A STUDY OF SAMPLING ERROR IN AN AREA-SEGMENT SAMPLE OF NEW YORK STATE FARMERS

by

D.S. Robson

BU-36-M April 1953

A survey study entitled "New York Farmers' Knowledge of, Participation in, and Suggestions on Agricultural Programs" was conducted in the Fall of 1951 through the cooperative efforts of the Extension Service, the Experiment Station of the New York State College of Agriculture, and the Bureau of Agricultural Education of the New York State Education Department. The general purpose of the study is indicated by its title and the main results now appear in Cornell Extension Bulletin 864 "New York Farmers' Opinions on Agricultural Programs" by Edward O. Moe. The present study, consisting of an investigation into the sampling variability inherent in a survey such as this is intended as an aid to investigators who may in the future conduct opinion surveys of New York State farmers. Most of the results which follow will apply only to surveys of similar design and only to studies of "full time farmers"; i.e., those who obtain at least half of their annual income from operating a farm.

The design employed in this survey is commonly known as the "stratified area-segment sample"; here the individual counties formed the strata and the the Master Sample segments formed the area-segments within each county. The number of sample segments for a county was determined by applying a constant sampling rate to the total number of Master Sample segments in the county; segments were then randomly selected with the aid of the Master Sample maps. Interviewers were given maps on which the sample segments were outlined and were instructed to interview all full time farmers in these designated areas. A total of 754 segments were selected in this manner with an aim to obtaining roughly 2000 interviews; earlier studies indicated that Master Sample segments in New York State contained an average of three full time farmers. The actual returns amounted to 1530 interviews and a known additional 179 eligible farmers were not interviewed. Information was also obtained on number of census farms and non-farm occupied dwelling units in each sample segment.

In the first presentation of the survey results the accuracy of the estimates was appraised by means of binomial probability theory; i.e., the stratified area-segment sample was regarded as equivalent to a simple random sample of fixed size 1530 from a single binomial population. It is not immediately clear whether this approximation would lead to an overestimate or an underestimate of sampling error; the stratification of the sample would tend to reduce sampling error below binomial variance; the clustering of the population elements within area segments represents an opposing force which tends to increase sampling error; the fact that sample size was in fact a chance quantity instead of fixed as in the binomial approximation has an unknown effect upon sampling error. An estimate of the amount and direction of bias in the binomial approximation was obtained by computing as a more precise estimate of sampling error the variance of a ratio of chance quantities. Table 1 and Figure 1 present a comparison of these two estimates of sampling error for 14 questions from different content areas of the questionnaire; the variances are compared on the basis of the confidence intervals which they generate under the normal approximation. Figure 1 reveals that the binomial approximation tended to underestimate sampling error to some extent, though the bias is negligible from a practical point of view. This is a heartening result in light of the fact that, due to its extreme simplicity, the binomial approximation is widely applied in practice.

The information on number of farm and non-farm occupied dwelling units in each sample segment provides a check on the present day accuracy of the Master Sample maps when the map count is compared to the observed count. Figures 2-6 show contrasts between the frequency distributions of observed counts and map counts, revealing that the indicated number of census farms on the Master Sample maps tends to be larger than the number of census farms actually found in the segments; likewise, the map count underestimates the number of non-farm occupied dwelling units while fairly close agreement exists between map count and observed count of the total number (farm and non-farm) of occupied dwelling units per segment. The means of these distributions are:

	Map	Survey
average number of census farms per segment	5.33	3.67
average number of non-farms per segment	3.63	6.18
average number of occupied dwelling units	8.96	9.85

Table 1

95 % Confidence Limits

Computed from

Estimated Sampling Error

Question	Estimated Percent Favorable p	Binomial Variance 1) $\frac{pq}{n}$	Variance of a Ratio 2)	Pooled Variance of a ratio 3)
11	70.78	68.51-73.06	68.38-73.19	68.32-73.24
13	29.41	27.13-31.70	26.83-31.99	26.95-31.87
16	44.51	42.02-47.00	41.90-47.12	41.82-47.20
34	83.27-	81.40-85.14	81.05-85.49	81.25-85.29
42	63.40	60.98-65.81	60.76-66.04	60.80-66.00
48	58.76	56.29-61.22	56.03-61.48	56.10-61.42
49	95.23	94.16-96.30	94.06-96.39	94.06-96.39
51	33.73	31.36-36.09	31.22-36.23	31.17-36.29
57	70.33	68.04-72.62	67.77-72.88	67.86-72.80
59	29.35	27.06-31.63	26.83-31.87	26.89-31.81
66	85.42	83.66-87.19	83.61-87.24	83.51-87.33
86	96.73	95.84-97.62	95.97-97.50	95.77-97.69
90	61.96	59.53-64.39	59.21-64.71	59.34-64.58
97	60.46	58.01-62.91	58.13-62.78	57.82-63.10

