

# Sensitivity Analysis for Stratified Comparisons in an Observational Study of the Effect of Smoking on Homocysteine Levels

Paul R. Rosenbaum<sup>1</sup>, University of Pennsylvania

Abstract: Sensitivity bounds for randomization inferences exist in several important cases, such as matched pairs with any type of outcome, or binary outcomes with any type of stratification, but computationally feasible bounds for any outcome in any stratification are not currently available. For instance, with 20 strata, some large, others small, there is no currently available, computationally feasible sensitivity bound testing the null hypothesis of no treatment effect in the presence of a bias from nonrandom treatment assignment of a specific magnitude. The current paper solves the general problem: it uses an inequality formed by taking a one-step Taylor approximation from a near extreme solution, known as the separable approximation, where the concavity of the underlying function ensures that the Taylor approximation is, at worst, conservative. In practice, the separable approximation and the one-step movement away from it provide computationally feasible lower and upper bounds, thereby providing both a usable, perhaps slightly conservative statement, together with a check that the conservative statement is not unduly conservative. In every example that I have tried, the upper and lower bounds barely differ, although with some effort one can construct examples in which the separable approximation gives a  $P$ -value of 0.0499 and the Taylor approximation gives 0.0501. The new inequality holds in finite samples, so it strengthens certain existing asymptotic results, additionally simplifying the proof of those results. The method is discussed in the context of an observational study of the effects of smoking on homocysteine levels, a possible risk factor for several diseases including cardiovascular disease, thrombosis, and Alzheimer's disease. This study contains two evidence factors, the comparison of smokers and nonsmokers, and the comparison of smokers to one another in terms of recent nicotine exposure. A new R package, `senstrat`, implements the procedure and illustrates it with the example from the current paper. Keywords: Causal inference; covariance adjustment; observational study; randomization inference; sensitivity analysis; stratification.

## 1 Smoking as a possible cause of elevated levels of homocysteine

### 1.1 A stratified observational study

Elevated levels of plasma homocysteine are widely believed to signify increased risk of various diseases, including cardiovascular disease, thrombosis, and Alzheimer's disease; see Hankey and Eikelboom (1999), Seshadri, Beiser, Selhub et al. (2002), Wald, Law and Morris (2002), and Welch and Loscalzo (1998). Bazzano, He, Muntner, et al. (2003) suggested that cigarette smoking may cause elevated levels of homocysteine. This possibility is examined here using more recent data from the 2005-2006 NHANES; see Pimentel et al. (2016) for a different analysis of the NHANES data.

---

<sup>1</sup>Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. rosenbaum@wharton.upenn.edu. 29 October 2017.

The comparison is restricted to adults, aged at least 20 years, in the 2005-2006 NHANES, and it compares daily smokers to never smokers. Homocysteine was measured in the 2005-2006 NHANES. Daily smokers smoked every day for the last 30 days, and smoked an average of at least 10 cigarettes each day. Never smokers smoked fewer than 100 cigarettes in their lives, do not smoke now and had no tobacco use in the previous five days. The outcome is the homocysteine level in blood plasma in  $\mu\text{mol/l}$ .

Individuals were grouped into  $S = 108 = 2 \times 3 \times 3 \times 3 \times 2$  strata,  $s = 1, \dots, S$ , based on five observed covariates: (i) gender, female or male, (ii) three age categories, 20-39, 40-50,  $\geq 60$  years, (iii) three education categories,  $<$  high school, high school,  $\geq$  some College, (iv) three categories of the body-mass index (BMI),  $< 30$ ,  $[30, 35)$ ,  $\geq 35$ , and (v) Federal poverty level, namely income  $< 2 \times$ poverty,  $\geq 2 \times$ poverty. There were 2475 individuals, consisting of 512 daily smokers and 1963 never smokers. The data are available as `homocyst` in the `senstrat` package in `R`, and the examples in that package reproduce several of the analyses reported in this paper. Of these 108 strata, 18 strata with a total of 124 individuals (5% of 2475 individuals) contained only treated subjects or only controls, and these strata do not affect randomization inferences. For instance, there are three daily smokers and no controls in the stratum for men under 40 with less than high school education, a BMI between 30 and 35, with an income more than twice the poverty level. One can avoid such uninformative strata, so that all 2475 individuals contribute to the comparison, using full matching in place of stratification; see Rosenbaum (1991) and Hansen (2004). In the analysis of this example in §6, the  $S = 108$  strata are used in conjunction with a robust covariance adjustment that makes a linear correction for continuous versions of age, BMI and income.

Figure 1 shows the 108 stratum sizes,  $n_s$ , in Tukey’s stem and leaf display, a histogram in which the final digit is used as a plotting symbol. For instance, there is one stratum with  $n_s = 1$ , six strata with  $n_s = 2$ , two strata with  $n_s = 3$ , and so on, and one stratum with  $n_s = 192$ . So Figure 1 shows that the 108 strata varied in size from one individual,  $n_s = 1$ , to  $n_s = 192$  individuals, with many small strata and a few large strata. The largest stratum with  $n_s = 192$  individuals had 10 daily smokers and 182 never smokers, and consisted of women under 40 with at least some college, BMI  $< 30$ , and incomes above twice the poverty line. Compared to all 107 other strata pooled, this stratum was five times more likely to not smoke, with an odds ratio of 5.1 and a 95% confidence interval of [2.7, 11.0]. The nine largest of 108 strata contain more than a third of 2475 individuals. As discussed in §3.1, there is no existing method for computing sensitivity bounds in this situation, with both large and small strata and continuous outcomes; however, a completely general method will be developed in §4.

Figure 2 shows homocysteine levels for each of the five stratifying variables and for smoking. Because homocysteine levels have a long right tail, they are plotted on the log scale. The base-2 log is used, so that  $\log_2(y) - \log_2(x) = 1$  means  $y = 2x$  and  $\log_2(y) - \log_2(x) = k$  means  $y = 2^k x$ , and differences signify doublings. Notably, homocysteine levels are lower for women than for men, increase with age, and are higher for smokers.

0		122222233444444555566666677777888999999
1		001233344444445555666677888999
2		000222333467899
3		0244455688
4		00126
5		169
6		58
7		8
8		
9		7
10		
11		1
12		3
13		
14		
15		
16		
17		
18		
19		2

Figure 1: Stratum sizes,  $n_s$ , for  $S = 108$  strata, ranging from  $\min(n_s) = 1$  to  $\max(n_s) = 192$ . The nine largest strata contain  $849 = 192 + 123 + 111 + \dots + 56$  individuals, or more than one third of the total sample,  $N = 2475$ .

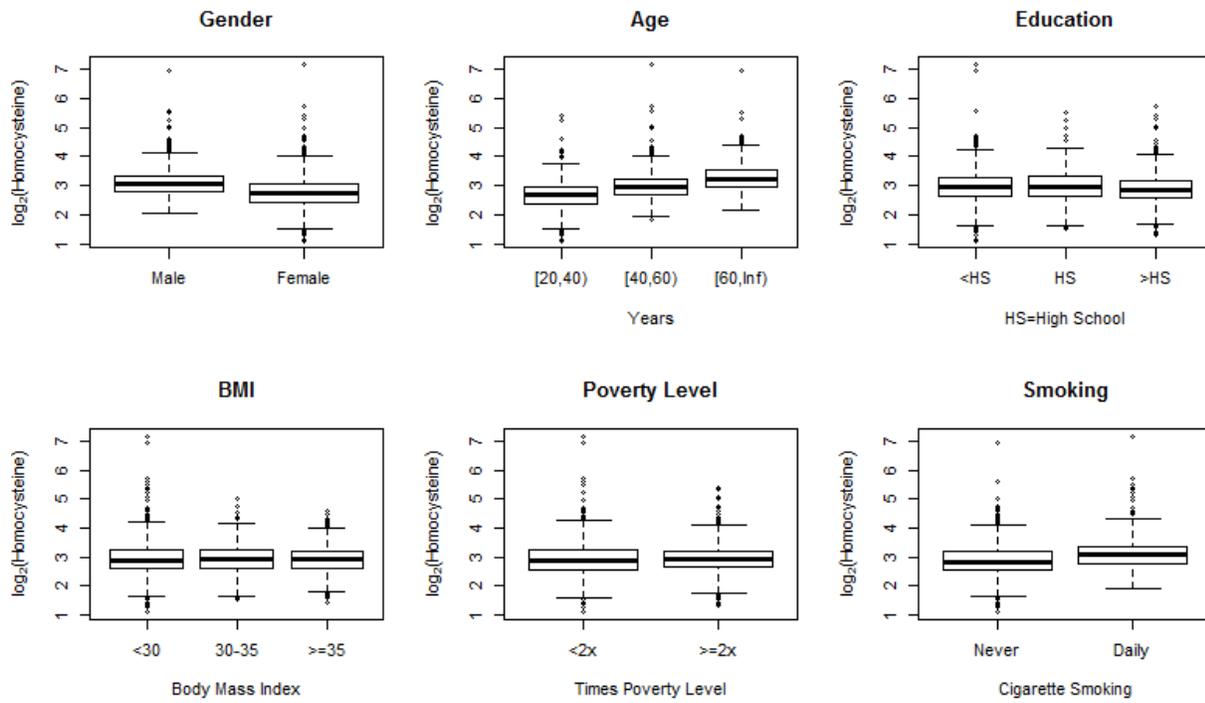


Figure 2: Plots of  $\log_2(\text{Homocysteine})$  levels for five stratifying variables and for the treatment, daily smoking of at least ten cigarettes or never smoking. A difference of  $k$  units on the  $\log_2(\cdot)$  means doubling  $k$  times, so  $\log_2(y) - \log_2(x) = 2$  if  $y = 2^2x = 4x$ .

If the Hodges and Lehmann (1962) aligned rank randomization test is used to compare logs of homocysteine levels of daily smokers and never smokers within the 108 strata, aligning using the Hodges-Lehmann estimate as suggested by the simulation of Mehrotra et al. (2010), then the one-sided  $P$ -value is  $3.3 \times 10^{-13}$ . However, this is a moderately large sample and randomization was not used to assign individuals to smoke or not, so a small  $P$ -value from a randomization test has little meaning here. How much bias in treatment assignment would need to be present to alter the naive impression from a randomization test that smoking causes an increase in homocysteine levels?

## 1.2 Outline: A Taylor correction makes a restricted method applicable in general

Notation for causal inference and sensitivity analysis is briefly reviewed in §2, and then known results and open problems about computing sensitivity bounds are discussed in §3. Briefly, computationally feasible sensitivity bounds exist for: (i) matched pairs, matched sets, and full matching with any outcome, (ii) treatment-control comparisons without strata, or with two or three strata, and any outcome, and (iii) any stratification with binary outcomes. Section 4 provides a computationally feasible general solution to the remaining open cases, that is, to any stratification with any outcome. The general method builds upon a solution, called the separable approximation, that was designed for many uniformly small but informative matched sets. Existing results do not justify use of the separable approximation in the stratification in §1.1, because in Figure 1 the largest stratum is 192 times larger than the smallest. Using a Taylor approximation to a concave function — that is, a Taylor approximation that always overestimates the function — a bound is obtained for the error of the separable approximation when used in cases in which it lacks a theoretical justification, like Figure 1. The Taylor approximation provides a (very slightly) conservative statement that may be used in all situations; that is, with a specified magnitude of bias in treatment assignment, a true null hypothesis about treatment effects is falsely rejected with probability at most  $\alpha$  when the nominal level of the test is  $\alpha$ . The procedure also provides a computable bound on how conservative the test might be. In all of the examples considered, the Taylor approximation is seen to be negligibly conservative, with the nominal level of the test very close to its actual size. Until §5, the only stratified permutation test that is considered is the Hodges-Lehmann (1962) aligned rank test. Maritz (1979) proposed exact permutation tests for matched pairs using Huber’s  $M$ -statistics, and §5 extends his method to sensitivity analyses in stratified comparisons. The example in §1.1 is then analyzed in detail in §6, using Hodges-Lehmann aligned ranks,  $M$ -statistics, and robust covariance-adjusted stratified permutation tests. The method and examples are available in an R package `senstrat` at `cran`.

