How Bad Can "Good" Data Really Be?

W. H. WILLIAMS*

Bias has different sources. Measurement errors create "bad" data and biased estimates. But selection biases occur even with "good" data and can be both subtle and large in magnitude.

Selection biases are not easily detected by internal examination of the data. Detection is more likely by comparison with external data sources.

KEY WORDS: Selection bias; Nonresponse; Confidence intervals; Sources of bias.

1. Introduction

Ninety percent confidence intervals are supposed to cover the true mean 90 percent of the time. In practice, it does not always work out this way; in fact, sometimes it may be closer to the truth to say that 90 percent of the time the true value lies *outside* of the 90 percent confidence limits! There seems to be no shortage of illustrations.

W.J. Youden (1972) discusses a number of examples from the physical sciences. In one, he lists 15 different values of the astronomical unit, which is the average distance to the sun, obtained over the period 1895–1961. Each estimated value is *outside* of the limits of the one immediately preceding it. The conclusion of systematic bias seems irresistible.

McNish (1962) presented a graphical representation of 24 measurements of the speed of light. The estimates are spread over a range of 3.5 km per second but half of the reported errors are well under 0.5 km per second. This certainly suggests that the individual scientists did not, or could not, set realistic error limits on their reported results. McNish in fact concluded, that in spite of his careful study of the subject, he was not able to put a quantitative measure of confidence on his own estimate of the speed of light.

Another extremely interesting example was described in a recent issue of *Sky and Telescope* (1975). Researchers at the University of Arizona recently released new measurements on the shape of the sun. They found it to be indistinguishable from a sphere; the difference between the equatorial and polar diameters was not found to be significantly different from zero. This result conflicts with an earlier one obtained at Princeton University which indicated a clear oblateness, with the equatorial diameter longer than the polar diameter.

While this difference may appear to be small, it is of extreme importance. If the Arizona researchers are correct that the sun is round, then Einstein's 1915 theory of general relativity is intact, and does not require replacement by a later (1961) theory proposed by some of the same Princeton researchers.

In the socioeconomic field, it has been observed (Williams and Goodman 1971) that forecasts of numbers of telephones were outside of the associated confidence limits far more frequently than they should be. More specifically, it was found that 95 percent confidence limits covered the subsequently observed true value about 80 percent of the time. Eighty percent confidence limits covered the true value about 65 percent of the time. This observation led to the development of empirical prediction intervals described by Williams and Goodman (1971).

The common problem in these examples is systematic bias combined with a failure to recognize that confidence intervals pertain only to sampling error; and I would suggest that the reason that the more famous of such examples are in the physical sciences is that we are not yet sufficiently suspicious about the nature of our socioeconomic survey data. So, for example, when the *New York Times* (July 21, 1975) says in describing the results of a survey on the financial plight of New York City, that "A total of 420 persons were interviewed, a random sample that statistical experts say yields 95 percent confidence that the results are within 5 percentage points of the attitudes of the population as a whole," should we believe it? Clearly we need not, at least not until any issues of potential bias are resolved. But in the *New York Times* report, no statements were made that would even indirectly help in the assessment of bias, possibly because it was felt that no bias exists, but we are nevertheless free to be appropriately suspicious.

2. Bias Effects

A bias exists in an estimator if its average value over all possible sample values is not equal to the true parameter. The effect of a bias on confidence limits is to shift them by an amount equal to the bias, so that they do not cover the true value the stated fraction of the time. Trouble can come quickly; if the sampling distribution is normal, a bias equal to one standard deviation changes 95 percent confidence intervals into 83 percent intervals. Larger biases cause even faster deterioration. In view of the fact that, in special studies, the Census Bureau found nonsampling errors that were 10 times the magnitude of sampling errors, it seems clear that we ought to be paying close attention to these factors, which can have

* W. H. Williams is at Bell Laboratories, Murray Hill, NJ 07974. This article is the written version of an invited talk, "The Seriousness of Selection Biases, Including Nonresponse," presented to the Social Statistics Section of the American Statistical Association, Atlanta, Georgia, August 1975. Thanks are due to R. W. Hamming for a number of interesting discussions on this topic.

such a substantial and disastrous effect on our assessment of estimates.

