{1/(1 + \Gamma)\} \times \sum_{i=1}^I q_i$, $\bar{\zeta} = \{\Gamma/(1 + \Gamma)\} \sum_{i=1}^I q_i$, and $\nu = \{\Gamma/(1 + \Gamma)^2\} \times \sum_{i=1}^I q_i^2$, which reduce to standard formulas when $\Gamma = 1$.

3.2 A Paired Example: Pollutants Found Attached to Human DNA

Concerned that aluminum production plant workers are exposed to polycyclic aromatic hydrocarbons that may damage DNA, Schoket et al. (1991) compared the blood of aluminum workers to that of unexposed controls recording aromatic DNA adducts. Table 1 describes 25 pairs of an aluminum worker and an unexposed control, matched for smoking and age. (Unlike all of the aluminum workers, 4 of 29 potential

controls smoked 40 or more cigarettes per day and were not matched.)

For a randomized experiment using $m = 4$, one finds $H = 11,435$, $E(\tilde{H}) = 6,325$, $\text{var}(\tilde{H}) = 3,894,303$, with approximate one-sided significance level 0.0048, so that the null hypothesis of no effect would not be plausible. In other words, there are $\binom{25}{4} = 12,650$ subsets consisting of $m = 4$ pairs, and in $H = 11,435$ of these subsets, the largest absolute treated-minus-control difference was positive, whereas this was expected by chance in only $12,650/2 = 6,325$ subsets, but we are only 95% confident that at least $\Delta \geq H - h = 1,854$ of these peak responses were actually caused by the treatment, where 1,854 is $11,435 - \{E(\tilde{H}) + \Phi(0.95)\sqrt{\text{var}(\tilde{H})}\}$. An exact calculation using the probability-generating function yields $\Pr(\tilde{H} \leq 9,585) = 0.95001$, or $\Delta \geq H - h = 11,435 - 9,585 = 1,850$, as opposed to the normal approximation $\Delta \geq 1,854$ above. That is, we observed $81\% = (11,435 - 6,325)/6,325$ more peak responses than expected by chance, but are 95% confident that at least $1,854/6,325 = 29\%$ were caused by the treatment. The results of this observational study become sensitive to unobserved bias at about $\Gamma = 2$, as the upper bound on the one-sided significance level goes from 0.047 at $\Gamma = 1.9$ to 0.053 at $\Gamma = 2$.

For $m = 2$, which is virtually the same as the signed rank statistic, there are $\binom{25}{2} = 300$ subsets of $m = 2$ pairs, and in $H = 235$ of these pairs, the larger of the two absolute differences is positive, whereas in a randomized

Table 1
DNA adducts per 10^8 nucleotides for 25 pairs of an aluminum production plant worker (W) and an unexposed control (C) matched for smoking (cigarettes/day) and age

W-ID	C-ID	W-smoke	C-smoke	W-age	C-age	W-adducts	C-adducts
8	104	0	0	26	27	1.83	0.65
31	101	0	0	33	29	1.06	1.38
5	110	0	0	34	34	0.82	1.48
40	103	0	0	43	44	1.05	1.39
10	102	0	0	45	45	0.87	1.26
18	106	0	0	45	46	0.99	0.23
24	108	0	0	46	47	1.15	0.40
23	105	0	0	49	50	0.31	0.71
13	111	0	0	51	53	1.91	1.32
4	112	0	0	54	54	1.05	1.24
45	113	5	4	28	24	2.05	1.12
20	125	10	10	31	23	1.80	2.22
15	126	10	10	49	34	0.74	1.31
11	123	15	12	36	38	3.02	1.59
7	118	20	20	32	32	1.75	0.79
33	119	20	20	33	36	6.37	1.20
2	120	20	20	36	36	0.80	0.62
16	124	20	20	37	37	4.12	2.03
43	121	20	20	39	40	2.08	1.30
26	122	20	20	43	43	2.95	1.30
38	116	20	25	49	51	2.32	1.55
12	114	20	20	50	58	1.11	2.42
27	115	30	30	37	42	3.90	1.42
29	109	30	30	38	46	2.94	2.26
39	117	30	30	43	50	0.36	0.93
Mean		11.6	11.6	40.3	40.8	1.9	1.3