### 1.3 Reading options

The main result is Proposition 1. It says: rejection of the null hypothesis of no treatment effect can or cannot be explained by a bias in treatment assignment of a certain magnitude. More precisely, Proposition 1 provides a very close upper bound on a numerical quantity that determines the outcome of a hypothesis test in the presence of a bias of a given magnitude. Unlike previous work, the method in Proposition 1 works for any outcome with any stratification; for instance, it does not require matched sets or binary outcomes. To understand Proposition 1 in all its detail, one needs the background and notation in §2, the known results and open questions in §3, the formal result in §4.1, and the technical remarks in §4.4.

Alternatively, one could initially focus on the examples and software implementation, trying the software on the examples, returning to the technical material later or not at all. The homocysteine example in §1.1 is analyzed in §4.3, §6.1 and §6.2. Additional examples are discussed in §5.2. Using the `senstrat` package in Remark 3 of §4.2, the analysis in §4.3 of the homocysteine data may be reproduced by typing: `data("homocyst"), attach(homocyst), sc<-hodgeslehmann(log2(homocysteine),z,stf,align="hl")` and `senstrat(sc,z,stf,gamma=1.95,detail=TRUE)`. Here, `hodgeslehmann(.)` computes the Hodges and Lehmann (1962) aligned ranks, and `senstrat(.)` does the sensitivity analysis at  $\Gamma = 1.95$ . The output is simpler with the default, `detail=FALSE`, and this suffices for data analysis. Two of the additional examples in §5.2 are analyzed in the examples in the documentation for the `senstrat(.)` function obtained by typing `help(senstrat)`.

## 2 Notation for randomization inference and sensitivity analysis

### 2.1 Causal inference in randomized experiments

There are  $S$  strata,  $s = 1, \dots, S$ , with  $n_s$  individuals in stratum  $s$ ,  $i = 1, \dots, n_s$ , of whom  $m_s$  received treatment, indicated by  $Z_{si} = 1$ , and  $n_s - m_s$  received control, indicated by  $Z_{si} = 0$ , so  $m_s = \sum_{i=1}^{n_s} Z_{si}$  for each  $s$ . Write  $N = \sum_{s=1}^S n_s$  for the total number of individuals, and  $\mathbf{Z} = (Z_{11}, Z_{12}, \dots, Z_{S,n_S})^T$  for the  $N$ -dimensional vector of treatment assignments. For a finite set  $\mathcal{S}$ , write  $|\mathcal{S}|$  for the number of elements of  $\mathcal{S}$ . Write  $\mathcal{Z}$  for the set containing the  $|\mathcal{Z}| = \prod_{s=1}^S \binom{n_s}{m_s}$  possible values  $\mathbf{z} = (z_{11}, \dots, z_{S,n_S})^T$  of  $\mathbf{Z}$ , so that  $z_{si} = 0$  or  $z_{si} = 1$  and  $m_s = \sum_{i=1}^{n_s} z_{si}$  for each  $s$ . Conditioning on the event  $\mathbf{Z} \in \mathcal{Z}$  is abbreviated as conditioning on  $\mathcal{Z}$ . Each stratum  $s$  is homogeneous in an observed covariate  $\mathbf{x}$ , so  $\mathbf{x}_{si} = \mathbf{x}_{si'}$  for all  $1 \leq i < i' \leq n_s$ , but individuals may differ in terms of an unmeasured covariate  $u$ , so possibly  $u_{si} \neq u_{si'}$  for many or all  $s, i, i'$ .

Each subject has two potential responses,  $r_{Tsi}$  if treated with  $Z_{si} = 1$  or  $r_{Csi}$  if control with  $Z_{si} = 0$ , so the response observed from the  $i$ th individual in stratum  $s$  is  $R_{si} = Z_{si} r_{Tsi} + (1 - Z_{si}) r_{Csi}$  and the effect caused by the treatment, namely  $r_{Tsi} - r_{Csi}$ , is not observed for any individual; see Neyman (1923) and Rubin (1973). Fisher's (1935) sharp null hypothesis of no treatment effect asserts  $H_0 : r_{Tsi} - r_{Csi} = 0, \forall s, i$ , and if  $H_0$  were

true then  $R_{si} = r_{Csi}$ . Write  $\mathbf{R} = (R_{11}, \dots, R_{S, n_S})^T$  and  $\mathbf{r}_C = (r_{C11}, \dots, r_{CS, n_S})^T$  for the  $N$ -dimensional vectors. Write  $\mathcal{F} = \{(r_{Tsi}, r_{Csi}, \mathbf{x}_{si}, u_{si}), s = 1, \dots, S, i = 1, \dots, n_s\}$ .

In a stratified randomized experiment, the treatment assignment  $\mathbf{Z}$  is picked at random from  $\mathcal{Z}$  so that  $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = |\mathcal{Z}|^{-1}$  for each  $\mathbf{z} \in \mathcal{Z}$ . If  $T = t(\mathbf{Z}, \mathbf{R})$  is a test statistic, then in a stratified randomized experiment under Fisher's null hypothesis  $H_0$ , the null distribution of  $t(\mathbf{Z}, \mathbf{R})$  is its permutation distribution,

$$\Pr\{t(\mathbf{Z}, \mathbf{R}) \geq k \mid \mathcal{F}, \mathcal{Z}\} = \Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq k \mid \mathcal{F}, \mathcal{Z}\} = \frac{|\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq k\}|}{|\mathcal{Z}|}, \quad (1)$$

because  $\mathbf{R} = \mathbf{r}_C$  if  $H_0$  is true,  $\mathbf{r}_C$  is fixed by conditioning on  $\mathcal{F}$ , and  $\mathbf{Z}$  is uniformly distributed on  $\mathcal{Z}$ . Randomization inference about the magnitude of a treatment effect involves inverting a test of Fisher's hypothesis of no effect; see Lehmann and Romano (2005, §5.12) or Rosenbaum (2002a, §5). As no new issues arise in this paper when inverting the test of  $H_0$ , it saves a considerable amount of otherwise unneeded notation if attention focuses on the test of Fisher's hypothesis  $H_0$ . Confidence intervals for a multiplicative effect of smoking on homocysteine levels — that is, for an additive effect on the log scale — are computed in §6.

## 2.2 Sensitivity analysis in stratified observational studies

In an observational study, randomization is not used to assign treatments, so there is no reason to expect the test statistic  $T = t(\mathbf{Z}, \mathbf{R})$  to have the randomization distribution (1) when  $H_0$  is true. A sensitivity analysis asks: How large would the departure from random assignment need to be to alter the qualitative conclusions reached on the basis of (1)? For instance, how much bias would need to be present to lead to acceptance of  $H_0$  at level  $\alpha$  when  $H_0$  would have been rejected at level  $\alpha$  by (1) in a randomized experiment?

A simple model for sensitivity analysis assumes that treatment assignments  $Z_{si}$  in the population are independent, and that two individuals,  $i$  and  $i'$ , with the same value of the observed covariate,  $\mathbf{x}_{si} = \mathbf{x}_{si'}$  — that is, two individuals in the same stratum  $s$  — may differ in their odds of treatment by at most a factor of  $\Gamma \geq 1$ ,

$$\frac{1}{\Gamma} \leq \frac{\Pr(Z_{si} = 1 \mid \mathcal{F}) \Pr(Z_{si'} = 0 \mid \mathcal{F})}{\Pr(Z_{si'} = 1 \mid \mathcal{F}) \Pr(Z_{si} = 0 \mid \mathcal{F})} \leq \Gamma \text{ whenever } \mathbf{x}_{si} = \mathbf{x}_{si'}, \quad (2)$$

and then returns the distribution of  $\mathbf{Z}$  to  $\mathcal{Z}$  by conditioning on  $\mathbf{Z} \in \mathcal{Z}$ . Write  $\mathcal{U} = [0, 1]^N$  for the  $N$ -dimensional unit cube. Also, write: (i)  $\mathbf{Z}_s = (Z_{s1}, \dots, Z_{s, n_s})^T$  for the treatment assignments in stratum  $s$ , (ii)  $\mathcal{Z}_s$  for the set containing the  $\binom{n_s}{m_s}$  possible values of  $\mathbf{Z}_s$ , and (iii)  $\mathbf{u}_s = (u_{s1}, \dots, u_{s, n_s})^T$  for the corresponding values of  $u_{si}$ . The set  $\mathcal{Z}$  is the direct product of the  $S$  sets  $\mathcal{Z}_s$ : that is,  $\mathbf{z} \in \mathcal{Z}$  if and only if  $\mathbf{z}^T = (\mathbf{z}_1^T, \dots, \mathbf{z}_S^T)$  with  $\mathbf{z}_s \in \mathcal{Z}_s$ , for  $s = 1, \dots, S$ . It is straightforward to show that (2) and conditioning on  $\mathbf{Z} \in \mathcal{Z}$  is

equivalent to assuming for  $\mathbf{z} \in \mathcal{Z}$ ,

$$\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \frac{\exp(\gamma \mathbf{z}^T \mathbf{u})}{\sum_{\mathbf{v} \in \mathcal{Z}} \exp(\gamma \mathbf{v}^T \mathbf{u})} = \prod_{s=1}^S \frac{\exp(\gamma \mathbf{z}_s^T \mathbf{u}_s)}{\sum_{\mathbf{v}_s \in \mathcal{Z}_s} \exp(\gamma \mathbf{v}_s^T \mathbf{u}_s)} \text{ with } \mathbf{u} \in \mathcal{U}, \quad (3)$$

where  $\gamma = \log(\Gamma) \geq 0$ . The equivalence of (2) and (3) is demonstrated by constructing  $\mathbf{u} \in \mathcal{U}$  from  $\Pr(Z_{si} = 1 \mid \mathcal{F})$  satisfying (2) and conversely; see Rosenbaum (2002a, §4.2.2). If  $\gamma = 0$  or equivalently if  $\Gamma = 1$ , then (3) becomes the randomization distribution,  $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = |\mathcal{Z}|^{-1}$  or each  $\mathbf{z} \in \mathcal{Z}$ . As  $\Gamma = e^\gamma$  increases, (3) permits progressively larger departures from randomization, and the question is: How large must  $\Gamma$  be to alter inferences obtained from (1)?

There is a close connection between (2) and omissions from the propensity score. In principle, we may define  $u_{si} = \Pr(Z_{si} = 1 \mid \mathbf{x}_{si}, r_{Tsi}, r_{Csi})$  where  $\Pr(Z_{si} = 1 \mid \mathbf{x}_{si})$  is the propensity score; then, (i)  $u_{si}$  may be ignored if treatment assignment is ignorable given  $\mathbf{x}_{si}$  in the sense that  $0 < \Pr(Z_{si} = 1 \mid \mathbf{x}_{si}, r_{Tsi}, r_{Csi}) = \Pr(Z_{si} = 1 \mid \mathbf{x}_{si}) < 1$ ; (ii) if treatment assignment is not ignorable given  $\mathbf{x}_{si}$  then it is ignorable given  $(\mathbf{x}_{si}, u_{si})$  providing  $0 < u_{si} = \Pr(Z_{si} = 1 \mid \mathbf{x}_{si}, r_{Tsi}, r_{Csi}) < 1$ . An unobserved covariate  $u_{si}$  defined in this way involves the relationship between treatment assignment  $Z_{si}$  and potential outcomes,  $(r_{Tsi}, r_{Csi})$  conditional on covariates,  $\mathbf{x}_{si}$ . For details of this view of (2), see Rosenbaum (2017, §9).

