A NEW U-STATISTIC WITH SUPERIOR DESIGN SENSITIVITY IN OBSERVATIONAL STUDIES

PAUL R. ROSENBAUM

ABSTRACT. Talk is based on a paper of the same title in *Biometrics* 2011; 67:1017-1027. R software and a worked example are at: http://www-stat.wharton.upenn.edu/~rosenbap/software.htm

1. NOTATION; REVIEW

1.1. **Treatment effects and treatment assignments.** There are I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated, $Z_{ij} = 1$, the other control, $Z_{ij} = 0$, with $Z_{i1} + Z_{i2} = 1$, matched for observed covariates, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ but possibly differing in term of an unmeasured covariate, $u_{i1} \neq u_{i2}$. Following Neyman (1923) and Rubin (1973), each subject ij has two potential responses, r_{Tij} if assigned to treatment, $Z_{ij} = 1$, or r_{Cij} if assigned to control, $Z_{ij} = 0$, so the observed response from ij is $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$, and the effect of the treatment, $r_{Tij} - r_{Cij}$, on ij is not observed for any subject. Fisher's (1935) sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tij} = r_{Cij}$, for $i = 1, \ldots, I$, $j = 1, 2$. Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}),\, i = 1, \ldots, I,\, j = 1, 2\}$. If the treatment has an additive effect, $r_{Tij} - r_{Cij} = \tau$ for all ij, then the ith treated-minus-control difference in observed responses, $Y_i = (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2})$, is

$$(1.1)\quad Y_i = (Z_{i1} - Z_{i2})\,(r_{Ci1} + Z_{i1}\tau - r_{Ci2} - Z_{i2}\tau) = \tau + \epsilon_i \text{ where } \epsilon_i = (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})$$

Write Ω for the set of possible values of $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{I2})^T$, so $\mathbf{z} \in \Omega$ if $\mathbf{z} = (z_{11}, z_{12}, \ldots, z_{I2})^T$ with $z_{ij} = 0$ or $z_{ij} = 1$ and $z_{i1} + z_{i2} = 1$ for every i. Finally, write $\mathcal{Z}$ for the event $\mathbf{Z} \in \Omega$.

1.2. **General signed rank statistics testing no effect in a randomized experiment.** In a randomized paired experiment, one subject in each pair is picked at random to receive treatment, the other receiving control, with independent assignments in distinct pairs, so $\Pr(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$ for all ij, and $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 2^{-I}$ for each $\mathbf{z} \in \Omega$. If Fisher's sharp null hypothesis H_0 of no effect were true, then $Y_i = Y_{Ci} = (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})$. Let $q_i \geq 0$ be a function of the $|Y_i|$'s with the property that $q_i = 0$ if $|Y_i| = 0$. Let $\mathrm{sgn}(y) = 1$ or 0 for, respectively $y > 0$ or $y \leq 0$. A general signed rank statistic is of the form $T = \sum_{i=1}^{I} \mathrm{sgn}(Y_i)\, q_i$. Wilcoxon's signed rank statistic takes q_i equal to the rank of $|Y_i|$ when $|Y_i| > 0$. The sign test takes $q_i = 1$ when $|Y_i| > 0$. Randomization creates the null distribution $\Pr(T \mid \mathcal{F}, \mathcal{Z})$ of T. Under H_0, the absolute difference $|Y_i| = |Y_{Ci}| = |r_{Ci1} - r_{Ci2}|$ is fixed by conditioning on $\mathcal{F}$, so q_i is also fixed, and $\mathrm{sgn}(Y_i) = 1$ or 0 each with equal probability $\frac{1}{2}$ if $|Y_i| > 0$, or $\mathrm{sgn}(Y_i) = 0$ if $|Y_i| = 0$; therefore, $\Pr(T \mid \mathcal{F}, \mathcal{Z})$ is the distribution of the sum of the I independent discrete random variables $\mathrm{sgn}\{(Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2})\}\, q_i$, taking values q_i or 0 with equal probabilities, with $\mathrm{E}(T \mid \mathcal{F}, \mathcal{Z}) = \sum q_i/2$ and $\mathrm{var}(T \mid \mathcal{F}, \mathcal{Z}) = \sum q_i^2/4$.

Date: July 2012.