1) The limits are computed from $p \pm 1.96\sqrt{\frac{pq}{n}}$

where n=1530 and $p = \frac{\text{number of favorable responses in the sample}}{1530}$

2) The limits are computed from $p \pm 1.96\sqrt{\hat{V}(p)}$

where $\hat{V}(p) = r(1-r)\,\frac{p^2}{n^2}\sum_{i=1}^{k} N_i\left(s^2_{yi} + \frac{s^2_{xi}}{p^2} - 2\,\frac{\hat{\rho}_i s_{yi} s_{xi}}{p}\right)$

where r = .0282 + = the sampling rate

k = 56 = the number of strata or counties in the sample

N_i = the number of Master Sample segments in the i'th county

y_{ij} = the number of farms, or interviews, in the j'th segment of the i'th county

x_{ij} = the number of favorable responses among the y_{ij} interviews in the ij'th segment

s^2_{yi} = the sample variance of y_{ij} within the i'th stratum

s^2_{xi} = the sample variance of x_{ij} within the i'th stratum

$\hat{\rho}_i$ = the sample correlation between y_{ij} and x_{ij} within the i'th stratum

3) The limits are computed from the least squares curve fitted to the points in Figure 1

Figure 1. A Comparison of Confidence Intervals Computed From the Binomial Variance and the Variance of a Ratio. $n = 1530$