For various methods of sensitivity analysis in observational studies, see Cornfield et al. (1959), Egleston et al. (2009), Fogarty and Small (2016), Gilbert et al. (2003), Hosman et al. (2010), Liu et al. (2013), and Yu and Gastwirth (2005). In particular, Fogarty and Small (2016) use the model (3) with multiple outcomes when matching with multiple controls,  $m_s = 1$  and  $n_s \geq 2$  for each  $s$ .

### 3 Computing sensitivity bounds: known results and open problems

#### 3.1 Known results about particular situations

In principal, under  $H_0$  and (3) with one fixed  $\Gamma = e^\gamma$  and  $\mathbf{u} \in \mathcal{U}$ , we may compute the tail probability  $\Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq k \mid \mathcal{F}, \mathcal{Z}\}$  for any  $k$  by summing (3) over  $\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq k\}$ , and we would reject  $H_0$  at level  $\alpha$  for this  $(\Gamma, \mathbf{u})$  if  $\Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq k \mid \mathcal{F}, \mathcal{Z}\} \leq \alpha$  when  $k$  is replaced by the observed value of the statistic,  $t(\mathbf{Z}, \mathbf{R})$ , which equals  $t(\mathbf{Z}, \mathbf{r}_C)$  under  $H_0$ . More usefully, we would reject  $H_0$  at level  $\alpha$  in the presence of a bias of at most  $\Gamma$  if

$$\max_{\mathbf{u} \in \mathcal{U}} \Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq k \mid \mathcal{F}, \mathcal{Z}\} \leq \alpha \text{ with } k \text{ replaced by } t(\mathbf{Z}, \mathbf{R}). \quad (4)$$

If the event (4) occurs, then a bias of magnitude  $\Gamma$  is too small to explain away rejection of  $H_0$  at level  $\alpha$ .

Although the mathematical problem in (4) is well defined, as stated it is not a feasible

computation in general because  $\mathbf{u}$  is  $N$ -dimensional and  $|\mathcal{Z}|$  is enormous. In simple cases, it is possible to determine the extreme  $\mathbf{u} \in \mathcal{U}$  by a mathematical argument without computation. These simple cases include: (i) matched pairs with  $n_s = 2$  and  $m_s = 1$  for every  $s$ , with any type of outcome, and (ii) any stratum sizes  $n_s$  and  $m_s$  with binary outcomes  $(r_{Tsi}, r_{Csi})$ ; see Rosenbaum (2002a, §4.3-§4.4) and Rosenbaum and Small (2017).

Many test statistics have the form  $t(\mathbf{Z}, \mathbf{R}) = \mathbf{Z}^T \mathbf{q} = \sum_{s=1}^S \sum_{i=1}^{n_s} Z_{si} q_{si}$  where  $\mathbf{q} = (q_{11}, \dots, q_{S, n_S})^T$  is a function of  $\mathbf{R}$  and hence of  $\mathbf{r}_C$  when  $H_0$  is true. For instance, the Mantel (1963) extension statistic, the Hodges-Lehmann (1962) aligned rank statistic and the stratified Wilcoxon rank sum statistic have this form; see Lehmann (1975). Many other test statistics that are not explicitly in this form may be replaced by statistics in this form without changing the permutational  $P$ -value; see §5 for  $M$ -statistics and means. In determining the null tail probability (4), we assume  $H_0$  is true so that the  $q_{si}$  are fixed, and then, without loss of generality, we may sort individuals in stratum  $s$  by their  $q_{si}$ , so  $q_{s1} \leq \dots \leq q_{s, n_s}$ .

For a large class of test statistics,  $t(\mathbf{Z}, \mathbf{R})$ , including  $t(\mathbf{Z}, \mathbf{R}) = \mathbf{Z}^T \mathbf{q}$ , Rosenbaum and Krieger (1990) showed that the  $\mathbf{u} \in \mathcal{U}$  that provides the exact upper bound in (4) is one of several corners of the cube  $\mathcal{U} = [0, 1]^N$ ; specifically,  $u_{si} = 0$  or  $u_{si} = 1$  for every  $si$  with  $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_S)^T$ ,  $\mathbf{u}_s = (u_{s1}, \dots, u_{s, n_s})$  and  $0 = u_{s1} \leq u_{s2} \leq \dots \leq u_{s, n_s} = 1$ . In other words,  $q_{si}$  and  $u_{si}$  are ordered in the same way within each stratum  $s$ ; that is,  $(u_{si} - u_{si'})(q_{si} - q_{si'}) \geq 0$  for all  $s, i, i'$ . In the unstratified two-sample problem with  $S = 1$ , this means there are only  $N - 1 = n_1 - 1$  candidate  $\mathbf{u} \in \mathcal{U}$  that need to be checked to determine whether (4) has occurred, often a feasible task. Alas, with several or many strata,  $S > 1$ , the number of candidates is  $\prod_{s=1}^S (n_s - 1)$ , so this direct approach is feasible only for fairly small  $S$ . For instance, with  $S = 20$  and each  $n_s = 101$ , there would be  $100^{20} = 10^{40}$  candidate  $\mathbf{u}$ 's. Write  $\mathcal{U}_+ \subset \mathcal{U} = [0, 1]^N$  for the set containing these  $|\mathcal{U}_+| = \prod_{s=1}^S (n_s - 1)$  candidate values of  $\mathbf{u}$ . Then, under  $H_0$ ,

$$\max_{\mathbf{u} \in \mathcal{U}} \Pr \{ \mathbf{Z}^T \mathbf{q} \geq k \mid \mathcal{F}, \mathcal{Z} \} \leq \alpha \text{ if and only if } \max_{\mathbf{u} \in \mathcal{U}_+} \Pr \{ \mathbf{Z}^T \mathbf{q} \geq k \mid \mathcal{F}, \mathcal{Z} \} \leq \alpha. \quad (5)$$

For a specific  $\Gamma = e^\gamma \geq 1$  in (3), Rosenbaum and Krieger (1990) give simple formula for  $\mu_{s\ell} = E(\sum_{i=1}^{n_s} Z_{si} q_{si} \mid \mathcal{F}, \mathcal{Z})$  and  $\nu_{s\ell} = \text{var}(\sum_{i=1}^{n_s} Z_{si} q_{si} \mid \mathcal{F}, \mathcal{Z})$  when  $\mathbf{u}_s$  consists of  $n_s - \ell$  zeros followed by  $\ell$  ones, for  $\ell = 1, \dots, n_s - 1$ . Both  $\mu_{s\ell}$  and  $\nu_{s\ell}$  depend on  $\Gamma$ , but the notation does not indicate this explicitly. The `ev` function in the `senstrat` package in R computes the expectation  $\mu_{s\ell}$  and variance  $\nu_{s\ell}$ .

For a fixed  $\Gamma = e^\gamma$  and a fixed  $\mathbf{u} \in \mathcal{U}$ , various central limit theorems imply that  $\{\mathbf{Z}^T \mathbf{q} - E(\mathbf{Z}^T \mathbf{q} \mid \mathcal{F}, \mathcal{Z})\} / \sqrt{\text{var}(\mathbf{Z}^T \mathbf{q} \mid \mathcal{F}, \mathcal{Z})}$  converges in distribution to the standard Normal distribution, providing that  $q_{si}$ 's are not too unstable. A conventional central limit theorem lets the number of strata increase,  $S \rightarrow \infty$ , with  $1 \leq m_s < n_s \leq \tilde{n}$  for some bound  $\tilde{n}$ , exploiting the fact that  $\mathbf{Z}^T \mathbf{q}$  is the sum of  $S$  independent random variables when (3) is true; see, for instance, Gastwirth, Krieger and Rosenbaum (2000). An alternative

central limit theorem fixes the number of strata,  $S$ , and lets  $n_s \rightarrow \infty$  for each  $s$ , and in this case  $\sum_{i=1}^{n_s} Z_{si} q_{si}$  converges to a Normal distribution for each  $s$  by Theorem 2.1 of Bickel and van Zwet (1978). Although Bickel and van Zwet's Theorem 2.1 nominally provides a Normal approximation and expansion for the nonnull distribution of a randomization test, their expression (2.9) is mathematically the same as the null sensitivity distribution  $\Pr\{t(\mathbf{Z}, \mathbf{r}_C) \leq k \mid \mathcal{F}, \mathcal{Z}\}$  when  $S = 1$  for one specific  $\mathbf{u} \in \mathcal{U}$ . Using such a Normal approximation in large samples, the task of checking (5) for fixed  $\Gamma = e^\gamma$  and  $\mathbf{u} \in \mathcal{U}_+$  is replaced by a comparison involving  $k$ , the various  $\mu_{s\ell}$  and  $\nu_{s\ell}$ , and a critical constant from the Normal distribution; see (6) below.

If there are many small strata,  $S \rightarrow \infty$  with  $1 \leq m_s < n_s \leq \tilde{n}$ , then Gastwirth, Krieger and Rosenbaum (2000) proposed a very fast approximate determination of whether (5) holds that avoids consideration of  $|\mathcal{U}_+| = \prod_{s=1}^S (n_s - 1)$  candidate values of  $\mathbf{u} \in \mathcal{U}_+$ . The method picks a single  $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_S)^T \in \mathcal{U}_+$  by picking one  $\mathbf{u}_s$  at a time, for  $s = 1, \dots, S$ . The idea is to put as much mass in the upper tail by maximizing the expectation  $\mu_{s\ell}$ , and if there is a tie in doing that, then also maximizing the variance  $\nu_{s\ell}$  among  $\mathbf{u}_s$  that maximize the expectation. Saying the same thing more precisely, let  $\mathcal{J}_s \subseteq \{1, \dots, n_s - 1\}$  be the set of values  $\ell$  such that  $\mu_{s\ell} = \max_{1 \leq d \leq n_s - 1} \mu_{sd}$ , and if  $|\mathcal{J}_s| > 1$  then pick any  $\ell \in \mathcal{J}_s$  that maximizes  $\nu_{s\ell}$ , so  $\nu_{s\ell} = \max_{d \in \mathcal{J}_s} \nu_{sd}$ . This picks a single  $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_S)^T \in \mathcal{U}_+$ . The tail probability  $\Pr\{\mathbf{Z}^T \mathbf{q} \geq k \mid \mathcal{F}, \mathcal{Z}\}$  for this single  $\mathbf{u} \in \mathcal{U}_+$  need not find  $\max_{\mathbf{u} \in \mathcal{U}_+} \Pr\{\mathbf{Z}^T \mathbf{q} \geq k \mid \mathcal{F}, \mathcal{Z}\}$  in (5) but, as  $S \rightarrow \infty$  with  $1 \leq m_s < n_s \leq \tilde{n}$ , the error it makes becomes negligible. The method works because, as  $S$  increases, the expectation term  $\mu_{s\ell}$  becomes more important than the variance term,  $\nu_{s\ell}$ , a fact that will also become evident in Proposition 1 below. This method is called an asymptotically separable approximation because (i) it separates the  $N$ -dimensional optimization problem in (5) into  $S$  very simple, smaller optimization problems, and (ii) shows that as  $S \rightarrow \infty$  the solution to the  $N$ -dimensional optimization problem and the solution to much easier piecewise optimization problem differ negligibly. A by-product of the inequality in Proposition 1 will be a new, much simpler, more general proof of the performance of the asymptotically separable approximation.

