Adjustment for Bias Caused by Non-Response in Mailed Surveys

By Walter A. Hendricks

This article is a companion to the one in the last issue which reported on the use of an enumerative survey to determine biases that may exist in mail data. This discussion presents some exploratory thinking on ways and means of using the mailed survey data, themselves, to determine and adjust for such biases.

INCOMPLETENESS of returns in a mail survey usually implies a certain degree of bias in the results because a respondent's willingness to return a schedule is generally related to the nature of the item to be estimated from the survey. The bias may be either positive or negative, depending upon whether prospective respondents with large or with small quantities of the item are the more willing to take the trouble to fill out and return the schedules. This does not mean that the amount he has of the item is the main influence in his decision to return or not to return the schedule. It means simply that the amount of the item is correlated with the factors that affect the decision. The actual amount of the item itself may be exerting no causal effect at all.

The fact that such biases exist has been rather generally known ever since mail surveys were first used by statisticians. Methods for dealing with these biases have been tested from time to time. At one extreme, there have been suggestions that mail surveys be abandoned as a sampling tool and that interview-sampling methods be used exclusively. That proposal has not been universally adopted by statistical agencies because interview sampling methods are usually expensive. Furthermore, some statisticians, including the author of this article, have clung rather tenaciously to the opinion that a careful analysis of the behavior and characteristics of the respondents to mail surveys would reveal some pertinent relationships that would make it possible to estimate the extent of the bias in any survey and to make the necessary adjustments.

For many problems, the application of scientific principles to the use of mail surveys would probably strengthen such surveys to the point where they would yield just as accurate results as do enumerative surveys. This is not an attempt to minimize the importance of enumerative surveys in an over-all statistical program; enumerative surveys are needed to provide the base information that must be available before mail surveys can be used scientifically. Furthermore, there will always be situations in which an enumerative survey is the most practicable method of getting data. It means, however, that a mail survey should be planned with as much attention to scientific principles as an enumerative survey. When that is done, the mail approach can be expected to yield satisfactory results in many situations in which its use has seemed undesirable.

Devices that have been used to adjust the results from mail surveys include (1) enumeration by interview of a subsample of the non-respondents to the mail surveys, (2) charting of historical data from mail surveys against more accurate data obtained later by complete enumeration or similar methods, and (3) using control information that is known for both respondents and non-respondents and that is also correlated with the item to be estimated, to "true-up" the returns received by mail. All these methods, together with the direct-interview type of survey itself, have been tested by statisticians of the BAE and other statistical agencies. Each seems to have its proper place in the over-all sampling program of a statistical organization. It is not the purpose of this article to give an appraisal of these methods; they are merely mentioned to provide some background for the discussion of a problem that has seemed hopeless of solution, but one that has intrigued the writer for some time.

Estimating the Bias

Suppose a statistical agency sends questionnaires to every individual in a universe, or in a well-designed sample of that universe, and only a fraction of those questionnaires are filled out and returned. Past experience with mail surveys in

general makes it plain that there is a good chance of bias being present in the results. But suppose that neither final check data nor control information is available. Under those conditions would it be possible to estimate the bias in the mailed returns without sending out some interviewers to visit a sample of the non-respondents? Some data assembled by the North Carolina Research Office, at Raleigh, indicate strongly that this question can be answered in the affirmative. All that seems to be necessary is to send a few follow-up requests to the non-respondents. Returns from at least two such follow-ups seem to be needed; more successive requests for information may be used to obtain results of greater precision.

Table 1 shows the results obtained by mail solicitation from universes of 3,241 North Carolina fruit growers and 1,189 North Carolina producers of Grade A milk. In the first case assume that we are trying to estimate the average number of trees per farm; in the second case we are trying to estimate the average number of cows per producer. Those two items are particularly well suited to this study because unbiased estimates of both averages are available from independent sources to test the accuracy of the method.

TABLE 1.—*Results from repeated mailings to fruit growers and milk producers in North Carolina*

Mailing	Fruit growers		Milk producers	
	Schedules returned	Av. trees per farm	Schedules returned	Av. cows per farm
	Number	*Number*	*Number*	*Number*
1	300	456	165	23. 03
2	543	382	170	23. 79
3	434	340	114	24. 23
Total mailing list or universe	3, 241	329	1, 189	24. 27

On an inspection of table 1 two things are immediately apparent. First, the per farm averages drawn from the schedules received from the three successive mailings show trends in opposite directions, in the two surveys. In these surveys large-scale fruit growers seem to be more willing to return their mailed schedules than are small-scale growers; but small-scale milk producers seem to be more willing to report than are the large-scale producers. The second striking feature of the table is the smoothness of the trend in the per farm averages, in both surveys.

Question To Be Solved

The problem to be solved is now clear: Is there a general mathematical law that will enable a statistician to project a trend based on results from three or more successive mailings to arrive at the correct universe average, corresponding to a 100-percent response?