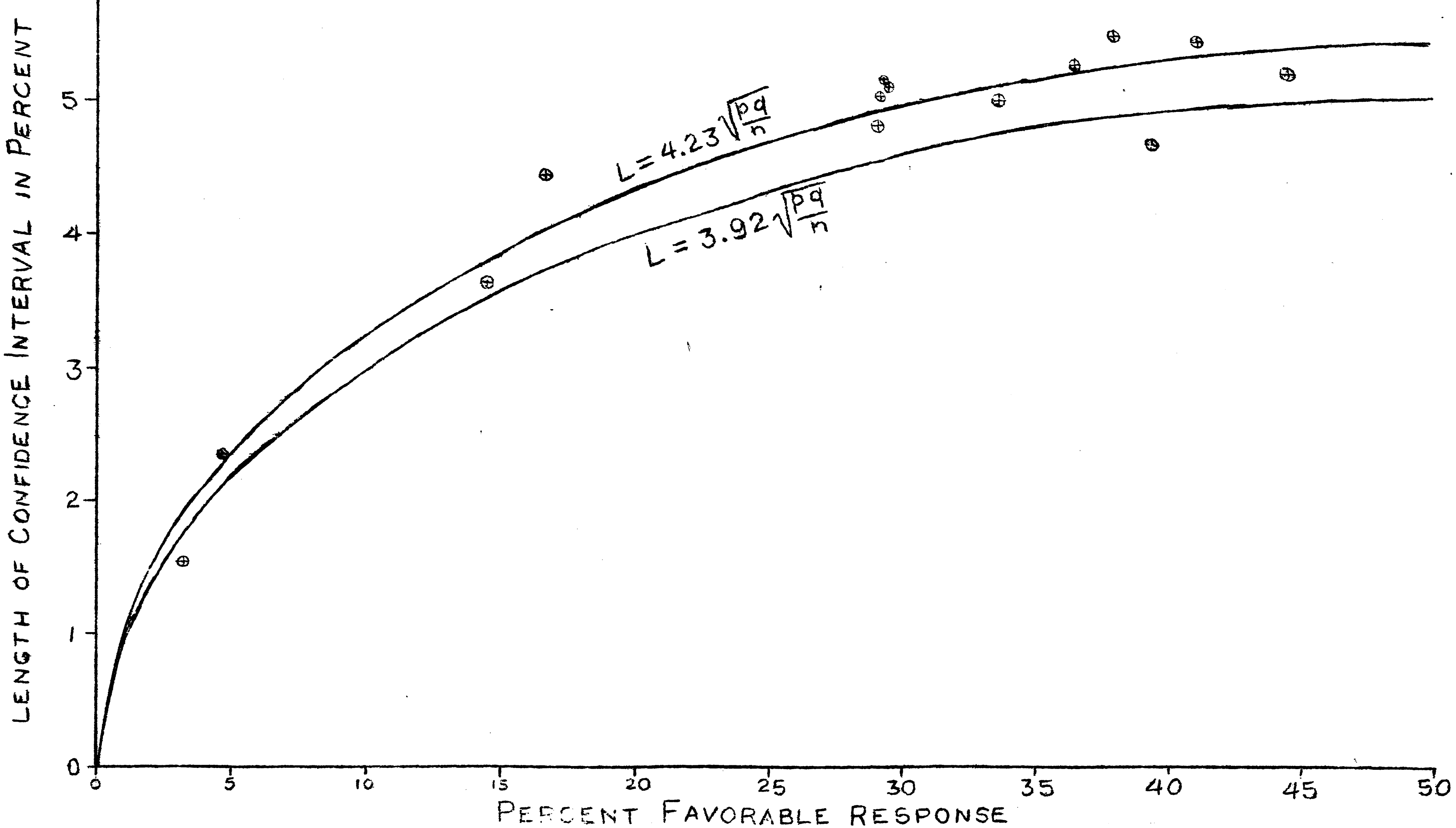


Figure 2. A comparison of the frequency distribution of the number of all farms per segment as indicated on the Master Sample maps and the number enumerated by a personal visit to the segment. Total number of segments = 407

