Observational Studies 1 (2015) 1-17 Submitted 7/14; Published 11/14

Two R Packages for Sensitivity Analysis in Observational Studies

Paul R. Rosenbaum **rosenbaum@wharton.upenn.edu**
Department of Statistics
Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6340 US

Abstract

Two `R` packages for sensitivity analysis in observational studies are described. Package `sensitivitymw` is for matched pairs with one treated subject and one control, or matched sets with one treated subject and a fixed number, $K \geq 2$, of controls. Package `sensitivitymv` is for matched sets with variable numbers of controls. The packages offer conventional statistics, such as the permutational t-test and M-statistics using Huber's weights, but they also offer less familiar test statistics that have higher power in sensitivity analyses. The packages provide several tools useful in sensitivity analyses, such as an aid, `amplify`, to the interpretation of the value of the sensitivity parameter, and a device for combining evidence from several independent sensitivity analyses, `truncatedP`, for instance, several evidence factors or several subgroups.

Keywords: M-test; observational study; permutational t-test; randomization inference; sensitivity analysis.

1. Introduction

1.1 `R` Packages `sensivitymv` and `sensitivitymw`

The two `R` packages `sensivitymv` and `sensitivitymw` perform sensitivity analyses for observational studies with matched pairs or matched sets containing multiple controls. Package `sensitivitymw` is for matched pairs or matching with a fixed number of controls, for instance matching each treated subject to two controls. In contrast, package `sensivitymv` is for matched sets with variable numbers of controls, perhaps some treatment-control pairs together with some triples containing a treated subject and two controls. Also, the packages contain several data sets and several additional functions useful in sensitivity analysis. The packages overlap considerably, but package `sensitivitymw` is faster with additional features for matched pairs and for matching with a fixed number of controls. Both packages are available at CRAN and contain documentation.

My purpose here is to present a gentle introduction to these `R` packages, with pointers to articles for technical detail and pointers to the software documentation for additional options.

1.2 Scope of the current discussion

In an observational study, a sensitivity analysis replaces qualitative claims about whether unmeasured biases are present with an objective quantitative statement about the magnitude of bias that would need to be present to change the conclusions. In this sense, a sensitivity analysis speaks to the assertion "it might be bias" in much the same way that a P-value speaks to the assertion "it might be bad luck". If someone asserted that the higher responses in the treated group in a randomized experiment "might be bad luck," an unlucky randomization with no treatment effect, then a P-value does not deny the logical possibility of bad luck, but objectively measures the quantity of bad luck that would need to be present to alter the impression that the treatment did have an effect. In parallel, a sensitivity analysis measures the magnitude of bias from nonrandom treatment assignment that would need to be present to alter the conclusions of an observational study.

A sensitivity analysis is one tool useful in the large task of designing and interpreting an observational study. The discussion here is rather narrowly focused on carrying out such a sensitivity analysis in `R`.

1.3 What do the packages do?

In an observational study, treated and control subjects may be matched to be similar in terms of observed or measured covariates, but people who look similar in terms of measured covariates may still differ in terms of unmeasured covariates. The packages perform a sensitivity analysis asking about the magnitude of bias from nonrandom treatment assignment that would need to be present to alter the qualitative conclusions of a naive analysis that presumes matching for observed covariates removes all bias.

In a matched randomized experiment, each subject in a matched set has the same chance of being assigned to treatment or control because randomization has ensured that this is so. Without randomization, two people who look similar may differ in their chances of receiving treatment because they differ in terms of an unmeasured covariate not controlled by matching for measured covariates. The sensitivity analysis assumes that one subject in a matched set may be $\Gamma \geq 1$ times more likely than another to receive treatment because they differ in terms of unobserved covariates. If $\Gamma = 1$, then subjects who look the same are the same: matched subjects have equal chances of treatment, as in a randomized experiment. For $\Gamma = 1$, the sensitivity analysis reports a single answer, for instance a single P-value testing the null hypothesis of no treatment effect, and that single answer is the P-value that would be appropriate in a matched randomized experiment. For $\Gamma > 1$, there is no longer a single P-value, but rather an interval of possible P-values. The sensitivity analysis asks: How large must Γ be before the interval is so long that it is inconclusive, perhaps both accepting and rejecting the null hypothesis of no effect at the 0.05 level? The interval of possible P-values would be inconclusive in this sense if it extended from below 0.05 to above 0.05. The `senmw` and `senmv` functions compute sensitivity bounds for P-values. Specifically, they compute the upper bound on the P-value, for a specific Γ, so if that upper bound is at most 0.05, then a bias of magnitude Γ is too small to lead to acceptance of the null hypothesis. The `senmwCI` function inverts bounds on P-values to obtain sensitivity bounds for confidence intervals and point estimates. For detailed discussion of this model, see Rosenbaum (2002, §4; 2007).

Both packages use M-tests, that is, the tests associated with Huber's (1981) M-estimates, including the permutational t-test. The permutational t-test is a permutation or randomization test that permutes the observations; see, for instance, Fisher(1935), Pitman (1937) and Welch (1937). In an M-statistic, the observations are slightly transformed, often with a view to preventing one or two observations from having overwhelming influence, and in an M-test the transformed observations are permuted. Maritz (1979, 1995) slightly adjusted Huber's M-statistics to make them more suitable for matched-pair permutation tests, and there is a straightforward extension to matching with multiple controls (Rosenbaum 2007, §4); see §3.

1.4 Outline

Section 2 discusses matched pairs with two familiar test statistics, the permutational t-test and an M-test using Huber's weights, similar to the method proposed by Maritz (1979). An aid to interpreting values of Γ, its amplification into two equivalent sensitivity parameters using `amplify`, is discussed in §2.4. Section 3 discusses matched sets with one treated subject matched to more than one control. An observational study may contain evidence factors, that is, two independent tests of no treatment effect that are likely to be affected by different unobserved biases. Evidence factors are discussed in §4 along with an analytic tool, `truncatedP`, that implements Zaykin et al. (2002)'s truncated product of independent P-values. Related issues of effect modification in subgroup analyses are discussed in §5. Where §2 and §3 used conventional test statistics, §6 considers test statistics with higher power when used in a sensitivity analysis. Examples from the `R` packages are discussed throughout. The examples may be used to reproduce analyses in several published articles. Two appendices provide more information about the `R` packages.