Source: Schoket et al. (1991).

experiment, $E(\tilde{H}) = 150$ were expected by chance, with one-sided significance level 0.008. This is $(235 - 150)/150 = 57\%$ more than expected by chance, but we are 95% confident that at least $\Delta \geq H - h = 235 - \{150 + 1.65\sqrt{1,225}\} = 27$ or $27/150 = 18\%$ were peak responses caused by the treatment. With $m = 2$, the results of this observational study become sensitive to unobserved bias at about $\Gamma = 1.5$, as the one-sided significance level goes from 0.041 at $\Gamma = 1.4$ to 0.054 at $\Gamma = 1.5$.

4. Summary

Conover and Salsburg (1988) developed a locally most powerful rank test when only a subset of treated subjects respond strongly to treatment. Motivated by different considerations, Stephenson (1981) and Stephenson and Ghosh (1985) proposed rank tests focused on peak performance in small subsets. It was noted that these two different ranks are similar, and that Stephenson's tests can be inverted to provide confidence intervals for the number of peak performances actually caused by exposure to the treatment. Sensitivity analysis in observational studies was also discussed.

In both examples, dramatic responses were more common among treated subjects. When compared to conventional rank tests, rank tests designed with this alternative in mind yielded: (i) smaller significance levels from the randomization distribution, (ii) higher point estimates and confidence intervals formed by inverting these tests, and (iii) less sensitivity to unobserved biases in observational studies.

5. Supplementary Materials

A web-based appendix referenced in Section 1.2 is available under the Paper Information line at the *Biometrics* website, <http://www.tibs.org/biometrics>.

ACKNOWLEDGEMENT

This study was supported by a grant from the U.S. National Science Foundation.

REFERENCES

- Conover, W. J. and Salsburg, D. S. (1988). Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment. *Biometrics* **44**, 189–196.
- Copas, J. and Eguchi, S. (2001). Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society B* **63**, 871–896.
- Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959). Smoking and lung cancer. *Journal of the National Cancer Institute* **22**, 173–203.
- Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.
- Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (2000). Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society B* **62**, 545–555.
- Hammond, E. C. (1964). Smoking in relation to mortality and morbidity. *Journal of the National Cancer Institute* **32**, 1161–1188.
- Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* **93**, 126–132.
- Lehmann, E. L. (1953). The power of rank tests. *Annals of Mathematical Statistics* **24**, 23–43.
- Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54**, 948–963.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. Reprinted in *Statistical Science* **1990**, **5**, 463–480.
- Pagano, M. and Tritchler, D. (1983). On obtaining permutation distributions in polynomial time. *Journal of the American Statistical Association* **78**, 435–440.
- Robins, J. M., Rotnitzky, A., and Scharfstein, D. (1999). Sensitivity analysis unmeasured confounding in missing data and causal inferenc. In *Statistical Models in Epidemiology*, E. Halloran and D. Berry (eds), 1–94. New York: Springer.
- Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74**, 13–26.
- Rosenbaum, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika* **88**, 219–231.
- Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.
- Salsburg, D. S. (1986). Alternative hypotheses for the effects of drugs in small-scale clinical studies. *Biometrics* **42**, 671–674.
- Salsburg, D. S. (1992). *The Use of Restricted Significance Tests in Clinical Trials*. New York: Springer.
- Schocket, B., Phillips, D. H., Hower, A., and Vincze, I. (1991). ³²P-Postlabelling detection of aromatic DNA adducts in peripheral blood lymphocytes from aluminum production plant workers. *Mutation Research* **260**, 89–98.
- Stephenson, W. R. (1981). A general class of one-sample non-parametric test statistics based on subsamples. *Journal of the American Statistical Association* **76**, 960–966.
- Stephenson, W. R. and Ghosh, M. (1985). Two sample non-parametric tests based on subsamples. *Communications in Statistics* **14**, 1669–1684.
- Weiss, L. (1955). A note on confidence sets for random variables. *Annals of Mathematical Statistics* **26**, 142–144.

Received September 2006. Revised December 2006.

Accepted December 2006.