1.3. **Sensitivity analysis in an observational study.** A sensitivity analysis asks about the magnitude of departure from $\Pr(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$ that would need to be present to alter the qualitative conclusions of a study. A simple model for sensitivity analysis begins by assuming that in the population prior to matching, subjects have independent treatment assignments with unknown probabilities, $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$, such that two subjects, say ij and ij', with the same observed covariates, $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$, may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}\,(1-\pi_{ij'})}{\pi_{ij'}\,(1-\pi_{ij})} \leq \Gamma \quad \text{whenever } \mathbf{x}_{ij} = \mathbf{x}_{ij'}, \tag{1.2}$$

and then restricts the distribution of $\mathbf{Z}$ to Ω by conditioning on the event $\mathcal{Z}$; see Rosenbaum (1987; 2002, §4). This is the same as assuming

$$\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \frac{\exp(\gamma\,\mathbf{z}^T\mathbf{u})}{\sum_{\mathbf{b}\in\Omega}\exp(\gamma\,\mathbf{b}^T\mathbf{u})} = \prod\nolimits_{i=1}^{I}\frac{\exp(\gamma\,z_{i1}\,u_{i1} + \gamma\,z_{i2}\,u_{i2})}{\exp(\gamma\,u_{i1}) + \exp(\gamma\,u_{i2})},\ \ \mathbf{u} \in [0,1]^{2I}, \tag{1.3}$$

for $\mathbf{z} \in \Omega$, where $\gamma = \log(\Gamma) \geq 0$, so the I terms in the product in (1.3), namely $\Pr(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}) = \exp(\gamma\,u_{ij}) / \{\exp(\gamma\,u_{i1}) + \exp(\gamma\,u_{i2})\}$, are bounded below by $1/(1+\Gamma)$ and above by $\Gamma/(1+\Gamma)$; see Rosenbaum (2002, §4.2) for the easy steps demonstrating equivalence and for generalizations beyond matched pairs. For $\Gamma = 1$ and $\gamma = 0$, in (1.2) $\pi_{ij} = \pi_{ij'}$ whenever $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$ and (1.3) equals the randomization distribution, $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 2^{-I}$. Let $\overline{\overline{T}}_\Gamma$ be the sum of I independent random variables where the ith random variable takes the value q_i with probability $\Gamma/(1+\Gamma)$ and the value 0 with probability $1/(1+\Gamma)$, and let $\overline{T}_\Gamma$ be defined in the same way except with the roles of $\Gamma/(1+\Gamma)$ and $1/(1+\Gamma)$ interchanged. It is straightforward to show (Rosenbaum 1987; 2002, §4) that, under Fisher's H_0 and (1.3), the null distribution of T satisfies

$$\Pr\left(\overline{T}_\Gamma \geq k \mid \mathcal{F}, \mathcal{Z}\right) \leq \Pr(T \geq k \mid \mathcal{F}, \mathcal{Z}) \leq \Pr\left(\overline{\overline{T}}_\Gamma \geq k \,\middle|\, \mathcal{F}, \mathcal{Z}\right) \text{ for all } \mathbf{u} \in [0,1]^{2I}, \tag{1.4}$$

and the bounds are sharp, being attained for particular $\mathbf{u} \in [0,1]^{2I}$, so the bounds cannot be improved without further information about $\mathbf{u}$. Under mild conditions on the score function q_i, as $I \to \infty$, the probability $\Pr\left(\overline{\overline{T}}_\Gamma \geq k \,\middle|\, \mathcal{F}, \mathcal{Z}\right)$ may be approximated using a Normal approximation to the distribution of $\overline{\overline{T}}_\Gamma$ with $\mathrm{E}\left(\overline{\overline{T}}_\Gamma \,\middle|\, \mathcal{F}, \mathcal{Z}\right) = \frac{\Gamma}{1+\Gamma}\sum_{i=1}^{I} q_i$ and $\mathrm{var}\left(\overline{\overline{T}}_\Gamma \,\middle|\, \mathcal{F}, \mathcal{Z}\right) = \frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I} q_i^2$ with an analogous approximation for $\overline{T}_\Gamma$.

