10. Statistical Conclusion Validity

It cannot be overemphasized that to be useful, each account of how the data are generated has to be a good approximation of what really happened.

The account derived from design-based sampling must correspond to how the probability sampling plan was implemented in the field. Information is often available to help make this assessment (e.g., reported response rates). Model-based accounts can be far more difficult to evaluate because thorough and accurate information on how nature actually proceeded is required. Sometimes there is useful information available. For example, interviews with experimental and control subjects may provide insights about why some individuals chose the intervention and why other individuals chose the comparison condition. But too often the information one would need to make confident assessments is not available. Then researchers often fall back on their disciplinary theory, which is too often insufficiently precise or convincing.

Given a strong case for the particular chance mechanism by which uncertainty in the data has been introduced, confidence intervals and formal statistical tests can follow. The tests need to be explicitly formulated before the data are examined. Hypotheses that are stated after a look at the data undermine the p-values that follow. Generally, the p-values will be too small; there is false power. Another common error is a failure to discount p-values after multiple statistical tests. The problem is that the researcher is capitalizing on chance. For example, one in twenty tests will on the average be "statistically significant" at the .05 level when the null hypothesis is true. There are many interesting proposals for how to constrain this "false discover rate," the details of which are beyond the scope of this discussion (See, for example, Banjamini and Hochberg, 1995).

10. Statistical Conclusion Validity

A deeper problem is the use of statistical tests after model selection procedures are applied.

Basically, the tests will have unknown properties and cannot be relied upon; the model winnowing process invalidates the tests. As Leeb and Pötscher (2006):2554 observe,

> "...a post-model-selection estimator here refers to the combined procedure resulting from first selecting a model (e.g., by a model selection criterion such as AIC or by a hypothesis testing procedure) and then estimating the parameters of the selected model (e.g., by least-squares or maximum likelihood), all based in the same data. We show that it is impossible to estimate this distribution with reasonable accuracy, even asymptotically."

The best response to this problem is to have a training data set with which to build the statistical model and a test data set with which to undertake any statistical inference. Ideally the two data sets would be random samples from the same population or random realizations of the same data generating process. If one has a large enough data set on hand, an equally good strategy is to randomly divide the data into two parts and treat one part as a training sample and the other part as a test sample.

Putting all this together leads to the following reporting suggestions. One should report:

1. The account being used to characterize the uncertainty (e.g., design-based sampling);
2. The names of any tests used;
3. The null and alternative hypotheses for any statistical tests;
4. Any distributional assumption being made and their credibility for the data on hand;
5. The actual p-values of any statistical tests;
6. The degrees of freedom for any statistical tests;
7. Any model selection procedures used before the reported tests; and
8. The methods used to adjust for multiple tests.

11. Summary

It is difficult to draw convincing conclusions from observational studies. There are many potential pitfalls and many research decisions for which clear methodological advice does not exist. The reporting burdens are, therefore, far heavier than for studies that can proceed largely by well-accepted recipes. Beyond the reporting suggestions discussed above, there will often be additional matters of a study-specific nature that will need to be disclosed. Good practice depends on reporting anything about the data collection and analysis that could materially affect the findings.

12. References

Angrist, J. (1990). "Lifetime earnings and the Vietnam era draft lottery: Evidence from social security records." *The American Economics Review.* 83(3): 313-336.

Benjamini, Y., and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society.* (Series B) 57(1): 289-300.

Berk, R.A. (2003). *Regression analysis: A constructive critique.* Sage Publications, Newbury Park, CA.

Berk, R.A., and 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(448): 1045-1052.

Cochran, W.G., (1983). *Planning & analysis of observational studies.* New York: John Wiley and Sons.

Cook, R.D. and Weisberg, S. (1999). *Applied regression including computing and graphics.* New York: John Wiley and Sons.

Freedman, D.A. (2005a). *Statistical models: Theory and practice.* Cambridge University Press, Cambridge.

