ON THE PROBLEM OF MATCHING LISTS BY SAMPLES

By

W. EDWARDS DEMING

AND

GERALD J. GLASSER

NEW YORK UNIVERSITY

Reprinted from
THE JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
Vol. 54, June 1959: pp. 403-415

ON THE PROBLEM OF MATCHING LISTS BY SAMPLES

W. Edwards Deming and Gerald J. Glasser
New York University

This paper presents theory for estimation of the proportions of names common to two or more lists of names, through use of samples drawn from the lists. The theory covers (a) the probability distributions, expected values, variances, and the third and fourth moments of the estimates of the proportions duplicated; (b) testing a hypothesis with respect to a proportion; (c) optimum allocation of the samples; (d) the, effect of duplicates within a list; (e) possible gains from stratification. Examples illustrate some of the theory.

Statement *of the problem.* There are 2 or more long lists of names. Some names may be common to some or all of the lists, and it is of some economic or scientific importance to discover how many. The lists may be very long: in practice they may run to several hundred thousand or millions of names. One example came up in Germany a few years ago where the government wished to know how many people receive regular cheques from several sources—for example, government payroll, social security, unemployment compensation, subsidy of one kind or another, ex-soldier's allowance, and possibly other sources. Another example is provided by a publisher of a magazine who wished to discover how many of his subscribers were on a list of executives, and on other special lists.

An advertising agency or the marketing department of a firm must decide whether to use one or both of 2 lists of names available to them for an advertising campaign. The number of names common to both lists would be, if they knew it, the key decision-parameter. A firm may wish to determine, by comparing two lists, how many of their present employees worked there in some past year. The number of shareholders common to two or more companies, and the number of companies that do business in each of two states, are additional problems which require the matching of lists. Library work affords other illustrations.

This paper presents some statistical theory for the solution of such problems. Several of the results (including estimators for relevant parameters and approximations to their variances) have already appeared in a note by Goodman.[1] Here we apply and extend his work.

The theory that we give here, like Goodman's, is based on probability samples from both lists. Incidentally, a sample from one list matched against the other in full (100%) presents only a simple case of random sampling from a finite population of attributes; it is also a limiting case of the general theory of matching two samples.

Notation. The accompanying table shows the scheme of notation. $a_1, a_2, \cdots, a_M$ are distinct and ordered names on one list; $b_1, b_2, \cdots, b_N$ are distinct and ordered names on the other list. D names are common to both lists. We assume that no name appears more than once on one list; however, we illustrate later

[1] Goodman, Leo A., "On the analysis of samples from k lists," *Annals of Mathematical Statistics*, 23 (1952), pp. 632–4.

the relaxation of this requirement. The number D is important: it is the number that the publisher (for example) wishes to know. Let

$$p = \frac{D}{M} \tag{1}$$

$$P = \frac{D}{N} \tag{2}$$

It will suffice to estimate either p or P.

	List 1	List 2
	a_1 a_2 a_3 a_M	b_1 b_2 b_3 b_N
Number on the list	M	N
Number common to both lists	D	D
Proportion common to both lists	p	P

The comparison of name a_i in List 1 with name b_j in List 2 gives

$$\begin{cases} a_ib_j = 1 & \text{if the 2 names are identical} \\ \quad\;\; = 0 & \text{otherwise.} \end{cases} \tag{3}$$

Then the number D of names common to both lists is

$$\sum a_ib_j = D \tag{4}$$

where i in the summation runs through List 1, and j runs through List 2. Let names in the samples be

$$x_1, x_2, \cdots, x_m \quad \text{from List 1}$$

$$y_1, y_2, \cdots, y_n \quad \text{from List 2}$$

$$\begin{cases} x_iy_j = 1 & \text{if the 2 names are identical} \\ \quad\;\; = 0 & \text{otherwise.} \end{cases} \tag{5}$$

We also define

$$d = \sum x_iy_j \qquad \lfloor i = 1, 2, \cdots, m; j = 1, 2, \cdots, n \rfloor \tag{6}$$

the number of names common to the 2 samples. d is a random variable.

