Adjustment for Non-Response Bias in a Rural Mailed Survey

By A. L. Finkner

Reasonable reliability of the mailed inquiry has been attained by Agricultural Estimates and other agencies whose restricted budgets require them to rely largely on this method of collecting data for their surveys. Reliability is achieved mainly by building up historical series of mail-survey results plotted against more accurate data obtained later. But as there are no historical series for some surveys that are desired, different techniques must be used to increase their accuracy. This paper is confined to the statistical analysis of one source of bias in estimates—the bias of non-response—in a survey conducted by the North Carolina Field Office of Agricultural Estimates. It is possible in some instances, according to the author, to estimate totals of agricultural items reliably by using information from successive waves of mailed inquiries. He is convinced from the accumulated evidence that a general law governing reliability is operating but that further research is needed to learn its precise character.

AMONG the possible sources of biases in estimates resulting from the use of mailed questionnaires three important ones are: (1) Bias of a selective list, (2) bias of interpretation, and (3) bias of non-response. Bias of a selective list arises from failure to use a probability sample in selecting the original mailing list. Bias of interpretation is the difference that may result from asking questions by mail and asking in personal interviews. This bias has not been investigated extensively as it involves putting questions to the same individuals, both in writing and orally. This source of bias may be serious, especially in the case of difficult questionnaires.

Ways of Treating Non-Response Bias

In the present report results and remarks are confined to the third source of bias—the bias of non-response. It is widely known that the characteristics of the respondent population may differ from those of the non-respondent population. Research workers have proposed four main ways of treating non-response bias: (1) use check data for the purpose of establishing historical series; (2) use control data for purposes of stratification, and regression estimation or both; (3) mail successive waves of questionnaires, which is the technique known by some as sampling in depth; and (4) make personal enumeration of a subsample of persons who refuse to respond by mail. Of these, only (4) will guarantee unbiased estimates, although good results are obtained from the other methods. In this investigation, we are concerned with a comparison of results obtained from (2) and (3).

In October and November 1948, mailed inquiries were sent to all farmers in three North Carolina counties, Caswell, Edgecombe, and Macon.[1] A farmer was defined as a rural-tract owner listed by the 1947 North Carolina Farm Census. The list was originally taken from the tax-scroll books of these counties. As the North Carolina Farm Census is reported on the basis of tracts, all tracts under one name in one township were combined and only one schedule was mailed to cover the operations on such combined tracts. A second request was sent to those farmers who failed to respond to the initial inquiry, and a third went to those who did not respond to either. The number and percentage responding by request in each of the three counties are summarized in table 1. The patterns of response in the three counties were similar to those in previous mailed surveys in North Carolina. In each county, the second request brought in a larger response than either the first or the third request.

Expansions of the data into county estimates were made by different methods and are designated in this report as the regression method (using control data) and the extrapolation

[1] The mailings were made by the North Carolina field office of Agricultural Estimates under the supervision of Frank Parker and R. P. Handy.

TABLE 1.—*Number and percentage of farm owners responding to three successive waves of a mailed inquiry in Caswell, Edgecombe, and Macon Counties.*

Request	Caswell		Edgecombe		Macon	
	No.	*Percent*	*No.*	*Percent*	*No.*	*Percent*
1	360	18.3	239	13.3	497	18.5
2	494	25.0	400	22.3	789	29.3
3	302	15.3	232	13.0	344	12.8
N.R.[1]	817	41.4	922	51.4	1,060	39.4
Total	1,973	100.0	1,793	100.0	2,690	100.0

[1] Non-respondents—those who failed to reply to any of the three mailed inquiries.

method (using successive waves). The regression approach was suggested by Hendricks (*2*) in which the regression coefficient b is computed by a simple method of averages rather than by least squares. To utilize this method, the sampling units are separated into approximately equal groups on the basis of size with respect to y, the item being estimated. For each sampling unit, there must be available the value of an auxiliary variable x. Further, population values for x must be known. Averages are then computed for x and y for each group. The estimate of b then becomes

$$b = \frac{\bar{y}_e - \bar{y}_s}{\bar{x}_e - \bar{x}_s}.$$ [2]