### 3.2 Open problems and a general solution to them

The specialized methods described above cover many useful cases but leave substantial gaps with no computationally feasible method. As noted above, for  $S = 10$  and  $n_s = 101$  for each  $s$ , no computationally feasible method exists, because the asymptotically separable approximation may be inapplicable, as  $S = 10$  may be unlike  $S \rightarrow \infty$ . Moreover, there are many marginal situations, say  $S = 30$  strata with half of individuals in the first two strata,  $n_1 + n_2 = N/2$ , and it is unclear whether the separable approximation, which assumes many uniformly small strata,  $n_s \leq \tilde{n}$ , is adequate in these marginal cases. See also the example in §1.1 and Figure 1.

The separable approximation is always a tad liberal: its solution is never larger than,

and is typically just a little smaller than the desired  $\max_{\mathbf{u} \in \mathcal{U}_+} \Pr(\mathbf{Z}^T \mathbf{q} \geq k \mid \mathcal{F}, \mathcal{Z})$  in (5). The main new result in the current paper yields a computationally feasible and entirely general approximation that is always a tad conservative: it is never smaller than, and is typically just a little larger than the desired  $\max_{\mathbf{u} \in \mathcal{U}_+} \Pr(\mathbf{Z}^T \mathbf{q} \geq k \mid \mathcal{F}, \mathcal{Z})$  in (5). The quantity we want but cannot calculate,  $\max_{\mathbf{u} \in \mathcal{U}_+} \Pr(\mathbf{Z}^T \mathbf{q} \geq k \mid \mathcal{F}, \mathcal{Z})$ , is always sandwiched between two quantities that we can easily compute, and, in all of the examples I have examined, these two quantities differ negligibly. The new upper bound is safe to use on its own: the test achieves its nominal level, falsely rejecting  $H_0$  with probability at most  $\alpha$  in the presence of a bias of at most  $\Gamma$ ; however, it may be slightly conservative, with the size of the test below its nominal level of  $\alpha$ . The existing separable approximation can provide reassurance that the size of the test is very close to its nominal level.

## 4 A general, computationally feasible method

### 4.1 An inequality

Section 4 provides an easily computed approximation to the general sensitivity analysis in (4), illustrating its use in §4.3 for the example from §1.1.

Let  $\mathbf{a} = (a_{11}, a_{12}, \dots, a_{S, n_S})^T$  be a vector of 1's and 0's such that  $a_{s1} = 0$  and  $1 = \sum_{i=2}^{n_s} a_{si}$  for each  $s$ , so that  $(a_{s1}, \dots, a_{s, n_s})$  contains exactly one 1 and  $n_s - 1$  0's, one of which is  $a_{s1} = 0$ . Each such  $\mathbf{a}$  corresponds with exactly one  $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_S)^T \in \mathcal{U}_+$ : specifically,  $a_{s\ell} = 1$  if  $\mathbf{u}_s$  consists of  $n_s - \ell$  zeros followed by  $\ell$  ones, so that  $a_{s\ell} = 1$  for the  $\mathbf{u}_s$  that yields the expectation  $\mu_{s\ell}$  and the variance  $\nu_{s\ell}$ . Let  $\mathcal{A}$  be the set containing the  $|\mathcal{A}| = \prod_{s=1}^S (n_s - 1)$  possible values of  $\mathbf{a}$ . In words,  $a_{s\ell} = 1$  signifies use of  $\mathbf{u}_s = (0, \dots, 0, 1, \dots, 1)$  with  $\ell$  ones. For instance,  $\sum_{j=1}^{n_s} a_{sj} \mu_{sj}$  is simply the one  $\mu_{s\ell}$  for which  $a_{s\ell} = 1$ , and it is the null expectation of  $\sum_{j=1}^{n_s} q_{sj} Z_{sj}$  when  $\mathbf{u}_s = (0, \dots, 0, 1, \dots, 1)$  has  $\ell$  ones.

Let  $\kappa = \Phi^{-1}(1 - \alpha)$  be the upper  $\alpha$  critical value of the standard Normal distribution. In (5), rejection of  $H_0$  at level  $\alpha$  is insensitive to a bias in treatment assignment of  $\Gamma$  if rejection occurs for every  $\mathbf{u} \in \mathcal{U}$ , or equivalently for every  $\mathbf{u} \in \mathcal{U}_+$ , or equivalently for every  $\mathbf{a} \in \mathcal{A}$ . In large samples, condition (5) holds under  $H_0$  in the presence of a bias of at most  $\Gamma$  if

$$\min_{\mathbf{a} \in \mathcal{A}} \frac{k - \sum_{s=1}^S \sum_{\ell=1}^{n_s} a_{s\ell} \mu_{s\ell}}{\sqrt{\sum_{s=1}^S \sum_{\ell=1}^{n_s} a_{s\ell} \nu_{s\ell}}} \geq \kappa, \quad (6)$$

or equivalently if

$$\max_{\mathbf{a} \in \mathcal{A}} \lambda(\mathbf{a}) \leq 0 \text{ where } \lambda(\mathbf{a}) = \sum_{s=1}^S \sum_{j=1}^{n_s} a_{sj} \mu_{sj} - k + \kappa \sqrt{\sum_{s=1}^S \sum_{j=1}^{n_s} a_{sj} \nu_{sj}}. \quad (7)$$

Proposition 1 is a key result. In Proposition 1,  $\mathbf{b} \in \mathcal{A}$  signifies one choice of

$\mathbf{u} \in \mathcal{U}_+$  and  $\mathbf{a} \in \mathcal{A}$  signifies another choice; then, the proposition places a bound on how different  $\lambda(\mathbf{a})$  and  $\lambda(\mathbf{b})$  can be. In particular, if we knew that  $\lambda(\mathbf{b})$  is large, Proposition 1 would place a bound on how much  $\max_{\mathbf{a} \in \mathcal{A}} \lambda(\mathbf{a})$  might exceed  $\lambda(\mathbf{b})$ . The separable approximation that maximizes expectations rather than tail probabilities produces a  $\mathbf{u} \in \mathcal{U}_+$  and a corresponding  $\mathbf{b} \in \mathcal{A}$  such that  $\lambda(\mathbf{b})$  is large, even though it may fall a bit short of  $\max_{\mathbf{a} \in \mathcal{A}} \lambda(\mathbf{a})$ . Following the statement and proof of Proposition 1, in §4.2, the practical role of Proposition 1 is discussed in three remarks. Proposition 1 assumes that, at the  $\mathbf{u} \in \mathcal{U}_+$  corresponding with  $\mathbf{b} \in \mathcal{A}$ , the null variance of the test statistic,  $T = \mathbf{Z}^T \mathbf{q}$ , namely  $\text{var}(\mathbf{Z}^T \mathbf{q} \mid \mathcal{F}, \mathcal{Z}) = \sum_{s=1}^S \sum_{\ell=1}^{n_s-1} b_{s\ell} \nu_{s\ell}$  is strictly positive. This assumption about the variance will fail only in pathological situations, such as all responses being equal in every stratum  $s$ . Remark 4 in §4.4 shows that we often have reason to expect most  $\eta_s$  in (8) to be either zero or small.

**Proposition 1** *Let  $\mathbf{b} \in \mathcal{A}$  with  $\sum_{s=1}^S \sum_{\ell=1}^{n_s-1} b_{s\ell} \nu_{s\ell} = \text{var}(\mathbf{Z}^T \mathbf{q} \mid \mathcal{F}, \mathcal{Z}) > 0$ . Then*

$$\lambda(\mathbf{b}) \leq \max_{\mathbf{a} \in \mathcal{A}} \lambda(\mathbf{a}) \leq \lambda(\mathbf{b}) + \sum_{s=1}^S \eta_s, \quad (8)$$

where  $\eta_s = (\max_{1 \leq \ell \leq n_s-1} \zeta_{s\ell}) - \sum_{\ell=1}^{n_s-1} b_{s\ell} \zeta_{s\ell}$  and

$$\zeta_{s\ell} = \mu_{s\ell} + \frac{\kappa \nu_{s\ell}}{2 \sqrt{\sum_{s=1}^S \sum_{p=1}^{n_s-1} b_{sp} \nu_{sp}}}. \quad (9)$$

**Proof.** The first inequality in (8) holds trivially because  $\mathbf{b} \in \mathcal{A}$ . The proof of the second inequality in (8) will demonstrate a slightly sharper result, namely for every  $\mathbf{a} \in \mathcal{A}$ :

$$\lambda(\mathbf{a}) \leq \lambda(\mathbf{b}) + \sum_{s=1}^S \sum_{\ell=1}^{n_s-1} (a_{s\ell} - b_{s\ell}) \zeta_{s\ell} \leq \lambda(\mathbf{b}) + \sum_{s=1}^S \eta_s. \quad (10)$$

We are interested in a large but finite set of values of  $\mathbf{a} \in \mathcal{A}$ , but the function  $\lambda(\mathbf{a})$  is well-defined on the convex set  $\mathcal{D} = \left\{ \mathbf{a} \in \mathbb{R}^N : \sum_{s=1}^S \sum_{\ell=1}^{n_s-1} a_{s\ell} \nu_{s\ell} \geq 0 \right\}$ . On  $\mathcal{D}$ , the function  $\lambda(\mathbf{a})$  is concave because it is the sum of an affine function, namely  $\sum_{s=1}^S \sum_{j=1}^{n_s} a_{sj} \mu_{sj} - k$ , and a concave function of a linear function,  $\kappa \sqrt{\sum_{s=1}^S \sum_{j=1}^{n_s} a_{sj} \nu_{sj}}$ ; see Bertsekas (2009, Propositions 1.1.4 and 1.1.5) or Boyd and Vandenberghe (2004, §3.2). Moreover,  $\lambda(\mathbf{a})$  is continuously differentiable on the open convex set  $\left\{ \mathbf{a} \in \mathbb{R}^N : \sum_{s=1}^S \sum_{\ell=1}^{n_s-1} a_{s\ell} \nu_{s\ell} > 0 \right\}$ , and in particular at  $\mathbf{b}$ . Then  $\partial \lambda(\mathbf{b}) / \partial b_{s\ell} = \mu_{s\ell} + \kappa \nu_{s\ell} / \left( 2 \sqrt{\sum_{s=1}^S \sum_{p=1}^{n_s-1} b_{sp} \nu_{sp}} \right) = \zeta_{s\ell}$  in (9). Because  $\lambda(\mathbf{a})$  is concave, the first order Taylor approximation evaluated at  $\mathbf{b}$  in the direction  $\mathbf{a}$  exceeds the value of  $\lambda(\mathbf{a})$ , yielding the first inequality in (10); see Bertsekas (2009, Proposition 1.1.7) or Boyd and Vandenberghe (2004, §3.1.3). Restricting attention

to  $\mathbf{a} \in \mathcal{A}$ , the quantity  $\sum_{s=1}^S \eta_s$  is simply the maximum value of  $\sum_{s=1}^S \sum_{\ell=1}^{n_s-1} (a_{s\ell} - b_{s\ell}) \zeta_{s\ell}$  for  $\mathbf{a} \in \mathcal{A}$ , proving (10) and hence also (8). ■

## 4.2 Use of Proposition 1 in data analysis

Remarks 1 through 3 discuss use of Proposition 1 in a stratified observational study. In particular, Remark 3 discusses a package in R that performs the required calculations and contains the homocysteine example from §1.1 and §4.3.