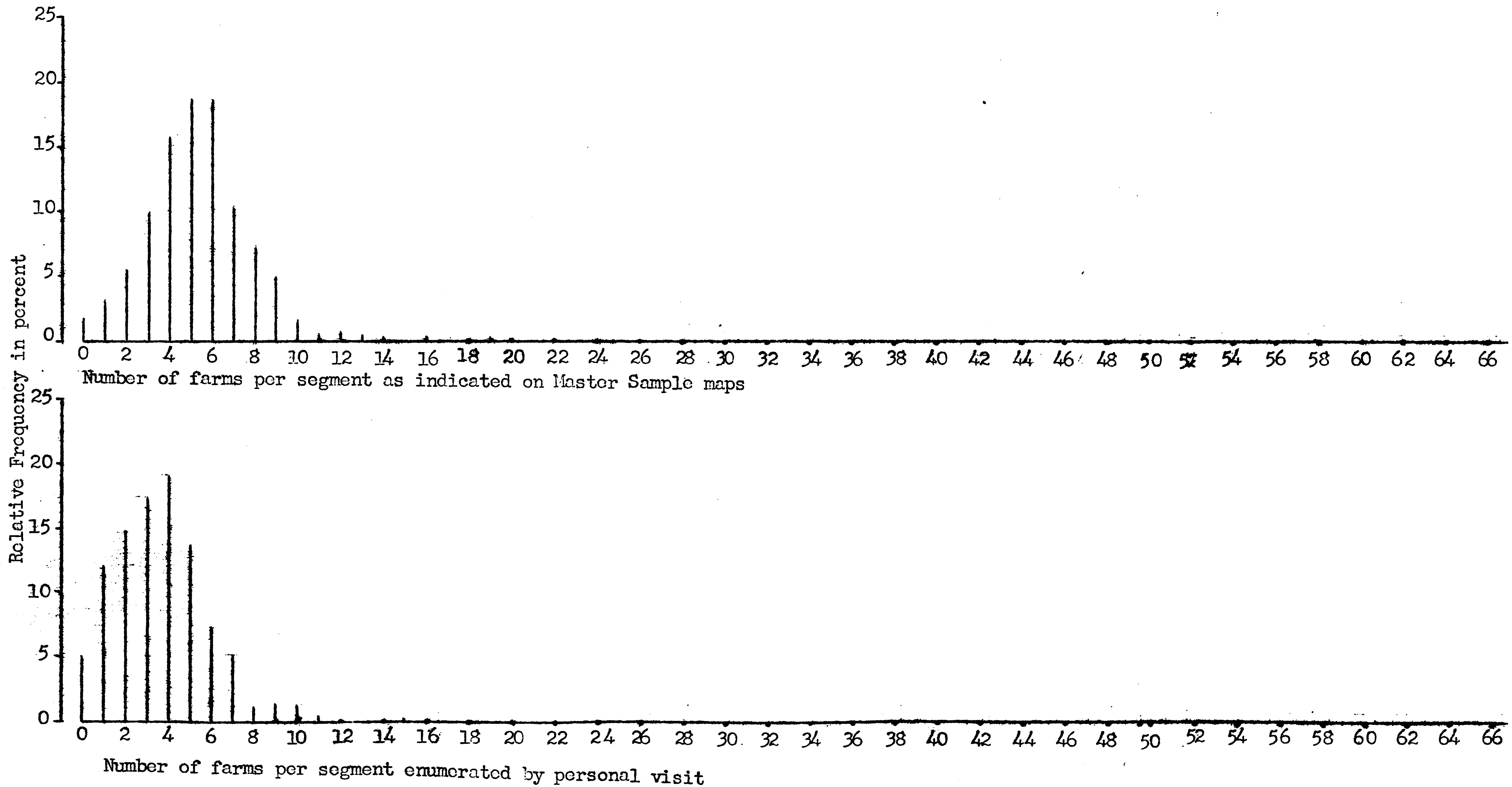


Figure 3. A comparison of the frequency distributions of the number of non-farm occupied dwelling units per segment as indicated on the Master Sample maps and the number enumerated by a personal visit to the segment.