The e subscript refers to the group of large sampling units and the s to the group of small units. The estimate of a, the y intercept, is given as

$$a = \bar{y}_e - b\bar{x}_e = \bar{y}_s - b\bar{x}_s$$

The estimate of the population mean for a given item is then

$$\bar{y} = a + b\mu_x$$

and the estimate of the population (county) total is

$t = N(\bar{y})$ where
μ_x is the known county mean for the auxiliary variable, and
N is the total number of farms in the given county.

As data were available by sample farm for all items in 1947 and 1948, this regression approach is applicable here. A complete Farm Census[3] was taken so that 1947 population values were known for each of the three counties. The measure of the item taken in 1947 was used as the x variable and the measure of the same item in 1948 was taken as the y value. In this analysis, the data from all three requests were combined. The county estimates based on the regression method are given in tables 3, 4, and 5.

In an estimation of county totals, using the information from successive waves of requests, an attempt was made to utilize the "resistance" method described by Hendricks (*3, 4*). In brief, Hendricks' method is based on two assumptions:

(1) That those responding in each successive wave do so under more pressure; that is, they have progressively more resistance to releasing their information to a surveying agency. The logs of these resistances are assumed to be normally distributed, which allows the mean resistance to be estimated from the median response.

(2) That some definite basic relationship exists between the average resistances within each wave and the item means for each wave. On the basis of this relation, the population mean can be estimated from the mean resistance. This relationship was assumed to be quadratic in (*3*) and cubic in (*4*).

The resistance method was developed empirically using data from two previous mail surveys in North Carolina. Each of the surveys was concerned with only one item, fruit trees in one, and milk cows in the other. The proposed method

[2] Bartlett (*1*) has shown that this procedure in estimating b has an efficiency equal to or greater than ¾. The efficiency can be increased equal to or greater than 8/9 by dividing the sampling units in the sample into three groups (approximately equal in number) on the basis of size with respect to the item being estimated. The averages of the first and third groups determine the slope of the line which is to be run through the mean coordinates of the three groups combined to obtain the y intercept.

[3] Although a complete enumeration was undertaken in each of the three counties in 1947, there was a small percentage of non-response. The figures adopted officially for the 1947 Farm Census were adjusted for this non-response and the adjusted figures were accepted in this analysis as being the population values.

worked well in both surveys. In the county surveys under investigation here, the assumptions do not hold. When the log of the resistance is plotted against the normal deviate corresponding to the cumulative percentage of response, there is considerable deviation from a straight line, indicating that the assumption of normality does not fit well in these circumstances. Similarly, neither a quadratic formula of the Gregory-Newton type (*3*) nor a cubic formula (*4*) seems to explain the relation between the item means of successive waves and the resistances.

One possible explanation for the discrepancy in our case is that resistance is a complex function of all items and length of schedule, whereas in the case Hendricks investigated resistance was coupled almost exclusively to the size of the operation of one particular item. This hypothesis is under study at the Research Office of the Bureau of Agricultural Economics at the Institute of Statistics in Raleigh, North Carolina.

TABLE 2.—*Cumulative totals of farmers responding in Caswell County and the cultivated land held by those responding.*

Request	Respondents		Cultivated Land	
	No.	Log of No.	Acres	Log of Acres
1	360	2.556	10,935	4.039
2	854	2.931	25,649	4.409
3	1,156	3.063	34,861	4.542
County	1,973	3.295	(59,293)	(4.773)

Some of the earlier exploratory work on successive waves indicated that a straight-line relationship often existed between the log of the cumulative number responding, and the log of the cumulative total for a given item. For example, in Caswell County, the pertinent data for

TABLE 3.—*Comparison of acreage of various agricultural item totals in Caswell County as estimated by different procedures.*