Sampling procedure

(1) Draw by random numbers between 1 and M and without replacement m names from List 1.

(2) Draw by random numbers between 1 and N and without replacement n names from List 2.

(3) Compare every name in the sample from List 1 with every name in the sample from List 2 to discover how many names are common to both samples. Let d be this number.

(4) Form the estimates

$$\hat{p} = \frac{N}{n}\,\frac{d}{m} \tag{7}$$

$$\hat{P} = \frac{M}{m}\,\frac{d}{n} \tag{8}$$

$$\hat{D} = \frac{NM}{nm}\,d = M\hat{p} = N\hat{P}. \tag{9}$$

For a problem that requires a statistical test, step 4 specifies regions of acceptance and rejection, and not estimators. Suppose that we wish to test the hypothesis $p = p_0$ against the alternative $p < p_0$. The region of rejection for a test at level α is $d < d^*$, where d^* is an integer for which

$$P\{d < d^* \mid p_0\} \doteq \alpha. \tag{10}$$

The critical value d^* may be determined by reference to the exact probability distribution of d, Eq. (11), or to the approximations afforded by either Eq. (12) or (13). An example appears later.

Results for 2 lists. The estimates just formed by the sampling procedure given above possess the following properties:

$\hat{p}$ is an unbiased estimate of p

$\hat{P}$ is an unbiased estimate of P

$\hat{D}$ is an unbiased estimate of D.

As Goodman pointed out, these estimators may under special conditions lead to impossible results. For example, with $N = M = 1000$, $n = m = 100$ and $d = 20$, Eq. (7) shows that $\hat{p} = 2$. However, the probability of an unreasonable estimate is generally small, unless p or P is close to 1. Thus, impossible values of $\hat{D}$, $\hat{p}$, or $\hat{P}$ may simply mean that practically all the names in the small list are duplicates of those in the larger list.

The probability distribution of d, the number of names common to the 2 samples, is

$$P(d) = \frac{\binom{D}{d}}{\binom{M}{m}\binom{N}{n}} \sum_{k=d}^{D} \binom{D-d}{k-d}\binom{M-D}{m-k}\binom{N-k}{n-d} \tag{11}$$

which one may use to determine critical values for a statistical test and to compute the power of the test. Alternative forms appear later in Eqs. (42) and (13).

We introduce 2 limiting cases. Case 1: M, N, m, n all increase without limit in such manner that D, m/M, n/N remain fixed. Case 2: M, N, m, n, and D all increase without limit in such manner that mnD/MN remains fixed at the value λ. In Case 1

$$P(d) \to \binom{D}{d} f^d (1-f)^{D-d} \tag{12}$$

where $f=mn/MN$. The limit in Case 1 is obviously a binomial with parameters D and f. It is comparable to the binomial limit for the hypergeometric distribution.[2] It gives a good approximation to the exact distribution Eq. (11) if M, N, m, and n are reasonably large. In Case 2

$$P(d) \to \frac{\lambda^d}{d!} e^{-\lambda} \tag{13}$$

which is a Poisson distribution. This equation also approximates the exact distribution in Eq. (11) if f is small. In addition

$$\mathrm{Var}\ \hat{p} = \frac{Np}{mn}\left\{1 + \frac{m-1}{M-1}\,\frac{n-1}{N-1}\,(D-1)\right\} - p^2 \tag{14}$$

$$\to \frac{Np}{mn}\left\{1 - \frac{nm}{NM}\right\} \qquad \text{Case 1} \tag{15}$$

$$\to \frac{Np}{mn} \qquad \text{Case 2.} \tag{16}$$

If the sample from List 2 is complete, then $n=N$, and the above formulas for the probability distribution of d and for Var $\hat{p}$ reduce to

$$P(d) = \binom{D}{d}\binom{M-D}{m-d}\Big/\binom{M}{m} \tag{17}$$

the hypergeometric distribution, 1 and

$$\mathrm{Var}\ \hat{p} = \frac{M-m}{M-1}\,\frac{pq}{m} \qquad [p+q=1]. \tag{18}$$