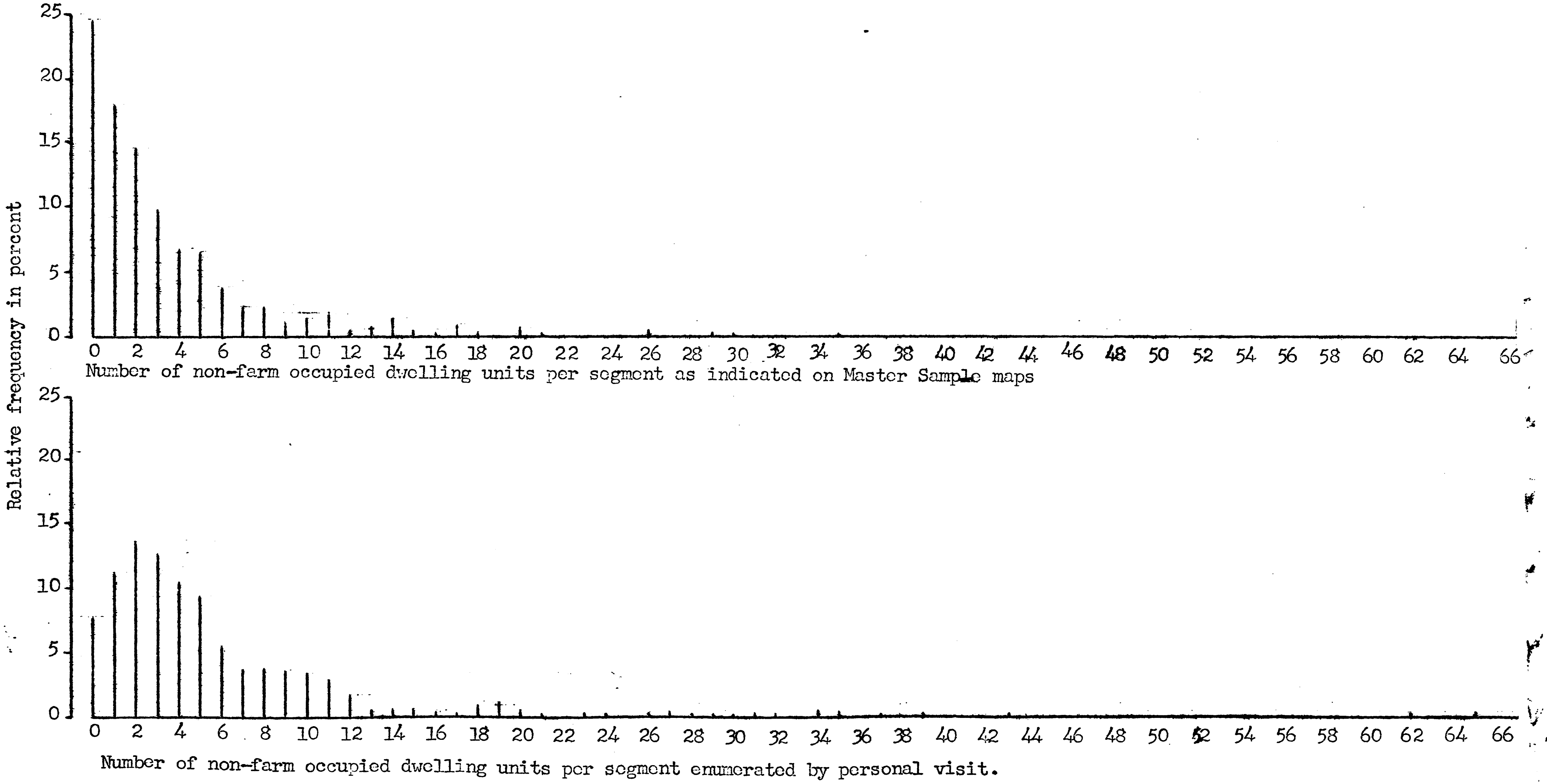


Figure 4. A comparison of the frequency distribution of the total number of occupied dwelling units per segment as indicated on the Master Sample maps and the number enumerated by personal visit.

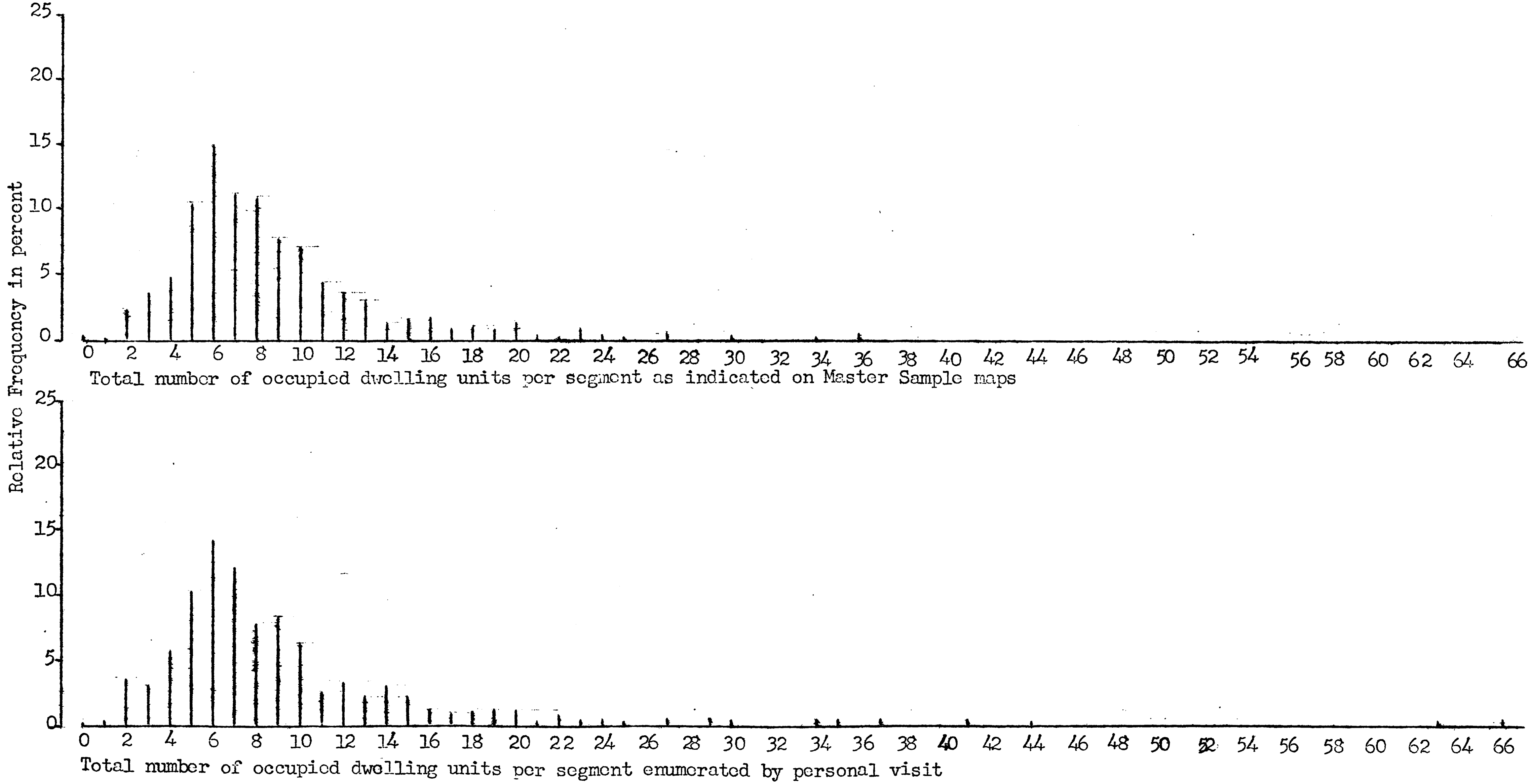


Figure 5. A comparison of the frequency distribution of the difference: number of farms per segment as indicated on Master Sample maps - number of farms unumerated by personal visit, and number of non-farms per segment as indicated on Master Sample maps - numbers of non-farms enumerated by a personal visit.