Item	Unit	1948 regression estimate	1948 extrapolation estimate	1947 [1] N. C. Farm Census	1945 U. S. Census	1950 U. S. Census
Farms	No.	1,973	1,973	1,973	2,689	3,051
All people living on farms	No.	15,513	16,218	15,597	12,428	[2]
All land in farms	Acres	267,653	268,540	259,776	218,239	244,036
Cultivated land	Acres	58,140	59,293	58,046	50,330	51,088
Idle land	Acres	29,089	29,174	29,722	20,989 [3]	29,251 [3]
Pasture land	Acres	17,729	19,907	13,492	17,697 [4]	21,202 [4]
Corn for all purposes	Acres	19,870	20,417	18,852	18,274	17,600
Tobacco	Acres	11,691	12,106	14,349	11,462	12,122
Soybeans, grown alone for all purposes	Acres	743	679	372	18	151
Wheat	Acres	5,065	5,470	6,200	5,811	5,726
Oats	Acres	2,311	2,618	636	902 [5]	1,579 [5]
Small-grain hay	Acres	3,591	3,622	1,641	131	912
Lespedeza for hay	Acres	12,903	13,428	8,735	10,622	11,156
All other hay	Acres	2,632	2,831	1,287	[2]	[2]
Potatoes	Acres	857	838	393	433	152
Sweetpotatoes	Acres	761	759	377	426	152
Fertilizer used	Tons	11,488	11,534	10,941	[2]	[2]
Milk cows and heifers 2 years old and over	No.	4,844	5,035	3,949	4,089	3,889 [6]

[1] Adjusted to 100 percent completeness.
[2] No comparable data available.
[3] Includes crop failure.
[4] Cropland used only for pasture plus "other" pasture. (Does not include woodland pastured.)
[5] Oats threshed or combined plus oats cut for feeding unthreshed.
[6] Cows including heifers that have calved.

TABLE 4.—*Comparison of acreage of various agricultural item totals in Edgecombe County as estimated by different procedures.*

Item	Unit	1948 regression estimate	1948 extrapolation estimate	1947 [1] N. C. Farm Census	1945 U. S. Census	1950 U. S. Census
Farms	No.	1,793	1,793	1,793	3,615	3,976
All people living on farms	No.	23,211	18,837	20,390	18,217	([2])
All land in farms	Acres	295,495	234,420	289,026	261,226	274,363
Cultivated land	Acres	130,097	103,510	128,387	116,563	120,485
Idle land	Acres	8,413	7,294	8,133	7,228 [3]	9,553 [3]
Pasture land	Acres	9,357	8,590	7,163	7,080 [4]	11,845 [4]
Corn for all purposes	Acres	43,098	36,307	40,192	39,125	44,840
Cotton	Acres	20,282	15,276	17,662	18,104	22,748
Tobacco	Acres	14,792	12,883	18,990	16,729	15,264
Peanuts, grown alone for all purposes	Acres	28,864	20,749	26,519	25,342	21,258
Soybeans, grown alone for all purposes	Acres	10,779	8,375	9,999	6,629	9,357
Wheat	Acres	1,021	565	830	1,679	574
Oats	Acres	3,031	2,748	1,644	3,702 [5]	2,621 [5]
Small-grain hay	Acres	1,981	1,652	1,795	48	1,942
Lespedeza for hay	Acres	2,622	2,250	2,068	1,631	1,861
All other hay	Acres	1,094	1,052	786	([2])	([2])
Potatoes	Acres	873	679	645	622	210
Sweetpotatoes	Acres	838	697	717	675	383
Fertilizer used	Tons	26,925	20,606	25,195	([2])	([2])
Milk cows and heifers 2 years old and over	No.	2,088	2,818	1,826	3,241	3,666 [6]

[1] Adjusted to 100 percent completeness.
[2] No comparable data available.
[3] Includes crop failure.
[4] Cropland used only for pasture plus "other" pasture. (Does not include woodland pastured.)
[5] Oats threshed or combined plus oats cut for feeding unthreshed.
[6] Cows including heifers that have calved.

establishing this relation for the item, cultivated land, is given in table 2.

