Sensitivity Analysis in Observational Studies

PAUL R. ROSENBAUM

Volume 4, pp. 1809–1814

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sensitivity Analysis in Observational Studies

Randomization Inference and Sensitivity Analysis

Randomized Experiments and Observational Studies

In a randomized experiment (*see* **Randomization**), subjects are assigned to treatment or control groups at random, perhaps by the flip of a coin or perhaps using random numbers generated by a computer [7]. Random assignment is the norm in **clinical trials** of treatments intended to benefit human subjects [21, 22]. Intuitively, randomization is an equitable way to construct treated and control groups, conferring no advantage to either group. At baseline before treatment in a randomized experiment, the groups differ only by chance, by the flip of the coin that assigned one subject to treatment, another to control. Therefore, comparing treated and control groups after treatment in a randomized experiment, if a common statistical test rejects the hypothesis that the difference in outcomes is due to chance, then a treatment effect is demonstrated. In contrast, in an observational study, subjects are not randomly assigned to groups, and outcomes may differ in treated and control groups for reasons other than effects caused by the treatment. Observational studies are the norm when treatments are harmful, unwanted, or simply beyond control by the investigator.

In the absence of random assignment, treated and control groups may not be comparable at baseline before treatment. Baseline differences that have been accurately measured in observed covariates can often be removed by **matching**, **stratification** or model based adjustments [2, 28, 29]. However, there is usually the concern that some important baseline differences were not measured, so that individuals who appear comparable may not be. A sensitivity analysis in an observational study addresses this possibility: it asks what the unmeasured covariate would have to be like to alter the conclusions of the study. Observational studies vary markedly in their sensitivity to hidden bias: some are sensitive to very small biases, while others are insensitive to quit large biases.

The First Sensitivity Analysis

The first **sensitivity analysis** in an observational study was conducted by Cornfield, et al. [6] for certain observational studies of cigarette smoking as a cause of lung cancer; see also [10]. Although the tobacco industry and others had often suggested that cigarettes might not be the cause of high rates of lung cancer among smokers, that some other difference between smokers and nonsmokers might be the cause, Cornfield, et al. found that such an unobserved characteristic would need to be a near perfect predictor of lung cancer and about nine times more common among smokers than among nonsmokers. While this sensitivity analysis does not rule out the possibility that such a characteristic might exist, it does clarify what a scientist must logically be prepared to assert in order to defend such a claim.

Methods of Sensitivity Analysis

Various methods of sensitivity analysis exist. The method of Cornfield, et al. [6] is perhaps the best known of these, but it is confined to binary responses; moreover, it ignores sampling variability, which is hazardous except in very large studies. A method of sensitivity analysis that is similar in spirit to the method of Cornfield et al. will be described here; however, this alternative method takes account of sampling variability and is applicable to any kind of response; see, for instance, [25, 26, 29], and Section 4 of [28] for detailed discussion. Alternative methods of sensitivity analysis are described in [1, 5, 8, 9, 14, 18, 19, 23, 24], and [33].

The sensitivity analysis imagines that in the population before matching or stratification, subjects are assigned to treatment or control independently with unknown probabilities. Specifically, two subjects who look the same at baseline before treatment – that is, two subjects with the same observed covariates – may nonetheless differ in terms of unobserved covariates, so that one subject has an odds of treatment that is up to $\Gamma \geq 1$ times greater than the odds for another subject. In the simplest randomized experiment, everyone has the same chance of receiving the treatment, so $\Gamma = 1$. If $\Gamma = 2$ in an observational study, one subject might be twice as likely as another to receive the treatment because of unobserved pretreatment differences. The sensitivity analysis asks how much hidden bias can be present – that is, how

large can Γ be – before the qualitative conclusions of the study begin to change. A study is highly sensitive to hidden bias if the conclusions change for Γ just barely larger than 1, and it is insensitive if the conclusions change only for quite large values of Γ.

If $\Gamma > 1$, the treatment assignments probabilities are unknown, but unknown only to a finite degree measured by Γ. For each fixed $\Gamma \geq 1$, the sensitivity analysis computes bounds on inference quantities, such as P values or **confidence intervals**. For $\Gamma = 1$, one obtains a single P value, namely the P value for a randomized experiment [7, 16, 17]. For each $\Gamma > 1$, one obtains not a single P value, but rather an interval of P values reflecting uncertainty due to hidden bias. As Γ increases, this interval becomes longer, and eventually it become uninformative, including both large and small P values. The point, Γ, at which the interval becomes uninformative is a measure of sensitivity to hidden bias. Computations are briefly described in Section titled 'Sensitivity Analysis Computations' and an example is discussed in detail in Section titled 'Sensitivity Analysis: Example'.

