JOURNAL OF RESEARCH of the National Bureau of Standards—C. Engineering and Instrumentation
Vol. 64C, No. 1, January–March 1960

A Statistical Chain-Ratio Method for Estimating Relative Volumes of Mail to Given Destinations

Norman C. Severo [1] and Arthur E. Newman

(August 31, 1959)

A sampling method, called the chain ratio method, is applied in estimating the distribution of mail by destination. Variances and coefficients of variation for the estimators are given. The details and results of three applications of this sampling method to outgoing first-class letter mail are given. These studies were conducted by the National Bureau of Standards in San Francisco, Los Angeles, and Baltimore.

1. Introduction

The National Bureau of Standards has been active in developing equipments and systems for improved letter sorting by automation. To develop design parameters it is necessary to determine the physical characteristics of mail and the proportion of mail going to various *destinations*.[2] Since the volume of mail is much too large for complete piece counts to be feasible, sampling methods of known and adequate accuracy must be used. The present paper is the first step by NBS in the effort to develop such methods as applied to mail distribution. Studies and results concerning letter-size characteristics are reported by Severo, Newman, Young, and Zelen in [1][3] and a general background to the mechanization program is given by I. Rotkin [2].

This paper discusses a sampling procedure designed to estimate the proportion of mail going to each *destination*. The sampling plan used in this study is referred to as the "chain-ratio" method because the nature of the formulas involved in the analyses resembles a chain of ratios. The method has been applied to outgoing first class letter-mail at the San Francisco, Los Angeles, and Baltimore Post Offices.

It was intended, initially, to study five cities: Baltimore, Washington, Philadelphia, Chicago, and Los Angeles. Philadelphia, Baltimore, and Washington were chosen because they would tend to give a pattern of postal operations on the East Coast. Chicago was chosen to show Midwest influence, and Los Angeles was selected to show the West Coast influence. San Francisco was added to the list in an effort to find out whether Los Angeles was atypical, because Los Angeles serves an unusually large area.

The Post Office Department made special studies in Philadelphia, Chicago, and New York, where in each case a complete count was made of the *total volume* of mail to each *destination* for either a 24- or 48-hour period of time. Actually this complete count was obtained by footage measurements of stacks of mail and a conversion factor of 290 letters per foot of mail was used. The NBS also made a modified version of the complete count on November 5, 1956, in Baltimore. In this count, only the *total volume* entering the system between 4 P.M. and 7 P.M. was included.

However, any complete count of large volumes of mail, even for short periods of time such as 3 hours, involves a considerable number of man hours and invariably tends to delay the normal function of sorting mail. Furthermore, any such complete counts are open to criticisms that may be leveled against complete enumeration methods. (The literature contains many examples [3, 4, 5, 6] comparing complete enumeration methods with statistically designed sampling procedures, and shows the desirability, from the economics and reliability point of view, of the sampling techniques.) A complete count of mail, properly done, say, for 24 hours, gives a good indication of what happens during a particular $^1/_{365}$ part of a year. If one wishes to enlarge this fraction then additional complete counts can be made. Thus to represent a particular $^5/_{365}$ part of a year one might take 5 consecutive days—e.g., Monday through Friday or Thursday through Monday depending upon whether or not the weekend is to be included. This is expensive and time consuming. Furthermore tremendous effort is needed on the part of all concerned to keep track of all the mail to each *destination*. Thus errors are bound to occur. Finally, the mail itself will tend to be delayed during such exhaustive counts. A sampling study, on the other hand, enables one to check the flow pattern of mail from time to time during any interval of time and with far less effort and disruption to routine operations than in a complete enumeration and hence may more accurately represent normal operations. Thus, for example, to obtain information about mail for some given week, samples may be taken for short intervals several times each day throughout the week. (Actually in the application discussed here, two samples a day were taken during a 5-day period excluding the weekends.) Or if one wanted to check

[1] Present address: University of Buffalo.
[2] Italicized terms have special meanings in this study and are defined in section 2.1 of this paper or in the *Postal Term Glossary*, U.S. Post Office Department, August 1956, P.O.D. Publication 18.
[3] Figures in brackets indicate the literature references at the end of this paper.

the behavior of mail for any other given time period, say some particular month or during the Christmas rush, then samples could be taken from time to time during that particular time period.

The destination data obtained by application of the chain-ratio method has been used as basic input for: (1) Simulation studies of the effectiveness of an NBS proposed sorting machine; (2) studies of comparative costs for various types of mechanized letter sorting systems, including the one embodied in the machine mentioned in (1); (3) analytic comparisons of suggested configurations for automatic mail sorting equipment [7]; and, (4) improvement of current sorting procedures.

Only some typical results of the San Francisco study are presented here. The reader is referred to [8] for detailed results of the San Francisco, Los Angeles, and Baltimore studies.

Section 2 gives the definitions as used in this paper and the model of the flow of mail that is studied. Section 3 presents in detail the sampling procedures, analysis, and the volume counts used for the particular applications discussed. Section 4 defines precisely the types of mail that were studied at San Francisco, Los Angeles, and Baltimore. Section 5 presents the details of the San Francisco study.

2. Definitions and the Model

2.1. Definitions

A list of definitions of terms, as used in this paper is given here for reference. These definitions are given in order to avoid misinterpretation and ambiguity because of postal language differences between post offices.

1. SEPARATION: a classification characterized by a labeled pigeonhole on a sorting case.
2. DESTINATION: a final *separation* made at a given post office. All *directs* and *residues* are included in this classification.[4]
3. DIRECT: a *destination* to a single given post office.
4. DISTRIBUTION: the function of physically sorting letters into their respective *separation* boxes.
5. PRIMARY: the first stage of *distribution* of outgoing mail.
6. SECONDARY: the second stage of *distribution* of outgoing mail.
7. TERTIARY: the third stage of *distribution* of outgoing mail.
8. BYPASS: mail which receives its first *distribution* in the *secondary* and *tertiary* cases. Also mail which goes directly to the city section.
9. RESIDUE: mail destined for post offices for which no direct *separation* is provided in a case or rack.
10. TOTAL VOLUME: the defined classes of mail studied. (*Total volume* is defined more explicitly as used in this study in section 4.)

The expression "off the *primary*, *secondary*, or *tertiary*" indicates mail which has just undergone that stage of *distribution*.

[4] *Air* mail and foreign mail off the *primary* are also considered *destinations* in this study.

