Black Ravens, White Shoes, and Case Selection

David A. Freedman

1. Introduction

How should qualitative researchers select cases? This is an important question, which has been widely canvassed. Mahoney and Goertz (2004) offer some principles to govern case selection, illustrating the argument by Hempel's raven paradox. In this paper, I suggest resolving the paradox by distinguishing between samples and populations. I also suggest that the Mahoney-Goertz rules have limited scope.

2. The paradox

The raven paradox is due to Carl Hempel (1945). To explain it, suppose that objects can be classified unambiguously as

(i) raven or not, and
(ii) black or not.

The data can then be presented in a 2×2 table, with columns corresponding to the first classification and rows to the second. For reference, the cells are labeled A, B, C, D. All four cells are observed.

	Raven	
Black	Yes	No
Yes	A	B
No	C	D

Now consider the time-honored proposition that all ravens are black. According to Jean Nicod (1930) and many scholars who followed him, data in cell A support the proposition. In other words, a black raven is evidence that all ravens are black. As Hempel notes, however, "all ravens are black" is logically equivalent to "all nonblack objects are nonravens."[1] Thus, by Nicod's rule, data in cell D—nonblack objects that are nonravens—also support the blackness of ravens.

In particular, white shoes provide evidence that ravens are black. Many of us find this paradoxical, although Hempel seems eventually to have accepted the idea. There is an extended philosophical literature on white shoes and ravens, including an exchange between I. J. Good (1967, 1968) and Hempel (1968):

"The White Shoe Is a Red Herring,"
"The White Shoe: No Red Herring,"
"The White Shoe *Qua* Herring Is Pink."

The debate has spilled over into the political science journals (see, for instance, *Political Analysis* Spring 2002; 10: 178–207). The paradox is also discussed by Taleb (2007), in a searching critique of current statistical methodology.[2]

I believe the paradox should be resolved by making the following distinction. The proposition "all ravens are black" can be advanced with respect to

(i) the data at hand; or,
(ii) some larger population of objects, the data at hand being viewed as a sample from the larger population.

In the first case, what matters is the raven-nonblack cell—C in the table. If this cell is empty, the proposition is correct; if this cell is nonempty, the proposition is incorrect. Other cells in the table are simply irrelevant.[3] Nicod's rule does not apply, and white shoes are beside the point.

On the other hand, if the assertion is about some larger population, and statistical inferences are to be made from the data to the population, then the nature of the sample and the population must be specified (the "sampling model"). In this scenario, "all" is defined relative to the larger population; so is the set of objects that are not ravens, as well as the set of objects that are not black.

Nicod's rule applies in some sampling models but not others. White shoes may be powerful evidence for the blackness of ravens, or against—

or shoes may be entirely irrelevant. Good (1967) has a cunning example where seeing a black raven increases the likelihood that white ravens will turn up later: see the Appendix below. Hempel (1968) and the rejoinder by Good (1968) gum up the works with herrings of various colors.

To summarize, the illusion of paradox is created by blurring the distinction between the sample and the population. The illusion is dispelled by deciding whether we are discussing the data at hand, or extrapolating from the data to a larger population—although, in the second case, a sampling model is needed.

2. Case selection

Enough about ravens, shoes, and herrings; what about principles for case selection? Mahoney and Goertz (2004, p. 653) claim their

> "Possibility Principle ... provides explicit, rigorous, and theoretically informed guidelines for choosing a set of negative cases .... The Possibility Principle holds that only cases where the outcome of interest is *possible* should be included in the set of negative cases; cases where the outcome is *impossible* should be relegated to a set of uninformative and hence irrelevant observations."

The possibility principle is elaborated into a rule of exclusion and a rule of inclusion, the former being primary (Mahoney and Goertz 2004, pp. 657–8). These rules will be explained below. They sometimes provide useful heuristics for case selection. However, if the principles are supposed to have general application, they leave something to be desired. In particular, claims of explicitness and rigor are not justified.