Eqs. (14), (15), and (16) take the form

$$C_{\hat{p}}^2 = \frac{N}{mnp}\left\{1 + \frac{m-1}{M-1}\,\frac{n-1}{N-1}\,(D-1)\right\} - 1 \tag{19}$$

$$\to \frac{N}{mnp}\left\{1 - \frac{nm}{NM}\right\} \qquad \text{Case 1} \tag{20}$$

$$\to \frac{N}{mnp} \qquad \text{Case 2} \tag{21}$$

[2] Coggins, Paul P., "Some general results of elementary sampling theory for engineering use," *Bell System Technical Journal*, 7 (1928), p. 44.

where $C_{\hat{p}}^2$ is the rel-variance of $\hat{p}$ (Var $\hat{p}$ divided by p^2). Var $\hat{P}$ and $C\hat{P}^2$ follow by symmetry. An unbiased estimate of Var $\hat{p}$ is

$$\text{Est Var } \hat{p} = \frac{\hat{p}}{M}\left\{\frac{M-1}{m-1}\,\frac{N-1}{n-1} - 1\right\} + \hat{p}^2\left\{1 - \frac{m}{M}\,\frac{n}{N}\,\frac{M-1}{m-1}\,\frac{N-1}{n-1}\right\} \tag{22}$$

$$\rightarrow \frac{\hat{p}}{M}\left\{\frac{MN}{mn} - 1\right\} \qquad \text{Case 1} \tag{23}$$

$$\rightarrow \frac{N\hat{p}}{mn} \qquad \text{Case 2.} \tag{24}$$

Est Var $\hat{P}$ follows by symmetry. For the higher central moment coefficients of d, put $f = mn/MN$ and define

$$\Delta_i = \frac{m-i}{M-i}\,\frac{n-i}{N-i}\,(D-i). \tag{25}$$

Then

$$E(d-Ed)^3 = \Delta_0\{1 - 3\Delta_0 + 3\Delta_1 + \Delta_1\Delta_2 + 2\Delta_0^2 - 3\Delta_0\Delta_1\} \tag{26}$$

$$\rightarrow \Delta_0(1-f)(1-2f) \qquad \text{Case 1} \tag{27}$$

$$\rightarrow \Delta_0 \qquad \text{Case 2} \tag{28}$$

$$E(d-Ed)^4 = \Delta_0\{1 - 4\Delta_0 + 7\Delta_1 + 6\Delta_0^2 + 6\Delta_1\Delta_2 - 12\Delta_0\Delta_1 - 3\Delta_0^3 + 6\Delta_0^2\Delta_1 - 4\Delta_0\Delta_1\Delta_2 + \Delta_1\Delta_2\Delta_3\} \tag{29}$$

$$\rightarrow 3\Delta_0^2(1-f)^2 + \Delta_0(1 - 6f + 6f^2) \qquad \text{Case 1} \tag{30}$$

$$\rightarrow 3\Delta_0^2 + \Delta_0. \qquad \text{Case 2.} \tag{31}$$

It will be observed that Eqs. (27) and (30) agree with the corresponding moment coefficients of the binomial of Eq. (12), while Eqs. (28) and (31) agree with those of the Poisson distribution in Eq. (13).

Examples for 2 lists

(1) Probability samples of 900 and 1,800 are selected from lists of 40,000 and 20,000 names. They contain 16 duplicates. Eqs. (7), (8), and (9) give the unbiased estimates

$$\hat{p} = \frac{20{,}000}{1800}\cdot\frac{16}{900} = .198$$

$$\hat{P} = \frac{40{,}000}{900}\cdot\frac{16}{1800} = .395$$

$$\hat{D} = \frac{20{,}000}{1800}\cdot\frac{40{,}000}{900}\,16 = 7901$$

Eq. (22) gives

$$\hat{\sigma}_{\hat{p}} = .049$$

$$\hat{\sigma}_{\hat{P}} = .098$$

$$\hat{\sigma}_{\hat{D}} = 1974$$

Eqs. (23) and (24) lead to practically the same numerical estimates, because m, n, M, and N are all big.