2. Power of a sensitivity analysis; design sensitivity

For each fixed $\Gamma \geq 1$, (1.4) yields an upper bound on the one-sided significance level. For fixed $\Gamma \geq 1$, the power of an α level sensitivity analysis is the probability that this upper bound will be less than or equal to α; see Rosenbaum (2004, 2010b). For $\Gamma = 1$, this is the power of a randomization test. Power is computed under some model for the generation of $\mathcal{F}$ and $\mathbf{Z}$. In the '*favorable situation*' there is a treatment effect and no bias from unmeasured covariates, and it is in this situation that we hope to report insensitivity to unmeasured bias. The power computed in the favorable situation is the probability that this hope will be realized. In the favorable situation, $\mathbf{Z}$ is randomized, $Z_{i1} - Z_{i2} = \pm 1$ with equal conditional probabilities of $\frac{1}{2}$ given $(\mathcal{F}, \mathcal{Z})$, and $\mathcal{F}$ is produced under some model for treatment effects. In the discussion here, the Y_i in (1.1) are independent and identically distributed with a distribution $G(\cdot)$ with density $g(\cdot)$; e.g., $Y_i \sim N(\tau, 1)$. Not knowing that we are in the favorable situation, we perform a sensitivity analysis that we hope will report a high degree of insensitivity when the favorable situation does arise. Symmetry about τ is one convenient alternative hypothesis, but it is not essential.

For a given test statistic and model for generating $\mathcal{F}$, there is a value $\widetilde{\Gamma}$ called the design sensitivity such that, as $I \to \infty$, the power of the sensitivity analysis tends to 1 if performed with $\Gamma < \widetilde{\Gamma}$ and to 0 if performed with $\Gamma > \widetilde{\Gamma}$. That is, in infinitely large sample sizes, this test statistic can distinguish this model for $\mathcal{F}$ from all biases smaller than $\widetilde{\Gamma}$ but not from some biases larger than $\widetilde{\Gamma}$.

Table 1. Simulated power of a one-sided 0.05 level sensitivity analysis conducted with $\Gamma = 3$, $I = 250$ pairs, and $Y_i = \tau + \epsilon_i$ where errors are standard Normal, standard logistic or t-distributed with 4 degrees of freedom.

Errors Statistic	Normal $\tau = 1/2$	Logistic $\tau = 1$	t with 4 df $\tau = 1$
Wilcoxon	0.08	0.40	0.43
(5,4,5)	0.34	0.67	**0.65**
(8,7,8)	**0.63**	**0.74**	0.57
(20,14,20)	0.53	**0.74**	**0.65**
(20,16,19)	0.52	0.69	0.61

3. A new U-statistic

Fix an integer m with $1 \le m \le I$, write $\mathcal{K}$ for the set containing the $\binom{I}{m}$ sequences $\mathcal{I} = \langle i_1, \ldots, i_m \rangle$ of m distinct integers $1 \le i_1 < \cdots < i_m \le I$, and write $\mathbf{Y}_{\mathcal{I}} = \langle Y_{i_1}, \ldots, Y_{i_m} \rangle$. A U-statistic (Hoeffding 1948) has the form

$$T = \binom{I}{m}^{-1} \sum_{\mathcal{I} \in \mathcal{K}} h\left(\mathbf{Y}_{\mathcal{I}}\right)$$

where the kernel, $h(\cdot)$, is a symmetric function of its m arguments $\langle Y_{i_1}, \ldots, Y_{i_m} \rangle$. For $\mathcal{I} = \langle i_1, \ldots, i_m \rangle \in \mathcal{K}$, sort $Y_{i_1}, \ldots, Y_{i_m}$, into increasing order by their absolute values, $0 < \left|Y_{[\mathcal{I},1]}\right| < \cdots < \left|Y_{[\mathcal{I},m]}\right|$. Fix two integers $\underline{m}$, $\overline{m}$ with $1 \le \underline{m} \le \overline{m} \le m$. In the new u-statistic, $h\left(\mathbf{Y}_{\mathcal{I}}\right)$ is the number of positive differences among $Y_{[\mathcal{I},\underline{m}]}, \ldots, Y_{[\mathcal{I},\overline{m}]}$, so $h\left(\mathbf{Y}_{\mathcal{I}}\right)$ is an integer in $\{0, 1, \ldots, \overline{m} - \underline{m} + 1\}$. If $m = \overline{m} = \underline{m} = 1$, then $h\left(\mathbf{Y}_{\mathcal{I}}\right) = \operatorname{sgn}\left(Y_{i_1}\right) = \operatorname{sgn}\left(Y_{[\mathcal{I},1]}\right)$ and T is the sign statistic, whereas if $m = \overline{m} = \underline{m} = 2$, then $h\left(\mathbf{Y}_{\mathcal{I}}\right) = \operatorname{sgn}\left(Y_{[\mathcal{I},2]}\right)$, and T is the U-statistic that closely approximates Wilcoxon's signed rank statistic. If $m = \overline{m} = \underline{m}$, then $h\left(\mathbf{Y}_{\mathcal{I}}\right) = \operatorname{sgn}\left(Y_{[\mathcal{I},m]}\right)$ and T is Stephenson's (1981) statistic which has excellent power when only a subset of treated subjects respond to treatment; see Conover and Salsburg (1988) and Rosenbaum (2007; 2010a, §16). With $m = 8$, the statistic $(m, \underline{m}, m) = (8, 7, 8)$ has $h\left(\mathbf{Y}_{\mathcal{I}}\right) = \operatorname{sgn}\left(Y_{[\mathcal{I},7]}\right) + \operatorname{sgn}\left(Y_{[\mathcal{I},8]}\right)$ with values 0, 1, 2. This U-statistic is also a signed rank statistic with $q_i = \binom{I}{m}^{-1} \sum_{\ell=\underline{m}}^{\overline{m}} \binom{a_i - 1}{\ell - 1}\binom{I - a_i}{m - \ell}$ where a_i is the rank of $|Y_i|$.