Freedman, D.A. (2005b). "Statistical models for causation: What inferential leverage do they provide?" *Evaluation Review.* 30(6): 691-713.

Freedman, D.A. (2008a) "On regression adjustments to experimental data." *Advances in Applied Mathematics.* 40(2): 180-193.

Freedman, D.A. (2008b). "Diagnostics cannot have much power against general alternatives." Working paper at Berkeley.

Freedman, D.A., and Berk, R.A. (2008). "Weighting regressions by propensity scores." *Evaluation Review.* 32(4): 392-409.

Galster, G., Temkin, K., Walker, C., and Sawyer, N. (2004). "Measuring the impacts of community development initiatives: A new application of the adjusted interrupted time-series method." *Evaluation Review.* 28(6): 502-538.

Hoenig, J.M. and Heisey, D.M. (2001). "The abuse of power: The pervasive fallacy of power calculation for data analysis." *The American Statistician.* 55: 19-24.

Holland, P. (1986). "Statistics and causal inference." *Journal of the American Statistical Association* 8: 945-60.

Imbens, G., (2004). "Nonparametric estimation of average treatment effects under exogeneity: A review." *Review of Economics and Statistics.* 86: 4-30.

Klein, S., Benjamin, R., and Bolus, R. (2007). "The collegiate learning assessment: facts and fantasies." *Evaluation Review.* 31(5): 415-439.

Layzer, J.I., and Goodson, B.D. (2006). "The 'Quality' of Early Care and Education Settings" *Evaluation Review.* 30(5): 556-576.

Leeb, H., and Potscher, B.M. (2006). "Can one Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *The Annals of Statistics.* 34(5): 2554-2591.

Mamun, A. (2003). *Life history of cardiovascular disease and Its risk factors - multistate life table approach and application to the framingham heart study.* Amsterdam: Rozenberg Publishers.

Morgan, S.L., and Winship, C. (2007). *Counterfactuals and causal inference: Methods and principle for social research.* Cambridge University Press, Cambridge.

Neyman, Jerzy. 1923 [1990]. “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles.” *Statistical Science.* 5 (4): 465–472. Translation by Dorota M. Dabrowska and Terence P. Speed.

Rosenbaum, P.R. (2002). *Observational studies, Second Edition.* New York: Springer.

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

Rush, B.R., Dennis, M.L., Scott, C.K., Castel, S., and Funk, R.R. (2008). "The interaction of co-occurring mental disorders and recovery management checkups on substance abuse treatment participation and recovery." *Evaluation Review.* 32(1): 7-38.

Vigdor, E.R., and Mercy, J.A. (2006). "Do laws restricting access to firearms by domestic violence offenders prevent intimate partner homicide." *Evaluation Review* .30(3): 313-346.

Wolf, E.M., and Wolf, D.A. (2008). "Mixed results in a transitional planning program for alternative school studies." *Evaluation Revie*w. 32 (2): 187-215.

13. Author Biography

Richard Berk, PhD, formerly a Distinguished Professor of Statistics at UCLA, is a Professor of Criminology and Statistics at the University of Pennsylvania. He works on a wide variety of issues in criminology: inmate classification and placement systems, law enforcement strategies for reducing domestic violence, the role of race in capital punishment, detecting violations of environmental regulations, claims that the death penalty serves as a general deterrent, and forecasting short-term changes in urban crime patterns. He is equally active on a range of methodological concerns: causal inference, statistical learning, and methods for evaluating social programs. Professor Berk is an elected fellow of the American Association for the Advancement of Science, The American Statistical Association, and the Academy of Experimental Criminology. He has published 13 books and over 150 book chapters and articles. His most recent book is the controversial *Regression Analysis: A Constructive Critique* (Sage Publications, 2004). He currently finished up another book, *Statistical Learning from a Regression Perspective*, published in the Springer Series in Statistics, 2008.