In some cases, it is possible to adjust interval estimates by use of the mean square error, but to discuss this aspect we need to classify the sources of possible bias.

3. Bias Sources

Technical Bias

Technical bias is the most familiar type of bias. This is the type most often discussed by mathematical statisticians. Technical biases occur when the functional form of the estimator is such that the average over all possible samples is not equal to the true parameter value. Ratio and regression estimators are generally biased this way. But if standard estimators are used and some attention is given to the usually known technical bias characteristics, this bias source should not present great difficulty.

Measurement Error

Measurement error is another source of bias. In this case, the effects can be substantial. In fact, there is virtually no limit to the difficulties that can be brought about by measurement error. Conceptually, these difficulties are usually easy to understand. The measurement process should measure a value y but manages to feed a completely different value into the analysis. But while the problem is conceptually simple, it can be difficult in practice, because the errors can be introduced in subtle ways. Furthermore, these errors can be introduced by either human or mechanical measuring devices. There is a large literature on this problem in the special case of variable human responses. The book by Sudman and Bradburn (1974) contains a description and a large number of references.

In general, we cannot analyze measurement bias in the same way that we do technical biases, that is, develop bounds and hence determine the effect directly on the estimators and also on any confidence limits. In most cases the appropriate response to measurement bias is to correct the errors directly. Unfortunately it is the measurement bias which, if undetected, can be the most serious survey problem and can lead to confidence limits which are substantially off target.

Measurement error creates bad data, which can be a very serious source of bias. But in this article, we choose to focus on a different type of bias.

Selection Bias

Selection biases can occur when sample units are thought to have been drawn into the sample with one set of probabilities but actually are unknowingly drawn in with a different set. Such a difficulty can be associated with a failure to implement a sampling design properly or with the specific problem of nonresponse. We give an example of each type.

In a statistical sampling design each unit in the population is associated with a probability of selection into the sample. These probabilities specify the sampling design and indeed are referred to by some authors as the design probabilities. The knowledge of these probabilities permits the analyst to create unbiased estimators. But sometimes operational difficulties result in units actually being selected with probabilities other than those specified by the design. For example, in a Bell System study of its business customers, each customer was thought to have entered the sample with equal probability, but it was subsequently discovered that the bigger the customer the higher the chance of entering the sample. This meant that the probabilities specified by the design had not been implemented but rather a second set had been inadvertently used. Furthermore, since the larger customers had a larger chance of selection, the actual design probabilities were correlated with the measurements made on the units. In this study complete accounting records were available so there was no problem of nonresponse.

Even if the design probabilities are properly implemented, selection biases can be associated with nonresponse. To illustrate, suppose that m units have been selected by a sampling process and it remains only to visit the m households and make the planned observations. But unfortunately, not all households will be interviewed, some will be nonrespondents. There are various reasons for this but these need not concern us here. So if r_i is the probability of actually getting a response from a household after it has been selected as a sample household, then ideally all $r_i = 1$; but in practice $r_i \leq 1$. Next if n is the actual number of respondents, then

$$E(n|\text{the sample realized}) = \sum_{i=1}^{m} r_i \leq m.$$

and

$$E(n) = E\left\{\sum_{i=1}^{m} r_i\right\} = mE(r_i).$$

Furthermore, it is known (Finkner 1950; Williams 1970) that these probabilities of response are often correlated with measurable characteristics of the units. For example, the probability of response is higher for households with children than without them.

In both the preceding examples the units actually enter the sample with probabilities other than those specified by the design, and further, these probabilities are correlated with the measurements on the units. This clearly changes the expectation of an estimator; previously unbiased estimators are now biased. We refer to biases achieved in this way as selection biases, which is consistent with the terminology used by Neyman (1969) and Yates (1960). Notice that there can be selection biases with estimators which

have no technical bias and with data which have no measurement errors. In practice, data with little or no measurement error is often referred to as "good" data, so we are led to the question "How bad can 'good' data really be?"

Finally, some additional remarks about selection biases are in order. First if the real probabilities are known, or subsequently determined, they may be used in place of the original weights to create unbiased (but perhaps less efficient) estimators. The problem with selection biases exists when the actual selection probabilities are unknown.