2. Matched pairs

2.1 The permutational t-test for matched pairs

There are I matched sets, $i = 1, \ldots, I$, and n_i subjects, $j = 1, \ldots, n_i$, in set i, where $n_i = 2$ for treatment-control pairs and $n_i = K + 1$ for matching every treated subject to K controls. One subject in each set is treated, denoted $Z_{ij} = 1$, the others are untreated controls, denoted $Z_{ij} = 0$, so $1 = \sum_{i=1}^{n_i} Z_{ij}$ for each i. The jth subject in set i would exhibit response r_{Tij} if assigned to treatment with $Z_{ij} = 1$ or response r_{Cij} if assigned to control with $Z_{ij} = 0$, so this subject actually exhibits response $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$ and the effect of treatment on this subject, namely $r_{Tij} - r_{Cij}$, is not observed; see Neyman (1923), Welch (1937) and Rubin (1974). Fisher's (1935) null hypothesis H_0 of no treatment effect asserts that $r_{Tij} = r_{Cij}$ for all i, j. The treatment has additive shift or constant effect if $r_{Tij} - r_{Cij} = \tau$ for all i, j.

Consider, first, matched pairs with $n_i = 2$ for all i. The treated-minus-control pair difference in outcomes is $Y_i = (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2})$. If the treatment has an additive effect τ, then $Y_i = \tau + \epsilon_i$ where $\epsilon_i = (Z_{i1} - Z_{i2})\,(r_{Ci1} - r_{Ci2}) = \pm\,|r_{Ci1} - r_{Ci2}|$, whereas, under Fisher's null hypothesis, H_0, of no effect, the pair difference is $Y_i = \epsilon_i$. Under H_0 in a paired randomized experiment, $Y_i = \pm\,|r_{Ci1} - r_{Ci2}|$ with equal probabilities.

In a randomized experiment, the permutational t-test is the randomization test that uses as its test statistic either the total, $T = \sum_{i=1}^{I} Y_i$, or the mean, $(1/I)\sum_{i=1}^{I} Y_i$, where these two statistics give the same permutational P-value. The permutation distribution of the mean, or the permutational t-test, is of historical and conceptual importance, in part because, in a randomized experiment, the expectation of $(1/I)\sum_{i=1}^{I} Y_i$ is the average treatment effect, $\{1/(2I)\}\sum_{i=1}^{I}\sum_{j=1}^{2}(r_{Tij} - r_{Cij})$.

2.2 Using the permutational t-test in matched pairs

Werfel et al. (1998) matched 39 welders exposed to chromium and nickel to 39 unexposed controls, measuring DNA damage in lymphocytes by DNA elution rates through polycarbonate filters with proteinase K (or ERPC+). Pairs were matched for age and smoking habits. The data frame `erpcp` in both packages has two columns, welder and control, and it contains the ERPC+ values for 39 pairs or rows.

The following calculations obtain the upper bound on the one-sided P-value testing the null hypothesis of no treatment effect using the permutational t-test (method="t"). For $\Gamma = 1$, this is the usual randomization P-value for the mean difference, namely 2.048×10^{-5}. For $\Gamma = 3$, the upper bound is 0.0228. For $\Gamma = 4$, the upper bound is 0.0579, so P-values well below and slightly above the conventional 0.05 level are possible under H_0 if the bias could be as large as $\Gamma = 4$. In other words, rejection of H_0 is sensitive to unmeasured biases of magnitude $\Gamma = 4$.

```
> library(sensitivitymw)
> data(erpcp)
> senmw(erpcp, gamma = 1, method = "t")$pval
[1] 2.048115e - 05
> senmw(erpcp, gamma = 3, method = "t")$pval
[1] 0.02275942
> senmw(erpcp, gamma = 4, method = "t")$pval
[1] 0.0579339
```

Association does not imply causation, and that is always true, but logical implication tells us less than sensitivity analysis of the data at hand. The sensitivity analysis says that the observed association between welding and DNA elution rates is too strong to be explained by a bias of $\Gamma = 3$, because the maximum possible P-value from a bias of $\Gamma = 3$ is 0.0228, so a bias of that magnitude would not make the null hypothesis of no effect plausible. However, a bias of $\Gamma = 4$ would make the null hypothesis barely plausible, because with a bias that large, the P-value could be as large as $0.0579 > 0.05$. Saying that association does not imply causation is essentially the same as saying that the upper bound on the P-value tends to 1 as $\Gamma \to \infty$.

The P-value bounds are one-sided. In a sensitivity analysis, it is safe though somewhat conservative to obtain a two-sided P-value by doubling the smaller of two one-sided P-values, reporting a two-sided bound of $0.02275942 \times 2 = 0.04551884$ for $\Gamma = 3$. The reason doubling the one-sided P-value is conservative in a sensitivity analysis is that the bias that pushes the test statistic T into the upper tail is different from the bias that pushes it into

the lower tail; see the related discussion of use of the Bonferroni inequality in sensitivity analyses in Rosenbaum and Silber (2009a, §4.5).