(2) An advertising agency has 2 lists of names A and B, but can not use them as they are if too many names are common to both lists. List A contains 40,000 names; list B contains 10,000 names. The director of the agency specifies that he wishes to take a risk no bigger than .01 of using the lists as they are if 1000 names or more are common to both lists. This number would make P, the proportions of duplicates in list B, equal to .1. If a test accepts the hypothesis that P may be .1 or bigger, he will purge the lists of duplicates by matching them 100%, or until tests of samples show that the duplicates have reached the required level. The costs of sampling the lists are equal, wherefore we select 2000 names from each list (cf. the later section on allocation).

Statistically, the problem is to test the hypothesis $P = .1$ against the alternative $P < .1$. As M, N, m, and n are all big, and as mn/MN is small, we may use the Poisson approximation Eq. (13) with

$$\lambda = \frac{mnP}{M} = \frac{2000 \times 2000 \times .1}{40{,}000} = 10. \tag{32}$$

The critical integer d^* is the nearest integer that satisfies the equation

$$\sum_{d=d^*}^{\infty} \frac{\lambda^d e^{-\lambda}}{d!} \doteq 1 - \alpha = .99 \qquad [\lambda = 10]. \tag{33}$$

One may use Molina's tables[3] to find the critical value d^*, which turns out to be 4. The exact distribution Eq. (11) and the binomial limit Eq. (12) give the same critical value. An easier way is to use the square-root-transformation with mean equal to $\sqrt{10}$ and with standard deviation $\frac{1}{2}$, noting that the area .01 under one tail corresponds to a standard deviate of 2.33, wherefore

$$\sqrt{10} - \sqrt{d^*} = 2.33 \times \tfrac{1}{2} = 1.165 \tag{34}$$

whence $\sqrt{d^*} = 2$ and $d^* = 4$. Hence the statistical rule for decision requires rejection of the hypothesis $P = .1$ and acceptance of the lists as they are if the number d of duplicates in the 2 samples of 2000 turns out to be less than 4.

Eqs. (11), (12), or (13) give the probabilities in the accompanying table for the power of the test, with samples of 2000 from each list.

P (proportion of duplicates in List B)	.005	.010	.025	.050	.075	.100	.125
Probability of rejection of the hypothesis $P = .1$	1.00−	.98	.76	.26	.06	.01	.00+

[3] Molina, E. C., *Poisson's Exponential Binomial Limit*, New York: D. Van Nostrand, 1942, Table II, p. 11.

Solution for 2 lists. Let $A_1, A_2, \cdots A_D$ be the D names common to both lists. Then the probability that a specified name A_i will fall into both samples is

$$P(A_i) = \frac{mn}{MN} \tag{35}$$

for all i. The probability that 2 specified names A_i and A_j will both fall into both samples is

$$P(A_i, A_j) = \frac{m}{M}\frac{m-1}{M-1}\frac{n}{N}\frac{n-1}{N-1} = \frac{\binom{m}{2}\binom{n}{2}}{\binom{M}{2}\binom{N}{2}} \tag{36}$$

for all $i \neq j$. Similarly, for any specified set of k names,

$$\begin{aligned} P(A_1, A_2, \cdots, A_k) &= \frac{m}{M}\frac{m-1}{M-1}\cdots\frac{m-k+1}{M-k+1}\frac{n}{N}\frac{n-1}{N-1}\cdots\frac{n-k+1}{N-k+1} \\ &= \frac{\binom{m}{k}\binom{n}{k}}{\binom{M}{k}\binom{N}{k}} \end{aligned} \tag{37}$$

for any set of k names, $k \leq D$, $k \leq m$, $k \leq n$.