The setting has a binary response variable Y, where $Y = 1$ indicates the presence of an outcome of interest; $Y = 0$ indicates its absence. There are binary independent variables, which may be causes of Y. Thus, $X = 1$ indicates the presence of a causal factor, whereas $X = 0$ indicates absence. Mahoney and Goertz are using language in a specialized way, because "impossible" things occur with some frequency. Impossibility, in their terminology, only means that the likelihood is below a selected cutpoint. Consequently, scholars who want to use the Mahoney-Goertz rules must assign likelihoods, choose cutpoints and then dichotomize. For example, "impossibility" might just mean that the likelihood is below the cutpoint of .5 (Mahoney and Goertz, pp. 659, 663).[4]

Claims for explicitness and rigor are therefore questionable. Quantifying likelihoods, even in large-N research, is fraught with difficulty. Logit models can of course be fitted to data, but rigorous justification for such models is rarely to be found.[5] Selecting cutpoints is another famous problem.[6] Smaller N does not make life easier.

With respect to defining likelihoods and cutpoints, Mahoney and Goertz (2004, p. 665) say only, "These tradeoffs underscore the importance of making substantively and theoretically informed choices about where to draw the line. . . ." This sound advice will not help when making hard choices. In short, quantifying likelihoods and choosing cutpoints is not an objective process; the claim to have formulated explicit and rigorous guidelines is not justified. Moreover, contrary to suggestions by Mahoney and Goertz, it would appear that the theory informing their guidelines must be supplied by the scholars who use those guidelines.

Another problem should be mentioned. Presence or absence of an outcome of interest seems clear enough in many circumstances. In other circumstances, however, difficulties abound. For example, consider a study showing that left-wing political power promotes economic growth. Scholars with another orientation will use the same data to prove that left-wing power promotes stagnation. Is the outcome of interest growth—or stagnation?

The answer determines which cases are positive and which are negative. The empirical relationship being tested is substantively the same, but different cases will be deemed relevant and irrelevant by the Mahoney-Goertz rules, according to the way the research hypothesis is framed (see the Appendix below for details). In short, if we follow the rules, the relevance of a case is likely to depend on arbitrary choices.

Suppose, however, that such ambiguities have been resolved. There is a binary response variable Y. The outcome of interest is coded as $Y = 1$; negative cases have $Y = 0$. There is one causal variable X, with $X = 0$ or 1. The data can be presented in the the following 2×2 table.

	X	
Y	1	0
1	A	B
0	C	D

Labels for the cells are shown in the body of the table. Our working hypothesis is that X and Y are positively related: setting X to 1 increases the likelihood that $Y = 1$.

Cases in cell D are irrelevant by the Mahoney-Goertz rule of exclusion:

> "Cases are irrelevant if their value on any eliminatory independent variable predicts the nonoccurrence of the outcome of interest" (658).

Indeed, cases in cell D (with $X = 0$ and $Y = 0$) are negative. Furthermore, an eliminatory independent variable predicts the nonoccurrence of the outcome of interest ($X = 0$ predicts $Y = 0$). Cell D is therefore irrelevant.

Moreover, cell D is also irrelevant by the rule of inclusion:

> "Cases are relevant if their value on at least one independent variable is positively related to the outcome of interest" (p. 657).

Indeed, $X = 0$ in cell D. Next, the value 0 for the independent variable X is not positively related to the outcome of interest ($Y = 1$). Finally, in our setup, there are no other variables to consider. Therefore, the Mahoney-Goertz rule of inclusion, like their rule of exclusion, says that cell D is irrelevant.[7]

Cell D may indeed be irrelevant under some circumstances. But a blanket assertion of irrelevance seems hasty. For example, most statisticians and epidemiologists would want to know about all four cells—if only to confirm that the association is positive, and to determine its magnitude.

We can make this more interesting (and more complicated). Suppose an observer claims there are two types of cases in cell D. For the first type of case, $X = 0$ causes $Y = 0$. For the second type, $Y = 0$ by necessity: in other words, Y would still have been 0 even if we had set X to 1. This is causal heterogeneity. The best way to test such a claim, absent other information, would seem to be scrutiny of cases with $X = 0$ and $Y = 0$. In this kind of scenario, far from being irrelevant, cell D can be critical.