The function `senmwCI` computes point estimates and confidence intervals for an additive effect τ. For $\Gamma = 1$, there is a single point estimate, which for `method`= "`t`" is the mean difference, `mean(erpcp$welder-erpcp$control) = 0.5739`. The default is a one-sided 0.05-level confidence interval. (The level is controlled by `alpha` and one-or-two sided is controlled by `one.sided`.)

```
> senmwCI(erpcp, gamma = 1, method = "t", one.sided = TRUE)
$PointEstimate
minimum maximum
 0.5739  0.5739

$Confidence.Interval
minimum maximum
  0.394     Inf
```

For $\Gamma = 2$, there is no longer a single point estimate, 0.5739, but rather an interval of point estimates, [0.4167, 0.7487] and a longer 95% confidence interval, $\tau \geq 0.2081$. Notably, with a bias of at most $\Gamma = 2$, the smallest possible point estimate of τ, namely 0.4167, is still fairly large.

```
> senmwCI(erpcp, gamma = 2, method = "t", one.sided = TRUE)
$PointEstimate
minimum maximum
0.4167  0.7487

$Confidence.Interval
minimum maximum
 0.2081     Inf
```

In a sensitivity analysis, it is safe but somewhat conservative to form a 95% two-sided confidence interval as the intersection of two one-sided 97.5% confidence intervals, for the same reason that two-sided P-values are safe but somewhat conservative; see Rosenbaum (1995, §2.1) for some details.

2.3 M-statistics for matched pairs

An M-statistic gives each Y_i a controlled degree of influence. Let s be the median of the $|Y_i| = |R_{i1} - R_{i2}|$, as in Maritz (1979). For matched pairs, the M-statistic is $T = \sum_{i=1}^{I} \psi(Y_i/s)$ where $\psi(\cdot)$ is a suitable function. Taking $\psi(y) = y$ yields the same P-values as the permutational t-test. Huber (1981) proposed a $\psi(\cdot)$ that tops out at a constant $h > 0$ and bottoms out at $-h$, specifically $\psi(y) = \max\{-h, \min(y, h)\} = \text{sign}(y) \cdot \min(|y|, h)$, thereby limiting to $\pm hs$ the influence one observation Y_i can have on the statistic T.

With the default settings (or `method`= "`h`") in the `erpcp` data, the upper bounds on P-values using Huber's weights are similar to those from the permutational t-test in §2.2,

but this will vary from one data set to another. In parallel, `senmwCI` may be used to obtain a sensitivity analysis for point estimates and confidence intervals.

```
> senmw(erpcp, gamma = 1, method = "h")$pval
[1] 6.402131e - 06
> senmw(erpcp, gamma = 2, method = "h")$pval
[1] 0.002410713
> senmw(erpcp, gamma = 3, method = "h")$pval
[1] 0.01859188
> senmw(erpcp, gamma = 4, method = "h")$pval
[1] 0.05304687
```

(Some comments about default settings follow. By default, `senmv`, `senmw` and `senmwCI` use the median of $|Y_i|$ to define s, but the user can select a different quantile by changing the value of `lambda`, the default being `lambda` $= 1/2$ for the median. By default, $h = 2.5$ in `senmv` and $h = 3$ in `senmw` and `senmwCI`, but the user can select different values by changing the value of `trim`. If the Y_i are discrete and most Y_i equal zero, the median $|Y_i|$ is not useful for scaling, and it may be reasonable to take `lambda` $= .90$ and $h =$ `trim` $= 1$, which resembles a trimmed mean.)

2.4 Amplification: an aid to interpreting Γ

When computing or reporting a sensitivity analysis, it is often convenient to have an analysis indexed by a single parameter, Γ. As discussed in §1.3, the sensitivity analysis reports the range of possible inferences when an unobserved bias alters the odds of treatment by a factor of at most Γ. The extremes of that range are produced by a bias strongly related to the outcome. An amplification interprets the single parameter Γ in terms of two parameters, one Λ controlling the relationship between the unobserved bias and treatment assignment Z_{ij}, the other Δ controlling the relationship between the unobserved bias and the outcome Y_i. Here, Λ is the maximum impact of the bias on the odds of treatment, $Z_{i1} - Z_{i2} = 1$, and Δ is the maximum impact of the unobserved bias on the odds of a positive response difference, $Y_i > 0$. A bias of Γ is equivalent to the curve defined by $\Gamma = (\Lambda\Delta + 1) \,/\, (\Lambda + \Delta)$. More precisely, under a certain semiparametric model for Y_i and $Z_{i1} - Z_{i2}$, a sensitivity analysis at Γ gives exactly the same P-value bounds as all sensitivity analyses at (Λ, Δ) such that $\Gamma = (\Lambda\Delta + 1) \,/\, (\Lambda + \Delta)$. In other words, one can calculate and report using one parameter Γ but have available the equivalent interpretations involving two parameters (Λ, Δ). See Rosenbaum and Silber (2009b) for a precise discussion.

The function `amplify` in the `sensitivitymv` package performs the required elementary calculations. Specifically, the call `amplify(gamma, lambda)` takes a scalar $\Gamma > 1$ and a vector of Λ's and computes the corresponding vector of Δ's. The analyses in §2.2 and §2.3 were insensitive to $\Gamma = 3$. The following call considers $\Lambda = (4, 5, 6, 7)$.

```
> library(sensitivitymv)
> amplify(3, c(4 : 7))
```

The result is:

4	5	6	7
11.00	7.00	5.67	5.00

For example, an unobserved covariate that increases the odds of treatment, $Z_{i1} - Z_{i2} = 1$, by at most $\Lambda = 5$ and the odds of a positive response difference, $Y_i > 0$, by at most $\Delta = 7$ is equivalent to $\Gamma = 3$. However, $\Gamma = 3$ is also equivalent to $(\Lambda, \Delta) = (7, 5)$, to $(\Lambda, \Delta) = (4, 11)$, to $(\Lambda, \Delta) = (11, 4)$, and to $(\Lambda, \Delta) = (6, 5.67)$. That is, a bias of $\Gamma = 3$ is quite a large bias, the omission of a covariate strongly related to both treatment assignment and response.