To derive the distribution of d, one may apply a general rule of addition of probabilities.[4] Thus, if

$$S_1 = \sum_{i=1}^{D} P(A_i) = D\frac{mn}{MN} \tag{38}$$

$$S_2 = \sum_{j \neq i}^{D} P(A_iA_j) = \binom{D}{2}\binom{m}{2}\binom{n}{2} \Big/ \binom{M}{2}\binom{N}{2} \tag{39}$$

and in general, if

$$S_k = \binom{D}{k}\binom{m}{k}\binom{n}{k} \Big/ \binom{M}{k}\binom{N}{k} \tag{40}$$

then the probability distribution of exactly d names common to both samples is

$$P(d) = S_d - \binom{d+1}{d}S_{d+1} + \binom{d+2}{d}S_{d+2} - + \cdots \pm \binom{D}{d}S_D \tag{41}$$

whence

$$P(d) = \sum_{k=d}^{D} (-1)^{k-d}\binom{k}{d}\binom{D}{k}\binom{m}{k}\binom{n}{k} \Big/ \binom{M}{k}\binom{N}{k} \tag{42}$$

[4] Feller, William, *An Introduction to Probability Theory and its Applications*, 2d ed., New York: John Wiley and 1957, Chap. 4.

$$= \frac{\binom{D}{d}}{\binom{M}{m}\binom{N}{n}} \sum_{k=d}^{D} (-1)^{k-d} \binom{D-d}{D-k} \binom{M-k}{M-m} \binom{N-k}{N-n} \tag{43}$$

$$= \frac{\binom{D}{d}}{\binom{M}{m}\binom{N}{n}} \sum_{k=d}^{D} \binom{D-d}{k-d} \binom{M-D}{m-k} \binom{N-k}{n-d} \tag{44}$$

as already given in Eq. (11).

To derive the expected values and variances of $\hat{p}$ and of $\hat{P}$ we note that

$$\begin{aligned} Ed &= E \sum x_i y_j = \frac{mn}{MN} \sum a_i b_j = \frac{mn}{MN} D \\ &= \frac{mn}{N} p = \frac{mn}{M} P \end{aligned} \tag{45}$$

whence $E\hat{p} = p$ and $E\hat{P} = P$, as already recorded. Next,

$$\begin{aligned} Ed^2 &= E[\sum x_i x_j]^2 = E[\sum x_i y_j x_i y_j + \sum x_i y_j x_{i'} y_{j'}] \qquad (i' \neq i,\ j' \neq j) \\ &= E \sum x_i y_j + E \sum x_i y_j x_{i'} y_{j'} \\ &= \frac{mn}{MN} D + \frac{m}{M} \frac{m-1}{M-1} \frac{n}{N} \frac{n-1}{N-1} \sum a_i a_{i'} b_j b_{j'} \\ &= \frac{mn}{MN} D + \frac{m}{M} \frac{m-1}{M-1} \frac{n}{N} \frac{n-1}{N-1} (D^2 - D) \\ &= \frac{mn}{N} p \left[1 + \frac{m-1}{M-1} \frac{n-1}{N-1} (D-1) \right]. \end{aligned} \tag{46}$$

It follows that

$$\begin{aligned} \text{Var } \hat{p} &= E(\hat{p} - p)^2 = E\hat{p}^2 - p^2 \\ &= \left[\frac{N}{mn}\right]^2 E[\sum x_i y_j]^2 - p^2 \\ &= \frac{Np}{mn} \left[1 + \frac{m-1}{M-1} \frac{n-1}{N-1} (D-1) \right] - p^2 \end{aligned}$$

as already recorded in Eq. (14).