**Remark 1: Select  $\mathbf{b}$  using the separable approximation.** Which  $\mathbf{b} \in \mathcal{A}$  should be used in (8)? If  $\Gamma = 1$ , then all  $\mathbf{a} \in \mathcal{A}$  are equivalent and there is equality throughout (8), so assume  $\Gamma > 1$ . As  $\max_{\mathbf{a} \in \mathcal{A}} \lambda(\mathbf{a})$  determines acceptance or rejection of  $H_0$  at level  $\alpha$  in the presence of a bias of at most  $\Gamma$ , we would like the interval (8) to be very short. If  $\mathbf{b}$  is picked so that  $\lambda(\mathbf{b})$  is close to  $\max_{\mathbf{a} \in \mathcal{A}} \lambda(\mathbf{a})$ , then because (10) is a Taylor approximation to  $\max_{\mathbf{a} \in \mathcal{A}} \lambda(\mathbf{a})$ , we expect (8) to be a short interval. We cannot compute  $\max_{\mathbf{a} \in \mathcal{A}} \lambda(\mathbf{a})$  directly in many cases because  $|\mathcal{A}| = \prod_{s=1}^S (n_s - 1)$  can be very large. The natural candidate for  $\mathbf{b} \in \mathcal{A}$  is the  $\mathbf{b}$  produced quickly by the separable approximation. By definition, the  $\mathbf{b}$  produced by the separable approximation has maximized the expectation,  $\sum_{s=1}^S \sum_{j=1}^{n_s} b_{sj} \mu_{sj} = \max_{\mathbf{a} \in \mathcal{A}} \sum_{s=1}^S \sum_{j=1}^{n_s} a_{sj} \mu_{sj}$ , and has maximized the variance  $\sum_{s=1}^S \sum_{j=1}^{n_s} a_{sj} \nu_{sj}$  among all  $\mathbf{a} \in \mathcal{A}$  that maximize the expectation, so it is an attempt to make  $\lambda(\mathbf{b})$  large, even though it may fall slightly short of  $\max_{\mathbf{a} \in \mathcal{A}} \lambda(\mathbf{a})$ .

**Remark 2. Testing  $H_0$  at level  $\alpha$  in the presence of a bias of at most  $\Gamma$ .** In large samples, we reject  $H_0$  at level  $\alpha$  in the presence of a bias of at most  $\Gamma > 1$  if (6) is true with the observed value of the test statistic  $T = \mathbf{q}^T \mathbf{Z}$  in place of  $k$ , or equivalently if (7) is true with the same substitution. Pick  $\mathbf{b}$  using the separable approximation, and check the (nearly trivial) condition that  $\text{var}(\mathbf{Z}^T \mathbf{q} \mid \mathcal{F}, \mathcal{Z}) > 0$  at this  $\mathbf{b}$ , that is, check that  $\sum_{s=1}^S \sum_{\ell=1}^{n_s-1} b_{s\ell} \nu_{s\ell} > 0$ . A sufficient condition for (7) — that is, a sufficient condition for rejection of  $H_0$  — is that  $\lambda(\mathbf{b}) + \sum_{s=1}^S \eta_s \leq 0$  in (10). A necessary condition for (7) is that  $\lambda(\mathbf{b}) \leq 0$  in (8). The computations required in (8) are straightforward and very fast, because they involve a single  $\mathbf{b} \in \mathcal{A}$  that corresponds with a single  $\mathbf{u} \in \mathcal{U}_+$ . More precisely, computing the interval (8) involves computing the  $\zeta_{s\ell}$ , and there are  $\sum_{s=1}^S (n_s - 1)$  of these, unlike (5) or (6) that involve  $\prod_{s=1}^S (n_s - 1)$  cases of  $\mathbf{u} \in \mathcal{U}_+$ .

**Remark 3. Software implementation.** The `senstrat` package in R implements the procedure by: (i) determining the separable approximation  $\mathbf{b}$ , (ii) calculating  $\lambda(\mathbf{b}) + \sum_{s=1}^S \eta_s$ , and if this quantity is nonpositive reporting that  $H_0$  has been rejected at level  $\alpha$  in the presence of a bias of at most  $\Gamma$ , (iii) optionally reporting that the separable approximation concurs if  $\lambda(\mathbf{b})$  and  $\lambda(\mathbf{b}) + \sum_{s=1}^S \eta_s$  have the same sign. If, as is common,  $\lambda(\mathbf{b})$  and  $\lambda(\mathbf{b}) + \sum_{s=1}^S \eta_s$  do concur, if they have the same sign, then

the slight conservatism of the inequality (10) could not have affected whether  $H_0$  was rejected. See §4.3 for numerical results in the smoking example.

### 4.3 Numerical illustration in the homocysteine example

To clarify Proposition 1, the current section briefly illustrates calculations using the smoking data, whereas §6 presents an analysis of the same data. That is, §4.3 is about Proposition 1, while §6 is about the effects of smoking on homocysteine. Consider again the 90 strata with at least one smoker and one control in §1.1. The separable approximation in §3.1 picks  $\mathbf{u} \in \mathcal{U}_+$  to maximize the null expectation of the test statistic,  $T$ , and to maximize its null variance among all  $\mathbf{u} \in \mathcal{U}_+$  that maximize its null expectation. Because expectations and variances from independent strata are additive, this maximization may be carried out one stratum at a time, with the results combined at the end — that is, this maximization problem for expectations and variances is separable — however, it is not quite the original optimization problem that we were trying to solve, namely the maximization of a tail probability. At  $\Gamma = 1.95$ , the separable approximation suggests an upper bound on the  $P$ -value from the Hodges-Lehmann aligned rank test of 0.04663. Because the separable approximation picks a  $\mathbf{u} \in \mathcal{U}_+$ , but may pick only a very bad  $\mathbf{u}$  rather than the absolute worst  $\mathbf{u} \in \mathcal{U}_+$ , its  $P$ -value bound is a tad liberal, a tad too small. Asymptotic results of Gastwirth et al. (2000) concerning the separable approximation say that it errs trivially if  $S \rightarrow \infty$  with  $n_s$  uniformly bounded, but it is unclear whether this approximation should be expected to work in §1.1 where the nine largest strata contain more than a third of 2475 individuals, and the largest stratum is 192 times larger than the smallest. The separable approximation determines the  $\mathbf{b} \in \mathcal{A}$  that will be used in Proposition 1 and, as discussed below, Proposition 1 then confirms rejection of the null hypothesis  $H_0$  of no effect at level  $\alpha = 0.05$  in the presence of a bias of at most  $\Gamma = 1.95$ . This statement from Proposition 1 is a tad conservative, unlike the separable approximation which is a tad liberal. Indeed, the approximate  $P$ -value bound from Proposition 1 is  $0.04688 > 0.04663$ , so the two bounds — the liberal and the conservative — differ trivially in this example. Both the separable approximation and Proposition 1 reject  $H_0$  at level  $\alpha = 0.05$  for  $\Gamma = 1.95$ , whereas they both accept  $H_0$  for  $\Gamma = 1.96$ . At  $\Gamma = 1.9578$ , the two approximations disagree about rejection at  $\alpha = 0.05$ , with the separable approximation quoting a slightly liberal  $P$ -value bound of 0.04990 and Proposition 1 quoting 0.05016; however, in most contexts, a difference of this magnitude is too small to be of practical concern. The practical point is that we lack any theoretical justification for using the separable approximation in an example like §1.1, but Proposition 1 provides an easily computed conservative bound on its error in any stratification, so we can always use the conservative statement with confidence, and we can often confirm that the separable approximation erred negligibly and the conservative statement is trivially conservative using  $\lambda(\mathbf{b})$ .

The remainder of this section performs calculations at  $\Gamma = 1.95$  using Hodges and Lehmann’s aligned ranks and is intended to provide a sense for how (10) and (10) work in

practice. Also,  $\mathbf{b} \in \mathcal{A}$  was picked by the separable approximation. So,  $\mathbf{b} \in \mathcal{A}$  is nearly the worst  $\mathbf{a} \in \mathcal{A}$  and the conservative Taylor approximation in Proposition 1 is seeking something a tad worse than  $\mathbf{b} \in \mathcal{A}$ , which in fact it finds. Finally, the test is at  $\alpha = 0.05$ , so  $\kappa = \Phi^{-1}(1 - \alpha) = 1.6449$ .

For 68 of the 90 informative strata,  $\eta_s = 0$  in (10), thereby making no adjustment to the separable  $\mathbf{b} \in \mathcal{A}$ . In these 68 strata, the separable approximation and the Taylor correction agree, picking the same  $\mathbf{u}_s$  as the worst. For instance, the largest stratum with  $n_s = 192$ , discussed in §1.1, had  $\eta_s = 0$ , requiring no correction. Also,  $\lambda(\mathbf{b}) = -452.54 < 0$ , suggesting rejection at  $\alpha = 0.05$ , while  $\lambda(\mathbf{b}) + \sum_{s=1}^S \eta_s = -419.92 < 0$  thereby confirming rejection at  $\alpha = 0.05$ . Strictly speaking, Proposition 1 makes a formal statement about rejection or acceptance of  $H_0$  at a fixed level  $\alpha$ , here the conventional  $\alpha = 0.05$ , and one would have to repeat the calculations for various  $\alpha$  to obtain a  $P$ -value. Typically, however, a sensitivity analysis fixes  $\alpha$ , perhaps  $\alpha = 0.05$ , and then asks: What is the largest bias in treatment assignment,  $\Gamma$ , that would lead to rejection at level  $\alpha$ ? In the example,  $H_0$  is rejected at level  $\alpha = 0.05$  for  $\Gamma = 1.95$  but not for  $\Gamma = 1.96$ . As noted above, at  $\Gamma = 1.9578$ , the two ends of the interval (8) do not concur, but the difference is trivially small. Whenever  $\sum_{s=1}^S \eta_s > 0$  in (10), the Taylor approximation in (10) has found an  $\mathbf{a} \in \mathcal{A}$  that it judges worse than the separable approximation  $\mathbf{b} \in \mathcal{A}$ , yielding a smaller minimum deviate in (6), and the **R** package informally reports a  $P$ -value at this  $\mathbf{a} \in \mathcal{A}$ , even though the formal content of Proposition 1 is confined to acceptance or rejection at a fixed  $\alpha$ .

#### 4.4 Understanding technical aspects of Proposition 1

Unlike Remarks 1-3 in §4.2, Remarks 4 through 6 in the current section discuss technical implications of Proposition 1. In particular, although it is possible that  $\eta_s > 0$  in (8), we often see that  $\eta_s = 0$  for many or all strata  $s$ . Remark 4 discusses why this happens, while Remarks 5 and 6 discuss some consequences.