Similarly, `amplify(1.5,2)` yields 4, so $\Gamma = 1.5$ corresponds with both $(\Lambda, \Delta) = (2, 4)$ and $(\Lambda, \Delta) = (4, 2)$, while `amplify(1.25,2)` yields 2, so $\Gamma = 1.25$ corresponds $(\Lambda, \Delta) = (2, 2)$. In words, $\Gamma = 1.25$ corresponds with a doubling of the odds of treatment and a doubling of the odds of a positive response difference, not a trivially small bias. In $\Gamma = (\Lambda\Delta + 1) / (\Lambda + \Delta)$, as $\Delta \to \infty$, the corresponding Λ approaches Γ.

3. Matched sets with multiple controls

3.1 M-statistics with multiple controls

With $n_i \geq 2$ subjects in set i, there are $n_i - 1$ treated-minus-control pair differences, Y_{ik}, $k = 1, \ldots, n_i - 1$, all with the same treated subject, $Z_{ij} = 1$, but each with a different control, $Z_{i\ell} = 0$. The scale factor, s, is now defined to be the median of the $\sum_{i=1}^{I} \binom{n_i}{2}$ absolute differences, $\left|R_{ij} - R_{ij'}\right|$ with $j < j'$. The M-statistic is then $T = \sum_{i=1}^{I} w_i \sum_{k=1}^{n_i - 1} \psi\left(Y_{ik}/s\right)$, summing over all $\sum_{i=1}^{I} (n_i - 1)$ pair differences Y_{ik}, where set i is given weight w_i. See Rosenbaum (2007, 2014) for technical discussion of sensitivity analyses using these statistics.

There are various ways to attach weights w_i to matched sets, and `senmv` and `senmw` provide several options. Before discussing weights, consider an example with constant weights, essentially an unweighted example, in which every treated subject is matched to $n_i - 1 = 2$ controls.

3.2 Example with two controls

Fish often contains mercury. Does eating large quantities of fish increase levels of mercury in the blood? Data set `mercury` in the `sensitivitymw` package is from the 2009-2010 National Health and Nutrition Examination Survey (NHANES) and is the example in Rosenbaum (2014). There are 397 rows or matched triples and three columns, one treated with two controls. The values are methylmercury levels in blood in μg/dL. Column 1, "Treated", describes an individual who had at least 15 servings of fish or shellfish in the previous month. Column 2, "Zero", describes an individual who had 0 servings of fish or shellfish in the previous month. Column 2, "One", describes an individual who had 1 serving of fish or shellfish in the previous month. In the comparison here, Zero and One are not distinguished; both are controls. Sets were matched for gender, age, education, household income, black race, Hispanic, and cigarette consumption; see Table 1 in Rosenbaum (2014). A description of the data follows.

```
> data(mercury)
>summary(mercury[,1])
Min.  1st Qu. Median Mean 3rd Qu. Max.
0.230  1.300  2.360  3.724  4.280 38.000
>summary(unlist(mercury[,2:3]))
Min.  1st Qu. Median Mean 3rd Qu. Max.
0.230  0.330  0.520  0.781  0.880 14.600
```

The upper bounds on the one-sided P-value testing the null hypothesis H_0 of no treatment effect are above 0.05 for $\Gamma = 16$ for both the permutational t-test and using Huber's ψ-function. In this example, rejection of H_0 is highly insensitive to unmeasured bias.

```
> senmw(mercury, gamma = 1, method = "t")$pval
[1] 0
> senmw(mercury, gamma = 1, method = "h")$pval
[1] 0
> senmw(mercury, gamma = 16, method = "t")$pval
[1] 0.05208875
> senmw(mercury, gamma = 16, method = "h")$pval
[1] 0.1187912
```

Consider the possible impact of a bias of magnitude $\Gamma = 5$ on estimates of an additive effect, τ. The interval of point estimates is [0.93, 4.09] and the (somewhat conservative) two-sided 95% confidence interval is [0.74, 4.66]. Notice that the confidence interval is only a little longer than the interval of point estimates, indicating that most of the uncertainty comes from the bias $\Gamma = 5$ rather than sampling variability.

```
> senmwCI(mercury, gamma = 5, method = "h", one.sided = FALSE)
$PointEstimate
minimum maximum
   0.93    4.09

$Confidence.Interval
minimum maximum
   0.74    4.66
```

3.3 Weighting matched sets of unequal sizes

Data set `tbmetaphase` in the `sensitivitymv` package contains both matched pairs and matched triples. It is the example in Rosenbaum (2007, §4.3), and the documentation for `senmv` shows how to reproduce the analyses in that paper.

When matched sets have different sizes, different n_i, it is natural to ask whether larger sets should receive more weight in the analysis. Should a matched triple receive more weight than a matched pair? The default in `senmv` aims for efficiency in a randomization test, $\Gamma = 1$. The default in `senmv` uses weights w_i that would be optimal with $\psi(y) = y$ under a Gaussian model with constant variance; see Rosenbaum (2007, §4.2) for specifics.

In some studies, all treated subjects in some well-defined population are matched to variable numbers of controls. In this case, one may wish to describe the actual population of treated subjects, to use weights that refer to average effect of the treatment on treated subjects. By setting `TonT = TRUE` in `senmv`, weights of this form are used. See the documentation for `senmv` for specifics.

4. Evidence factors

4.1 What are evidence factors?

Some observational studies permit two nearly independent tests of the null hypothesis of no treatment effect, tests that are likely to be influenced by different unmeasured biases; see Rosenbaum (2010b, 2011), Zhang et al. (2011), and Zubizarreta et al. (2012). Although the same data are used twice, when carefully designed the two tests are nonetheless nearly independent, so the study provides an independent replicate of itself with a different design subject to different biases. Two such tests are called two "evidence factors." If the two evidence factors support each other, then they provide somewhat stronger evidence of effect. In Zhang et al. (2011), two evidence factors supported one-another, strengthening evidence of effect, but in Zubizarreta et al. (2012) they contradicted each other, weakening evidence of effect.