Sensitivity Analysis Computations

The straightforward computations involved in a sensitivity analysis will be indicated briefly in the case of one standard test, namely Wilcoxon's signed rank test for matched pairs (*see* **Distribution-free Inference, an Overview**) [17]. For details in this case [25] and many others, see Section 4 of [28]. The null hypothesis asserts that the treatment is without effect, that each subject would have the same response under the alternative treatment. There are S pairs, $s = 1, \ldots, S$ of two subjects, one treated, one control, matched for observed covariates. The distribution of treatment assignments within pairs is simply the conditional distribution for the model in Section titled 'Methods of Sensitivity Analysis' given that each pair includes one treated subject and one control. Each pair yields a treated-minus-control difference in outcomes, say D_s. For brevity in the discussion here, the D_s will be assumed to be untied, but ties are not a problem, requiring only slight change to formulas. The absolute differences, $|D_s|$, are ranked from 1 to S, and Wilcoxon's signed rank statistic, W, is the sum of the ranks of the positive differences, $D_s > 0$.

For the signed rank statistic, the elementary computations for a sensitivity analysis closely parallel the elementary computations for a conventional analysis. This paragraph illustrates the computations and may be skipped. In a moderately large randomized experiment, under the null hypothesis of no effect, W is approximately normally distributed with expectation $S(S+1)/4$ and variance $S(S+1)(2S+1)/24$; see Chapter 3 of [17]. If one observed $W = 300$ with $S = 25$ pairs in a randomized experiment, one would compute $S(S+1)/4 = 162.5$ and $S(S+1)(2S+1)/24 = 1381.25$, and the deviate $Z = (300 - 162.5)/\sqrt{(1381.25)} = 3.70$ would be compared to a Normal distribution to yield a one-sided P value of 0.0001. In a moderately large observational study, under the null hypothesis of no effect, the distribution of W is approximately bounded between two Normal distributions, with expectations $\mu_{\max} = \lambda S(S+1)/2$ and $\mu_{\min} = (1-\lambda) S(S+1)/2$, and the same variance $\sigma^2 = \lambda(1-\lambda) S(S+1)(2S+1)/6$, where $\lambda = \Gamma/(1+\Gamma)$. Notice that if $\Gamma = 1$, these expressions are the same as in the randomized experiment. For $\Gamma = 2$ and $W = 300$ with $S = 25$ pairs, one computes $\lambda = 2/(1+2) = 2/3$, $\mu_{\max} = (2/3)\,25\,(25+1)/2 = 216.67$, $\mu_{\min} = (1/3)\,25\,(25+1)/2 = 108.33$, and $\sigma^2 = (2/3)(1/3)\,25\,(25+1)\,(2\times 25+1)/6 = 1227.78$; then two deviates are calculated, $Z_1 = (300 - 108.33)/\sqrt{(1227.78)} = 5.47$ and $Z_2 = (300 - 108.33)/\sqrt{(1227.78)} = 2.38$, which are compared to a Normal distribution, yielding a range of P values from 0.00000002 to 0.009. In other words, a bias of magnitude $\Gamma = 2$ creates some uncertainty about the correct P value, but it would leave no doubt that the difference is significant at the conventional 0.05 level.

Just as W has an exact randomization distribution useful for small S, so too there are exact sensitivity bounds. See [31] for details including S-Plus code.

Sensitivity Analysis: Example

A Matched Observational Study of an Occupational Hazard

Studies of occupational health usually focus on workers, but Morton, Saah, Silberg, Owens, Roberts and Saah [20] were worried about the workers' children. Specifically, they were concerned that workers exposed to lead might bring lead home in clothes and hair, thereby exposing their children as well. They matched 33 children whose fathers worked in a battery factory to 33 unexposed control children of the

Table 1 Blood lead levels, in micrograms of lead per decaliter of blood, of exposed children whose fathers worked in a battery factory and age-matched control children from the neighborhood. Exposed father's lead exposure at work (high, medium, low) and hygiene upon leaving the factory (poor, moderate, good) are also given. Adapted for illustration from Tables 1, 2 and 3 of Morton, et al. (1982). Lead absorption in children of employees in a lead-related industry, *American Journal of Epidemiology* **115**, 549–555. [20]