One of the preceding examples had to do with nonresponse and in connection with it some remarks about terminology are useful. In that example we might have used the term "nonresponse bias." The reader should be aware, however, that nonresponse bias does not usually have the same meaning as the term "response bias." Response bias is a result of the failure, for whatever reason, to get an accurate response or measurement, usually in a live interview situation. There is a large literature on this latter subject; see Sudman and Bradburn (1974) and the many references given there. More recent work has been described by Bailar, Bailey, and Corby (1977), Cowan (1977), and Goldfield et al. (1978). Response bias is a measurement problem and is not the subject of this article.

4. What Are the Characteristics of Selection Bias?

The Magnitude Can Be Serious

A Bell System study (Williams 1970) specified that successive (rotation) groups, made up of homes in geographical urban areas, were to be brought into the survey and retained for three consecutive monthly interviews. It was found that the average number of children per family, for rotation groups appearing in the sample for the first time, was 3.2, and for rotation groups appearing in the sample for the second and third times, the averages were 2.5 and 2.4, respectively. The average within rotation group standard error of the monthly estimates was 0.1. Consequently, it appeared that the first month estimate was significantly different from the second and third.

Analysis revealed that the cause was selection bias. The first time the panel was observed, households with children were more readily interviewed than those households with no children. As the interviewers became more familiar with the habits of the households in their areas, this bias became less pronounced. Nevertheless, such biases are very serious.

Finkner (1950) presented an interesting example of a different type. His study was a multiple mail survey of fruit growers, for whom a complete census was available. There was a major systematic characteristic in the response of the growers. Big growers responded to the mailing much more readily than small growers. Experienced practitioners will of course recognize this kind of behavior in both call-back and mail surveys.

As a final illustration of the magnitude of selection biases, Williams (1970) discussed the effects of differential response rates for employed and unemployed persons. It is possible to have a four percent relative bias in the estimate of the unemployment rate even if the response rate was 98 percent for employed persons and 95 percent for unemployed persons, with an overall response rate of almost 98 percent! It takes little imagination to anticipate the magnitude of biases that are possible with 50 or even 60 or 70 percent responses.

So, in summary, the first point about selection biases seems clear, specifically that the magnitude of selection biases can be large indeed.

The Effects Are Subtle

The second point to be made about selection biases is that their effects can be subtle. In an article by Williams and Mallows (1970), it was shown that estimates of change through time can be badly biased even though the study is based on a completely identical set of sampled persons. It has been almost axiomatic in sampling that fixed panel surveys are the best way to design studies for maximum information on changes through time. This statement can certainly be found in many sampling texts, see, e.g., Cochran (1977) and Stephan and McCarthy (1958). The conflict is that these statements have been based solely on variance and not at all on bias. If bias is included, quite different design conclusions can emerge. In fact, it is not hard to construct examples where, by including both bias and variance, the best information on change through time comes from a complete replacement design and not a fixed panel! For example, if the average age of the U.S. population is estimated from a fixed panel, it is apt to be badly biased. The fixed panel ages one year every year while the average age of the U.S. population does not necessarily increase and may even decrease. While the bias is immediately obvious in this example, others are much more subtle and difficult to detect.

A second interesting and unanticipated result of Williams and Mallows's (1970) study has to do with population mobility. Response rates which are not the same for employed and unemployed persons can make it appear that far more employed people are "found" at later observation periods than the number of employed persons who are "lost" from the survey. This characteristic has been interpreted as a population mobility phenomenon, in which unemployed persons move, and show up somewhere else employed, that is, they move to get a job. Unfortunately, this perplexing result can arise even when the population is completely static.

A third possible example was presented by Prais (1958), who described two matched geographical areas that had been drawn into a survey of consumer habits. The treatment of these groups was not identical however, one group was retained in the survey for four consecutive weeks and the other for only two. The puzzling result was that at the end of the first two weeks the estimates derived from the two groups were substantially different. If the response rates for the two groups were truly different, presumably as a result of the differential duration of inclusion, it may be plausible to ascribe the differences in the estimates to selection bias.

Relationships Are Not Invariant

Selection biases can also change relationships. The modern theory of finance is built upon the assumption of a correlation between risk, as measured by variance, and rate of return. This assumption has been examined empirically in the literature and correlations in the range of 0.4 to 0.6 found.

