Investigators led astray by nonrandom samples.

A Selection of Selection Anomalies

Howard Wainer, Samuel Palmer, and Eric T. Bradlow

During the 1984 U.S. presidential campaign, the Democratic vice-presidential nominee, Geraldine Ferraro, was asked how she could persevere in the face of very discouraging poll results. She said, "I don't believe those polls. If you could see the enthusiasm for our candidacy out there, you wouldn't believe them either." Of course, part of her response must have been political hyperbole, but even after the election, when the polls' predictions proved to be accurate, she remained dismayed by the results. Why? It was difficult for Ferraro to get an accurate reading of her popularity in the general population from the enthusiasm she saw at Democratic gatherings. The reason for this difficulty is that the individuals who showed up at such gatherings chose to do so. They were a long way from a random sample from the population of voters. Errors of inference obtained from a nonrandom sample did not originate with Ferraro. The telephone poll taken by *The Literary Digest* in 1948 predicting a Dewey victory over Truman is a well-known precursor; at that time many more Republican voters than Democrats had phones.

In this essay we illustrate four circumstances in which nonrandom samples could lead investigators far astray. These investigators range from a 19th-century Swiss physician to modern educational theorists. The fifth example we present illustrates one way to draw correct inferences from a nonrandom sample with Abraham Wald's ingenious model for aircraft armoring. We conclude with a discussion of multiple imputations as a tool for assessing the uncertainty due to nonrandom selection.

Example 1: The Most Dangerous Profession

In 1835 the Swiss physician H. C. Lombard published the results of a study on the longevity of various professions. His data were very extensive, consisting of death certificates gathered over more than a half century in Geneva. Each certificate contained the name of the deceased, his profession, and age at death. Lombard used these data to calculate the mean longevity associated with each profession. Lombard's methodology was not original with him, but instead was merely an extension of a study carried out by R. R. Madden, published two years earlier. Lombard found that the average age of death for the various professions mostly ranged from the

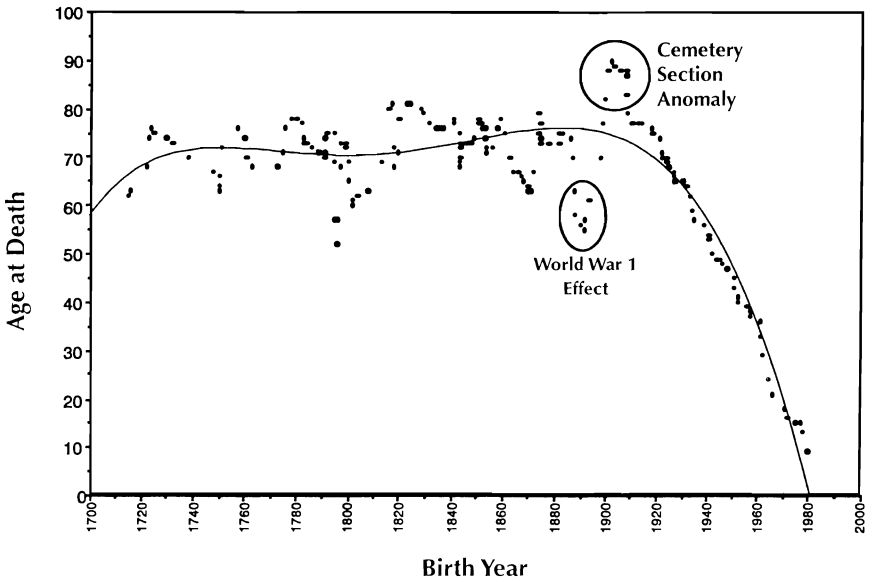


Figure 1. Death data from the Princeton cemetery. The longevities of 204 people buried in Princeton Cemetery shown as a function of the year of their birth. The data points were smoothed using '53h twice' (Tukey 1977), an iterative procedure of running medians.

early 50s to the mid 60s. These were somewhat younger than those found by Madden, but this was expected because Lombard was dealing with ordinary people rather than the "geniuses" of Madden (the positive correlation between fame and longevity was well known even then) But Lombard's study yielded one surprise: the most dangerous profession —the one with the lowest longevity— was that of "student" with an average age of death of only 20.7! Lombard recognized the reason for the anomaly but apparently did not connect it to his other results.