4.2 Combining P-values using `truncatedP`

Because two evidence factors are nearly independent, their P-values may be combined by methods for combining independent P-values, for instance, by Fisher's product of P-values or by Zaykin et al. (2002)'s truncated product of P-values. The truncated product is the product of those P-values less than or equal to a certain cutpoint, κ, $0 < \kappa \leq 1$. For $\kappa = 1$, the truncated product is the same as Fisher's method. When some null hypotheses are composite, some null P-values will be much larger than the uniform distribution, and the truncated product eliminates these. As discussed by Hsu et al. (2013), the truncated product is often much better than Fisher's method when combining P-value bounds obtained by sensitivity analyses.

The two equivalent functions, `truncatedP` and `truncatedPbg` in the `sensitivitymv` package, take as input a sequence of independent P-values (or independent P-value bounds) and return the P-value (or P-value bound) for the truncated product. By default, the truncation point is $\kappa =$ `trunc` $= 0.2$, but see Hsu et al. (2013) for comparisons of different κ's in sensitivity analyses.

4.3 Example: two comparisons in matched triples

Meibian et al. (2008) used the mean tail moment (mtm) of the comet assay to study possible DNA damage among workers exposed to chromium at a tannery. They compared 30 unexposed controls to 60 tannery workers, where the tannery workers divided into two groups of 30, one with higher direct exposure to chromium, the other with lower indirect exposure. The two evidence factors compare controls to tannery workers and low to high exposure. Each comparison may be biased, but presumably the process that leads some people to work at a tannery is a different process than the process that assigns jobs within

the tannery. Weiss (1981) distinguishes doses varied by a single process, hence perhaps biased by a single bias, and doses varied by more than one process, hence less easily biased by a single bias. The claim that exposure to chromium damages DNA would be strengthened by an observation that mtm is lower with no exposure than with exposure at the tannery, and by an observation that mtm is lower with low exposure at the tannery than with high exposure, whereas other patterns would weaken that claim. The data were used as an example in Rosenbaum (2011), where technical specifics may be found. The documentation for `truncatedP` and `truncatedPbg` shows how to reproduce the analyses in that paper.

The data set `mtm` in the `sensitivitymv` package has 30 rows for 30 matched triples, and three columns for control (cmtm), low exposure (e2mtm), high exposure (e1mtm). The first lines of `mtm` are:

```
     cmtm  e2mtm  e1mtm
[1,] 1.45   1.79   3.02
[2,] 2.48   2.15   6.60
              ⋮
```

In the first factor (`mtm`), there is one control and two treated subjects. In the second factor (the last two columns, `mtm[,2:3]`), there are low-versus-high dose matched pairs. The P-values computed by `senmv` are one-sided, upper tail, so to look in the opposite tail for low mean tail moments among controls or among low dose exposed individuals, we replace `mtm` by its negative. We see that the comparison of tannery workers and controls has a P-value bound of 0.049 at $\Gamma = 11$, whereas the comparison of low-versus-high exposure among tannery workers has a P-value bound of 0.076 at $\Gamma = 2$. These two tests are nearly independent; see Rosenbaum (2011) for the definition of "nearly".

```
> nmtm < -(-mtm)
> senmv(nmtm, method = "h", gamma = 11)$pval
[1] 0.04883432
> senmv(nmtm[, 2 : 3], method = "h", gamma = 2)$pval
[1] 0.07641043
```

Combining the two P-value bounds gives a P-value bound for the combination of 0.019. If the first factor is biased at $\Gamma = 11$ and the second factor is biased at $\Gamma = 2$, those two biases are insufficient to prompt acceptance of the null hypothesis of no effect as judged by the combined P-value bound.

```
> truncatedP(c(0.07641043, 0.04883432))
[1] 0.01855308
```

5. Effect modification

Effect modification means that the magnitude of a treatment effect is not constant, but rather is larger for certain values of an observed covariate and smaller for other values. If the covariate has been controlled by matching, the matched pairs may be split into

subgroups of pairs, where one subgroup experiences larger effects than another. Hsu et al. (2013) show that when there is effect modification, subgroup analyses that allow for multiple testing may be much less sensitive to unmeasured biases than a single summary analysis. In particular, they study combining P-value bounds using Zaykin et al. (2002)'s truncated product for two or more nonoverlapping subgroups of pairs. In parallel with §4.3, the calculations for such subgroup analyses may be performed using `senmv`, `senmw`, and `truncatedP`.

6. Better choices of test statistics for sensitivity analyses

6.1 Power of a sensitivity analysis

The power of a sensitivity analysis — its ability to distinguish treatment effects from biases — is affected by the choice of test statistic. Packages `sensitivitymw` and `sensitivitymv` offer several options for statistics with high power in a sensitivity analysis. After reviewing some conceptual issues, §6.2 discusses statistics for matched pairs and §6.3 discusses statistics for matched sets with $K \geq 2$ controls. The discussion here is a brief summary of material from Rosenbaum (2010c, 2013, 2014).

In an observational study, if the treatment has an effect and there is no bias from unmeasured covariates, then the investigator will not know this. The best the investigator can hope to say in this favorable situation is that the study's conclusions are insensitive to small and moderate biases. The power of a sensitivity analysis is the probability that the investigator will be able to say this. More precisely, for a specific Γ, the power of a sensitivity analysis is the probability that the upper bound on the P-value is at most α, conventionally $\alpha = 0.05$. The power is computed under a model in which there is a treatment effect but — unknown to the investigator — no bias from unmeasured covariates. When computed with $\Gamma = 1$, this is the power of a randomized experiment.