2.2. The Model

The model for the operation of outgoing mail consists of a three stage sorting scheme which can be represented by a flow chart as given in figure 1. The *total volume* in the top box consists of those types of mail indicated in section 4. This volume then divides into two parts, that which goes to the *primary* and that which bypasses the *primary*. The *bypass* mail is sent either to the city section or to the *secondary*. Mail leaving the *primary* may go either to its *destinations* or to the *secondary*. The *secondary* consists of sections which can be numbered 1, 2, 3, . . . and which correspond to *primary separations* needing further *distribution*. We call the i-th section the "i-th *secondary*." From any section in the *secondary*, mail can go either to its *destination* or to one of the *tertiary* sections which can be numbered 1, 2, 3, . . . Therefore sections in the *tertiary*, corresponding to *separations* from the i-th *secondary* can be numbered $i1$, $i2$, $i3$, . . . The ij-th section is called the "ij-th *tertiary*." Mail leaving the *tertiary* goes directly to its *destinations*. A more detailed description of how a letter flows through this system is given in [9].

Since the model for incoming mail is similar to that for outgoing mail, the procedures discussed below may also be applied in studies of incoming mail.

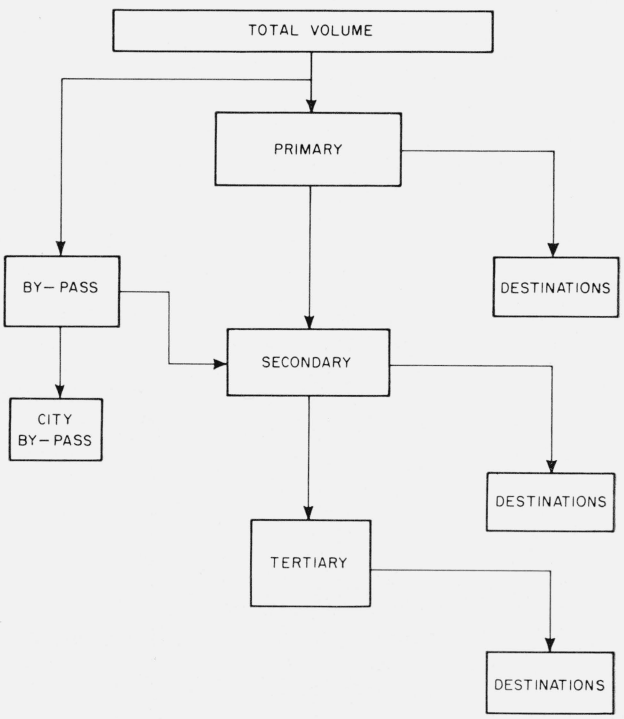


FIGURE 1. *Flow chart model for the distribution of outgoing mail.*

3. Chain-Ratio Estimates

In this section we discuss the estimation formulas and their associated variances and coefficients of variation for estimating the proportion of mail to a given *destination*. We also present a list of notations and specific formulas used in the applications given in section 5 and in [8].

3.1. The General Method

The basic idea involved in the estimation formulas consists of multiplying together a chain of ratios. Two conditions are required for setting up the chain. The first is that each ratio must be one that can be estimated conveniently. This can often be done by using volume count data customarily recorded by the particular post office. If such records are not kept then it must be possible either to arrange that they be kept or to devise appropriate sampling plans that would provide estimates of each ratio. It is essential that such plans be simple to implement and not interrupt the flow of the mail.

The second requirement is that the ratios must be linked together in chain form so that the desired ratio is all that remains after "canceling." This is similar to the usual chain differentiation carried out in the calculus. There if it is desired to obtain $\delta f/\delta x$, where $f=f[z[y[x]]]$, then one writes

$$\frac{\delta f}{\delta z}\times\frac{\delta z}{\delta y}\times\frac{\delta y}{\delta x}$$

and a "cancellation" check gives the desired results; i.e.,

$$\frac{\delta f}{\delta z}\times\frac{\delta z}{\delta y}\times\frac{\delta y}{\delta x}=\frac{\delta f}{\delta x}.$$

Such a "cancellation" is, of course, only a convenient artifice. It must be proved that the multiplication of such a chain of derivatives actually does yield the desired derivative $\delta f/\delta x$.

Here we have a similar situation. Suppose we are interested in estimating the ratio of mail to a *primary destination* to the *total volume*. Let us denote this ratio by the parameter D_P/T. Suppose that (1) the particular post office under study keeps records which enable us to obtain an unbiased estimate of the ratio of *primary* mail to the *total volume*, call this estimate T_P/T, and (2) it is possible to set up a simple sampling plan which yields an unbiased estimate of the ratio of mail to the *primary destination* to the *primary volume*, call this estimate D_P/T_P. Then we write $(D_P/T_P)(T_P/T)$, cancel, as in the calculus, and obtain the desired ratio estimate D_p/T. That such "cancellation" is permitted is seen as a special case of the following:

Let $R_1, R_2, \ldots, R_K$ be a set of K statistically independent random variables such that

$$E(R_i)=\frac{r_{i-1}}{r_i}, \quad i=1,2,\ldots,K.$$

Then for any $j\leq K$

$$E(R_1R_2\ldots R_j)=E(R_1)E(R_2)\ldots E(R_j)$$

$$=\frac{r_0}{r_1}\times\frac{r_1}{r_2}\times\ldots\times\frac{r_{j-1}}{r_j}$$

$$=\frac{r_0}{r_j}.$$

Thus $R_1R_2\ldots R_j$ is an unbiased estimate of the ratio r_o/r_j.

It is important to point out that the chain of ratios used for a particular application should be devised with that application in mind. By so doing, optimum use may be made of records already being kept by that post office. No single set of formulas can be used easily in every case because not all post offices maintain the same volume count records.