Example 2: The Twentieth Century Is a Dangerous Time

In a revisitation of this methodology, we gathered 204 birth and death dates from the Princeton (NJ) Cemetery. This cemetery opened in the mid 1700s and has people buried in it born in the early part of that century. Those interred include Grover Cleveland, John von Neumann, and Kurt Gödel.

When age at death was plotted as a function of birth year (after suitable smoothing to make the picture coherent), we see the result shown as Fig. 1. We find that the age of death stays relatively constant until 1920, when the longevity of the people in the cemetery begins to decline rapidly. The average age of death decreases from around 70 years of age in the 1900s to as low as 10 in the 1980s. It becomes obvious immediately that there must be a reason for the anomaly in the data (what we might call the "Lombard Surprise"), but what? Was it a war or a plague that caused the rapid decline? Has a neonatal section been added to the cemetery? Was it only opened to poor people after 1920? Obviously, the reason for the decline is nonrandom sampling. People cannot be buried in the cemetery if they are not already dead. Relatively few people born in the 1980s are buried in the cemetery and no one born in the 1980s that we found in Princeton Cemetery could be older than 17.

Example 3: Scientific Publishing Is Getting Faster

This sort of anomaly shows up in many other circumstances. For example in an earlier study we (Bradlow and Wainer 1998) examined publication delays in statistical journals. We did this by compiling a database of hundreds of articles from seven statistics journals that consisted of the date of an article's publication as well as the date it was first submitted. By subtracting the latter from the former we could obtain the publication delay. In Fig. 2 are sequential boxplots showing the distribution of delay for all the articles published in the journal *Psychometrika* over the 10 year period 1988–97. Once again the Lombard Surprise pops up, suggesting that the publication delays in this journal have been diminishing lately. A more careful study of these data, however, suggested that delays are actually increasing.

There are many examples of situations in which this anomaly arises. Four of these are the following:

1. In 100 autopsies, a significant relationship was found between age at death and the length of the line on the palm (Newrick, Affie, and Corrall 1990). What they actually discovered was that wrinkles and old age go together.
2. In 90% of all deaths resulting from barroom brawls, the victim was the one who instigated the fight. One questions the wit of the remaining 10% who didn't point at the body on floor when the police asked, "Who started this?"
3. *The New York Times* reported the results of data gathered by the American Society of Podiatry, which stated that 88% of all women wear shoes at least one size too small. One wonders who would be most likely to participate in such a poll.
4. In testimony before a committee of the Hawaii State Senate, then considering a law requiring all motorcyclists to wear a helmet, one witness declared that despite having been in several accidents during his 20 years of motorcycle riding, a helmet would not have prevented any of the injuries he received. Who was unable to testify? Why?

Selection, or more generally nonrandom sampling, is often as subtle in its manifestation as it is substantial in its effect. We have so far emphasized the size of their effects. Next let us consider an instance of selection in education whose interpretation has yielded substantial debate among experts.