**Remark 4. Behavior of the Taylor formula at the separable approximation.** In examples, with  $\mathbf{b}$  picked by the separable approximation, many  $\eta_s$  in (10) are not just small but are actually zero. Why does this happen? For  $\Gamma > 1$ , consider (10) with  $\mathbf{b}$  produced by the separable approximation, so  $\mathbf{b}$  maximizes the expectation  $\mu_{sl}$  in each stratum  $s$ , and among all  $\mathbf{a}$  that achieve this maximum expectation,  $\mathbf{b}$  maximizes the variance  $\nu_{sl}$ ; that is, for all  $\mathbf{a} \in \mathcal{A}$ ,

$$\sum_{\ell=1}^{n_s-1} (a_{s\ell} - b_{s\ell}) \mu_{s\ell} \leq 0 \text{ with equality if and only if } \sum_{\ell=1}^{n_s-1} (a_{s\ell} - b_{s\ell}) \nu_{s\ell} \leq 0. \quad (11)$$

Recalling formula (9), it follows that in (10),  $\sum_{\ell=1}^{n_s-1} (a_{s\ell} - b_{s\ell}) \zeta_{s\ell} > 0$  if and only if

$$\sum_{\ell=1}^{n_s-1} (a_{s\ell} - b_{s\ell}) \mu_{s\ell} < 0 \text{ and } \sum_{\ell=1}^{n_s-1} (a_{s\ell} - b_{s\ell}) (\mu_{s\ell} + h_s \nu_{s\ell}) > 0 \quad (12)$$

where

$$h_s = \frac{\kappa}{2\sqrt{\sum_{s=1}^S \sum_{\ell=1}^{n_s-1} b_{s\ell} \nu_{s\ell}}}.$$

In finite samples condition (12) can occur. Suppose, however, that the total sample size increases while leaving stratum  $s$  unchanged, so that  $\sum_{s=1}^S \sum_{\ell=1}^{n_s-1} b_{s\ell} \nu_{s\ell} = \text{var}(\mathbf{Z}^T \mathbf{q} \mid \mathcal{F}, \mathcal{Z}) \rightarrow \infty$ ; then,  $h_s \rightarrow 0$  and eventually condition (12) cannot occur, so that eventually  $\eta_s = 0$ . Indeed, as seen in §4.3, 68 of the 90 nondegenerate strata  $s$  in the smoking example in §1.1 had  $\eta_s = 0$  in (10).

**Remark 5. An alternative justification for the separable approximation.** Gastwirth

et al. (2000) proposed the separable approximation  $\mathbf{b} \in \mathcal{A}$  as an asymptotic approximation for the situation with a growing number of uniformly small but informative strata,  $S \rightarrow \infty$  with  $1 \leq m_s < n_s \leq \tilde{n}$  and  $\text{var}(\mathbf{Z}^T \mathbf{q} \mid \mathcal{F}, \mathcal{Z}) = \sum_{s=1}^S \sum_{\ell=1}^{n_s-1} b_{s\ell} \nu_{s\ell} \rightarrow \infty$ . In light of Remark 4, Proposition 1 is an alternative proof of the result in Gastwirth et al. (2000) that the separable approximation errs negligibly with many uniformly small but informative strata. Indeed, Proposition 1 is a simpler proof under weaker conditions. In §5.2, an example is considered with  $S = 397$  matched sets in which one treated individual is matched to two controls,  $n_s = 3$  and  $m_s = 1$ , and in this example,  $\eta_s = 0$  for  $s = 1, \dots, S = 397$ , so that there is equality throughout (8). To repeat Remark 4,  $\eta_s > 0$  is the same as condition (12), so  $\eta_s > 0$  eventually becomes impossible if stratum  $s$  remains unchanged while  $\sum_{s=1}^S \sum_{\ell=1}^{n_s-1} b_{s\ell} \nu_{s\ell} \rightarrow \infty$ .

**Remark 6: Uniformly small strata are not needed.**

Remark 5 observed that each  $\eta_s$  in (8) is driven to zero with many uniformly small strata providing  $\text{var}(\mathbf{Z}^T \mathbf{q} \mid \mathcal{F}, \mathcal{Z}) = \sum_{s=1}^S \sum_{\ell=1}^{n_s-1} b_{s\ell} \nu_{s\ell} \rightarrow \infty$ . Actually,  $\eta_s$  is driven toward zero much more generally. For instance, the stratum sizes in the smoking study in Figure 1 are quite varied, not uniformly small, yet as discussed in §4.3,  $\eta_s = 0$  for the stratum in Figure 1 with  $n_s = 192$ . In another illustration in §5.2 with just  $S = 2$  strata, (8) holds as an equality with  $\eta_1 = \eta_2 = 0$ . Although  $\eta_s > 0$  does occur in many instances, so (8) is not an equality in general, and although the Taylor bound on the right of (8) can be conservative in the sense that there can be strict inequality,  $\max_{\mathbf{a} \in \mathcal{A}} \lambda(\mathbf{a}) < \lambda(\mathbf{b}) + \sum_{s=1}^S \eta_s$ , nonetheless the degree of conservatism has been trivially small in every example I have tried. In any actual study, an investigator can check whether the degree of conservatism is trivially small by computing both  $\lambda(\mathbf{b})$  and  $\lambda(\mathbf{b}) + \sum_{s=1}^S \eta_s$  when finding the largest  $\Gamma$  that leads to rejection at level  $\alpha$ . See §4.3 and §5.2 for several examples.

## 5 Stratified permutation tests using $M$ -statistics

### 5.1 Expressing $M$ -statistics as a sum of scores, $T = \mathbf{Z}^T \mathbf{q}$

Maritz (1979) proposed exact permutation tests of  $H_0$  based on Huber's (1981)  $M$ -statistics for matched pairs, and there is a straightforward extension to matching with multiple controls; see Rosenbaum (2014). The current section briefly extends this method to stratified permutation inferences. In the current section,  $H_0$  is assumed to be true for the purpose of testing it, so  $R_{si} = r_{Csi}$ .

Let  $\varsigma$  be a quantile, typically the median, of the  $\sum_{s=1}^S \binom{n_s}{2}$  pairwise absolute differences  $|r_{Csi} - r_{Csi'}|$ ,  $1 \leq i < i' \leq n_s$ , within the  $S$  strata. Under  $H_0$ , this scale factor  $\varsigma$  is fixed by conditioning on  $\mathcal{F}$  in (3), not changing with the treatment assignment  $\mathbf{Z}$ . Define the  $M$ -statistic to be  $T = \sum_{s=1}^S w_s \sum_{i=1}^{n_s} \sum_{i'=1}^{n_s} Z_{si} (1 - Z_{si'}) \psi \{(r_{Csi} - r_{Csi'}) / \varsigma\}$ , where  $w_s \geq 0$  is a weight for stratum  $s$  and  $\psi(\cdot)$  is a monotone increasing odd function,  $\psi(y) = -\psi(-y)$ , so that, in particular,  $\psi(0) = 0$ . Huber (1981) favored  $\psi_{\text{hu}}(y) = \max\{-\varkappa, \min(y, \varkappa)\}$  for some  $\varkappa > 0$ , which is analogous to a trimmed mean, but  $\psi_t(y) = y$  is analogous to a mean. In the examples,  $\varkappa = 3$ . The median of  $|y|$  is 0.674 if  $y$  is standard Normal, so  $\psi_{\text{hu}}(y/0.674)$  trims a standard Normal  $y$  at about  $y = \pm 2.02$  for  $\varkappa = 3$ . The weight  $w_s$  for stratum  $s$  is, under  $H_0$ , a function of the  $(n_s, m_s)$  and  $\mathcal{F}$ , so it is fixed by conditioning on  $(\mathcal{F}, \mathcal{Z})$  in (3).

Within each stratum  $s$ , the statistic  $T$  compares every treated subject,  $Z_{si} = 1$  to every control,  $(1 - Z_{si'}) = 1$ . It will now be shown that  $T$  equals a sum of fixed scores  $\mathbf{q}$  for treated individuals,  $T = \mathbf{Z}^T \mathbf{q}$ , for suitable  $\mathbf{q}$ . Because  $\psi(\cdot)$  is odd,  $\psi \{(r_{Csi} - r_{Csi'}) / \varsigma\} = -\psi \{(r_{Csi'} - r_{Csi}) / \varsigma\}$  and  $0 = \psi \{(r_{Csi} - r_{Csi}) / \varsigma\}$ . It follows that for every  $\mathbf{Z} \in \mathcal{Z}$ , we have  $0 = \sum_{i=1}^{n_s} \sum_{i'=1}^{n_s} Z_{si} Z_{si'} \psi \{(r_{Csi} - r_{Csi'}) / \varsigma\}$ , because if  $Z_{si} Z_{si'} = 1$  then  $\psi \{(r_{Csi} - r_{Csi'}) / \varsigma\}$  and  $\psi \{(r_{Csi'} - r_{Csi}) / \varsigma\}$  both appear once in this sum and they cancel. So,  $T = \sum_{s=1}^S \sum_{i=1}^{n_s} Z_{si} q_{si}$  where

$$q_{si} = w_s \sum_{i' \in \{1, \dots, i-1, i+1, \dots, n_s\}} \psi \{(r_{Csi} - r_{Csi'}) / \varsigma\}. \quad (13)$$

Also,  $0 = \sum_{i=1}^{n_s} q_{si}$  for each  $s$ , because  $\psi \{(r_{Csi} - r_{Csi'}) / \varsigma\}$  and  $\psi \{(r_{Csi'} - r_{Csi}) / \varsigma\}$  both appear once in  $\sum_{i=1}^{n_s} q_{si}$  and they cancel. Under  $H_0$ , the score  $q_{si}$  in (13) is a function of  $\mathcal{F}$  and hence is fixed by conditioning on  $\mathcal{F}$  in (2). So  $T = \sum_{s=1}^S \sum_{i=1}^{n_s} Z_{si} q_{si}$  is a statistic of the general form discussed in §3.1 despite initially appearing to have a different form. If  $H_0$  were true, if  $r_{Csi} = a_{si} - \beta_s$  for arbitrary block parameters  $\beta_s$ , then  $q_{si}$  in (13) would be unchanged by changes in the  $\beta_s$ , so in this specific respect the scores in (13) resemble Hodges and Lehmann's (1962) aligned ranks.

The score  $q_{si}$  in (13) compares individual  $i$  and the other  $n_s - 1$  individuals in stratum  $s$  measured by  $\psi \{(r_{Csi} - r_{Csi'}) / \varsigma\}$ . If  $w_s = (n_s - 1)^{-1}$ , then  $q_{si}$  in (13) is an average of  $\psi \{(r_{Csi} - r_{Csi'}) / \varsigma\}$ , and these weights are used in the examples. For varied objectives and methods for weighting strata, see Hodges and Lehmann (1962), Puri (1965), Mehrotra et al. (2010) and Rosenbaum (2014).

## 5.2 Further numerical examples

Three additional numerical examples shed light on the method in Remark 2. For more about these examples, see the documentation in the `senstrat` package in R. In two of the examples, the extreme  $\mathbf{u} \in \mathcal{U}$  can be determined theoretically. Does the method in Remark 2 find this known extreme  $\mathbf{u} \in \mathcal{U}$ ? The third example satisfies the conditions required for use of the separable approximation, that is, many small strata. Is the method in Remark 2 unduly conservative when the separable approximation is adequate? In each example, the value of  $\Gamma$  is set to a round number such that  $H_0$  is barely rejected at level  $\alpha = 0.05$  in the presence of a bias of at most  $\Gamma$ .

The first example is from Werfel et al. (1988) and is contained in the `sensitivitymw` package in R. There are  $S = 39$  matched pairs,  $n_s = 2$  and  $m_s = 1$ , of a welder and a control and the outcome is a measure of DNA damage. In this case, using  $M$ -scores, at  $\Gamma = 3$  and  $\alpha = 0.05$ , both the separable approximation and the Taylor correction method in Remark 2 yield exactly the same answer as the extreme  $\mathbf{u} \in \mathcal{U}$  determined theoretically. Indeed,  $\eta_s = 0$  for every  $s$  and there is equality in (8) essentially because both  $\mathcal{U}_+$  and  $\mathcal{A}$  contain only one vector, and this will be true in every case involving matched pairs.