An example with only one important causal variable may seem unusual, but the reasoning about the rule of exclusion continues to apply if there are several variables. For the rule of inclusion, condition on all the covariates but one; then use the argument given above to conclude that some of the cells in the multi-dimensional cross-tab are irrelevant. This is not a sensible conclusion; the reasons stay the same, no matter how many variables are in play. Therefore, the rules of exclusion and inclusion are not good general rules.

Mahoney and Goertz may be thinking of necessary and sufficient causation, although this is not made clear. Let us assume, which would be highly favorable to the enterprise, that there is only one causal variable and no cases in cell B or cell C. If cell D is empty, there is no variance on X or on Y, which will affect the interpretation of the data for some observers. If cells A and D are both nonempty, qualitative researchers will want to examine some cases in each cell, in order to check that the association is causal, and to discern the mechanisms by which $X = 1$ causes $Y = 1$, whereas $X = 0$ causes $Y = 0$. So, the cell with $X = 0$ and $Y = 0$ is worth considering even for necessary and sufficient causation.

A real example might be useful. In their multi-methods research on the probabilistic causes of civil war, Fearon and Laitin (2008) found it illuminating to examine cases in the analog of cell D (low probability of civil war according to the model, and no civil war in historical fact).

Fearon and Laitin contradict the Mahoney-Goertz rules. In summary, general advice to disregard any particular cell in the 2×2 table is bad advice.

Appendix

Good's example. We begin by sketching Good's construction. With probability 1/2, the population comprises 100 black ravens and 1,000,000 birds that are not ravens; with probability 1/2, the population comprises 1000 black ravens, 1 white raven, and 1,000,000 birds that are not ravens. The population is chosen at random, then a bird is selected at random from the chosen population. If the bird is a black raven, it is likely to have come from the second population. In short, a black raven is evidence that there is a white raven to be seen (eventually).

Simple random samples. We turn to more familiar sampling models. Suppose that a sample is chosen at random without replacement from a much larger population, each object in the population being classified as U or not-U. For instance, the U's might be the sought-after white ravens, so the not-U's comprise red ravens, green ravens, blue ravens, ..., and black ravens, together with nonravens.

From a Bayesian perspective, it is easy to test the hypothesis that there are no U's in the population. However, much depends on the prior that is used, and justifying the choice can be difficult (Freedman 1995; Freedman and Stark 2003).

Now take the frequentist perspective. If the fraction of U's in the sample is small, that proves U is rare in the population (modulo the usual qualifications). However, unless we make further assumptions, it is impossible to demonstrate by sampling theory that there are no U's in the population. For instance, if the sample size is 1000 and the fraction of U's in the population is 1/1000, there is a substantial chance that no U's will turn up in the sample: the chance is $\left(1 - \frac{1}{1000}\right)^{1000} \doteq 0.37$. So, if there are no U's in the sample, we are entitled to conclude that U is rare—but we cannot conclude that there no U's in the population.

Other possibilities. The two examples below indicate other logical possibilities. For the sake of variety, white shoes are replaced by red herrings. In the first example, *pace* Hempel, a red herring is decisive evidence that not all ravens are black. In the second, by contrast, a red herring is decisive evidence that all ravens are indeed black.

A "population" consists of objects classified as white ravens, black ravens, red herrings, and other things (neither raven nor herring). Different populations have different compositions; however, there are black ravens and things that are neither raven nor herring in every population.

Each example consists of two populations, labeled Population I and Population II. A sample is drawn at random from one of the two populations. It is unknown which population is being sampled. It is required to decide whether, in the population being sampled, all ravens are black.

Example 1. In Population I, there are both white ravens and red herrings. In Population II, there are neither white ravens nor red herrings. If a red herring turns up in the sample, you must be sampling from Population I containing white ravens. This is a useful clue if there are a lot of red herrings and few white ravens.