Extension to L *lists.* Let d be the number of names common to samples of sizes $n_1, n_2, \cdots, n_L$ drawn at random from lists of size $N_1, N_2, \cdots, N_L$ in which D names are common to all L Lists. The distribution of d, the number of names common to all L samples, is

$$P(d) = \sum_{k=d}^{D} (-1)^{k-d} \binom{k}{d} \binom{D}{k} \prod_{i=1}^{L} \binom{n_i}{k} \Big/ \prod_{i=1}^{L} \binom{N_i}{k} \tag{47}$$

and asymptotic results analogous to Eq. (12) and Eq. (13) include

$$P(d) \to \binom{D}{d} f^d (1-f)^{D-d} \qquad \text{Case 1} \tag{48}$$

$$P(d) \to \frac{\lambda^d}{d!} e^{-\lambda} \qquad \text{Case 2} \tag{49}$$

where

$$f = \prod_{i=1}^{L} \frac{n_i}{N_i} \tag{50}$$

$$\lambda = Df. \tag{51}$$

Put now

$$\hat{D} = d \prod \frac{N_i}{n_i} \tag{52}$$

wherein i runs here and hereafter from 1 to L. Then $\hat{D}$ is an unbiased estimate of D, and

$$\text{Var } \hat{D} = D \prod \frac{N_i}{n_i} \left\{ 1 + (D-1) \prod \frac{n_i - 1}{N_i - 1} - D \prod \frac{n_i}{N_i} \right\} \tag{53}$$

$$\to D \prod \frac{N_i}{n_i} \left\{ 1 - \prod \frac{n_i}{N_i} \right\} \qquad \text{Case 1} \tag{54}$$

$$\to D \prod \frac{N_i}{n_i} \qquad \text{Case 2.} \tag{55}$$

An unbiased estimate of this variance is

$$\text{Est Var } \hat{D} = \hat{D} \left\{ \prod \frac{N_i - 1}{n_i - 1} - 1 \right\} + \hat{D}^2 \left\{ 1 - \prod \frac{n_i}{N_i} \frac{N_i - 1}{n_i - 1} \right\} \tag{56}$$

$$\to \hat{D} \left\{ \prod \frac{N_i}{n_i} - 1 \right\} \qquad \text{Case 1} \tag{57}$$

$$\to D \prod \frac{N_i}{n_i} \qquad \text{Case 2.} \tag{58}$$

Optimum sample-sizes. For matching 2 samples let the costs be:

c_1 to draw a name from List 1, and to write it down or to prepare a card therefor, in preparation to compare it with the sample from List 2. c_1 includes also a proper share of the cost of sorting the cards of the sample to put them in alphabetic order.

c_2 the same for List 2.

c_3 to compare a name in one sample with a name in the other sample, and to record the comparison as 0 or 1.

Then the total cost of the job will be

$$K = mc_1 + nc_2 + mnc_3 \tag{59}$$

the optimum sizes m and n being such that

$$mc_1 = nc_2 \qquad \text{[Approximately]} \tag{60}$$

which means that to get the most precision for our money, we should choose mn big enough to yield the required precision in $\hat{p}$ or in $\hat{P}$, and then equate the costs of drawing the 2 samples. To derive this result, we satisfy ourselves with Eq. (21) for Case 2, for which

$$C_{\hat{p}}^{\ 2} = N/mnp.$$

This equation fixes the product mn, also the cost mnc_3 of matching. We now write

$$mc_1 \cdot nc_2 = Nc_1c_2/pC_{\hat{p}}^{\ 2}. \tag{61}$$

The right-hand side of this last equation is a number, once we fix N, c_1, c_2, $c_{\hat{p}}^2$ and insert plausible value of p. We may treat mc_1 as one variable, nc_2 as another. If now we were to plot mc_1 on one axis of rectangular coordinates, and nc_2 on the other, the graph of the last equation would be a hyperbola. The coordinates of any point thereon are merely the costs of drawing the 2 samples. The sum of these 2 costs, and hence also the total cost K, is a minimum where the hyperbola meets the 45°-line $mc_1 = nc_2$. If the costs of drawing names from the lists are equal ($c_1 = c_2$), an exact result for the optimum sizes is $m = n$.

For L lists, the optimum sizes of the samples would satisfy approximately the equations

$$n_1c_1 = n_2c_2 = \cdots = n_Lc_L. \tag{62}$$

An example in allocation of 2 samples. The procedure to find the optimum sizes of the samples could then be this:

1. Choose a plausible value of p.
2. Choose the desired coefficient of variation, $C_{\hat{p}}$.
3. Find $mn = N/pC_{\hat{p}}^2$.
4. Find $m = \sqrt{mnc_2/c_1}$
   $n = mc_1/c_2$.