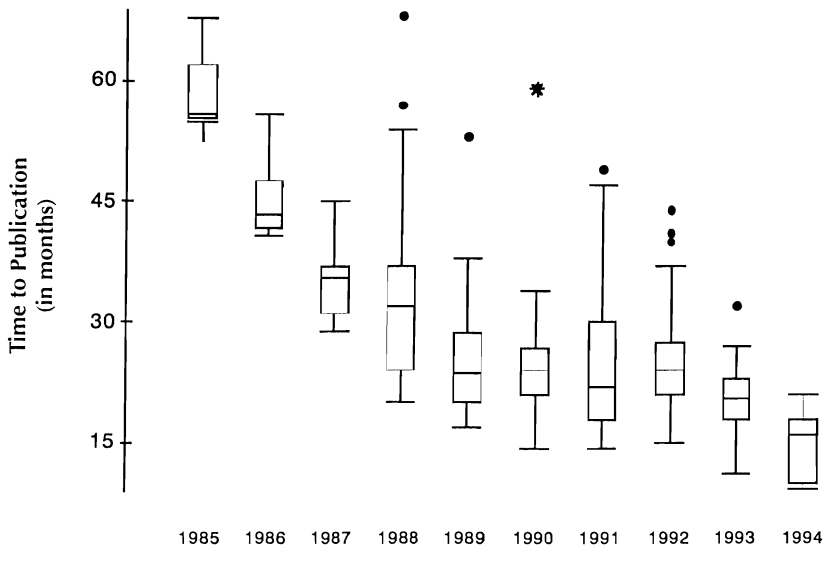


Figure 2. The distribution of publication delays, shown as box-and-whisker plots, for 283 articles that appeared in the journal *Psychometrika* between 1988 and 1997, shown as a function of the year they were submitted.

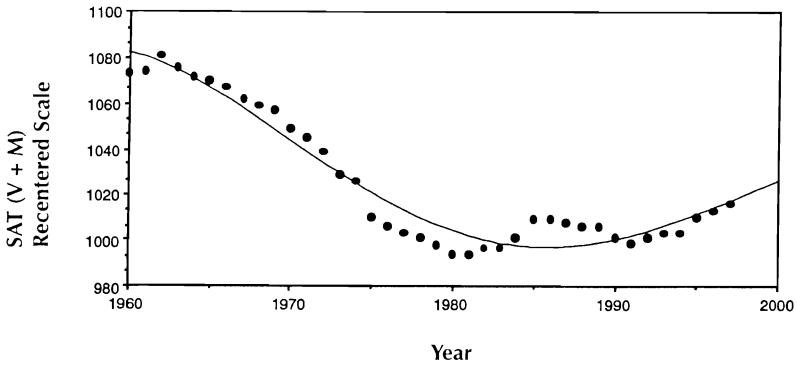


Figure 3. SAT scores began their decline in the early 1960s and hit their nadir about 1980. The mean SAT score (verbal plus Mathematical) for all high school seniors in the United States who took the test since 1960. These scores are shown on the newly (April 1996) recentered scale.

Example 4: What Do Changing SAT Scores Mean?

The Scholastic Assessment Test (SAT) is taken by more than a million high school seniors annually. Since 1962 the average SAT score has declined (see Fig. 3). Many have interpreted this decline as an indicator of the failure of the American educational system, although its principal causes remain in question. A national panel created by the College Board in 1977 placed most of the blame on students taking too many watered-down courses.

Gene Maeroff in a front-page story in *The New York Times* (September 22, 1982) suggested that the test results were further "evidence that a loosening of high school requirements in the 1960s had led to a deterioration of educational standards." Maeroff also cited a 1980 article in the *Phi Delta Kappan*, a respected educational journal, that "attributed the score decline to the effect on the young of fallout from above-ground nuclear tests of the 1950s and early 1960s."

Time magazine (October 4, 1982) suggested that the score decline was caused by a mix of "social factors, including television, the frequency of divorce and the softening of high school curriculums."

Laura Durkin (September 22, 1982) in *Newsday* understood the effects of selection and attributed the decline to "the much larger pool of students taking the test in recent years."

In the early to mid 1980s a consensus began to emerge to explain this trend. The essence of this argument was that an increasing proportion of students in the senior class around the nation took the SAT and that this group included many minority students who historically have not done well on standardized tests. This opinion was borne out by such evidence as that shown in Fig. 4 in which the declines in SAT scores are paralleled by declines in the size of the majority population.

Of course, this interpretation falls flat when data since 1985 are examined and shows exactly the opposite result (see Fig. 5).

How can we draw valid inferences from nonrandomly sampled data? The answer is "not easily" and certainly not without risk. The only way to draw inferences is if we have a model for the mechanism by which the data were sampled. Let us consider one well-known example of such a model.

