THE SERIOUSNESS OF SELECTION BIASES, INCLUDING NONRESPONSE

W. H. Williams, Bell Telephone Laboratories

I. Introduction

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

W. J. Youden in a paper Enduring Values [13] discusses a number of examples from the physical sciences. He lists 15 values of the Astronomical Unit, which is the average distance to the sun, over the period 1895-1961. Each estimated value is outside of the limits of the one immediately preceeding it. The conclusion of systematic bias seems irresistable.

McNish [4] presented a graphical representation of 24 measurements of the speed of light. The estimators 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 quantitive measure of confidence on his own estimate of the speed of light.

Another extremely interesting example was described in the August 1975 issue of Sky and Telescope, [6]. 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 were not found to be significantly different from zero. This result conflicts with an earlier one obtained at Princeton University which indicated a clear oblatedness, 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 [9] 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% confidence limits covered the true value about 80% of the time. Eighty per cent confidence limits covered the true mean about 65% of the time. This observation led to the development of empirical prediction intervals described by Williams and Goodman in 1971, [9].

The common problem in these examples is systematic bias, and I would like to 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 sociological and economic 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 not, at least not until the issues of bias are resolved. But in the New York Times report no statements about bias were made, possibly because it was felt that no bias exists, but we are nevertheless free to be appropriately suspicious.

II. Bias Effects

A bias exists in an estimator if its average value over all possible 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 10 times the magnitude of sampling errors, it seems clear that we ought to be paying close attention to these factors, which can have such a substantial and disasterous 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.

III. Bias Sources

(1) The most familiar kind of bias is the technical bias. These are the biases most often discussed by mathematical statisticians and are the result of the functional form of the estimator not averaging over all possible samples to the true population 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.

(2) 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 measurement error. Conceptually, these difficulties are usually easy to understand; the measurement process should measure y but somehow manages to feed x 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 no practical way of analyzing measurement bias in the same way that we do with technical biases, that is to develop bounds and hence determine the effect both directly on the estimators and also on any confidence limits. The appropriate response to measure bias is to correct the errors directly. Unfortunately, it is the measurement bias, which if undetected may be the most serious survey problem and may lead to confidence limits which are substantially off target.

Measurement error is an important problem, and most analysts realize that the data must be good. But in this paper, I choose to focus, not on the problem of how bad can bad data be, but rather on the question how bad can good data be?

(3) Selection Bias. Selection biases occur when the population units are to be selected into the sample with one set of probabilities but are actually selected with a different set. This changes the expectation of any estimator; previously unbiased estimators are now biased. There are no technical biases and no measurement errors. In this case, the analysis is being made of "good" data and is the reason for asking the question "How Bad Can Good Data Really Be?

Two remarks about selection biases are in order. First if the real probabilities are known, or subsequently determined, they can be used in place of the original weights to create unbiased estimators. The problem only exists when the probabilities are unknown. The second point is that selection biases can be thought of as including nonresponse. In this case, units which have already been selected to appear in the sample, actually appear in the sample with probabilities somewhat less than one.

IV. What Are The Characteristics Of Selection Bias?

(1) The Magnitude Can Be Serious. In a Bell System study, [10], the average number of children per family for rotation groups appearing in the sample for the first time was 3.2. 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 appears that the first month estimate is significantly different from the second and third.

Analysis revealed that the cause was a 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. But such biases are nevertheless very serious.

Finkner [3] 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 callback and mail surveys.

As a final illustration of the magnitude of selection biases, Williams [10] discussed the effects of differential response rates for employed and unemployed people. It was shown that it is possible to have a four percent bias in the unemployment rate even if the response rate was 98 percent for employed persons and 95 percent for unemployed persons, giving an overall response rate of 98%! It takes little immagination 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.

(2) The Effects Are Subtle. The second point to be made about selection biases is that their effects can be subtle.

In a paper by Williams and Mallows [11], 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. 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!

A second interesting and unanticipated result of this same paper [11] has to do with population mobility. Different response rates 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 which umemployed persons move and show up elsewhere as employed persons, i.e., they move to get a job. This perplexing result can arise even when the population is completely static.

A third example was presented by Prais [5], 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 in that one group was to be retained in the survey for four consecutive weeks and the other for only two. The puzzling result was that at the end of the two weeks the estimates derived from the two groups were substantially different. Since the response rates for the two groups were different, presumably as a result of the differential duration of inclusion, it seems entirely possible to ascribe the differences in the estimates to selection bias.

(3) 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 [12] revealed that the data sets used by the reporters of the empirical results fell into three classes. First the data upon which the studies were based were for corporations with matched sets of data for a fixed time period. The Compustat tapes is such a set, they present data only for corporations with 20 year data histories.

The second class included those papers which used data from corporations which made up 80% of the studied industry. And the third class were those papers in which no background information was given on the data at all.

Now the question arises as to which companies are missing. It turns out that these are the high risk-low return and the low risk-high return companies because these are the companies involved in mergers and which, as a result, do not have easy-to-analyze data histories. Williams and Hwang [12] 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.

(4) Increasing Response May Not Help. A fourth important characteristic that needs to be listed is that increases in response rate do not necessarily bring an improvement. This is easily seen in the unemployment example, [10].