Example 2. In Population I, there are white ravens but no red herrings. In Population II, there are no white ravens but there are red herrings. If a red herring turns up in the sample, you must be sampling from Population II, where all ravens are black.

So far, we have considered simple random samples. Different kinds of samples are often used, including convenience samples. Procedures that favor some cells at the expense of others can easily skew the data. Sample design is a crucial piece of the puzzle. If you do not look, you will not find evidence against your hypothesis.

Samples and inductive inference. I have focused on inductive inference by sampling, without meaning to imply that statistical theory is the only basis for induction. On the contrary, I believe that in most cases, statistical theory—whether frequentist or Bayesian—permits inductive inference only by imposing artifical assumptions. The frequentist incantation is "independent and identically distributed." The Bayesian denounces frequentists for incoherence, requiring instead that observations are exchangeable—a distinction of Talmudic subtlety (Freedman 1995; Freedman and Stark 2003). How then are scientists to make inductive inferences? That is a topic for another lifetime, but maybe we could start by thinking about what they actually do.

The ravens and causal inference. As I see it, the paradox of the ravens has to do with description and inductive reasoning. Others may see the paradox as being about logic and semantics. What should be blatantly obvious is that the paradox has nothing to do with causal inference per se—which is not to deny that causal reasoning depends on description, classification, induction, logic, and ordinary language.

Ambiguity in the rules. Finally, let us consider the example of left-wing political power and economic growth. Cases can be arrayed in the familiar 2×2 table:

	Growth	Stagnation
Left-wing power	A	B
Right-wing power	C	D

One perspective is that left-wing power causes growth. Then growth is the outcome of interest. As argued above, the Mahoney-Goertz rules imply that cell D is irrelevant. Another perspective is that left-wing power causes stagnation. Now stagnation is the outcome of interest, and it is cell C (negative on outcome, negative on left-wing power) that is irrelevant. This is untidy at best.

Mahoney and Goertz might agree that positive cases are generally relevant. Now there is something of a contradiction. If the research hypothesis is formulated to please the left wing, cell C is relevant, because it is positive. If the hypothesis is formulated to humor the right, cell C is irrelevant, as shown in the previous paragraph.

The odds ratio. Epidemiologists would use the "odds ratio" to summarize the data in a 2×2 table of the kind we have been considering. Let a denote the number of elements in cell A, and so forth. If there are cases in all four cells, the odds ratio is $(a/c)/(b/d) = (a/b)/(c/d) = (ad)/(bc)$. You need all four numbers to compute the odds ratio. The association is positive when the odds ratio is above 1.0; the association is negative when the odds ratio is below 1.0. For additional information, see Gordis (2008).

If ρ denotes the odds ratio, the causal interpretation is this: setting X to 1 rather than 0 multiplies the odds that $Y = 1$, by the factor ρ. Equivalently, if $Y = 1$ rather than 0, the odds that $X = 1$ are multiplied by the factor ρ. In the present context, given a, b, and c, it is cell D that determine whether X causes Y or X prevents Y—a substantial difference. Cell D is not be ignored.

Notes

1. Suppose A and B are sets. Write A^c for the complement of A, i.e., the set of things that are not in A. The logical principle is this:

$$A \text{ is a subset of } B$$

if and only if

$$B^c \text{ is a subset of } A^c.$$

2. Taleb argues that rare events ("Black Swans") have major consequences, and conventional statistical models are ill-suited for analyzing such matters. Efforts by statisticians to refute him have so far been unconvincing (*The American Statistician* August 2007; 61: 189–200).

3. We can either assume there is at least one black raven, or rely on an irritating logical technicality—an empty set is a subset of all sets. In particular, if there are no ravens, they must all be black (as well as any other color of interest).

4. As Mahoney and Goertz (2004, p. 662) explain,"the impossible . . . is very likely to happen in large-N research," that is, with enough cases. To rephrase the rules in terms of the possible rather than the impossible, you have to quantify the probability that $Y = 1$, then choose a cutpoint, and then declare that $Y = 1$ is "possible" if the probability falls above that cutpoint. Compare Mahoney and Goertz (2004, pp. 659–660, 663–65). "[T]he analyst must decide and justify the exact threshold at which the outcome is considered possible" (659). There are similar considerations for the explanatory variables.