Example 5: Bullet Holes and a Model for Missing Data

Abraham Wald in some work he did during World War II (Mangel and Samaniego 1984; Wald 1980) was trying to determine where to add extra armor to planes on the basis of the pattern of bullet holes in returning aircraft. His conclusion was to determine carefully where returning planes had been shot and *put extra armor every place else!*

Wald made his discovery by drawing an outline of a plane (crudely shown in Fig. 6 and then putting a mark on it where a returning aircraft had been shot. Soon the entire plane had been covered with marks *except* for a few key areas. It was at this point that he interposed a model for the missing data, the planes that did not return. He assumed that planes had been hit more or less

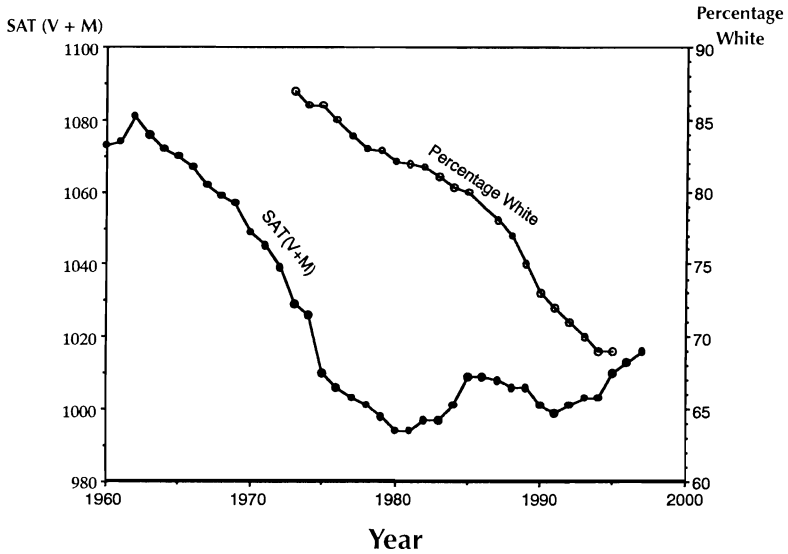


Figure 4. The decline in SAT scores parallels the decrease in the size of the white population. The mean SAT score shown on the same graph as the percentage of the U.S. population that classified themselves as "white" on the Current Population Survey. The decline of the SAT parallels the decline of the white population.

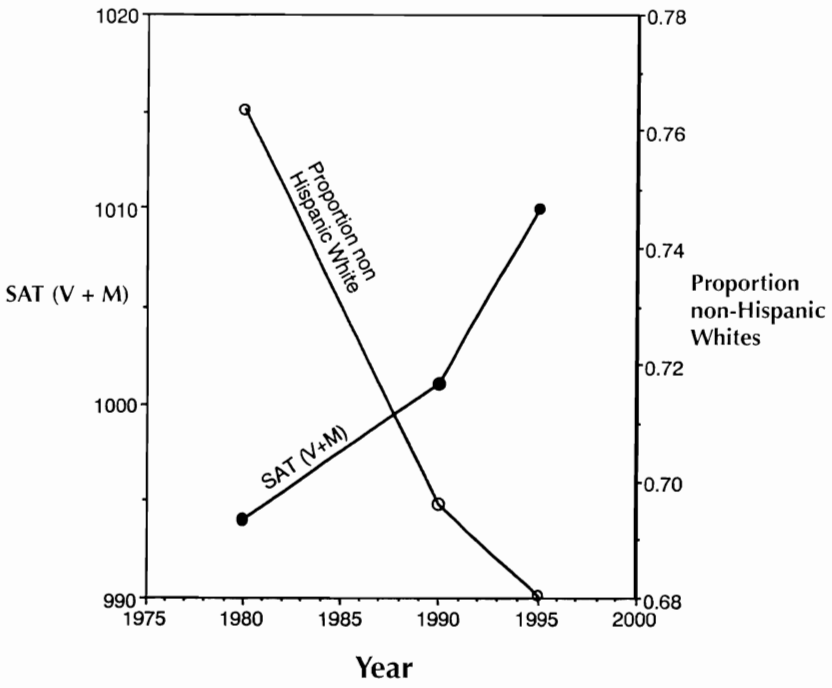


Figure 5. As the proportion of non-hispanic whites in the general population of 15–19 year-olds fell, SAT scores rose. The mean SAT score for 1980–1995 shown on the same graph as the proportion of the U.S. population, ages 15–19, that classified themselves as "white" and "non-Hispanic" on the Current Population Survey. The increase of the SAT stands in stark contrast to the decline of the white population.

uniformly, and hence those aircraft hit in the unmarked places had been unable to return, and thus those were the areas that required more armor.