A study of these empirical papers by Williams and Hwang (1971) revealed that the data sets used by the reporters of the empirical results fall into three classes. In the first class are papers in which the studies were based on corporations with matched sets of data for a specified time period. The Compustat tapes are one such data set. These tapes contain data for corporations with complete 20-year data histories.

The second class includes those papers which used data from corporations making up 80 percent of the studied industry. Finally, the third class contains those papers in which no background information at all is given on the data.

Now the question can be raised "Which companies are missing?" It turns out that these are likely to be the high-risk–low-return, and the low-risk–high-return companies. These are the companies involved in mergers and which, as a result, do not have easy-to-analyze data histories. Williams and Hwang (1971) used this information to develop models in which selection biases easily generate correlations between risk and rate of return of the order of magnitude of those found in the empirical papers.

Increasing Response May Not Help

A fourth important characteristic is that increases in response rate do not necessarily bring a reduction in bias. This is easily seen in the unemployment example (Williams 1970).

In the literature of call-backs, it seems to be generally agreed that at the first go-around, unemployed persons are easier to find and interview than employed persons. That is, the probability of response is greater for an unemployed person than an employed one.

But there also exists a belief[1] that, even after many go-arounds, a hard core of unemployed persons will remain unobserved. In technical terms, the probability of getting a response from an employed person is now higher than from an unemployed person. So the bias has shifted from overrepresenting unemployed persons to underrepresenting them. The most accurate estimate would have occurred at the go-around at which the probability of a response from an employed person was approximately equal to the probability of response from an unemployed person. The trouble is that we may not know at which go-around this equality actually occurred.

So in summary this far, we have seen that the effects of selection bias can (1) be serious in magnitude, (2) be subtle, (3) change apparent relationships, and (4) react unpredictably to an increasing response rate.

5. Detection and Correction

Selection biases can be detected by reviewing the sampling process to determine if the sample was actually selected according to the design specifications. Unfortunately, such reviews often lead to what should have been done rather than what was actually done. Consequently, as in so many statistical procedures, the best detection is actually prevention of selection bias, by holding tight controls over the sampling process originally, rather than by trying to reach back for verification that the sampling was done correctly.

The best analytic method to seek out possible selection biases is the comparison of the sample with outside data. As a simple example, if a sample turns up with 75 percent male and 25 percent female respondents, the sample is clearly out of line with the approximately 50-50 sex ratio in the overall population. The sample is weighting the males too heavily. With the knowledge of the true population sex ratio in hand, the correction technique is simple; specifically, weight the male and female estimates separately and equally. Such a procedure was used in the Bell System data described earlier in which families with small children were overrepresented. Known population weights were used in place of the existing sample weights. This technique is called post-stratification (Cochran 1977; Williams 1962 and 1964).

But while reweighting data is fairly straightforward, it is not always possible because appropriate data may not be available. Furthermore, before any correction is possible, the analyst must be suspicious enough to look for a selection bias in the first place.

Another method for seeking selection biases is to compare sample estimates and the response rates. The object is to find a correlation between the estimates and the levels of response. This procedure is certainly possible when the survey involves more than one call-back. It is also possible in panel surveys in which sections of the population are inter-

[1] There is substantial advance planning being put into the 1980 census to try to reduce such problems.

viewed on repeated occasions (usually with a changing response rate). With some ingenuity, there are other situations in which this same approach can be taken, for example, in a mail survey, estimates can be recalculated continually as a function of the time of arrival.

In household surveys, it seems to be consistently true that the estimated number of children is related to the response rate. People with fewer children are home less. In the Current Population Survey (CPS) unemployment estimates are related to response levels, which leads to the conjecture (Williams 1970) that the observed rotation bias is actually a selection bias.

Operationally, it is sometimes possible to use the relationship between the response rates and the estimates to extrapolate to a "100 percent response estimate." This of course is not a new suggestion, it has been used in practice and has been discussed by Deming (1960).

In practice, the worst situation of all occurs where there is no apparent overall response problem and there are no useful external data available for comparison. Then there may be no reason to be suspicious and there are no analytic procedures of comparison available. This has certainly happened in quota sampling where the sample may be balanced on specified quota variables, such as political party and sex, but badly out of balance on other important variables.