The variance of $R_1R_2\ldots R_K$, which we symbolize by $\sigma^2_{R_1\ldots R_K}$ may be easily obtained. Denote the mean and variance of R_i by m_i and σ_i^2, respectively. Then, for $K=2$,

$$\sigma^2_{R_1R_2}=E(R_1^2R_2^2)-[E(R_1R_2)]^2$$

$$=(\sigma_1^2+m_1^2)(\sigma_2^2+m_2^2)-m_1^2m_2^2$$

$$=\sigma_1^2\sigma_2^2+\sigma_1^2m_2^2+m_1^2\sigma_2^2.$$

Similarly, for $K=3$ and 4 we obtain

$$\sigma^2_{R_1R_2R_3}=\sigma_1^2\sigma_2^2\sigma_3^2+m_1^2\sigma_2^2\sigma_3^2+\sigma_1^2m_2^2\sigma_3^2$$
$$+\sigma_1^2\sigma_2^2m_3^2+m_1^2m_2^2\sigma_3^2$$
$$+m_1^2\sigma_2^2m_3^2+\sigma_1^2m_2^2m_3^2,$$

and

$$\sigma^2_{R_1R_2R_3R_4}=\sigma_1^2\sigma_2^2\sigma_3^2\sigma_4^2+m_1^2\sigma_2^2\sigma_3^2\sigma_4^2+\sigma_1^2m_2^2\sigma_3^2\sigma_4^2$$
$$+\sigma_1^2\sigma_2^2m_3^2\sigma_4^2+\sigma_1^2\sigma_2^2\sigma_3^2m_4^2+m_1^2m_2^2\sigma_3^2\sigma_4^2$$
$$+m_1^2\sigma_2^2m_3^2\sigma_4^2+m_1^2\sigma_2^2\sigma_3^2m_4^2+\sigma_1^2m_2^2m_3^2\sigma_4^2$$
$$+\sigma_1^2m_2^2\sigma_3^2m_4^2+\sigma_1^2\sigma_2^2m_3^2m_4^2+m_1^2m_2^2m_3^2\sigma_4^2$$
$$+m_1^2m_2^2\sigma_3^2m_4^2+m_1^2\sigma_2^2m_3^2m_4^2+\sigma_1^2m_2^2m_3^2m_4^2.$$

In general

$$\sigma^2_{R_1\ldots R_K}=\sum\prod_{i=1}^{K}t_i-\prod_{i=1}^{K}m_i^2, \tag{1}$$

where the summation is over all possible 2^K combinations obtained by letting each t_i take on either the value m_i^2 or σ_i^2.

Let k_i denote the coefficient of variation of R_i $\left(\text{i.e., } k_i=\frac{\sigma_i}{m_i}\right)$ and let $k_{R_1\ldots R_K}$ denote the coefficient of variation of $R_1R_2\ldots R_K$. Then it follows from

eq (1) that

$$k^2_{R_1 \cdots R_K} = \sum \prod_{i=1}^{K} t_i - 1 \qquad (2)$$

where now the summation is over all possible 2^K combinations obtained by letting each t_i take on either the value k_i^2 or 1. Thus for the case $K=2$,

$$k^2_{R_1 \cdots R_2} = k_1^2 k_2^2 + k_1^2 + k_2^2.$$

For k_1 and k_2 small, we obtain by neglecting terms of higher order

$$k^2_{R_1 R_2} \doteq k_1^2 + k_2^2 \leq 2 \max\,(k_1^2, k_2^2).$$

Therefore, for k_1 and k_2 sufficiently small

$$k_{R_1 R_2} \leq \sqrt{2} \max\,(k_1, k_2).$$

In a similar way it is easy to show that

$$k_{R_1 \ldots R_K} \leq \sqrt{K} \max\,(k_1, k_2, \ldots, k_K) \qquad (3)$$

for the k_i sufficiently small. This says, essentially, what we would intuitively expect; namely that the coefficient of variation of the chain is bounded by a multiple of the coefficient of variation of the "weakest link." This weakest link is that ratio which has the greatest percent variability.

The estimates of the R_i used in the applications in later sections are of the form X/n where X is either a binomial or a multinomial random variable and n is the sample size. Such being the case, the coefficient of variation of any ratio estimate is $k = \sqrt{(1-p)/np}$ where p is the expected value of X/n. Since the sample sizes used here are large, the value k is indeed small. For example if $p=0.05$ and $n=5{,}000$ then the *relative* standard error of X/n is $k=0.0616$ or 6.2 percent and the *absolute* standard error is $0.062 \times 0.05 = 0.0031$. Thus the overall uncertainty is of the order of $3 \times 0.0031 = 0.0093$ so that in repeated drawings of 5,000 samples we would expect almost all of the estimates X/n to be between 0.05 ± 0.0093.

3.2. Notations and Formulas Used in the Applications

In the preceding section we presented the general method for setting up a chain-ratio estimate for the percentage of mail to a given *destination*. In this section we give the specific chain-ratio formulas used in the San Francisco, Los Angeles, and Baltimore studies. We list the notations of the ratios involved, and the related formulas for determining the percentage of mail to a *destination* off the *Primary*, *Secondary*, and *Tertiary* stages.

a. Notations

In the discussion of the general method in section 3.1, ratios appear with and without parentheses. In the list of notations that follows, all ratios appear within parentheses. Throughout this paper we shall always denote parameters by ratios without parentheses, and unbiased estimates of these parameters will always be denoted by ratios within parentheses.

$\left(\frac{D_P}{T_P}\right)$ = Ratio of mail to a *primary destination* to the *primary volume*. Obtained from *primary* samples.

$\left(\frac{S_i'}{T_P}\right)$ = Ratio of mail to an i-th *secondary* to the *total primary volume*. Obtained from *primary* samples.

$\sum\left(\frac{S_i'}{T_P}\right)$ = Sum of ratios of mail to all *secondaries* to *total primary volume*. Obtained from *primary* samples.

$\left(\frac{S_i}{S_i'}\right)$ = Ratio of mail to an i-th *secondary* including *bypass* mail to the i-th *secondary* excluding *bypass* mail. Obtained from volume counts.

$\left(\frac{D_{S_i}}{S_i}\right)$ = Ratio of mail to an i-th *secondary destination* to the i-th *secondary*. Obtained from i-th *secondary* samples.

$\left(\frac{t_{ij}}{S_i}\right)$ = Ratio of mail to a j-th *tertiary* (off i-th *secondary*) to i-th *secondary*. Obtained from i-th *secondary* samples.

$\left(\frac{D_{t_{ij}}}{t_{ij}}\right)$ = Ratio of mail to a j-th *tertiary destination* (off i-th *secondary*) to the j-th *tertiary*. Obtained from the ij-th tertiary samples.

$\left(\frac{D_P}{T}\right)$ = Ratio of mail to a *primary destination* to the *total volume*. Obtained from chain-ratio formula.

$\left(\frac{D_{S_i}}{T}\right)$ = Ratio of mail to an i-th *secondary destination* to the *total volume*. Obtained from chain-ratio formula.

$\left(\frac{D_{t_{ij}}}{T}\right)$ = Ratio of mail to a j-th *tertiary destination* (off i-th *secondary*) to the *total volume*. Obtained from chain-ratio formula.

$\left(\frac{T_P}{T}\right)$ = Ratio of *primary* mail to the *total volume*. Obtained from volume counts.