In a randomized experiment, some tests are better than others in that they have more power. In parallel, in a sensitivity analysis in an observational study, some tests are better than others in that they have more power for interesting values of Γ. As the sample size, I, increases, $I \rightarrow \infty$, there is a value, $\widetilde{\Gamma}$, called the design sensitivity, such that the power tends to one if the sensitivity analysis is done with $\Gamma < \widetilde{\Gamma}$ and the power tends to zero if $\Gamma > \widetilde{\Gamma}$, so $\widetilde{\Gamma}$ is the limiting sensitivity to unmeasured biases in large samples. For example, if matched pair differences Y_i were independent with Normal distributions having mean $\tau = 1/2$ and variance 1, then Wilcoxon's signed rank statistic has design sensitivity $\widetilde{\Gamma} = 3.17$, the M-test with Huber's ψ-function and $h = 2$ has $\widetilde{\Gamma} = 3.3$, the permutational t-test has $\widetilde{\Gamma} = 3.5$, but an M-statistic with a different ψ-function (ψ_{in} in §6.2) has $\widetilde{\Gamma} = 4.0$; see Rosenbaum (2013, Table 3). This means that if the sensitivity analysis were performed with $\Gamma = 3.6$, then the power of the first three statistics is declining to zero with increasing sample size, $I \rightarrow \infty$, while the power of the fourth statistic is increasing to one. The fourth test statistic ignores Y_i with small $|Y_i|$. Simulations show that the pattern suggested by $\widetilde{\Gamma}$ as $I \rightarrow \infty$ is quite visible in the finite-sample power for $I = 200$.

6.2 Test statistics for matched pairs: inner trimming

In matched pairs, inner trimming means ignoring pairs Y_i with small $|Y_i|$. The ψ-function with inner trimming, ψ_{in}, ignores those Y_i whose $|Y_i|$ is small compared to the scaling constant s. By default in `senmv` and `senmw`, s is the median $|Y_i|$, and by default ψ_{in} ignores Y_i if $|Y_i| < s/2$. A case can be made for ignoring Y_i if $|Y_i| < s$, rather than the default of $s/2$, but this means ignoring half of the pairs, and that will be advantageous only if I is fairly large.

Before describing ψ_{in}, some intuition is helpful. Suppose that the Y_i were independent with Normal distributions having expectation $\tau = 1/2$ and variance 1, and let $A_i = |Y_i|$. If $|Y_i| = 0.01$, then the chance that $Y_i > 0$ is 0.5025, almost a coin flip despite a substantial treatment effect; that is, $\Pr(Y_i > 0 \mid A_i = 0.01) = 0.5025$. A very small bias could explain such a faint signal, in fact a bias of $\Gamma = 0.5025/(1 - 0.5025) = 1.01$ would do it. Now consider $|Y_i| = 2$, so Y_i is either 1.5 standard deviations above its expectation, $\tau = 1/2$, or 2.5 standard deviations below its expectation. Then the chance that $Y_i > 0$ given $|Y_i| = 2$ is $\Pr(Y_i > 0 \mid A_i = 2) = 0.8808$ and the bias that would be needed to explain this is $\Gamma = 0.8808/(1 - 0.8808) = 7.39$. So in distinguishing a treatment effect of $\tau = 1/2$ with $Y_i \sim N(\tau, 1)$ from bias Γ in treatment assignment, a pair with $|Y_i| = 2$ is much more helpful than a pair with $|Y_i| = 0.01$. For more intuition along these lines, see the heuristic graph of the abz-function in Rosenbaum (2010c). The abz-function (for Albers, Bickel and van Zwet 1976) is $\text{abz}(y) = \Pr(Y_i > 0 \mid |Y_i| = y)$ viewed as a function of y. To have high power in a sensitivity analysis, a test statistic should pay close attention to values of y for which $\text{abz}(y)$ is large.

What precisely is ψ_{in}? The permutational t-test has a ψ-function of $\psi_{\text{t}}(y) = y$ and Huber's ψ-function performs outer trimming at h for resistance to outliers with $\psi_{\text{h}}(y) = \text{sign}(y) \cdot \min(|y|, h)$. Let ι be a nonnegative number below h, $0 \leq \iota < h$. The default in `method` = "`i`" of `senmv` and `method` = "`p`" of `senmw` is $\iota =$ `inner` $= 1/2$. The ψ-function with inner trimming, ψ_{in}, is proportional to $\text{sign}(y) \cdot \max\{0, \min(|y|, h) - \iota\}$. Multiplying a ψ-function by a positive constant has no effect on P-values or estimates; however, it may make it easier to compare two ψ-functions. In particular, the constant $h/(h - \iota)$ is helpful. So $\psi_{\text{in}}(y) = \{h/(h - \iota)\} \cdot \text{sign}(y) \cdot \max\{0, \min(|y|, h) - \iota\}$. Then $\psi_{\text{in}}(y) = 0$ for $y \in [-\iota, \iota]$. Also, $\psi_{\text{in}}(y) = h$ for $y \geq h$ and $\psi_{\text{in}}(y) = -h$ for $y \leq -h$; that is, $\psi_{\text{h}}(y) = \psi_{\text{in}}(y)$ for $|y| \geq h$. Finally, between $-h$ and $-\iota$ and between ι and h, $\psi_{\text{in}}(y)$ increases linearly.

Returning to the `mercury` data in §3.2, consider the matched pairs `mercury[,1:2]` formed by comparing column `treated` (i.e., column 1) with ≥ 15 servings of fish in the previous month and column `zero` (i.e., column 2) with 0 servings of fish. At $\Gamma = 17$, the upper bound on the P-value using $\psi_{\text{h}}(y)$ with $h =$ `trim` $= 3$ is 0.064. If $\psi_{\text{in}}(y)$ is used instead with $\iota =$ `inner` $= 1/2$ and $h =$ `trim` $= 3$, the upper bound on the P-value is 0.027 at $\Gamma = 17$ and 0.045 at $\Gamma = 19$. Theory and simulations suggest inner trimming increases the power of a sensitivity analysis. In this one example, rejection of H_0 is insensitive to larger biases if inner trimming is used.

```
> senmw(mercury[, 1 : 2], inner = 0, trim = 3, gamma = 17)$pval
[1] 0.0641723
> senmw(mercury[, 1 : 2], inner = 1/2, trim = 3, gamma = 17)$pval
[1] 0.02658181
> senmw(mercury[, 1 : 2], inner = 1/2, trim = 3, gamma = 19)$pval
[1] 0.04549254
```

6.3 Test statistics for matching with $K \geq 2$ controls: dispersion weighting

To increase the power of a randomization test in randomized block experiments, Tukey (1957), Quade (1979) and Tardiff (1987) suggested giving greater weight w_i to blocks in which the responses R_{ij} are more dispersed. This is analogous to giving greater weight to matched pairs in which $|Y_i| = |R_{i1} - R_{i2}|$ is larger, as in §6.2.