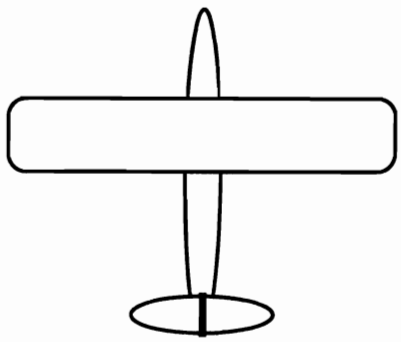
Wald's key insight was his model for the nonresponse. From his observation that planes hit in certain areas were still able to return to base, Wald inferred that the planes that didn't return must've been hit somewhere else. Note that if he used a different model analogous to "those lying within Princeton Cemetery have the same longevity as those without" (i.e., that the planes that returned were hit about the same as those that didn't return) he would have arrived at exactly the opposite (and wrong) conclusion.

To test Wald's model requires heroic efforts. Planes that did not return must be found and the patterns of bullet holes in them must be recorded. In short, to test the validity of Wald's model for missing data requires that we sample from the unselected population. In other words we must try to get a random sample, even if it is a small one. This strategy remains the basis for the only empirical solution to making inferences from nonrandom samples.

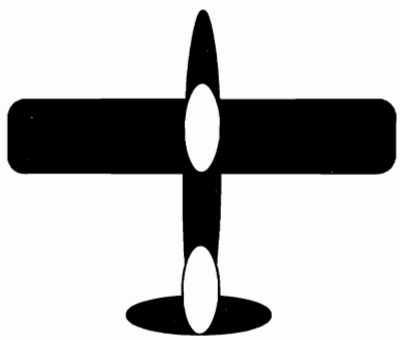
In our cemetery example, if we want to get an unbiased estimate of longevities we might halt our data series at a birth date of 1900. In the publishing-delay example, we might only consider articles submitted at least a decade earlier. For the SAT, several selection models have been tried (Dynarski 1987; Edwards and Beckworth 1990; Page and Feifs 1985· Powell and Steelman 1984; Steelman and Powell 1985; Taube and Linden 1989) and all have failed to be accurate enough for the purposes they were intended. We are thus drawn inexorably toward the conclusion that the best way to get an accurate indicator of average student performance is through a survey constituted from a well-designed, rational sampling process

Conclusion and Some Stupid Birds

We do not mean to suggest that it is impossible to gain useful insights from nonrandomly selected data, only that it is difficult and great care must be taken in drawing inferences. James Thurber's (1939) *Fables for Our Time*, tells the story of "The Glass in the Field." It seems that a builder left a huge pane of window glass standing upright in a field one day. Flying at high speed, a goldfinch struck the glass and was struck senseless. Later, upon recovering his wits, he told a sea gull, a hawk, an eagle, and a swallow about his injuries caused by crystallized air. The gull, the hawk, and the eagle laughed and bet the goldfinch a dozen worms that they could fly the same route without encountering crystal-



An outline of a plane.



A depiction of a plane with shading indicating where returning planes had been shot.

Figure 6. A schematic representation of Abraham Wald's ingenious scheme to investigate where to armor aircraft.

lized air, but the swallow declined and was alone in escaping injury. Thurber's moral: "He who hesitates is sometimes saved." This is our main point — that a degree of safety can exist when one makes inferences from nonrandomly selected data if those inferences are made with caution. There are some simple methods available that help us draw inferences when caution is warranted; they ought to be used.