The second example is a  $2 \times 2 \times 2$  table from Satagopan et al. (2001) linking BRCA mutations and breast cancer, adjusting for  $S = 2$  age strata, as reanalyzed by Rosenbaum and Small (2017) and in the `sensitivity2x2xk` package in R. Under  $H_0$ ,  $R_{si} = r_{Csi}$  and the stratified permutation statistic is the Mantel-Haenszel statistic,  $T = \sum_{s=1}^S \sum_{i=1}^{n_s} Z_{si} r_{Csi}$  where  $r_{Csi} = 0$  or  $r_{Csi} = 1$ . In this  $2 \times 2 \times 2$  table,  $\mathcal{U}_+$  and  $\mathcal{A}$  each contain  $(n_1 - 1) \times (n_2 - 1) = (753 - 1) \times (3411 - 1) = 2,564,320$  candidate  $\mathbf{u} \in \mathcal{U}_+$  or  $\mathbf{a} \in \mathcal{A}$  to consider. With binary responses, a theoretical argument identifies the one extreme  $\mathbf{u} \in \mathcal{U}_+$  as  $u_{si} = r_{Csi}$ ; see Rosenbaum (1995, §5; 2002, §4.4.1). Not knowing this theoretical argument, at  $\Gamma = 7$  and  $\alpha = 0.05$ , both the separable approximation and the method in Remark 2 identify this extreme  $\mathbf{u}$ , again yielding equality in (8) with  $\eta_1 = \eta_2 = 0$ . Importantly, neither the separable approximation nor the method in Remark 2 considered the 2,564,320 candidate  $\mathbf{u} \in \mathcal{U}_+$ .

The third example from Rosenbaum (2014) uses NHANES data to compare methylmercury in the blood of each of 397 heavy consumers of fish to two matched controls who consume little fish, so  $n_s = 3$  and  $m_s = 1$  for  $s = 1, \dots, S = 397$ . The separable approximation of Gastwirth et al. (2000) has good asymptotic properties in a situation like this, as there are many small strata. In this example,  $\mathcal{U}_+$  and  $\mathcal{A}$  contain  $\prod_{s=1}^S (n_s - 1) = 2^{397} = 3.2 \times 10^{119}$  candidate  $\mathbf{u} \in \mathcal{U}_+$  or  $\mathbf{a} \in \mathcal{A}$ . Without considering these candidates, at  $\Gamma = 15$  and  $\alpha = 0.05$ , the separable approximation and the method in Remark 2 find the extreme  $\mathbf{u}$  with  $\lambda(\mathbf{b}) = \max_{\mathbf{a} \in \mathcal{A}} \lambda(\mathbf{a}) = \lambda(\mathbf{b}) + \sum_{s=1}^S \eta_s$  with  $\eta_s = 0$  for every  $s$ .

In brief, in these three examples, there was no gap between the slightly liberal separable approximation,  $\lambda(\mathbf{b})$ , and the slightly conservative Taylor adjustment,  $\lambda(\mathbf{b}) + \sum_{s=1}^S \eta_s$ . In the first two examples, the unique extreme  $\mathbf{u} \in \mathcal{U}_+$  is known from theory, and both

the separable approximation and the Taylor adjustment found that known extreme  $\mathbf{u} \in \mathcal{U}_+$ . Because the  $\prod_{s=1}^S (n_s - 1)$  candidates were not considered, the computations were extremely fast.

## 6 Effects of smoking on homocysteine levels

### 6.1 Sensitivity analysis for the effects of smoking on homocysteine levels

Using Hodges-Lehmann (1962) aligned ranks, the null hypothesis  $H_0$  of no effect of smoking on the  $\log_2$  of homocysteine levels is rejected at level  $3.3 \times 10^{-13}$  in a one-sided, stratified randomization test that assumes no bias in treatment assignment,  $\Gamma = 1$ , and the 95% one-sided confidence interval for a multiplicative effect  $\beta$  is  $\beta \geq 1.086$ , or an 8.6% increase in homocysteine from smoking. Rejection of  $H_0$  at one-tailed level  $\alpha = 0.05$  is insensitive to a bias of  $\Gamma = 1.95$ , but not  $\Gamma = 1.96$ . In a matched pair, a bias of  $\Gamma = 1.95$  is equivalent to a covariate that increased the odds of smoking by 4-fold and increased the odds of a positive pair difference in homocysteine levels by 3.3-fold; see Rosenbaum (2017a, Table 9.1) or the `amplify` function in the R package `sensitivitymult`. Using the  $M$ -statistic in §5 instead, rejection of  $H_0$  is insensitive to a bias of  $\Gamma = 2.1$  but not  $\Gamma = 2.2$ . The  $M$ -scores and the Hodges-Lehmann aligned ranks have correlation 0.97 in this example, but the  $M$ -scores do a little more to limit the influence of the most extreme values.

Instead of permuting the responses, a randomization test may permute residuals from a robust regression of the response,  $R_{si}$ , on covariates,  $\mathbf{x}_{si}$ , providing the regression does not include the treatment  $Z_{si}$ ; see Rosenbaum (2002b). Figure 3 resembles Figure 2, using the same covariates that define the 108 strata, but Figure 3 depicts residuals of  $\log_2$  homocysteine from a robust regression on: (i) the binary indicator of gender, (ii) age in years as a continuous variable, (iii) education as a five-level integer score, (iv) BMI as a continuous variable, and (v) the ratio of income to the poverty level as a continuous variable. The robust regression used  $M$ -estimation as implemented in the `r1m` function of the `MASS` package in R, using the default settings. Notably, in comparison with Figure 2, in Figure 3 homocysteine no longer appears associated with gender or age, but continues to be associated with smoking. The usual multiple correlation coefficient in a least squares regression is the Pearson correlation between the outcome and the predicted values. The Spearman and Pearson correlations between  $\log_2$  homocysteine levels and their predicted values in the robust regression are 0.612 and 0.572, respectively, with squares 0.37 and 0.33. Although coarser than the robust linear adjustment, the 108 strata attend to interactions and nonlinearities involving the covariates.

Applied to the residuals in Figure 3, the stratified Hodges-Lehmann aligned rank test rejects  $H_0$  at level  $\alpha = 0.05$  in the presence of a bias of  $\Gamma = 2.1$  but not  $\Gamma = 2.2$ . The analogous test on residuals using  $M$ -scores rejects  $H_0$  in the presence of a bias of  $\Gamma = 2.35$ . In a matched pair, a bias of  $\Gamma = 2.35$  is equivalent to a covariate that increased the odds of smoking by 4-fold and increased the odds of a positive pair difference in homocysteine

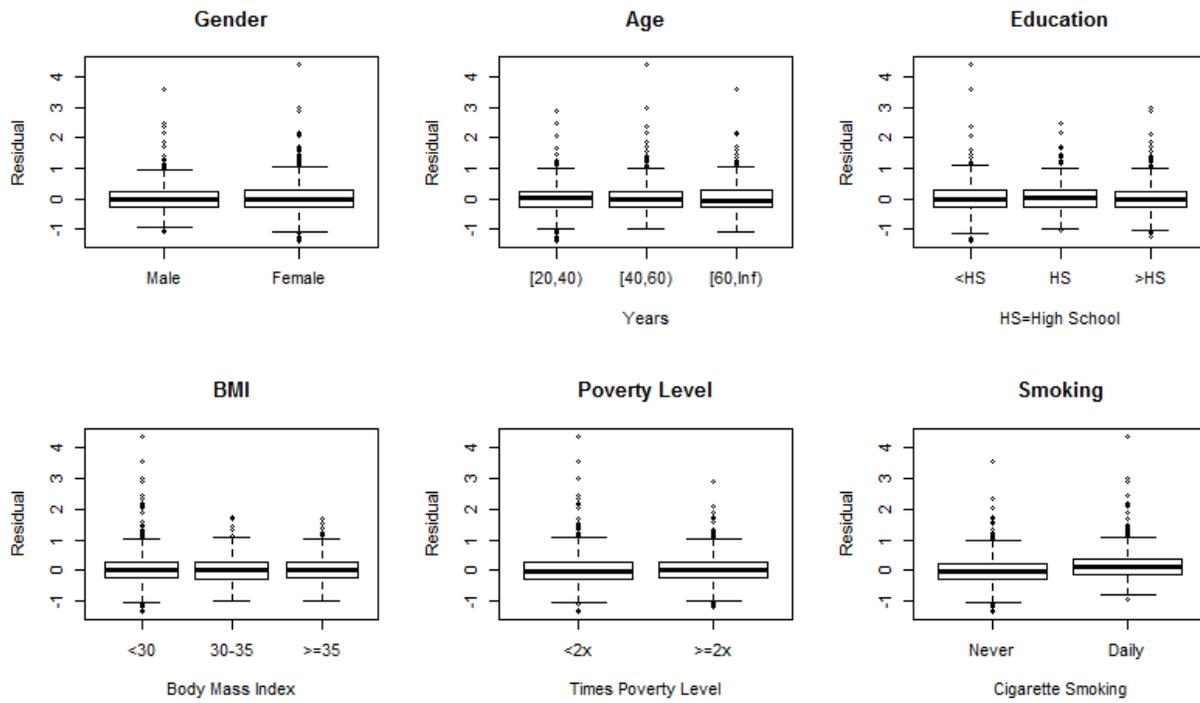


Figure 3: Plots of residuals from a robust regression of  $\log_2(\text{Homocysteine})$  levels on a binary indicator of gender, three continuous variables, namely age in years, BMI, and ratio of income to the poverty level, and an integer variable recording 5 levels of education. Smoking is not included in the regression.

levels by 5.1-fold, noticeably higher than for  $\Gamma = 1.95$  above; again see Rosenbaum (2017a, Table 9.1).

By removing the hypothesized treatment effect before robust residuals are computed, a confidence interval may be computed; see Rosenbaum (2002b). At  $\Gamma = 1.25$ , using  $M$ -scores applied to residuals, the one-sided 95% confidence interval for a multiplicative effect  $\beta$  on homocysteine levels is  $\beta \geq 1.077$  or a 7.7% increase in homocysteine levels. In a matched pair, a bias  $\Gamma = 1.25$  corresponds with a covariate that doubles the odds of smoking and doubles the odds of a positive pair difference in outcomes.

In brief, a 7.7% increase in homocysteine levels caused by smoking is insensitive to a nontrivial bias in treatment assignment of  $\Gamma = 1.25$ , and rejection of the null hypothesis  $H_0$  of no effect of smoking is insensitive to a moderately large bias of  $\Gamma = 2.35$ .

## 6.2 A second, independent consideration in evaluating the effects of smoking on homocysteine levels

Self-reported doses of exposure to addictive substances may be inaccurate, perhaps biased toward understatement. Cotinine in the blood is a marker of the extent of recent exposure to nicotine, in effect an objective dose of exposure. The Centers for Disease Control (2016) write:

Once absorbed, nicotine has a half-life in blood plasma of several hours. Cotinine, the primary metabolite of nicotine, is currently regarded as the best biomarker of tobacco smoke exposure. Measuring cotinine is preferable to measuring nicotine because cotinine persists longer in the body with a plasma half-life of about 16 hours.

If everyone metabolized nicotine in the same way, then the level of cotinine in the blood would be a measure of the dose of nicotine recently consumed. In fact, people vary somewhat in how they metabolize nicotine, so cotinine is an imperfect dose of treatment, but let us set that concern aside for a moment, revisiting the concern after examining the data.