Pair s	Exposure	Hygiene	Exposed child's Lead level μg/dl	Control child's Lead level μg/dl	Dose Score
1	high	good	14	13	1.0
2	high	moderate	41	18	1.5
3	high	poor	43	11	2.0
⋮	⋮	⋮	⋮	⋮	⋮
33	low	poor	10	13	1.0
Median			34	16	

same age and neighborhood, and used Wilcoxon's signed rank test to compare the level of lead found in the children's blood, measured in μg of lead per decaliter (dl) of blood. They also measured the father's level of exposure to lead at the factory, classified as high, medium, or low, and the father's hygiene upon leaving the factory, classified as poor, moderate, or good. Table 1 is adapted for illustration from Tables 1, 2, and 3 of Morton, et al. (1982) [20]. The median lead level for children of exposed fathers was more than twice that of control children, 34 μg/dl versus 16 μg/dl.

If Wilcoxon's signed rank test W is applied to the exposed-minus-control differences in Table 1, then the difference is highly significant in a one-sided test, $P < 0.0001$. This significance level would be appropriate in a randomized experiment, in which children were picked at random for lead exposure. Table 2 presents the sensitivity analysis, computed as in Section titled 'Sensitivity Analysis Computations'. Table 2 gives the range of possible one-sided significance levels for several possible magnitudes of hidden bias, measured by Γ. Even if the matching of exposed and control children had failed to control an unobserved characteristic strongly related to blood lead levels and $\Gamma = 4.25$ times more common among exposed children, this would still not explain the higher lead levels found among exposed children.

Where Table 2 focused on significance levels, Table 3 considers the one sided 95% confidence interval, $[\widehat{\tau}_{\text{low}}, \infty)$, for an additive effect obtained by inverting the signed rank test [28]. If the data in Table 1 had come from a randomized experiment

Table 2 Sensitivity analysis for one-sided significance levels in the lead data. For unobserved biases of various magnitudes, the table gives the range of possible significance levels

Γ	min	max
1	<0.0001	<0.0001
2	<0.0001	0.0018
3	<0.0001	0.0136
4	<0.0001	0.0388
4.25	<0.0001	0.0468
5	<0.0001	0.0740

Table 3 Sensitivity analysis for one-sided confidence intervals for an additive effect in the lead data. For unobserved biases of biases of various magnitudes, the table gives smallest possible endpoint for the one-sided confidence interval

Γ	$\min \widehat{\tau}_{\text{low}}$
1	10.5
2	5.5
3	2.5
4	0.5
4.25	0.0
5	−1.0

($\Gamma = 1$) with an additive treatment effect τ, then we would be 95% confident that father's lead exposure had increased his child's lead level by $\widehat{\tau}_{\text{low}} = 10.5$ μg/dl [17]. In an observational study with $\Gamma > 1$, there is a range of possible endpoints for the 95% confidence interval, and Table 3 reports the smallest value in the range. Even if $\Gamma = 3$, we would

Table 4 Sensitivity to hidden bias in four observational studies. The randomization test assuming no hidden bias is highly significant in all four studies, but the magnitude of hidden bias that could alter this conclusion varies markedly between the four studies

Treatment	$\Gamma = 1$	(Γ, max P value)
Smoking/Lung Cancer Hammond [11]	<0.0001	(5, 0.03)
Diethylstilbestrol/ vaginal cancer Herbst, et al. [12]	<0.0001	(7, 0.054)
Lead/Blood lead Morton, et al. [20]	<0.0001	(4.25, 0.047)
Coffee/MI Jick, et al. [15]	0.0038	(1.3, 0.056)

be 95% confident exposure increased lead levels by 2.5 μg/dl.

Studies Vary in Their Sensitivity to Hidden Bias

Studies vary markedly in their sensitivity to hidden bias. As an illustration, Table 4 compares the sensitivity of four studies, a study of smoking as a cause of lung cancer [11], a study of prenatal exposure to diethylstilbestrol as a cause of vaginal cancer [12], the lead exposure study [20], and a study of coffee as a cause of myocardial infarction [15].