This is an inappropriate vehicle to discuss these special methods for inference in detail. For such details the interested reader is referred to Little and Rubin (1987), Rosenbaum (1995), and Wainer (1986) for a beginning. Instead let us describe the general character of any "solution." First, no one should delude oneself into thinking that when there is a nonrandom sample unambiguous inferences can be made. They can't. The magic of statistics cannot create information when there is none. We cannot know for sure the longevity of those who are still alive, the publishing delay for papers that have not yet appeared, or the SAT scores for those who didn't take the test. Any inferences that involve such information are doomed to be equivocal. What can we do? One approach is to make up data that might plausibly have come from the unsampled population (i.e., from some hypothesized selection model) and include them with our sample as if they were real. Then see what inferences we would draw. Next make up some other data and see what inferences are suggested. Continue making up data until all plausible possibilities are covered. When this is done, see how stable the inferences were that we drew over the entire range of these data imputations. The multiple imputations may not give us a good answer, but they can provide us with an estimate of how sensitive our inferences are to the unknown. If we do not do this, we have not dealt with possible selection biases, only ignored them.

Data obtained via nonrandom sampling occurs often in practice. The specific form we address here are data that have been obtained through self-selection; that is, inclusion in the sample is determined by the units themselves and not the data gatherer. In such cases, valid inferences require careful thought about the character of the selection process. We present four examples which illustrate our findings, mostly, that ignoring self-selection can lead to flawed, often ridiculous findings. We also suggest a strategy to guard against flawed inferences.

References and Further Reading

Bradlow, E. T., and Wainer, H. (1998), "Publication Delays in Statistics Journals," *Chance*, 11(1), 42–45.

Dynarski, M. (1987), "The Scholastic Aptitude Test: Participation and Performance," *Economics of Education Review*, 6, 263–273.

Edwards, D., and Beckworth, C. M. (1990), Comment on Holland and Wainer's "Sources of Uncertainty Often Ignored in Adjusting State Mean SAT Scores for Differential Participation Rates: The Rules of the Game," *Applied Measurement in Education*, 3, 369–376.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley.

Lombard, H. C. (1835), "De l'Influence des Professions sur la Durée de la Vie," in *Annales d'Hygiène Publique et de Médecine Légale* (vol. 14), 88–131.

Madden, R. R. (1833), *The Infirmities of Genius, Illustrated by Referring the Anomalies in Literary Character to the Habits and Constitutional Peculiarities of Men of Genius*, London: Saunders and Otley.

Mangel, M., and Samaniego, F. J. (1984), "Abraham Wald's Work on Aircraft Survivability," *Journal of the American Statistical Association*, 79, 259–267.

Newrick, P. G., Affie, E., and Corrall, R. J. M. (1990), "Relationship Between Longevity and Lifeline: A Manual Study of 100 Patients," *Journal of the Royal Society of Medicine*, 83, 499–501.

Page, E. B., and Feifs, H. (1985), "SAT Scores and American States. Seeking for Useful Meaning," *Journal of Educational Measurement*, 22, 305–312.

Powell, B., and Steelman, L. C. (1984), "Variations in State SAT Performance: Meaningful or Misleading?" *Harvard Educational Review*, 54, 389–412.

Rosenbaum, P. R. (1989), "Safety in Caution," *Journal of Educational Statistics*, 14, 169–173.

——— (1995), *Observational Studies*, New York: Springer-Verlag.

Steelman, L. C., and Powell, B. (1985), "Appraising the Implications of the SAT for Educational Policy," *Phi Delta Kappan*, 67, 603–606.

Stigler, S. M. (1996), "Adolphe Quetelet: Statistician, Scientist, Builder of Intellectual Institutions," unpublished talk given at the Quetelet Bicentenary: Brussels, Belgium, 10/24/96.

Taube, K. T., and Linden, K. W. (1989), "State Mean SAT Score as a Function of Participation Rate and Other Educational and Demographic Variables," *Applied Measurement in Education*, 2, 143–159.

Thurber, J. (1939), *Fables for Our Time*, New York: Harper and Row.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Wainer, H. (1986), *Drawing Inferences From Self-selected Samples*, New York: Springer-Verlag.

Wald, A. (1980), "A Method of Estimating Plane Vulnerability Based on Damage of Survivors," CRC 432, July 1980. (These are reprints of work done by Wald while a member of Columbia's Statistics Research Group during the period 1942–45. Copies can be obtained from the Document Center, Center for Naval Analyses, 2000 N. Beauregard St., Alexandria, VA 22311.)