Figure 4 compares the homocysteine levels of smokers with high or low levels of cotinine. The right panel of Figure 4 uses the same residuals as in Figure 3, but Figure 4 is confined to smokers. Tentatively, if perhaps naively, viewing the level of cotinine as the recent dose of nicotine, we may conduct stratified analyses of the effects of high versus low doses of nicotine on the homocysteine levels of smokers. Using the 108 strata from §1.1 together with  $M$ -scores, at level  $\alpha = 0.05$ , the null hypothesis of no effect of dose on  $\log_2(\text{homocysteine})$  is rejected in the presence of a bias in dose assignments of  $\Gamma = 2.4$ , whereas the null hypothesis of no effect on the residuals is rejected at  $\Gamma = 2.5$ . In brief, there is evidence insensitive to small biases that higher recent doses of nicotine produce higher levels of homocysteine in smokers.

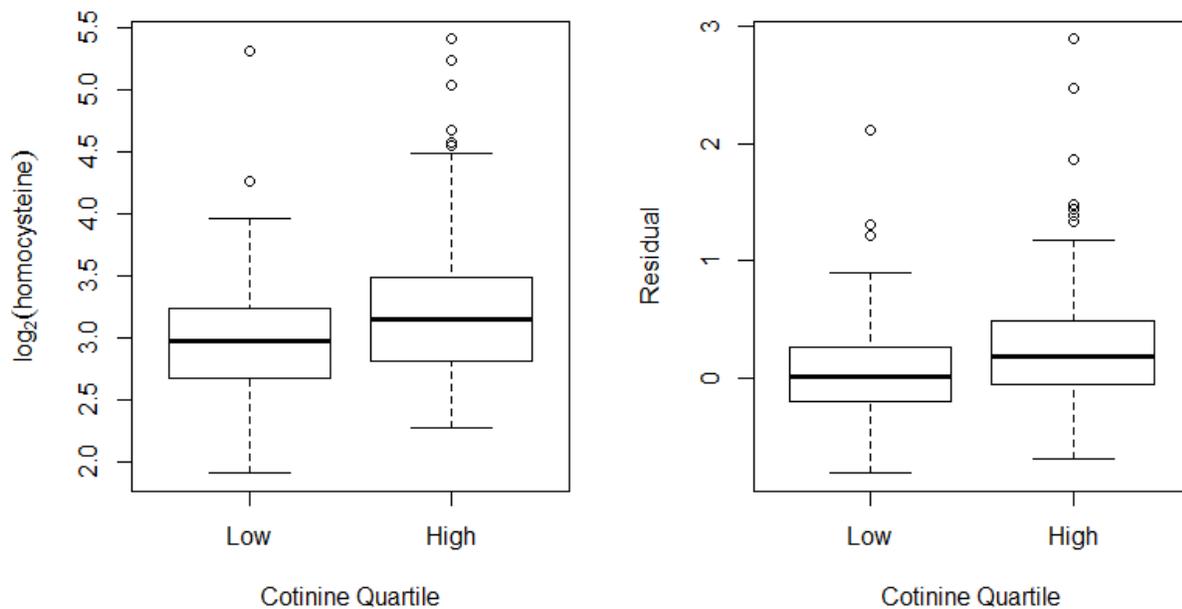


Figure 4: Homocysteine levels among smokers with high or low levels of serum cotinine. The left panel depicts  $\log_2(\text{homocysteine})$  levels, whereas the right panel depicts their residuals from a robust regression on five covariates. The boxplot for low cotinine describes 128 smokers with cotinine levels at or below the lower quartile of cotinine for smokers, while the boxplot for high cotinine describes 129 smokers with cotinine levels at or above the upper quartile.

Sections 6.1 and 6.2 have twice analyzed the same data from smokers, have twice tested hypotheses of no treatment effect, and have twice considered the magnitude of bias needed to explain rejection as a bias rather than a treatment effect. What is the relationship between these two analyses of the same data? The two analyses would be nearly independent if there were no treatment effect, and enormous biases affecting either analysis alone would have no impact on the other analysis. That is, these two analyses are two evidence factors, as developed informally in Rosenbaum (2017a, §7) and formally in Rosenbaum (2017b).

Returning to an issue mentioned above, strictly speaking cotinine is imperfect as a dose of nicotine. It is possible, in principle, that two people who received the same dose of nicotine would metabolize it differently, yielding different levels of cotinine. In that sense, Figure 4 has two possible interpretations. Figure 4 may indicate that higher doses of nicotine produce higher levels of homocysteine in smokers. Alternatively, Figure 4 may indicate that the metabolism of nicotine into cotinine and the level of homocysteine are interrelated in such a way that smokers who produce higher cotinine levels also produce higher homocysteine levels. As noted above, the Centers for Disease Control regard cotinine as “the best biomarker of tobacco smoke exposure,” a view that favors the first interpretation.

## References

- BAZZANO, L. A., HE, J., MUNTNER, P., VUPPUTURI, S. AND WHELTON, P. K. (2003). Relationship between cigarette smoking and novel risk factors for cardiovascular disease in the United States. *Ann. Int. Med.* **138** 891-897.
- BERTSEKAS, D. P. (2009). *Convex Optimization Theory*. Nashua, NH: Athena Scientific.
- BICKEL, P. J. AND W. R. VAN ZWET. (1978). Asymptotic expansions for the power of distribution free tests in the two-sample problem. *Ann. Statist.* **6** 937-1004.
- BOYD, S. AND VANDENBERGHE, L. (2004). *Convex Optimization*. New York: Cambridge University Press.
- CENTERS FOR DISEASE CONTROL (2016) Biomonitoring Summary: Cotinine, CAS No. 486-56-6, [https://www.cdc.gov/biomonitoring/Cotinine\\_BiomonitoringSummary.html](https://www.cdc.gov/biomonitoring/Cotinine_BiomonitoringSummary.html), dated December 27, 2016.
- CORNFIELD, J., HAENSZEL, W., HAMMOND, E., LILIENFELD, A., SHIMKIN, M., AND WYNDER, E. (1959). Smoking and lung cancer. *J. Nat. Cancer Inst.* **22** 173-203.
- EGLSTON, B. L., SCHARFSTEIN, D. O. AND MACKENZIE, E. (2009). On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics* **65** 497-504.
- FISHER, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.
- FOGARTY, C. B. AND SMALL, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *J. Am. Statist. Assoc.* **111** 1820-1830.

- GASTWIRTH, J. L., KRIEGER, A. M., AND ROSENBAUM, P. R. (2000) Asymptotic separability in sensitivity analysis. *J. Roy. Statist. Soc. B* **62** 545-556.
- GILBERT, P., BOSCH, R., HUDGENS, M. (2003). Sensitivity analysis for the assessment of the causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59** 531-41.
- HANKEY, G. J. AND EIKELBOOM, J. W. (1999). Homocysteine and vascular disease. *Lancet* **354** 407-413.
- HANSEN, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Am. Statist. Assoc.*, 99, 609-618. (R packages `optmatch` and `sensitivityfull`)
- HODGES, J. L. AND LEHMANN, E. L. (1962) Rank methods for combination of independent experiments in analysis of variance. *Ann. Math. Statist.* **33** 482-497.
- HOSMAN, C. A., HANSEN, B. B. AND HOLLAND, P. W. H. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Statist.* **4** 849-870.
- HUBER, P. (1981). *Robust Statistics*. New York: Wiley.
- LEHMANN, E. L. (1975) Nonparametrics. San Francisco: Holden-Day.
- LEHMANN, E. L. AND ROMANO, J. (2005). *Testing Statistical Hypotheses* (3rd edition). New York: Springer.
- LIU, W., KURAMOTO, J. AND STUART, E. (2013). Sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevent. Sci.* **14** 570-580.
- MANTEL, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *J. Am. Statist. Assoc.* **58** 690-700.
- MARITZ, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66** 163-166.
- MEHROTRA, D. V., LU, X., AND XIAOMING, L. (2010). Rank-based analysis of stratified experiments: alternatives to the van Elteren test. *Am. Statist.* **64** 121-130.
- NEYMAN, J. (1923, 1990). On the application of probability theory to agricultural experiments. Reprinted in: *Statist. Sci.* **5** 463-480.
- PIMENTEL, S. D., SMALL, D. S. AND ROSENBAUM, P. R. (2016). Constructed second control groups and attenuation of unmeasured biases. *J. Am. Statist. Assoc.* **111** 1157-1167.
- PURI, M. L. (1965). On the combination of independent two sample tests of a general class. *Rev. Int. Statist. Inst.* **33** 229-241.
- ROSENBAUM, P. R. AND KRIEGER, A. M. (1990). Sensitivity of two-sample permutation inferences in observational studies. *J. Am. Statist. Assoc.*, 85, 493-498.
- ROSENBAUM, P. R. (1991). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. B* **53** 597-610. (R packages `optmatch` and `sensitivityfull`)
- ROSENBAUM, P. R. (1995). Quantiles in nonrandom samples and observational studies. *J. Am. Statist. Assoc.*, 90, 1424-1431.
- ROSENBAUM, P. R. (2002a). *Observational Studies* (2nd edition). New York: Springer.
- ROSENBAUM, P. R. (2002b). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286-327.

- ROSENBAUM, P. R. (2014) Weighted M-statistics with superior design sensitivity in matched observational studies with multiple controls. *J. Am. Statist. Assoc.* 109, 1145-1158. (R package `sensitivitymw`)
- ROSENBAUM, P. R. (2017a). *Observation and Experiment: An Introduction to Causal Inference*. Cambridge, MA: Harvard University Press.
- ROSENBAUM, P. R. (2017b). The general structure of evidence factors in observational studies. *Statist. Sci.*, to appear.
- ROSENBAUM, P. R. AND SMALL, D. S. (2017). An adaptive Mantel-Haenszel test for sensitivity analysis in observational studies. *Biometrics* **73** 422-430. (R package `sensitivity2x2xk`)
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Ed. Psych.* **66** 688-701.
- SATAGOPAN, J. M., OFFIT, K., FOULKES, W., ROBSON, M. E. WACHOLDER, S., ENG, C. M., KARP, S. E. AND BEGG, C. B. (2001). The lifetime risks of breast cancer in Ashkenazi Jewish carriers of *brca1* and *brca2* mutations. *Cancer Epidemiol. Biomark. Prev.* **10** 467-473.
- SESHADRI, S., BEISER, A., SELHUB, J., JACQUES, P. F., ROSENBERG, I. H., D'AGOSTINO, R. B., WILSON, P. W. AND WOLF, P. A. (2002). Plasma homocysteine as a risk factor for dementia and Alzheimer's disease. *New Eng. J. Med.* **346** 476-483.
- WALD, D. S., LAW, M. AND MORRIS, J. K. (2002). Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis. *Brit. Med. J.* **325** 1202-1209.
- WELCH, G. N., & LOSCALZO, J. (1998). Homocysteine and atherothrombosis. *New Eng. J. Med.* **338** 1042-1050.
- WERFEL, U., V. LANGEN, I. EICKHOFF, J. SCHOONBROOD, C. VAHRENHOLZ, A. BRAUKSIEPE, W. POPP, AND NORPOTH, K. (1998). Elevated DNA single-strand breakage frequencies in lymphocytes of welders exposed to chromium and nickel. *Carcinogenesis* **19** 413-418.
- YU, B. B. AND GASTWIRTH, J. L. (2005). Sensitivity analysis for trend tests: application to the risk of radiation exposure. *Biostatistics* **6** 201-209.