If no effect is tested using a conventional test appropriate for a randomized experiment ($\Gamma = 1$), the results are highly significant in all four studies. The last column of Table 4 indicates sensitivity to hidden bias, quoting the magnitude of hidden bias $\Gamma \geq 1$ needed to produce an upper bound on the P value close to the conventional 0.05 level. The study [12] of the effects of diethylstilbestrol becomes sensitive at about $\Gamma = 7$, while the study [15] of the effects of coffee becomes sensitive at $\Gamma = 1.3$. A small bias could explain away the effects of coffee, but only an enormous bias could explain away the effects of diethylstilbestrol. The lead exposure study, although quite insensitive to hidden bias, is about halfway between these two other studies, and is slightly more sensitive to hidden bias than the study of the effects of smoking.

Reducing Sensitivity to Hidden Bias

Accurately predicting a highly specific pattern of associations between treatment and response is often said to strengthen the evidence that the effects of the treatment caused the association. For instance, Cook, Campbell, and Peracchio [3] write: 'Conclusions are more plausible if they are based on evidence that corroborates numerous, complex, or numerically precise predictions drawn from a descriptive causal hypothesis.' Hill [13] and Weiss [34] emphasized the role of a dose response relationship. Cook and Shadish [4] write: 'Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations.'

Does successful prediction of a complex pattern of associations affect sensitivity to hidden bias? It may, or it may not, and the degree to which it has done so can be appraised using methods similar to those in Section titled 'Sensitivity Analysis Computations'. See [27] and [30] for methods of analysis, and [32] for issues in research design. The issues will be illustrated using the example in Table 1.

Recall that exposed fathers were classified by their degree of exposure and their hygiene upon leaving the factory. If the fathers' exposure to lead at work were the cause of the higher lead levels among exposed children, then one would expect more lead in the blood of children whose fathers had higher exposure and poorer hygiene. Here, exposed children are divided into three groups of roughly similar size. The 13 exposed children in the category (high exposure, poor hygiene) were assigned a score of 2.0. Low exposure with any hygiene was assigned a score of 1, as was good hygiene with any exposure, and there were 12 such exposed children. The remaining 8 exposed children in intermediate situations were assigned a score of 1.5; they had either high exposure with moderate hygiene or medium exposure with poor hygiene. (None of the 33 matched children had medium exposure with moderate hygiene, although one unmatched child not used here fell into this category.)

The coherent or dose signed rank statistic D gives greater weight to matched pairs with higher doses [27, 30]. Table 5 compares the sensitivity to hidden bias of the usual Wilcoxon signed rank test W, which ignores doses, to the sensitivity of the coherent signed rank statistic. In particular, Table 5 gives the upper bound on the one-sided significance level for testing no effect. For W, this is the same computation as in Table 2. In fact, the coherent pattern of associations is somewhat less sensitive to hidden bias in this example: the upper bound on the

Table 5 Coherent patterns of associations can reduce sensitivity to hidden bias. Upper bounds on one-sided significance levels in the lead data, ignoring and using dose information

Γ	Wilcoxon W	Coherent D
1	<0.0001	<0.0001
3	0.0136	0.0119
4.35	0.0502	0.0398
4.75	0.0645	0.0503

P value for W ignoring doses is just above 0.05 at $\Gamma = 4.35$, but using doses with D the corresponding value is $\Gamma = 4.75$.

Exposed children had higher lead levels than unexposed controls, and also exposed children with higher exposures had higher lead levels than exposed children with lower lead levels. A larger hidden bias is required to explain this pattern of associations than is required to explain the difference between exposed and control children. In short, accurate prediction of a pattern of associations may reduce sensitivity to hidden bias, and whether this has happened, and the degree to which it has happened, may be appraised by a sensitivity analysis.

Acknowledgment

This work was supported by grant SES-0345113 from the US National Science Foundation.

References

[1] Berk, R.A. & De Leeuw, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design, *Journal of the American Statistical Association* **94**, 1045–1052.

[2] Cochran, W.G. (1965). The planning of observational studies of human populations (with Discussion), *Journal of the Royal Statistical Society, Series A* **128**, 134–155.

[3] Cook, T.D., Campbell, D.T. & Peracchio, L. (1990). Quasi-experimentation, in *Handbook of Industrial and Organizational Psychology*, M. Dunnette & L. Hough, eds, Consulting Psychologists Press, Palo Alto, pp. 491–576.