$\left(\frac{B_S}{T}\right)$ = Ratio of *by-pass* mail entering at the *secondary* to the *total volume*. Obtained from volume counts.

$\left(\frac{T_S}{T}\right)$ = Ratio of total *secondary* mail to *total volume*. Obtained from volume counts.

$\left(\frac{\Sigma D_P}{T}\right)$=Sum of ratios of mail to all *destinations* off the *primary* to the *total volume*. Obtained from volume counts.

$\left(\frac{S_i}{T_S}\right)$=Ratio of mail to an i-th *secondary* to the total *secondary volume*. Obtained from volume counts.

$\left(\frac{D_P}{\Sigma D_P}\right)$=Ratio of mail to a *destination* off the *primary* to the sum of all *destinations* off the *primary*. Obtained from the *primary* samples.

b. Related Formulas

Two essentially different sets of formulas were used. The choice between the two depended upon whether or not the percentage of *secondary* mail that entered the system at each specific *secondary* case was readily available. This often entailed setting up special and difficult procedures for obtaining this ratio. In Baltimore we made special volume counts. In all cases the aim was to estimate the ratio of mail going to a given *destination* to the sum of *primary* and all *bypass* mail.

(1) For Baltimore, where the percentage of *bypass* mail entering the system at the *secondary* was large, the following formulas were used:

a. For a *destination* off the *primary:*

$$\left(\frac{D_P}{T}\right)=\left(\frac{D_P}{\Sigma D_P}\right)\times\left(\frac{\Sigma D_P}{T}\right). \tag{4}$$

b. For a *destination* off the *secondary:*

$$\left(\frac{D_{S_i}}{T}\right)=\left(\frac{D_{S_i}}{S_i}\right)\times\left(\frac{S_i}{T_S}\right)\times\left(\frac{T_S}{T}\right). \tag{5}$$

c. For a *destination* off the *tertiary:*

$$\left(\frac{D_{t_{ij}}}{T}\right)=\left(\frac{D_{t_{ij}}}{t_{ij}}\right)\times\left(\frac{t_{ij}}{S_i}\right)\times\left(\frac{S_i}{T_S}\right)\times\left(\frac{T_S}{T}\right). \tag{6}$$

It is to be noted that formulas (5) and (6) of this section depend upon special volume count data that give (S_i/T_S).

For examples worked out in detail, see the San Francisco study, section 5.

(2) For both San Francisco and Los Angeles, where the percentage of *bypass* mail entering the system at the *secondary* was very small, no special volume counts of mail into the *secondary* were made. Instead, the following formulas were used:

a. For a *destination* off the *primary:*

$$\left(\frac{D_P}{T}\right)=\left(\frac{D_P}{T_P}\right)\times\left(\frac{T_P}{T}\right). \tag{7}$$

b. For a *destination* off the *secondary:*

$$\left(\frac{D_{S_i}}{T}\right)=\left(\frac{D_{S_i}}{S_i}\right)\times\left(\frac{S'_i}{T_P}\right)\times\left(\frac{T_P}{T}\right)\times\frac{\left(\frac{T_P}{T}\right)\Sigma\left(\frac{S'_i}{T_P}\right)+\left(\frac{B_S}{T}\right)}{\left(\frac{T_P}{T}\right)\Sigma\left(\frac{S'_i}{T_P}\right)}. \tag{8}$$

c. For a *destination* off the *tertiary:*

$$\left(\frac{D_{t_{ij}}}{T}\right)=\left(\frac{D_{t_{ij}}}{t_{ij}}\right)\times\left(\frac{t_{ij}}{S_i}\right)\times\left(\frac{S'_i}{T_P}\right)\times\left(\frac{T_P}{T}\right)\times\frac{\left(\frac{T_P}{T}\right)\Sigma\left(\frac{S'_i}{T_P}\right)+\left(\frac{B_S}{T}\right)}{\left(\frac{T_P}{T}\right)\Sigma\left(\frac{S'_i}{T_P}\right)}. \tag{9}$$

The justification for eqs (8) and (9) is the following: Set up the chain

$$\frac{D_{S_i}}{S_i}\times\frac{S_i}{S'_i}\times\frac{S'_i}{T_P}\times\frac{T_P}{T}.$$

Note that S_i/S'_i involves obtaining an estimate of the ratio of mail to an i-th *secondary* including *bypass* mail to the i-th *secondary* excluding *bypass* mail. As mentioned above this was difficult to accomplish in practice. However if $S_i/S'_i=S_j/S'_j$ for all i and j, then

$$\frac{S_i}{S'_i}=\frac{\sum S_i}{\sum S'_i}.$$

The quantity $\sum S_i/\sum S'_i$ can be written as

$$\frac{\frac{T_P}{T}\sum\frac{S'_i}{T_P}+\frac{B_S}{T}}{\frac{T_P}{T}\sum\frac{S'_i}{T_P}}. \tag{10}$$

Using the "propagation of error" formula, we obtain an estimate of (10) as

$$\frac{\left(\frac{T_P}{T}\right)\Sigma\left(\frac{S'_i}{T_P}\right)+\left(\frac{B_S}{T}\right)}{\left(\frac{T_P}{T}\right)\Sigma\left(\frac{S'_i}{T_P}\right)}. \tag{11}$$

Each of the estimates involved in (11) could be easily obtained. Thus we have eq (8). Justification for eq (9) follows similarly.

If the assumption $S_i/S'_i=S_j/S'_j$, for all i and j, is not true then eq (8) and (9) still apply approximately providing the ratio $\sum S_i/\sum S'_i$ is close to one. We confined the use of these formulas to those ap-

plications where the ratio of *bypass* mail to the *secondary* to the total *secondary* mail was small (San Francisco, 0.8 percent; Los Angeles, 2.5 percent).

3.3. Methods of Collecting Data

a. Volume Data

Certain ratios needed to be established in order to relate the pieces of mail counted in each *separation* of the sample to the *total volume* of mail. It was therefore necessary to acquire from volume counts in the post office the following data.

Daily volume information expressed in footage for: *a*. All mail into the *primary*; *b*. all mail bypassing the *primary* and entering the *secondary; c*. all *bypass* mail to the city; *d*. all mail into each individual type *secondary* case. (This count may not be necessary, see section 3.2(b).)