A similar strategy increases the power of a sensitivity analysis in matched sets with a fixed number $K \geq 2$ of controls, as discussed in detail in Rosenbaum (2014). The documentation for `senmw` shows how to reproduce the examples in that paper.

The matched triples in the mercury data were analyzed without weights in §3.2, in particular using `method` = "`h`", obtaining an upper bound on the P-value testing no effect of 0.1188 at $\Gamma = 16$. If weights are used, largely ignoring matched sets i in which R_{i1}, R_{i2}, and R_{i3} are almost the same, using method `method` = "`w`", the upper bound on the P-value is 0.0091 at $\Gamma = 16$ and is 0.0364 at $\Gamma = 19$.

```
> senmw(mercury, method = "h", gamma = 16)$pval
[1] 0.1187912
> senmw(mercury, method = "w", gamma = 16)$pval
[1] 0.009092439
> senmw(mercury, method = "w", gamma = 19)$pval
[1] 0.03644606
```

Again, theory and simulations suggest that a sensitivity analysis will be more powerful if matched sets with little variability are given little weight. In the example, a sensitivity analysis that gave greater weight w_i to matched sets with more dispersed responses reported greater insensitivity to unmeasured biases.

7. Using the data to select a test statistic

As seen above in several examples, the choice of test statistic affects the reported degree of sensitivity to unmeasured biases. One can select a test statistic based on a priori considerations, and advice about how to do that is given in Rosenbaum (2013, 2014). One cannot, however, perform many analyses searching for the least sensitive result, unless one properly takes account of multiple testing.

Heller et al. (2009) suggest splitting the sample at random into a 10% planning sample and a 90% analysis sample, planning the study using the planning sample, discarding the planning sample, and basing the analysis on the untouched analysis sample. In particular, the planning sample may be used to guide the choice of test statistic.

Another approach uses several tests and all of the data, but corrects for multiple testing using the joint distribution of the several tests; see Rosenbaum (2012a,b). This is called "testing one hypothesis twice" or briefly "testing twice". The method in Rosenbaum (2012b) is applicable to M-tests for matched pairs, and can be implemented using the `mvtnorm` package in `R`. Small (2013) developed an `R` package `SensitivityCaseControl` that performs several types of adaptive inference in observational studies, and in particular the function `adaptive.noether.brown` in that package implements the method in Rosenbaum (2012a).

Finally, several tests may be used with adjustment for multiple testing by the Bonferroni inequality. Use of the Bonferroni inequality is easy to implement, because the joint distribution of the several test statistics is not needed, but it is also quite conservative in this context. One reason it is conservative is that several tests of the same null hypothesis using the same data are typically highly positively correlated, and in this case the Bonferroni inequality is quite conservative; see Rosenbaum (2012b). Another reason, relevant to matching with multiple controls, is that bounds on individual P-values may not be jointly attainable; see Rosenbaum and Silber (2009a, §4.5).

Sample splitting, testing twice and use of the Bonferroni inequality are attractive in this context in that they attain the design sensitivity of the best of the several tests under consideration; see Heller et al. (2009) and Rosenbaum (2012b). Testing twice also achieves the best Bahadur efficiency of the several tests; see Berk and Jones (1978) and Rosenbaum (forthcoming).

8. Related articles and packages

The methods used in the two `R` packages are described in Rosenbaum (2007, 2014). For M-statistics, design sensitivity, the power of a sensitivity analysis, and inner trimming (§6) are discussed in Rosenbaum (2013). Weighting matched sets to increase design sensitivity (§6.3) is discussed in Rosenbaum (2014). Using M-statistics with evidence factors is discussed in Rosenbaum (2011).

Two other `R` packages that perform sensitivity analyses in observational studies are Keele's (2014) `rbounds` package and Small's (2013) `SensitivityCaseControl`. Specifically, `rbounds` performs a sensitivity analysis for matched pairs using Wilcoxon's signed rank statistic and the associated Hodges-Lehmann estimates and confidence intervals. Also, `SensitivityCaseControl` implements the adaptive method in Small et al. (2013) for case-control studies with more than one case-definition and the adaptive method in Rosenbaum (2012a).

Appendix I: Comments about Default Settings

The default settings in `senmw` and `senmv` are intended to be safe and familiar, rather than recommended. Theory and simulations favor inner trimming for matched pairs (§6.2) and dispersion weighting for matched sets with $K \geq 2$ controls (§6.3), and I recommend these methods, e.g., `method` = "`p`" for pairs and `method` = "`w`" for sets in `senmw`. Both approaches are less useful, perhaps harmful, if I is small, say $I < 50$, and dispersion weighting is useful with small numbers of controls in each set, say $K = 2$, 3, or 4 controls. The `method` parameter defines useful, reasonable combinations of trimming, inner trimming, scaling,

weighting, etc. The user can take `method = NA` and then specify trimming, inner trimming, etc; however, some combinations are not reasonable, so thought is required.

Appendix II: Contents of the Packages

The main function in package `sensivitymv` is `senmv`. The functions `amplify`, `truncatedP`, and `truncatedPbg` are also directly useful, and may be used in conjunction with functions in package `sensitivitymw`. The data sets in package `sensivitymv` are `erpcp`, `lead150`, `lead250`, `mercury`, `mtm`, and `tbmetaphase`.

The main functions in package `sensitivitymw` are `senmw` and `senmwCI`. The data sets in package `sensitivitymw` are `erpcp` and `mercury`.