[4] Cook, T.D. & Shadish, W.R. (1994). Social experiments: some developments over the past fifteen years, *Annual Review of Psychology* **45**, 545–580.

[5] Copas, J.B. & Li, H.G. (1997). Inference for non-random samples (with discussion), *Journal of the Royal Statistical Society* **B 59**, 55–96.

[6] Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. & Wynder, E. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions, *Journal of the National Cancer Institute* **22**, 173–203.

[7] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.

[8] Gastwirth, J.L. (1992). Methods for assessing the sensitivity of statistical comparisons used in title VII cases to omitted variables, *Jurimetrics* **33**, 19–34.

[9] Gastwirth, J.L., Krieger, A.M. & Rosenbaum, P.R. (1998b). Dual and simultaneous sensitivity analysis for matched pairs, *Biometrika* **85**, 907–920.

[10] Greenhouse, S. (1982). Jerome Cornfield's contributions to epidemiology, *Supplement of Biometrics* **38**, 33–45.

[11] Hammond, E.C. (1964). Smoking in relation to mortality and morbidity: findings in first thirty-four months of follow-up in a prospective study started in 1959, *Journal of the National Cancer Institute* **32**, 1161–1188.

[12] Herbst, A., Ulfelder, H. & Poskanzer, D. (1971). Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women, *New England Journal of Medicine* **284**, 878–881.

[13] Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[14] Imbens, G.W. (2003). Sensitivity to exogeneity assumptions in program evaluation, *American Economic Review* **93**, 126–132.

[15] Jick, H., Miettinen, O., Neff, R., Jick, H., Miettinen, O.S., Neff, R.K., Shapiro, S., Heinonen, O.P., Slone, D. (1973). Coffee and myocardial infarction, *New England Journal of Medicine*, **289**, 63–77.

[16] Kempthorne, O. (1952). *Design and Analysis of Experiments*, John Wiley & Sons, New York.

[17] Lehmann, E.L. (1998). *Nonparametrics: Statistical Methods Based on Ranks*, Prentice Hall, Upper Saddle River.

[18] Lin, D.Y., Psaty, B.M. & Kronmal, R.A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies, *Biometrics* **54**, 948–963.

[19] Manski, C. (1995). *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge.

[20] Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M. & Saah, M. (1982). Lead absorption in children of employees in a lead-related industry, *American Journal of Epidemiology* **115**, 549–555.

[21] Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J. & Smith, P. (1976). Design and analysis of randomised clinical trials requiring prolonged observation of each patient, I, *British Journal of Cancer* **34**, 585–612.

[22] Piantadosi, S. (1997). *Clinical Trials*, Wiley, New York.

[23] Robins, J.M., Rotnitzkey, A. & Scharfstein, D. (1999). Sensitivity analysis for selection bias and unmeasured

confounding in missing data and causal inference models, in *Statistical Models in Epidemiology*, E. Halloran & D. Berry, eds, Springer, New York, pp. 1–94.

[24] Rosenbaum, P.R. (1986). Dropping out of high school in the United States: an observational study, *Journal of Educational Statistics* **11**, 207–224.

[25] Rosenbaum, P.R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies, *Biometrika* **74**, 13–26.

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

[27] Rosenbaum, P.R. (1997). Signed rank statistics for coherent predictions, *Biometrics* **53**, 556–566.

[28] Rosenbaum, P.R. (2002a). *Observational Studies*, Springer, New York.

[29] Rosenbaum, P.R. (2002b). Covariance adjustment in randomized experiments and observational studies (with Discussion), *Statistical Science* **17**, 286–327.

[30] Rosenbaum, P.R. (2003a). Does a dose-response relationship reduce sensitivity to hidden bias? *Biostatistics* **4**, 1–10.

[31] Rosenbaum, P.R. (2003b). Exact confidence intervals for nonconstant effects by inverting the signed rank test, *American Statistician* **57**, 132–138.

[32] Rosenbaum, P.R. (2004). Design sensitivity in observational studies, *Biometrika* **91**, 153–164.

[33] Rosenbaum, P.R. & Rubin, D.B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome, *Journal of the Royal Statistical Society Series B* **45**, 212–218.

[34] Weiss, N. (1981). Inferring causal relationships: elaboration of the criterion of 'dose-response.', *American Journal of Epidemiology* **113**, 487–90.

PAUL R. ROSENBAUM