Thus, suppose that N is 20,000, and that p may be about 5%. The client says that $C_{\hat{p}} = 50\%$ will be sufficient for his purpose. The costs, we suppose, are:

$$c_1 = 50¢, \qquad c_2 = 25¢, \qquad c_3 = .1¢.$$

Then

$$\begin{aligned} mn &= N/pC_{\hat{p}}^{\ 2} = 20{,}000/.05 \times .25 = 1{,}600{,}000 \\ m &= \sqrt{mnc_2/c_1} = \sqrt{1{,}600{,}000 \times \tfrac{1}{2}} = \sqrt{800{,}000} \doteq 900 \\ n &= mc_1/c_2 = 1800 \end{aligned}$$

and the total cost of the job would be

$$K = mc_1 + nc_2 + mnc_3$$
$$= .50 \times 900 + .25 \times 1800 + 1{,}600{,}000 \times .001 = \$2500. \qquad (63)$$

To compare this cost with proportionate allocation, we keep $mn = 1{,}600{,}000$, but before we can go further, we must assume some value for M: let $M = 2N$, and $m = 2n$, $mn = 2n^2$. Then by Eq. (21),

$$C_{\hat{p}}^2 = N/mnp = N/2n^2p$$
$$n^2 = \tfrac{1}{2}N/pC_{\hat{p}}^2\, C_{\hat{p}}^2$$
$$= 10{,}000/.05 \times 5^2 = 800{,}000$$
$$n = 895$$
$$m = 1790$$
$$mn = 1{,}600{,}000 \text{ as before.}$$

The total cost would be

$$K = mc_1 + nc_2 + mnc_3 = \$895 + \$447.50 + \$1600$$
$$= \$2942.50 \qquad (64)$$

to compare with \$2500 by the optimum allocation.

We compute now also, for comparison, the cost to attain the same precision with equal allocation, $m = n$:

$$C_{\hat{p}}^2 = N/mnp = N/n^2p \qquad \text{(as before, from Eq. 21)}$$
$$n^2 = N/pC_{\hat{p}}^2$$
$$= 20{,}000/.05 \times .5^2 = 1{,}600{,}000$$
$$n = 1265 = m.$$

The total cost would be in this case

$$K = mc_1 + nc_2 + mnc_3$$
$$= .50 \times 1265 + .25 \times 1265 + \$1600 = \$2548.75 \qquad (65)$$

which exceeds only slightly the cost for optimum allocation.

Duplicates within lists. We now drop the requirement that no name appear more than once on a list.[5] We restrict this excursion to 2 lists, and to the possibility that some names occur twice on a list, but not thrice nor more. Let D_{ij} be the number of names that appear i times on List 1 and j times on List 2. Both i and j may be 0, 1, 2. Then if M' is the number of distinct names on List 1, likewise N' for List 2, and if D' is the number of distinct names common to the 2 lists, then

$$M' = M - (D_{20} + D_{21} + D_{22}) \qquad (66)$$
$$N' = N - (D_{02} + D_{12} + D_{22}) \qquad (67)$$
$$D' = D_{11} + D_{12} + D_{21} + D_{22}. \qquad (68)$$

[5] Cf. Mosteller, Frederick (ed.), "Questions and answers," *American Statistician* (1949), no. 3, pp. 12–3; and Goodman, Leo A., "On the estimation of the number of classes in a population," *Annals of Mathematical Statistics*, 20 (1949), pp. 572–9.