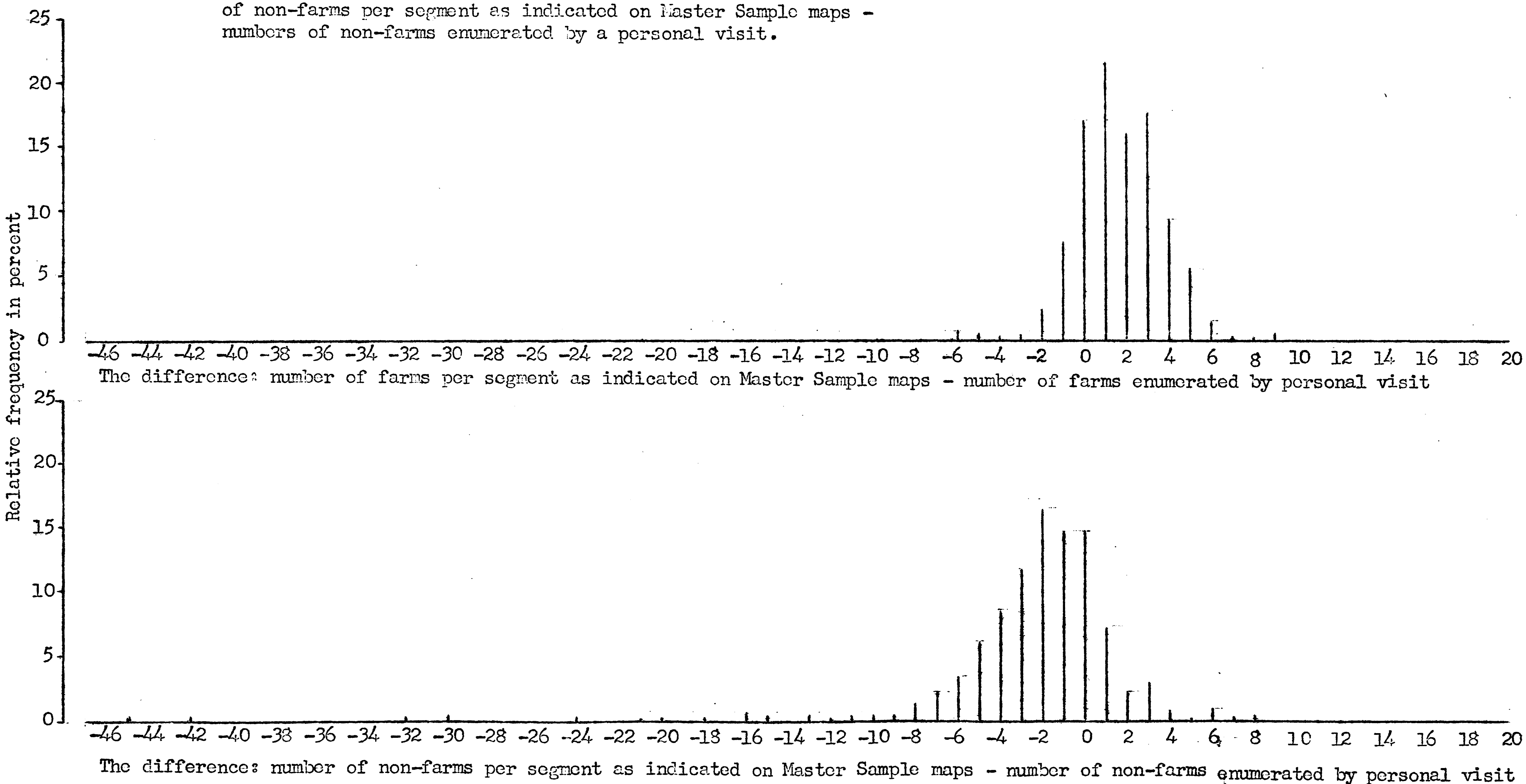
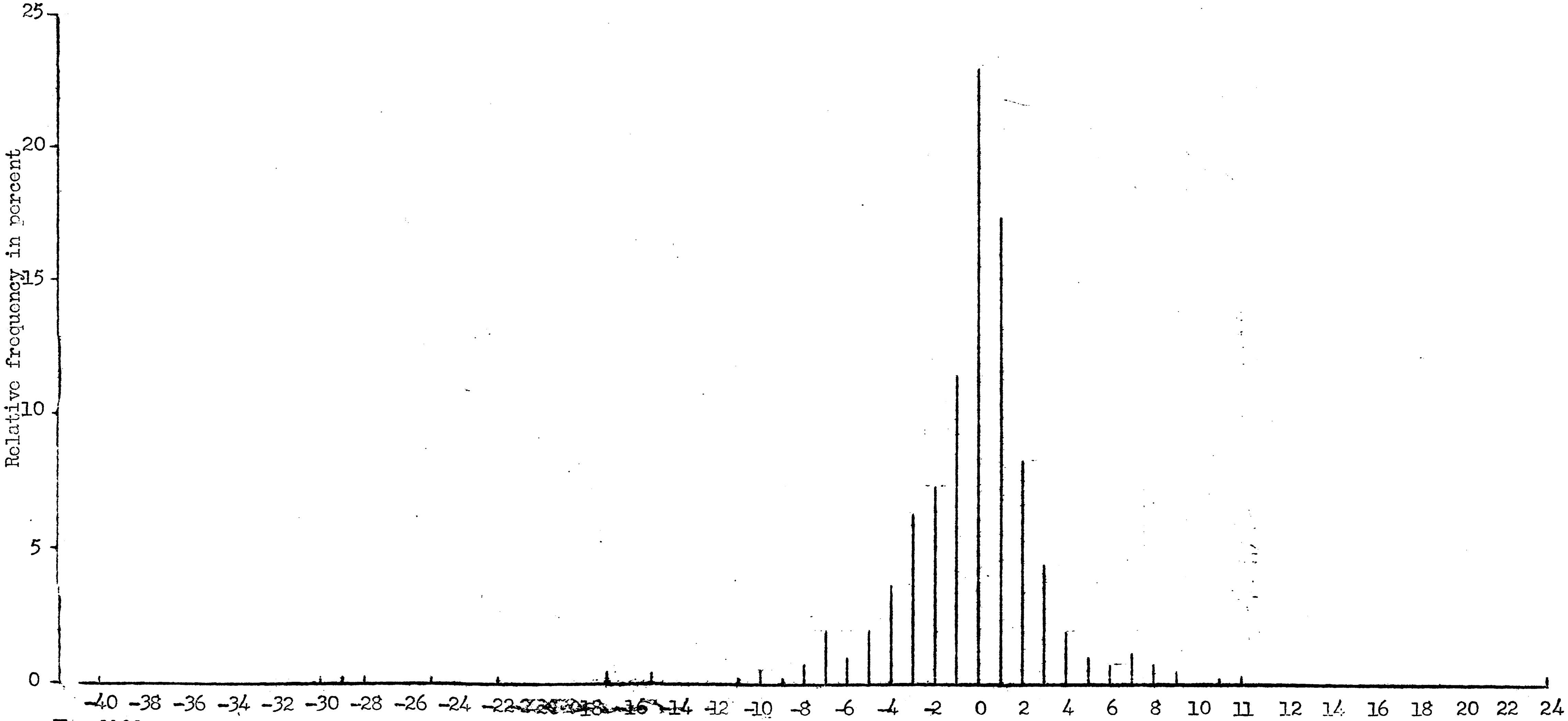