Items *a*, *b*, and *c* above are normally maintained daily by the post office. Item *d* usually involves special volume counts. From the data listed above it is possible to determine the ratio of each class and type processed to the *total volume* of mail. Several of these ratios are then utilized in the formulas of section 3.2(b) to estimate the percentage of the *total volume* going to each *destination*. These volume figures were obtained at least 1 day prior to drawing the sample so that decisions regarding the type of analysis to be used could be made early. Very often the analysis did not make use of certain volume ratios, such as those of *d* above, and therefore the particular volume counts could be discontinued. (See section 5.1 for example.)

b. Sample Data

(1) *Primary*. Two feet of mail was selected as it flowed into the *primary* cases from the canceling machines. It was placed on the ledge of the "test" case and distributed by a clerk. Special care was taken to make sure that no mail was added to or subtracted from the sample. After *distribution* had been made, the contents of each separation box were counted by the distributor and recorded by the supervising clerk (e.g., see fig. 3).

Special care was given to the choice of the sample. The randomness of the selection of the 2-ft tray was assured by choosing the first 2 feet flowing into the *primary* from the canceling machines at the predetermined time for drawing the sample. The mail accumulating in the stackers of a cancellation machine is fed from a moving conveyor belt that passes 7 or 8 persons, each of whom faces and places on the belt letters selected from those within his reach. Thus the letters undergo a fairly thorough mixing as they are being stacked so that the letters in any tray of mail sampled at this point would tend to have the property of randomness which is necessary in sampling studies. This method of sampling was selected in order to help eliminate the possibility of personal bias, conscious or unconscious, or personal responsibility for actual allocations.

However, *metered* mail and *patron segregated* mail, which does not undergo this mixing process at the facing table, was sampled differently. Any "bite" or "bunch" of this kind of mail may be addressed to the same *destination* and therefore would not have the required property of randomness. In this case successive letters were selected every few inches apart from each tier of mail until the required 2 feet was obtained. The distance between successive letters was predetermined and constant.

Two samples, each of which consists of about 580 letters, were drawn during the morning peak period and 2 during the evening peak period. Samples were taken for 5 successive days, exclusive of Saturday and Sunday, in order to obtain a fairly representative picture of the mail throughout the sampling period.

(2) *Secondary*. Mail flowing into the *secondary* comes either from the *primary* or from *bypass* mail. *Secondary* cases do not continuously generate enough mail to be sampled at any given moment. Each sample was drawn when enough mail was generated. In each case the sample used in the study was the first 2 feet of mail that accumulated after a case had been selected for sampling. After *distribution* had been made, the contents of each separation box was counted by the distributor and recorded by the supervising clerk. One sample was taken in the morning peak and one in the evening peak periods throughout the week.

(3) *Tertiary*. Mail flowing into the *tertiary* cases usually comes from the *secondary*. Therefore, it was possible to make counts on these cases only when enough mail was generated.

However, in cases where the required 2 ft did not generate, then smaller samples (i.e., whatever was available) were counted. Here again, after *distribution* had been made the contents of each separation box were counted by the distributor and recorded by the supervising clerk. Samples were taken once in the morning and once in the evening at peak periods throughout the week.

In order to satisfy the condition of statistical independence, we avoided, as much as possible, having the same letters represented in samples from more than one stage. Care was taken to record any mail dispatched during the sample period prior to the final count of each *destination* on *primary*, *secondary*, and *tertiary* cases. Thus missing observations were avoided.

4. Type of Mail Studied at San Francisco, Los Angeles, and Baltimore

The *total volume* of mail studied in the San Francisco, Los Angeles, and Baltimore Post Offices may be classified as outgoing first class letter mail of the following types:

1. Cancellation mail (machine and hand).
   a. *Stamped* mail to *primary*
   b. *Air* mail to *primary*
   c. *Specials* to *primary*

d. *Stamped* mail to *secondary* bypassing *primary*
e. *Stamped bypass* mail to city.

2. Noncancellation mail
   a. *Metered* to *primary*
   b. *Metered* to *secondary* bypassing *primary*
   c. *Air* mail to *primary*
   d. *Specials* to *primary*
   e. *Permit* to *primary*
   f. *Permit* to *secondary* bypassing *primary*
   g. *Penalty* to *primary*
   h. *Metered* and *permit bypass* to city.
3. Transit mail [5]
   a. *Transit* to *secondary*
   b. *Transit* to city.

Not included in this study is any type of incoming letter mail nor outgoing first class letter mail of the following types:

1. All mail to *air* mail and *special delivery* sections bypassing *primary*.
2. *Transit* mail receiving no *distribution*.
3. Large special mailings which would tend to bias the sample.

5. San Francisco Study

In this section we present a rather detailed description of the application of the chain-ratio method in the study conducted in San Francisco.

5.1. Volume Count Data

Volume counts made in San Francisco enabled us to determine what percentage of the *total volume* flowed into the *primary*, how much bypassed the *primary* and flowed either into the city section for local *distribution* or into the *secondary*. These counts were made on 6 days, June 21, 24, 25, 26, 27, and 28, 1957, between the hours of 10 a.m. and 10 p.m. Control counts were begun one day prior to drawing samples, so that decisions regarding sample size and optimum sampling periods and areas could be made. Volume control counts showed that mail flowing into the *secondary* that bypassed the *primary* was less than 1 percent. Thus San Francisco was analyzed according to part 2 of section 3.2(b). Therefore, it was established early that a footage count of mail flowing into the *secondary* could be discontinued.

Percentages corresponding to the *total volume* figures are summarized in table 1. The flow chart given in figure 2 contains the basic proportion figures which are then applied in the appropriate formula, as well as certain other summary figures that are a result of the sampling study.

5.2. Sampling Procedure

The sampling procedure adopted for San Francisco is the same as that described in section 3.3(b) with the modification that, wherever possible, the samples were made to consist of equal parts of the following: Stamped long; stamped short; metered long; and metered short letters. This was done because San Francisco makes a separation between long and short letters which is maintained throughout the *primary* and *secondary* cases but not, however, in the *tertiary* cases. Furthermore, metered and nonmetered mail are worked separately throughout the *primary* and *secondary* cases. Moreover the volume of the different classifications were relatively equal. The volume of mail generated in the *tertiary* cases was very small during the morning sampling period. Therefore, no *tertiary* samples were taken during this period.

Figure 3 shows copies of sample field data for the *primary*, a typical *secondary*, and a typical *tertiary* at the San Francisco post office. Each column represents samples taken on each of the 5 consecutive sampling days. Application of the formulas to an example from each stage is shown in section 5.4.