Denote by d_{ij} the number of names that appear i times in the sample from List 1 and j times in the sample from List 2. Then the random variables

$$\hat{M}' = M - \frac{M}{m}\frac{M-1}{m-1}(d_{20} + d_{21} + d_{22}) \tag{69}$$

$$\hat{N}' = N - \frac{N}{n}\frac{N-1}{n-1}(d_{02} + d_{12} + d_{22}) \tag{70}$$

$$\hat{D}' = \frac{MN}{mn}\left\{d_{11} - \left(\frac{N-1}{n-1} - 2\right)d_{12} - \left(\frac{M-1}{m-1} - 2\right)d_{21} + \left(\frac{M-1}{m-1}\frac{N-1}{n-1} - 2\frac{M-1}{m-1} - 2\frac{N-1}{n-1} - 4\right)d_{22}\right\} \tag{71}$$

are unbiased estimates of M', N', D', and $\hat{M}'+\hat{N}'-\hat{D}'$ is an unbiased estimate of $M'+N'-D'$, which is the number of distinct names on the 2 lists combined.

Eq. (71) reduces to Eq. (9) if $d_{12}=d_{21}=d_{22}=0$; that is, if no duplication within lists appears in the sample. This indicates that unless such duplication appears, thereby invalidating our assumption that no name appear more than once on a list, the theory of estimation presented up to this point is sufficient. The same is true in the more general cases.

Stratified random sampling. In some applications it may be possible to increase the efficiency of the sample results by the judicious use of stratification. To go about it we divide each list into R strata, with the strata in one list in a one-to-one correspondence with the strata in the other. The theory presented assumes that any duplicates occur only in the corresponding strata. If the lists are lists of names, we may accomplish stratification by reference to last initial, geographical location, or by some other criterion.

With M_i, N_i, D_i, m_i, n_i and d_i representing the appropriate characteristics of the i-th stratum and the sample selected from it, an unbiased estimate of D is

$$D_s = \sum_{i=1}^{R} \frac{M_iN_i}{m_in_i} d_i. \tag{72}$$

The variance of this estimate is

$$\operatorname{Var} \hat{D}_s = \sum_{i=1}^{R} \frac{M_iN_i}{m_in_i} D_i\left\{1 + \frac{(m_i-1)(n_i-1)}{(M_i-1)(N_i-1)}(D_i - 1) - \frac{m_in_i}{M_iN_i}D_i\right\} \tag{73}$$

and an unbiased estimate of this variance is

$$\operatorname{Est\,Var} \hat{D}_s = \sum_{i=1}^{R}\left(\frac{M_iN_i}{m_in_i}\right)^2\left\{d_i\left(1 - \frac{m_in_i}{M_iN_i}\right) + d_i(d_i-1)\left(1 - \frac{m_i}{m_i-1}\frac{n_i}{n_i-1}\frac{M_i-1}{M_i}\frac{N_i-1}{N_i}\right)\right\}. \tag{74}$$

The optimum allocation of a sample of fixed size, to minimize the variance of $\hat{D}_s$, involves the requirement that $m_i=n_i$ for all strata, and that

$$m_i = n_i \doteq m \frac{\sqrt[3]{D_i M_i N_i}}{\sum_{i=1}^{R} \sqrt[3]{D_i M_i N_i}} \qquad [i = 1, 2, \cdots, R] \tag{75}$$

where $m(=n)$ is the size of the sample to select from all strata combined in List 1 (or in List 2). The accompanying table illustrates the optimum allocation with a hypothetical example.

Stratum	M_i	N_i	D_i	$\sqrt[3]{D_i M_i N_i}$	m_i	n_i
1	100	200	12	62	27	27
2	200	100	13	64	28	28
3	300	350	36	156	68	68
4	400	350	39	176	77	77
Total	1000	1000	100	458	200	200

$$\begin{aligned} \text{Var } \hat{D} &= 2321 \qquad [\text{by Eq. } (21)] \\ \text{Var } \hat{D}_s &= 2223 \qquad [\text{by Eq. } (73)] \\ \frac{\text{Var } \hat{D}_s}{\text{Var } \hat{D}} &= .958. \end{aligned} \tag{76}$$

One requires some assumption about the unknown values of the D_i in order to apply Eq. (75). In the absence of any other hints, we might in some applications assume each D_i proportional to the smaller of the M_i and N_i.