Packages `sensivitymv` and `sensitivitymw` share several functions that are called by other functions, specifically `mscorev`, `multrnks`, `newurks`. If you load both packages, during the second load you will see a harmless warning saying that you already have these functions.

Package `sensivitymv` uses function `separable1v` while package `sensitivitymw` uses `separable1k`. When there are matched pairs or a fixed number of controls, `separable1k` is faster than `separable1v`, and this is important primarily when building confidence intervals, because `separable1k` is then called iteratively. Both `separable1k` and `separable1v` perform the asymptotically separable calculation discussed in Gastwirth, Krieger and Rosenbaum (2000). Function `senmwCI` in `sensitivitymw` produces confidence intervals for pairs and for matching with a fixed number of controls, but there is no corresponding function for variable controls in the `sensivitymv` package.

All of the functions and data sets mentioned above are documented in the packages.

References

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

Berk, R. H. and Jones, D. H. (1978). Relatively optimal combinations of test statistics. *Scandinavian Journal of Statistics*, 5:158-162.

Fisher, R. A. (1935). *Design of Experiments*, Edinburgh: Oliver & Boyd.

Gastwirth, J. L., Krieger, A. M. and Rosenbaum, P. R. (2000). Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society* B, 62:545-55.

Heller, R., Rosenbaum, P. R., Small, D. S. (2009). Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association* 104:1090-1101.

Hsu, J. Y., Small, D. S. and Rosenbaum, P. R. (2013). Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association*, 108:135-48.

Huber, P. (1981). *Robust Statistics*, New York: Wiley.

Keele, L. J. (2014) `rbounds`, Version 2.0. (An `R` package).

Maritz, J. S. (1979). Exact robust confidence intervals for location. *Biometrika*, 66:163-166.

Maritz, J. S. (1995). *Distribution-Free Statistical Methods*, London: Chapman & Hall.

Meibian, Z., Zhijian, C., Qing, C. et al. (2008). Investigating DNA damage in tannery workers occupationally exposed to tivalent chromium using the comet assay. *Mutation Research*, 654:45-51.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Statistical Science*, 5:463-480, 1990. (Reprint in English translation of Neyman 1923).

Pitman, E. J. (1937). Statistical tests applicable to samples from any population. *Journal of the Royal Statistical Society*, Supplement 4:119-130.

Quade, D. (1979). Using weighted rankings in the analysis of complete blocks with additive block effects. *Journal of the American Statistical Association*, 74:680-683.

Rosenbaum, P. R. (1995). Quantiles in nonrandom samples and observational studies. *Journal of the American Statistical Association*, 90:1424-1431.

Rosenbaum, P. R. (2002). *Observational Studies* (2^{nd} edition). New York: Springer.

Rosenbaum, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics*, 63:456-464.

Rosenbaum, P. R. and Silber, J. H. (2009a). Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *Journal of the American Statistical Association*, 104:501-511.

Rosenbaum, P. R. and Silber, J. H. (2009b). Amplification of sensitivity analysis in observational studies. *Journal of the American Statistical Association*, 104:1398-1405.

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

Rosenbaum, P. R. (2010b). Evidence factors in observational studies. *Biometrika*, 97:333-345.

Rosenbaum, P. R. (2010c). Design sensitivity and efficiency in observational studies. *Journal of the American Statistical Association*, 105:692-702.

Rosenbaum, P. R. (2011). Some approximate evidence factors in observational studies. *Journal of the American Statistical Association*, 106:285-295.

Rosenbaum, P. R. (2012a). An exact, adaptive test with superior design sensitivity in an observational study of treatments for ovarian cancer. *Annals of Applied Statistics* 6:83-105.

Rosenbaum, P. R. (2012b). Testing one hypothesis twice in observational studies. *Biometrika* 99:763-774.

Rosenbaum, P. R. (2013). Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics*, 69:118-127.

Rosenbaum, P. R. (2014). Weighted M-statistics with superior design sensitivity in matched observational studies with multiple controls. *Journal of the American Statistical Association*, 109:1145-1158.

Rosenbaum, P. R. (forthcoming). Bahadur efficiency of sensitivity analyses in observational studies. *Journal of the American Statistical Association*, published online on 10/1/14, doi: 10.1080/01621459.2014.960968.

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

Small, D. S. (2013). `adaptive.noether.brown`. (A function in R package `SensitivityCaseControl`).

Small, D. S., Cheng, J., Halloran, M. E. and Rosenbaum, P. R. (2013). Case definition and design sensitivity. *Journal of the American Statistical Association*, 108:1457-1468.

Tardif, S. (1987). Efficiency and optimality results for tests based on weighted rankings. *Journal of the American Statistical Association*, 82:637-644.

Tukey, J. W. (1957). Sums of random partitions of ranks. *Annals of Mathematical Statistics*, 28:987-992.

Weiss, N. S. (1981). Inferring causal relationships: elaboration of the criterion of dose-response. *American Journal of Epidemiology*, 113(5):487-490.

Welch, B. L. (1937). On the z-test in randomized blocks. *Biometrika*, 29:21-52.

Werful, U., Langen, V., Eickhoff, I. et al. (1998). Elevated DNA strand breakage frequencies in lymphocytes of welders exposed to chromium and nickel. *Carcinogenesis*, 19:413-418.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. and Weir, B. S. (2002). Truncated product method of combining P-values. *Genetic Epidemiology*, 22:170-85.

Zhang, K., Small, D. S., Lorch, S., Srinivas, S. and Rosenbaum, P. R. (2011). Using split samples and evidence factors in an observational study of neonatal outcomes. *Journal of the American Statistical Association*, 106:511-524.

Zubizarreta, J. R., Neuman, M., Silber, J. H. and Rosenbaum, P. R. (2012). Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia. *Journal of the American Statistical Association*, 107:901-915.

Acknowledgments

Supported by a grant from the US National Science Foundation.