TABLE 1. *Percentages obtained from volume count data supplied by the San Francisco Post Office during the test period*

Date	Primary	City bypass	Secondary bypass
	%	%	%
6-21-57	84.13	15.87	0.00
24	89.44	10.42	0.14
25	89.59	9.97	.46
26	85.66	14.03	.31
27	85.55	14.04	.41
28	86.34	13.40	.26
Average %-------	86.74	13.00	.26

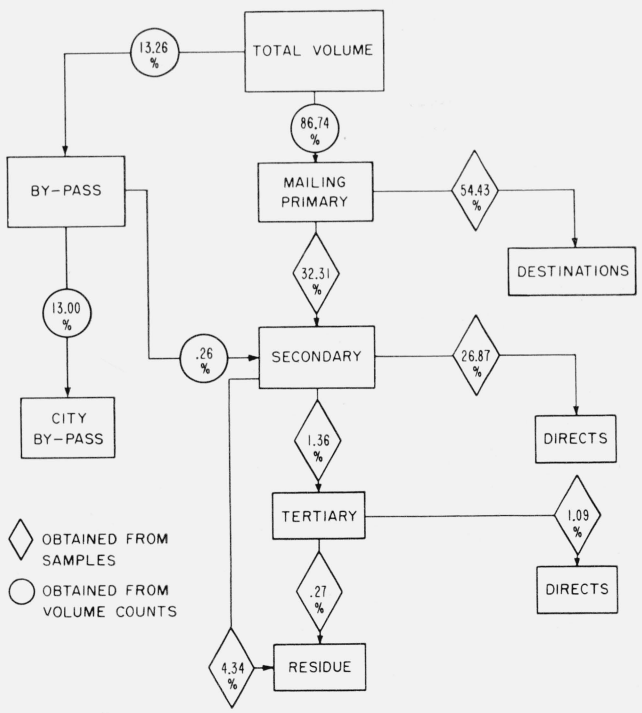
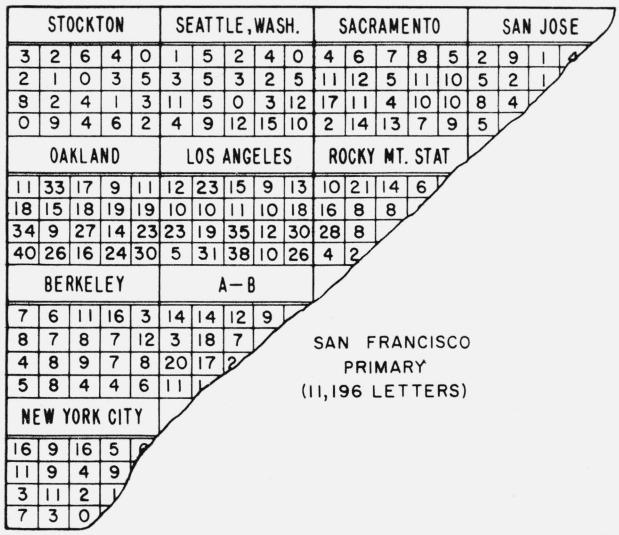
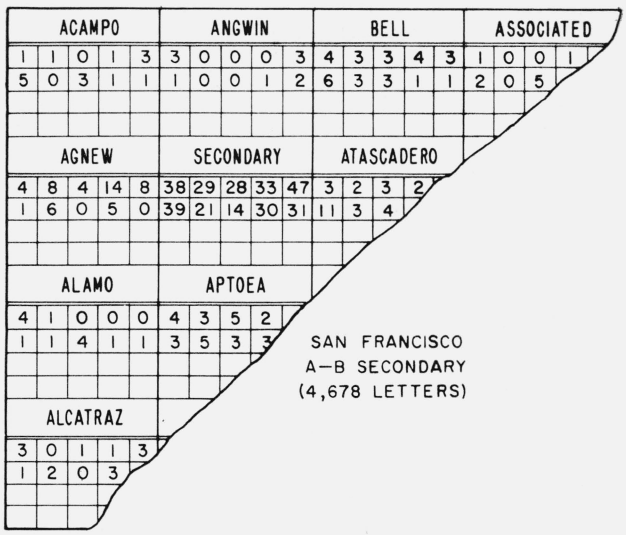


FIGURE 2. *San Francisco flow chart.*

[5] Mail received from another post office for outgoing processing.





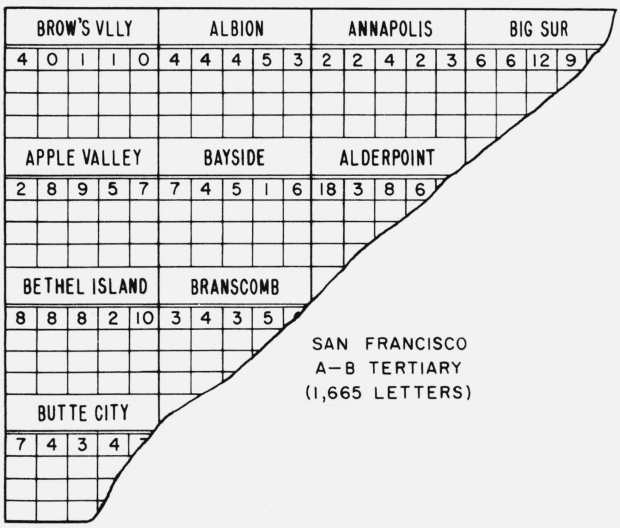


FIGURE 3. *Partial view of sample data for three San Francisco cases (worksheets).*

5.3. Computational Formulas

In this section the computational formulas used to estimate the percentage of the *total volume* of mail going to any given *destination* are given. As indicated above the eq (7), (8), and (9) are appropriate to the San Francisco study.

a. Primary

From figure 2 the value of $(T_p/T)=0.8674$ and therefore the appropriate formula becomes:

$$\left(\frac{D_P}{T}\right)=\left(\frac{D_P}{T_P}\right)\times\left(\frac{T_P}{T}\right)=\left(\frac{D_P}{T_P}\right)\times 0.8674.$$

(The total number of letters in the samples off the *primary* was 11,196.)

b. Secondary

The computational formula for *destinations* off the *secondary* depends upon the ratios obtained at the *primary* as well as the volume counts. Using such ratios gives the formula:

$$\left(\frac{D_{S_i}}{T}\right)=\left(\frac{D_{S_i}}{S_i}\right)\times\left\{\left(\frac{S'_i}{T_P}\right)\times\left(\frac{T_P}{T}\right)\times\frac{\left(\frac{T_P}{T}\right)\sum\left(\frac{S'_i}{T_P}\right)+\left(\frac{B_S}{T}\right)}{\left(\frac{T_P}{T}\right)\sum\left(\frac{S'_i}{T_P}\right)}\right\}=\left(\frac{D_{S_i}}{S_i}\right)\times c_i,$$

where the c_i are the quantities in brackets which depend upon the particular *secondary*. Values of c_i corresponding to particular *secondaries* are listed in table 2.