Working Toward the Solution

To arrive at the mathematical form of the law that seems to be suggested by the data at hand, the first fact that seems pertinent is that the number of the mailing measures the resistance of the respondents to returning the schedules. It will be assumed that each of the 300 fruit growers who responded to the first mailing has a resistance of 1 unit; each of the 543 who responded to the second request has a resistance of 2 units, and so on. As these resistance units, which may be represented by X, lie on a scale ranging from zero to infinity it seems reasonable to assume that $\log \frac{X}{\bar{x}}$ is Normally distributed about zero in the universe. In this expression $\bar{x}$ represents the average resistance of the individuals on the mailing list.

This assumption can be readily tested. The fraction of farms responding to the first request represents the area under the tail of the log X frequency curve extending from log (0) to log (1); the total fraction responding to the first and second mailings combined represents the area under the frequency curve from log (0) to log (2), and so on. The Normal deviates corresponding to these fractions can be found in any table of the Normal Probability Integral. If log X is Normally distributed, the values of log X should be linearly related to those Normal deviates. The Normal deviates corresponding to the three values of X are compared with values of log X in table 2, for the two sets of data given in table 1.

When values of log X are plotted against the Normal deviates shown in table 2, the points lie sufficiently close to a straight line to verify the

TABLE 2.—*Normal deviates compared with logarithms of resistance to returning a schedule*

Resistance X	Log X	Fruit growers		Milk producers	
		Total fraction responding	Normal deviate	Total fraction responding	Normal deviate
1	0.000	0.093	−1.323	0.139	−1.085
2	.301	.260	−.643	.282	−.577
3	.477	.394	−.269	.378	−.311

assumption that log X is Normally distributed (fig. 1). This linear relationship also makes it possible to determine the average resistance of all individuals in the population to returning a schedule. If the Normal deviate is represented by D and the standard deviation of the logarithms of the resistances is σ, we have the equation

$$\log \frac{X}{\bar{x}} = \sigma D$$

or

$$\log X = \sigma D + \log \bar{x} \quad . \quad . \quad . \quad (1)$$

This equation shows that when log X is plotted against D, the slope of the line represents the standard deviation of the logarithms of the resistances in the universe and the intercept on the vertical axis at D=0 represents the logarithm of the average resistance.

Using regression equations fitted by eye, we find that $\log \bar{x}$ for the universe is equal to 0.600 for the fruit growers and 0.665 for the milk producers. This shows that the average resistances are 3.98 and 4.62 respectively. The slopes of the lines, representing the standard deviations of the logarithms of the resistances, are 0.454 and 0.613. These standard deviations are of no particular concern in the problem at hand, but they provide some useful side information. For example, they enable one to predict the number of successive mailings that would be required to achieve any specified degree of completeness in the coverage of the universe. But at the moment we are more interested in the average resistances.

The next step in the analysis involves studying the relationship that is present between the resistance to returning a schedule and the farmer's scale of operations. If we let Y represent the

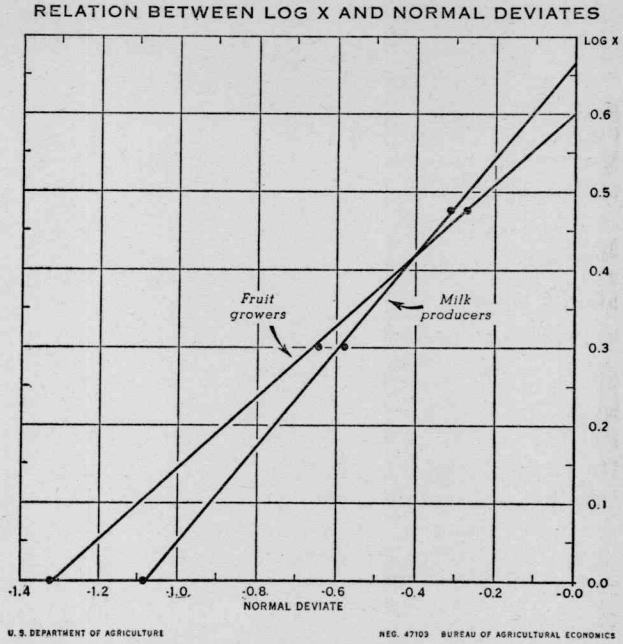

FIGURE 1.—RELATION BETWEEN LOG X AND NORMAL DEVIATES.

average number of fruit trees per farm for growers who have a specified resistance X, the relationship between Y and X can be used to estimate the universe average $\bar{y}$. But this relationship is as yet unknown. We can proceed under the assumption that a second-degree interpolation formula of the Gregory-Newton type will provide a satisfactory approximation. This is simply a quadratic equation with its constants so determined that it will fit exactly the three points on a chart that represent the data for the three mailings. It is assumed that this quadratic equation will represent the true relationship between Y and X, with a fair degree of accuracy, over a range of values of X that does not extend too far beyond the values used in fitting the equation. The relationship can then be represented by the equation.

$$Y = F(a+x) = F(a) + x\Delta F(a) + \frac{x(x-1)}{2}\Delta^2 F(a) \quad . \quad . \quad . \quad (2)$$

in which $\Delta F(a)$ and $\Delta^2 F(a)$ are the first and second differences of the number of trees per farm shown in table 1, and $F(a)=456$. We have

a+x	x	F	ΔF	Δ²F
1	0	456		
2	1	382	−74	
3	2	340	−42	+32

This yields the equation

$$Y=456-74x+\frac{32x(x-1)}{2} \quad . \quad . \quad . \quad (3)$$

in which $x=X-1$.