Figure 6. The frequency distribution of the differences number of occupied dwelling units per segment as indicated on Master Sample maps - number enumerated by a personal visit.



The close agreement between the binomial variance and the variance of a ratio suggests that the binomial approximation may be used satisfactorily in evaluating the accuracy of similar survey studies in the future. Likewise, binomial probability theory might be used to determine the sample size necessary to insure any specified degree of accuracy in the survey results. The accuracy of an estimate $\hat{p}$ is, however, measured by the variance pq/n which cannot be known in advance even when the sample size n is specified. Furthermore, the sample size n, measured in terms of number of interviews, is not under the complete control of the investigator; the number of sample segments may be specified in advance but the resulting number of interviews is a chance quantity and hence cannot be predicted with certainty. In general, however, one may safely assume that among the items on his questionnaire there is at least one for which the population splits roughly 50-50, where the quantity pq is maximized. Choosing a sample size to insure a specified degree of accuracy for such a question automatically insures an even greater degree of accuracy for other questions where the population split is different from 50-50. Thus, for example, the investigator may wish to know the number of sample segments to use in order to insure that when the population proportion is $p = 1/2$ his estimate $\hat{p}$ will lie within the interval $.475 < \hat{p} < .525$ with probability at least .95; in other words, he might wish to know the number of sample segments to use in order to insure with probability at least .95 that his estimate will lie within 2 1/2 percentage points of the population percentage which he is estimating. This required number of sample segments may be estimated quite accurately with the aid of the distribution of number of interviews per segment (Figure 7) obtained in this study.

Let k denote number of segments in the sample and N_k denote the number of interviews obtained from k sample segments; N_k is then a chance quantity, and we shall approximate its distribution by the normal distribution with mean $2.11k$ and standard deviation $1.78\sqrt{k}$. Likewise, under large sample theory we have that the sample proportion $\hat{p}$ is normally distributed with mean 1/2 and standard deviation $\frac{1}{2\sqrt{n}}$. Then we choose the smallest k for which

$$P\,[.475 < \hat{p} < .525 \mid k] \;>\; P\,[.475 < \hat{p} < .525 \mid N_k = n,\ k] \cdot P\,[N_k > n \mid k] = .95$$

where

$$P[.475 < \hat{p} < .525 | N_k = n, k] = P[2\sqrt{n}\,(.475 - .5) < t < 2\sqrt{n}\,(.525 - .5)] = a_n$$

where

$$t = \frac{\hat{p} - 1/2}{1/2}\sqrt{n}$$

is normally distributed with mean zero and variance 1; and, likewise,

$$P[N_k > n | k] = P\left[t > \frac{n - 2.11k}{1.78\sqrt{k}}\right] = a_k;\quad a_n a_k = .95.$$

Letting t_{a_n}, t_{a_k} be such that

$$P[-t_{a_n} < t < t_{a_n}] = a_n$$

$$P[t > t_{a_k}] = a_k$$

we have that

$$2\sqrt{n}\,(.525 - .5) = t_{a_n}$$

or

$$n = \frac{t_{a_n}^2}{.0025}$$

and

$$\frac{n - 2.11k}{1.78\sqrt{k}} = t_{a_k}.$$

hence

$$2.11k + 1.78t_{a_k}\sqrt{k} - n = 0$$

or

$$k = \left[\frac{-1.78t_{a_k} + \sqrt{3.17t_{a_k}^2 + .0211t_{a_n}^2}}{4.22}\right]^2.$$

The problem then is to determine the values of a_n and a_k which produce the smallest value of k. Perhaps the simplest procedure to follow is the iterative method which in this case gives the minimum k=808 for $a_n = .952$ and $a_k = .9979$.

We may in addition, present the investigator with a range on the number of interviews he may expect if he uses k = 808 sample segments; we may calculate two numbers $\underline{n}$ and $\bar{n}$ such that

$$P[\underline{n} < N_k < \bar{n} | k] = .95$$

which for k = 808 has a solution

$$\underline{n} = 2.11\,(808) - 1.96\,(1.78)\,\sqrt{808} = 1606$$

$$\bar{n} = 2.11\,(808) + 1.96\,(1.78)\,\sqrt{808} = 1804.$$

Table 2 presents additional results for various degrees of accuracy. The table applies only to survey studies of full time farmers in New York State where the survey design is identical to the one described here.

Table 2

Minimum number of sample segments required to assure with probability at least .95 that the estimate $\hat{p}$ lies within $\alpha/2$ percentage points of the population proportion p.

α	Number of Sample Segments	95% Range on Number of Interviews	Expected Number of Interviews
1%	18,747	39,078-40,034	39,556
2	4,776	9,836-10,320	10,077
3	2,164	4,403- 4,729	4,566
4	1,241	2,496- 2,742	2,618
5	808	1,606- 1,804	1,704
6	573	1,125- 1,293	1,209
7	427	829- 973	901
8	334	641- 769	705
9	269	511- 625	568
10	221	414- 518	466
11	186	345- 440	393
12	159	292- 380	336
13	138	251- 332	291
14	121	217- 294	255
15	107	190- 262	226
16	96	169- 237	203
17	86	149- 214	182
18	78	134- 195	165
19	71	121- 179	150
20	65	109- 165	137