These constants actually represent the ratio, as estimated by using volume and *primary* sample counts, of a *secondary* volume of mail to the total *volume*.

TABLE 2. *Number of pieces in sample and constants used in computational formula for destinations off the secondaries for San Francisco*

i	S_i	Number of pieces	c_i
1	Ariz.-N. Mex.-Tex	5,519	0.01290
2	Ill.-Ind.-Iowa-Mass.-Mich.-Minn	5,739	.01774
3	Southern States	5,865	.01468
4	Rocky Mountain States	5,252	.02266
5	N.Y.-N.J.-Ohio-Pa	6,286	.02289
6	Canada-Eastern	5,535	.01797
7	Calif. A–B	4,676	.02180
8	Calif. C–D	4,945	.02367
9	Calif. E–G	4,499	.01351
10	Calif. H–L	4,989	.02383
11	Calif. M–O	4,994	.02702
12	Calif. P–R	5,049	.03024
13	Calif. S	4,759	.02031
14	Calif. San Santa	4,893	.03446
15	Calif. T–Z	4,596	.02203
	Total	77,596	0.32571

C. Tertiary

The computational formula for *destinations* off the *tertiary* depends upon ratios obtained at the *primary* and *secondary*, as well as the volume counts. Using such ratios gives the formula:

$$\left(\frac{D_{t_{ij}}}{T}\right)=\left(\frac{D_{t_{ij}}}{t_{ij}}\right)$$

$$\times\left\{\left(\frac{t_{ij}}{S_i}\right)\times\left(\frac{S'_i}{T_P}\right)\times\left(\frac{T_P}{T}\right)\times\frac{\left(\frac{T_P}{T}\right)\sum\left(\frac{S'_i}{T_P}\right)+\left(\frac{B_S}{T}\right)}{\left(\frac{T_P}{T}\right)\sum\left(\frac{S'_i}{T_P}\right)}\right\}$$

$$=\left(\frac{D_{t_{ij}}}{t_{ij}}\right)\times k_{ij},$$

where the k_{ij} are the quantities in brackets which depend upon the particular *tertiary*. Values of k_{ij} corresponding to particular *tertiaries* are listed in table 3.

These constants actually represent the ratio, as estimated by using volume counts and *primary* and *secondary* sample counts, of a *tertiary* volume of mail to the *total volume*.

TABLE 3. *Number of pieces in sample and constants used in computational formula for destinations off the tertiaries for San Francisco*

i,j	t_{ij}	Number of pieces	k_{ij}
7,1	Calif. A–B	1,665	0.00145
8,1	Calif. C–D	2,507	.00277
9,1	Calif. E–G	1,727	.00081
10,1	Calif. H–L	2,648	.00229
11,1	Calif. M–O	2,086	.00185
12,1	Calif. P–R	2,262	.00135
13+14,1	Calif. S	1,118	.00107
15,1	Calif. T–Z	2,152	.00202
	Total	16,165	.01361

5.4. Examples

Applications of the formulas for each stage are given here.

Primary: (Seattle, Wash.)

D_P=111 pieces—Seattle, Wash.

T_P=11,196 pieces—Total primary

where the numbers are taken from figure 3. Thus,

$$\left(\frac{D_P}{T}\right)=\left(\frac{D_P}{T_P}\right)\times 0.8674=\frac{111}{11{,}196}\times 0.8674=0.0085996.$$

Secondary: (Bell, Calif.)

D_{S_7}=31 pieces—Bell, Calif.

S_7=4,676 pieces—Total Calif. A–B Secondary

where the numbers are taken from figure 3. Thus,

$$\left(\frac{D_{S_7}}{T}\right)=\left(\frac{D_{S_7}}{S_7}\right)\times c_7=\frac{31}{4{,}676}\times 0.02180=0.0001445$$

where the constant c_7 is taken from table 2.

Tertiary: (Albion, Calif.)

$D_{t_{7,1}}$=20 pieces—Albion, Calif.

$t_{7,1}$=1,665 pieces—Total Calif. A–B Tertiary

where the numbers are taken from figure 3. Thus,

$$\left(\frac{D_{t_{7,1}}}{T}\right)=\left(\frac{D_{t_{7,1}}}{t_{7,1}}\right)\times k_{7,1}=\frac{20}{1{,}665}\times 0.00145=0.0000174$$

where $k_{7,1}$ is taken from table 3.

5.5. Tabulation of Estimated Distribution and Observations

Part of the tabulation of the estimated proportions of the *total volume* mail going to each *destination* is given in table 4. Figure 4 graphically portrays the

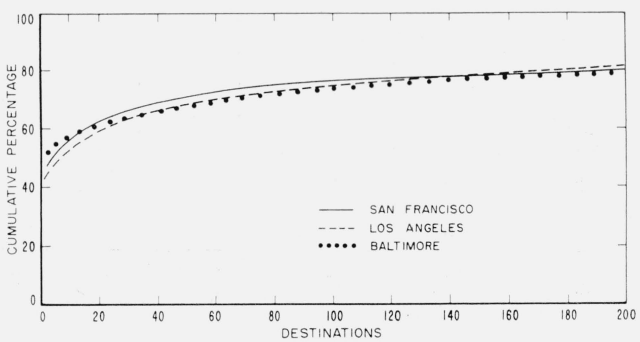


FIGURE 4. *Graph of largest 200 destinations for San Francisco, Los Angeles, and Baltimore post offices.*

TABLE 4.—*Tabulation of estimated percentages of the total volume to each direct destination for San Francisco*