In the literature of callbacks, 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 unemployeds than employeds.

But there also exists a belief that, even after many go-arounds, a hard core of unemployed persons will remain unobserved. That is, the probability of a response from an employed person is now higher than from an unemployed person. So the bias has shifted from overrepresenting employed persons to underrepresenting them. The closest estimate would have occurred at that go-around at which the probability of a response from an employed was closest to the probability of response from those unemployed. The trouble is that we may not know at which go-around this equality 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) can change apparent relationships, (4) are not necessarily responsive to increasing the response rate.

V. Detection and Correction

The best way to detect selection biases is to review the sampling process to determine if the sample was actually selected according to the design specifications. Unfortunately, it is not always possible to do this in a revealing way because a review is very likely to 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 trying to reach back for verification.

The best analytic method of detection of selection bias is the comparison of the sample with outside data. As a simple example, if a population sample turns up with 75% male and 25% female respondents, the sample is clearly out of line with the approximately 50-50 sex-ratio in the overall population. In this case the sample is weighting the males too heavily and the correction technique is clear, namely to weight the male and female estimates separately and equally. This procedure was used in the Bell System data described earlier in which families with small children were overrepresented. New weights were given to the families which were the population proportions and not the sample proportions. This technique is called post-stratification [7], [8].

While this correction is fairly straightforward, it is not always possible because appropriate data may not be available. And even before any possible correction, the analyst must be suspicious enough to seek out a selection bias in the first place.

Another method of seeking out selection biases is to relate the estimates to the response rate. The object is to find a correlation between the estimate and the level of response. It is certainly possible to do this when the survey involves more than one call back. It is also possible to do this in panel surveys in which sections of the population are interviewed on repeated occasions, usually with an increasing 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 calculated continually as a function of the time of arrival.

In household surveys, it seems to be consistently true that the number of children goes down as the response rate goes up. This characteristic is also true of the CPS unemployment estimates, which leads to the conjecture that the observed rotation bias [10] is actually a selection bias.

Operationally, it may be possible to use a developed relationship between the response rate and the value of the estimates to extrapolate to a "100% response estimate." This of course is not a new suggestion, it has certainly been used by Deming, [2].

Probably the worst situation of all occurs when there is no apparent nonresponse problem but a bias towards one particular kind of population unit exists. This has certainly happened in quota sampling.

VI. A Warning

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

For repeated surveys, it is necessary to deny that the selection probabilities change from one interview period to the next, because otherwise the estimates of change from time period to time period will be biased [10]. This denial is also very difficult because it is common for the response rate to change as the survey progresses, usually it increases, and the only way the response rate can change is for the selection probabilities to change.

So all in all, it would appear to be highly dangerous to ignore the possibility of selection bias.

VII. Summary

We have shown, that selection biases can be serious in magnitude, that they be subtle, that they can change relationships, and that 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.

VIII. Literature Cited

(1) Bailar, Barbara A., The Effects of Rotation Group Bias on Estimates from Panel Surveys. Journal of the American Statistical Association, Vol. 70, Number 69, 1975, pp. 23-30.

(2) Deming, W. E., Sample Design in Business Research. John Wiley and Sons, New York, 1960.

(3) Finkner, A. L., Methods of Sampling for Estimating Commercial Peach Production in North Carolina. North Carolina Agric. Exp. Stat. Tech. Bull., 91, 1950.

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

(5) Prais, S. J., Some Problems in the Measurement of Price Changes with Special Reference to the Cost of Living. Journal of the Royal Statistical Society, A, Vol. 121, 3, 1958, pp. 312-323.

(6) Editorial, Sky and Telescope, Vol. 50, No. 2, August 1975, p. 7.

(7) Williams, W. H., The Variance of an Estimator with Post-Stratified Weighting, Journal of the American Statistical Assoc., Vol. 57, 297, 1962, pp. 622-627.

(8) Williams, W. H., Sample Selection and the Choice of Estimator in Two-Way Stratified Populations. Journal of the American Statistical Assoc., Vol. 59, 308, 1964, pp. 1054-1062.

(9) Williams, W. H., and Goodman, M. L., A Simple Method for the Construction of Empirical Confidence Limits for Economic Forecasts. Journal of the American Statistical Association, Vol. 66, No. 336, Dec. 1971, pp. 752-754.

(10) Williams, W. H., The Systematic Bias Effects of Incomplete Responses. Public Opinion Quarterly, Vol. XXXIII, No. 4, 1970, pp. 593-602.

(11) Williams, W. H., and Mallows, C. L., Biases in Panel Surveys Due to Differential Nonresponse. Journal of the American Statistical Association, Vol. 65, No. 3, Sept. 1970, pp. 1338-1349.

(12) Williams, W. H., and Hwang, F., Estimation Biases in the Analysis of Risk and Return. Proceedings of the Business and Economic Statistics Section, American Statistical Assoc., Fort Collins, Colorado, August 1971, pp. 512-515.

(13) Youden, W. J., Enduring Values, Technometrics, Vol. 14, No. 1, February 1972, pp. 1-10.