5. See Berk (2004), Brady and Collier (2004), Duncan (1984), Freedman (2005), Lieberson and Lynn (2002), Mahoney and Rueschemeyer (2003), Sobel (1998).

6. Cournot (1843) discusses the impact of choosing categories. See Stigler (1986, 199) for a summary, or Shaffer (1995).

7. Mahoney and Goertz (2004, p. 658) might suggest that X is not an eliminatory variable in their sense. This is far from clear, especially in view of the claim that "observations with a zero for all the independent variables will always satisfy causal sufficiency and thus artificially inflate the number of cases where the theory works. . ." (p. 664). In any event, this suggestion would not explain the paradoxical implications of the rule of inclusion.

Acknowlegments

I would like to thank David Collier (Berkeley), Thad Dunning (New Haven), Paul Humphreys (Charlottesville), Janet Macher (Berkeley), Jay Seawright (Evanston), Philip Stark (Berkeley), and Jas Sekhon (Berkeley) for useful comments.

References

Berk (2004), R. A. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage Publications.

Brady, H. E. and Collier, D., eds. (2004). *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield Publishers, Inc.

Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilités*. Paris: Hachette. Reprinted in B. Bru, ed. (1984). *A. A. Cournot, Oeuvres complètes*. Vol. 1. Paris: J. Vrin.

Duncan, O. D. (1984). *Notes on Social Measurement*. New York: Russell Sage.

Fearon, J. D. and Laitin, D. D. (2008). Integrating qualitative and quantitative methods. In J. M. Box-Steffensmeier, H. E. Brady, and D. Collier, eds. *The Oxford Handbook of Political Methodology*. Oxford University Press, pp. 756–76.

D. A. Freedman (1995). Some issues in the foundation of statistics. *Foundations of Science* 1: 19–83 (with discussion). Reprinted in B. C. van Fraassen, ed. (1997). *Some Issues in the Foundation of Statistics*. Dordrecht, The Netherlands: Kluwer.

Freedman, D. A. (2005). *Statistical Models: Theory and Practice*. New York: Cambridge University Press.

Freedman, D. A. and Stark, P. B. (2003). What is the probability of an earthquake? In F. Mulargia and R. J. Geller, eds. *Earthquake Science and Seismic Risk Reduction*. NATO Science Series IV: Earth and Environmental Sciences, vol. 32. Dordrecht, The Netherlands: Kluwer, pp. 201–13.

Good, I. J. (1967). The white shoe is a red herring. *The British Journal for the Philosophy of Science* 17: 322.

Good, I. J. (1968). The white shoe *qua* herring is pink. *The British Journal for the Philosophy of Science* 19: 156–57.

Gordis, L. (2008). *Epidemiology*. 4th ed. Philadelphia: Elsevier-Saunders.

Hempel, C. G. (1945). Studies in the logic of confirmation. *Mind* 54: 1–26, 97–121.

Hempel, C. G. (1967). The white shoe: No red herring. *The British Journal for the Philosophy of Science* 18: 239–40.

Lieberson, S. and Lynn, F. B. (2002). Barking up the wrong branch: Alternatives to the current model of sociological science. *Annual Review of Sociology* 28: 1–19.

Mahoney, J. and Goertz, G. (2004). The possibility principle: Choosing negative cases in comparative research. *The American Political Science Review* 98: 653–69.

Mahoney, J. and Rueschemeyer, D. (2003). *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press.

Nicod, J. (1930). *Foundations of Geometry and Induction*. Translated from the French by P. P. Wiener. New York: Harcourt Brace.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology* 46: 561–84.

Sobel, M. E. (1998). Causal inference in statistical models of the process of socioeconomic achievement—A case study. *Sociological Methods & Research* 27: 318–48.

Stigler, S. M. (1986). *The History of Statistics*. Cambridge, MA: Harvard University Press.

Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable*. New York: Random House.