Largest 200 direct destinations	Percent [a]	Cumulative percent
1. San Francisco Inc. City bypass	38.501	38.501
2. Oakland, Calif	8.158	46.659
3. Los Angeles, Calif	2.789	49.448
4. Sacramento, Calif	1.364	50.812
5. Washington State	1.155	51.967
6. Berkeley, Calif	1.147	53.114
7. New York City, N.Y	1.116	54.230
8. San Jose, Calif	0.961	55.191
9. Seattle, Wash	.860	56.051
10. Oregon State	.775	56.826
11. San Mateo, Calif	.759	57.585
12. Redwood City, Calif	.679	58.264
13. Daly City, Calif	.670	58.934
14. Palo Alto, Calif	.654	59.588
15. Fresno, Calif	.612	60.200
16. Portland, Oreg	.605	60.805
17. South San Francisco	.574	61.379
18. Chicago, Ill	.566	61.945
19. San Rafael, Calif	.521	62.466
20. Stockton, Calif	.504	62.970
21. Burlingame, Calif	.396	63.366
22. Menlo Park, Calif	.394	63.760
23. Santa Rosa, Calif	.352	64.112
24. San Diego, Calif	.349	64.461
25. Vallejo, Calif	.295	64.756
* * *	*	*
* * *	*	*
* * *	*	*
196. Wilmington, Calif	0.030	79.907
197. Lakeport, Calif	.030	79.937
198. Willits, Calif	.029	79.966
199. Porterville, Calif	.029	79.995
200. Placerville, Calif	.029	80.024

Rank	Number in group	Individual percent	Group percent	Cumulative percent
201–204	4	0.029	0.116	80.140
205–207	3	.028	.084	80.224
208–214	7	.027	.189	80.413
215–220	6	.026	.156	80.569
221–225	5	.025	.125	80.694
226–231	6	.024	.144	80.838
232–239	8	.023	.184	81.022
240–249	10	.022	.220	81.242
250–256	7	.021	.147	81.389
257–264	8	.020	.160	81.549
265–281	17	.019	.323	81.872
282–292	11	.018	.198	82.070
293–304	12	.017	.204	82.274
305–321	17	.016	.272	82.546
322–335	14	.015	.210	82.756
336–360	25	.014	.350	83.106
361–380	20	.013	.260	83.366
381–401	21	.012	.252	83.618
402–429	28	.011	.308	83.926
430–467	38	.010	.380	84.306
468–505	38	.009	.342	84.648
506–550	45	.008	.360	85.008
551–604	54	.007	.378	85.386
605–667	63	.006	.378	85.764
668–729	62	.005	.310	86.074
730–798	69	.004	.276	86.350
799–919	121	.003	.363	86.713
920–1087	168	.002	.336	87.049
1088–1271	184	.001	.184	87.233
1272–1296	25	<.001	.006	87.239
Air mail			3.200	90.439
Foreign			0.201	90.640
Residues			4.617	95.257
Miscellaneous			4.743	100.000

[a] The standard error of the estimated percentages, expressed as percents of the estimates, are between 10 and 15 percent for most of the first 200 *destinations*. For the very small percents the standard error may increase to as high as 35 percent.

largest 200 *destinations* by percentage for the Los Angeles and Baltimore studies as well as for San Francisco. Several observations, based on the tabulation, are given here:

1. The largest 200 *destinations* received 80 percent of the *total volume*.
2. Seventy-six percent of the *total volume* remained in the State of California (not including *air* mail).
3. Thirty-nine percent of the *total volume* remained in San Francisco.
4. Seven *destinations:* San Francisco, Oakland, Los Angeles, Sacramento, Washington State, Berkeley, and New York City were the only *destinations* to receive more than one percent of the *total volume*.
5. Eighty percent of the *total volume* remained on the West Coast (not including *air* mail).

An outstanding feature of the chain-ratio method of sampling is that emphasis may be placed on estimating relatively small percentages. Adaptation of the formulas of section 3.1 shows that the standard errors of the estimated percentages of mail to the various *destinations* considered in table 4 expressed as percents of the estimates, are between 10 and 15 percent for most of the first 200 *destinations*. Thus for Oakland, the estimated *relative* standard error is 10.4 percent so that the *absolute* standard error of the percentage of San Francisco mail having Oakland for its destination is 0.104×8.158 percent$=0.85$ percent so that the overall uncertainty is of the order of 3×0.85 percent$=2.6$ percent and there is very little likelihood that it has been misranked in order of volume.

For the examples in the "tail" of the distribution cited in section 5.4, the relative standard errors are somewhat larger. Thus for Bell, Calif., which ranks about 350, the relative standard error is 21 percent. Likewise for Albion, Calif., which ranks in the 920 to 1087 group, the relative standard error is 34 percent, or 0.0007 percent on an absolute basis, so that its overall uncertainty is of the order of ± 0.002 percent and its "true" ranking position may be as high as 730.

Examination of the complete listings of the San Francisco study given here and of the Los Angeles and Baltimore studies presented in [8] suggests that the proportion of mail to any given *destination* is related to (a) some measure of the "size" of the *destination*, and (b) the distance of the *destination* from the point of origin. Finally, there appears to be rather strong evidence that the distribution of mail plotted against the ranked *destinations* is rather close to a straight line on log-log paper.

Among the many colleagues who assisted in various ways toward the completion of this study, the authors particularly thank Marvin Zelen of NBS for his enthusiastic encouragement and helpful suggestions and Inspector John Falconer of the Post Office Department for assistance in implementing and supervising the collection of the data involved in the sampling procedures.

6. References

[1] N. C. Severo, A. E. Newman, S. Young, and M. Zelen, Some applications of statistical sampling methods to outgoing letter mail characteristics, NBS Tech. Note No. 16 (PB151375).

[2] I. Rotkin, The mechanization of letter mail sorting, Proc. Eastern Joint Computer Conf., Dec. 1957.

[3] W. A. Wallis and H. V. Roberts, Statistics, A new approach (The Free Press, Glencoe, Ill., 1956).

[4] H. A. Freeman, M. Friedman, F. Mosteller, and W. A. Wallis, Sampling Inspection, p. 10 (McGraw-Hill Book Co., New York, N.Y., 1948).

[5] W. E. Deming, Some theory of sampling, p. 40, Chapters 2 and 7 (John Wiley & Sons, Inc., New York, N.Y., 1950).

[6] B. Epstein, R. Bacon, G. Prittham, L. Lewis, F. Grossman, and G. Miller, Jr., No. 561 Memorandum for Record, Eng. Branch, S.A.A. Division, Frankford Arsenal (August 1943).

[7] B. K. Bender and A. J. Goldman, Analytic comparison of suggested configurations for automatic mail sorting equipment, J. Research NBS, **63B**, 83 (1959).

[8] N. C. Severo and A. E. Newman, Distribution of mail by destination at the San Francisco, Los Angeles, and Baltimore Post Offices, NBS Tech. Note 27 (PB151392).

[9] N. C. Severo, A statistician and the post office; A case history in operations research, Transactions of the 1959 Conference of the Administrative Division of ASQC (March 1959).

WASHINGTON, D.C. (Paper 64C1–22)