When the log of the acres of cultivated land, y, is plotted against the log of the number responding, x, a straight line results. By extrapolation, the acres of cultivated land for Caswell County can be estimated. That is, the value of y (4.773) corresponding to an x value of 3.295 is determined from the straight-line relationship and the anti log of 4.773 is 59,293, the estimate of cultivated acres in Caswell County.[4] These extrapolation estimates are given, along with the regression estimates in tables 3, 4, and 5. Also shown for purposes of comparison are figures for the adjusted 1947 North Carolina Farm Census, the 1945 United States Census, and preliminary 1950 United States Census figures that are available. As sampling errors cannot be computed for either of these estimates, their accuracy must be judged by a comparison with other information—not too satisfactory a procedure. Check data are available on certain items, such as tobacco acreage. The study might have been materially improved by interviewing a subsample of the non-respondents so that unbiased estimates with calculable sampling errors could have been obtained.

[4] If all three points fell on a straight line the line was merely extended to obtain the estimate. If there were any perceptible departures from linearity, a least-squares estimate was computed.

With the possible exception of acres of all land in farms, which should be fairly stable from year to year, both types of estimates in Caswell County appear to be reasonable. Although the two estimates exceed the United States Census figures considerably, it should be kept in mind in making the comparisons that the same definitions may not operate in the United States and the North Carolina Farm Censuses. The estimates exceed the 1947 farm census figure by approximately 3.0-3.5 percent. The two estimates of tobacco acreage are close to the actual Production and Marketing Administration measured acreage in Caswell County in 1948. Considering a 15-percent cut in acreage from 1947 to 1948, they also agree closely with the 1947 North Carolina Farm Census figure.

The two estimates themselves are consistent; the extrapolation estimate exceeds the regression estimate slightly in 14 items, and is less in three.

About the same conclusions can be drawn from the estimates for Macon County. Here the only bad estimate appears to be the extrapolation estimate of acres of total land in farms. Again, this discrepancy may arise from differences in definition. The tax-scroll books list all rural tracts; many of these in the mountains are entirely wooded and ordinarily would not be classified as farms. But the owners may have replied to the inquiry and listed the acres of land owned as land in farms even though no farming was done. The over-estimate would not be reflected in other items, as zero would be recorded for them on the schedule. In Macon County, the regression estimate was larger than the extrapolation estimate in 11 items, and less in 5. Differences were slight, with the exception of all land in farms.

TABLE 5.—*Comparison of acreage of various agricultural item totals in Macon County as estimated by different procedures.*

Item	Unit	1948 regression estimate	1948 extrapolation estimate	1947 [1] N. C. Farm Census	1945 U. S. Census	1950 U. S. Census
Farms	No.	2,690	2,690	2,690	2,350	2,276
All people living on farms	No.	10,656	10,691	10,316	10,360	([2])
All land in farms	Acres	138,124	151,010	138,031	136,238	131,712
Cultivated land	Acres	22,172	22,131	23,000	21,624	17,986
Idle land	Acres	7,816	7,431	9,823	6,447 [3]	6,983 [3]
Pasture land	Acres	23,389	23,174	27,263	25,841 [4]	24,547 [4]
Corn for all purposes	Acres	10,530	10,544	9,086	11,437	8,721
Soybeans, grown alone for all purposes	Acres	627	564	855	392	331
Wheat	Acres	316	266	223	436	197
Oats	Acres	838	781	149	905 [5]	209 [5]
Small-grain hay	Acres	1,523	1,340	1,369	327	1,050
Lespedeza for hay	Acres	2,359	2,238	1,479	1,153	938
All other hay	Acres	3,685	4,457	3,031	([2])	([2])
Potatoes	Acres	972	933	937	862	510
Sweetpotatoes	Acres	322	324	84	220	38
Fertilizer used	Tons	3,476	3,404	2,768	([2])	([2])
Milk cows and heifers 2 years old and over	No.	4,228	4,217	3,458	4,469	3,877 [6]

[1] Adjusted to 100 percent completeness.
[2] No comparable data available.
[3] Includes crop failure.
[4] Cropland used only for pasture plus "other" pasture. (Does not include woodland pastured.)
[5] Oats threshed or combined plus oats cut for feeding unthreshed.
[6] Cows including heifers that have calved.