Substituting the universe average $\bar{x}=3.98$ for X, or $x=2.98$, in this equation, should give the universe average number of trees per farm. Making the necessary computations yields the following result:

$$\bar{y}=456-(74)(2.98)+(16)(2.98)(1.98)=329.9$$

This value is approximately equal to the value 329 which happens to be known in advance in this case.

Applying the same analysis to the milk-producer data gives:

a+x	x	F	ΔF	Δ²F
1	0	23.03		
2	1	23.79	+0.76	
3	2	24.23	+0.44	−0.32

$$Y=23.03+0.76x-\frac{0.32x(x-1)}{2}=$$

$$23.03+(0.76)(3.62)-(0.16)(3.62)(2.62)=24.26$$

This also closely approximates the known universe figure of 24.27.

Discussion of Results

Everything considered, this method shows considerable promise as a basis for adjusting the results of a mail survey for incompleteness. It is necessary, of course, to have a representative sample of the universe for the original mailing list, such as might be obtained from a basic enumerative survey, and to have data from at least three mailings before the method can be applied. But this is not an unsurmountable obstacle. It certainly involves much less work and expense than an enumerative survey of even a small sample of non-respondents.

In many ways the behavior of data obtained from successive mailings is analagous to the behavior of data obtained from successive "call-backs" in an enumerative survey, although a different set of factors is operating in the two situations. But the methods described here are probably not to be recommended for the analysis of call-back data because of the high degree of completeness that is attained both before and after the call-backs are made. Under such conditions, another technique of adjusting for incompleteness, when the call-backs fail to attain 100-percent completeness, would probably be more satisfactory. It might be mentioned as a point of interest that the possibility of using the present approach in mail surveys first occurred to the writer in connection with discussions of the call-back problem in enumerative surveys.

Data from more mailed surveys need to be investigated from this viewpoint to learn whether relationships of the kind found here represent the general rule. There is reason to believe that this will be the case, but the point should certainly not be accepted without further investigation. The relationships found with different kinds of subject matter would be of particular interest.

The mathematical form of the relationship between Y and X also needs further investigation with more extensive data, particularly for data in which more than 50 percent of the schedules are returned. A quadratic equation cannot represent the relationship for values of X covering a range that gives cumulative responses ranging from a figure of less than 50 percent to one of more than 50 percent. If Y is a function of X represented by F (X), we must be dealing with a function of such a form that $F[-(X-x)]$ is at least approximately equal to $-F(X-\bar{x})$. The data at hand do not cover a sufficiently wide range of potential responses to permit the determination of the true nature of this relationship. Another refinement that seems to be called for would be to use a value of $X=0.5$ rather than 1.0 for the first mailing, $X=1.5$ rather than 2.0 for the second, and $X=2.5$ rather than 3.0 for the third. The average resistance of the respondents in each of the three categories returning the schedules would be represented more accurately by those numbers. It may be pointed out that in both sets of data discussed in this paper $Y=F(X)$ could be represented accurately over the range of values of X with which we are dealing by a simpler quadratic equation of the form $Y=\bar{y}+b(X-\bar{x})^2$. It would be interesting to learn whether this is accidental or whether it is characteristic of the behavior of mail-survey data over this range.

In general, the results obtained here indicate that the resistance of a potential respondent to returning a schedule changes rather rapidly as we move a short distance in either direction from the average amount of the item on hand, but that

the resistance tends to stabilize when we reach respondents who have relatively large or relatively small quantities. It is known, for example, that a large-scale fruit grower has about the same psychological attitude toward returning his schedule regardless of whether he has 5,000 trees or 50,000. Similarly, small-scale producers are about the same sort of individuals, on the average, regardless of whether they have 10 trees or 50. The big differences in the kind of people who form the universe of potential respondents are found somewhere between an upper limit on the small-scale producers and a lower limit on large-scale producers.

The methods discussed here do not seem to work well when the universe is small. Table 3 shows the results obtained by sending four successive requests to a universe of 253 chick hatcheries in North Carolina. The hatching capacity of the hatchery is the variable under study.

The schedules returned on the successive mailings represent such small samples that a few unusually small or large hatcheries make the resulting average of the capacities rather erratic. Although there was an obvious tendency for a greater proportion of the larger hatcheries to respond to the survey, no clear-cut trend in the average capacity from one mailing to the next can be seen. It should also be borne in mind that no method of sampling is very efficient when the universe is small and subject to a high degree of variability.

TABLE 3.—*Results from repeated mailings to chick hatcheries in North Carolina*

Mailing	Schedules returned	Average egg capacity
	Number	*Thousands*
1	71	52. 5
2	43	52. 3
3	14	61. 5
4	14	56. 5
Universe	253	46. 3