3.1. A formula for the design sensitivity.

Assume Y_i are *iid* from some distribution $G(\cdot)$ and there is no unobserved bias, $\Pr\left(Z_{ij} \mid \mathcal{F}, \mathcal{Z}\right) = \frac{1}{2}$. Let $\theta = \mathrm{E}\left\{h\left(\mathbf{Y}_{\mathcal{I}}\right)\right\}$.

Proposition: Under these assumptions, the design sensitivity of the U-statistic $(m, \underline{m}, \overline{m})$ is $\widetilde{\Gamma} = \theta / \left(\overline{m} - \underline{m} + 1 - \theta\right)$.

TABLE 2. Design sensitivities $\widetilde{\Gamma}$ with $Y_i = \tau + \epsilon_i$ where errors are standard Normal, standard logistic or t-distributed with 3 or 4 df.

Errors Statistic	Normal $\tau = 1/2$	Logistic $\tau = 1$	t with 4 df $\tau = 1$	t with 3 df $\tau = 1$
Wilcoxon	3.2	3.9	6.8	6.0
(5,4,5)	3.9	4.7	8.4	6.8
(8,7,8)	**5.1**	5.5	9.1	6.8
(20,14,20)	4.6	5.3	9.4	7.3
(20,16,19)	4.9	**5.6**	**10.1**	**7.8**

4. References

Albers, W., Bickel, P. J., van Zwet, W. R. (1976), "Asymptotic expansions for the power of distribution free tests in the one sample problem," *Ann. Stat.*, 4, 108-156.

Conover, W. J. & Salsburg, D. S. (1988), "Locally most powerful tests for treatment effects when only a subset can be expected to 'respond' to treatment," *Biometrics*, 44, 189-196.

Cornfield, J., et al. (1959), "Smoking and lung cancer," *JNCI*, 22, 173-203.

Heller, R., Rosenbaum, P. R., Small, D. S. (2009), "Split samples and design sensitivity in observational studies," *JASA*, 104, 1090-1101.

Hoeffding, W. (1948), "A class of statistics with asymptotically normal distribution," *Ann. Math. Statist.*, 19, 293-325.(Introduces U-statistics.)

Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments," *Stat. Sci.*, 5, 463-480.

Rosenbaum, P. R. (1987), "Sensitivity analysis for certain permutation inferences in matched observational studies," *Biometrika*, 74, 13-26. SA

Rosenbaum, P. R. (2002), *Observational Studies* (2^{nd} ed.). NY: Springer.

Rosenbaum, P. R. (2004), "Design sensitivity in observational studies," *Biometrika*, 91, 153-64.

Rosenbaum, P. R. (2007), "Confidence intervals for uncommon but dramatic responses to treatment," *Biometrics*, 63, 1164–1171.

Rosenbaum, P. R. (2010a), *Design of Observational Studies*, NY: Springer.

Rosenbaum, P. R. (2010b), "Design sensitivity and efficiency in observational studies," *JASA*, 105, 692-702.

Rosenbaum, P. R. (2011), "A new U-statistic with superior design sensitivity in matched observational studies," *Biometrics*, 67, 1017-1027.

Rosenbaum, P. R. (2012), "Testing one hypothesis twice in observational studies," *Biometrika*, to appear, available on-line early.

Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Ed. Psych.*, 66, 688-701.

Stephenson, W. R. (1981), "A general class of one-sample nonparametric test statistics based on subsamples," *JASA*, 76, 960-966.

DEPARTMENT OF STATISTICS, UNIVERSITY OF PENNSYLVANIA
E-mail address: rosenbaum@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~rosenbap/index.html