6. A Warning

To deny the existence of a selection bias is a substantial undertaking. For one-time surveys, it requires a denial that the actual selection probabilities are not correlated with the measurements. But as we have seen, such correlations can happen easily and frequently. The average number of children and the Finkner fruit-tree data are clear-cut cases.

To deny the existence of a selection bias in repeated surveys, it is also necessary to deny that the selection probabilities change from one interview period to the next. Otherwise, the estimates of change from time period to time period will be biased (Williams 1970). This denial is also very difficult because it is common for the response rate to change as the survey progresses and the only way the response rate can change is for the selection probabilities to change.

It is not easy to ignore the possibility of selection bias.

7. Summary

We have shown that selection biases can be serious in magnitude, they can be subtle, they can change relationships, and they do not necessarily react to increasing response rates in a desirable way.

In the physical sciences, biases have appeared in highly focussed areas of research. It seems likely that in the social sciences we have not yet paid enough attention to the devastating possibilities of this kind of bias.

[*Received September 20, 1976. Revised January 26, 1978.*]

References

Bailar, Barbara A. (1975), "The Effects of Rotation Group Bias on Estimates from Panel Surveys," *Journal of the American Statistical Association*, 70, 23–30.

———, Bailey, Leroy, and Corby, Carol (1977), "A Comparison of Some Adjustment and Weighting Procedures for Survey Data," presented at the 1977 Sampling Symposium, Chapel Hill, North Carolina.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd ed., New York: John Wiley & Sons.

Cowan, Charles D. (1978), "Incentive Effects on Amounts Reported in an Expenditure Diary Survey," *1977 Proceedings of the American Statistical Association, Social Statistics Section*, Washington, D.C.: American Statistical Association, 498–503.

Deming, W.E. (1960), *Sample Design in Business Research*, New York: John Wiley & Sons.

Finkner, A.L. (1950), "Methods of Sampling for Estimating Commercial Peach Production in North Carolina," *North Carolina Agricultural Experiment Station Technical Bulletin*, 91.

Goldfield, Edwin D., Turner, Anthony G., Cowan, Charles D., and Scott, John C. (1978), "Privacy and Confidentiality as Factors in Survey Response," *1977 Proceedings of the American Statistical Association, Social Statistics Section*, Washington, D.C.: American Statistical Association, 219–231.

McNish, A.G. (1962), *IRE Transactions on Instrumentation* I-11, No. 3 and 4, 138–148.

Neyman, Jerzy (1969), "Bias in Surveys Due to Nonresponse," in *New Developments in Survey Sampling*, eds. Norman L. Johnson and Harry Smith, Jr., New York: John Wiley & Sons.

Prais, S.J. (1958), "Some Problems in the Measurement of Price Changes with Special Reference to the Cost of Living," *Journal of the Royal Statistical Society*, ser. A, 121, 312–323.

Stephan, Frederick F., and McCarthy, Philip J. (1958), *Sampling Opinions*, New York: John Wiley & Sons.

Sudman, Seymour, and Bradburn, Norman M. (1974), *Response Effects in Surveys*, Chicago: Aldine Publishing.

Sky and Telescope (1975), Editorial, 50, No. 2, 7.

Williams, W.H. (1962), "The Variance of an Estimator with Post-Stratified Weighting," *Journal of the American Statistical Association*, 57, 622–627.

——— (1964), "Sample Selection and the Choice of Estimator in Two-Way Stratified Populations," *Journal of the American Statistical Association*, 59, 1054–1062.

——— (1970), "The Systematic Bias Effects of Incomplete Responses," *Public Opinion Quarterly*, 33, 593–602.

———, and Mallows, C.L. (1970), "Biases in Panel Surveys Due to Differential Nonresponse," *Journal of the American Statistical Association*, 65, 1338–1349.

——— and Goodman, M.L. (1971), "A Simple Method for the Construction of Empirical Confidence Limits for Economic Forecasts," *Journal of the American Statistical Association*, 66, 752–754.

———, and Hwang, F. (1971), "Estimation Biases in the Analysis of Risk and Return," *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, Washington, D.C.: American Statistical Association, 512–515.

Yates, Frank (1960), *Sampling Methods for Censuses and Surveys*, New York: Hafner Publishing Company.

Youden, W.J. (1972), "Enduring Values," *Technometrics*, 14, 1–10.