Results in Edgecombe County gave an entirely different picture. For most items, the discrepancy between the extrapolation estimates and the regression estimates was considerable. Except in the case of milk cows, the regression estimate was greater than the extrapolation estimate. In general, the regression estimates appear to be a little too high and the extrapolation estimates much too low.

Discrepancies in Edgecombe County Estimates Noted

The logs of the cumulative number responding by request were again plotted against the logs of the cumulative total for the item, by request. A straight line was then drawn to connect the first and third points. The items were classified into three groups on the basis of the relationship between the middle point and the line drawn between the first and third points. Group 1 consists of two items, idle land and milk cows, whose middle point was noticeably above the line; group 2 included two items, "all other hay" and sweetpotatoes, whose middle points fell on the line; and group 3 contained the remaining 15 items whose middle points were considerably below the line. A least-squares regression estimate was used to obtain the best fitting straight line through the three points for those items falling in groups 1 and 3. Estimates based on these lines are the extrapolation estimates given in table 4.

The estimates of the group-3 items might be slightly improved by using the straight line between the two end points instead of the least-squares line for extrapolation. Similarly, the estimate of milk cows in group 1 would be improved slightly, but the group 2 items would not be affected. The estimate of idle land in group 1 would become poorer as it is already underestimated. A curve drawn through the three points and extended to the population number of farms resulted in over-compensation for groups 1 and 3, with group 2 again not affected. In other words, a curvilinear adjustment now resulted in an over-estimate of all group-3 items and under-estimate of group-1 items.

Several possible causes for the discrepancies may be assigned. For one, the mean for the second-response group was, in most instances, lower than the means for the other two request groups. This was true, with the exception of idle land and milk cows. Although milk cows were over-estimated, idle land was under-estimated. Hence it appears that one or more other factors also operated. Their exact nature is not easily discernible, but these factors may be tied up with the response rate.

Conclusions

In conclusion, these data furnish additional evidence that information from successive waves of requests can be used in some instances to estimate totals of agricultural items with reliability. It is also clear that no general relationship yet established will always hold. If control data, such as previous information on the same item for the same sampling units, are available, a regression approach as described herein will provide satisfactory estimates. Evidence continues to indicate that a general law operates, but further research is needed to ascertain its exact character.

References

(1) BARTLETT, M. S.
FITTING A STRAIGHT LINE WHEN BOTH VARIABLES ARE SUBJECT TO ERROR. Biometrics 5 (3) : 207-212, 1949.

(2) HENDRICKS, W. A.
A REGRESSION METHOD FOR EXPANDING SAMPLE INDICATIONS TO STATE ESTIMATES. U. S. Bur. Agr. Econ. CRP No. 7, 1942 (Processed)

(3) HENDRICKS, W. A.
ADJUSTMENT FOR BIAS BY NON-RESPONSE IN MAILED SURVEYS. U. S. Bur. Agr. Econ. Agricultural Economics Research 1 (2) 52-56, 1949.

(4) HENDRICKS, W. A.
ADJUSTMENT OF DATA FOR NON-RESPONSE IN MAIL SURVEYS. IN THE AGRICULTURAL ESTIMATING AND REPORTING SERVICE OF THE UNITED STATES DEPARTMENT OF AGRICULTURE. U. S. Dept. Agr. Misc. Pub. 703, pp. 31-34. 1949.