3-1976

Random Juror Selection from Multiple Lists

John P. Lehoczky
Carnegie Mellon University, jpl@stat.cmu.edu

Joseph B. Kadane
Carnegie Mellon University, kadane@stat.cmu.edu

Published In
Operations Research, 24, 2, 207- 219.

OPERATIONS RESEARCH, Vol. 24, No. 2, March-April 1976

Random Juror Selection from Multiple Lists

JOSEPH B. KADANE and JOHN P. LEHOCZKY

Carnegie-Mellon University, Pittsburgh, Pennsylvania

(Received original April 3, 1975; final, August 22, 1975)

We examine the selection of jurors' names from multiple source lists, using statistical and optimization methodology. Five plans for sampling at random from overlapping lists of names are analyzed for their probabilistic and cost properties. In each plan the probability of a name being selected is independent of which and how many lists it appears on. We consider the optimal ordering of the frames to minimize cost and develop a heuristic for solving this problem. Although the methods are discussed in terms of juror selection, the results apply to sampling from overlapping frames in any context. For instance, if lists of equipment are kept according to possible uses, with versatile equipment listed many times, the methods of this paper can be used to draw a random sample of equipment to check for readiness.

THE FEDERAL Jury Selection and Service Act of 1968 (P.L. 90-274, 82 Stat. 53) provides methods for the selection of citizens to serve on federal juries. The law provides that

> It is the policy of the United States that all litigants in Federal Courts entitled to trial by jury shall have the right to grand and petit juries selected at random from a fair cross-section of the community (Section 1861).

and

> Each United States district court shall devise and place into operation a written plan for random selection of grand and petit juries that shall . . . specify whether the names of prospective jurors shall be selected from the voter registration list or the lists of actual voters. . . . The plan shall prescribe some other source or sources of names in addition to voter lists when necessary to foster the policy and protect the rights secured by sections 1861 and 1862 of this title (Section 1863).

The law clearly provides for the use of supplemental lists when the voter registration roll is not representative of a fair cross-section of the community. However, the law does not specify how unrepresentative the voter list must be before supplemental lists are required, nor does it mention what lists should be used as supplements. The common implementation of this law is to use only the voter registration roll. One dis-

advantage of using only the voter registration list for selecting jurors is that some citizens do not register in order to avoid jury duty.[13]

In 1970 the National Conference of Commissioners on Uniform State Laws drafted the Uniform Jury Selection and Service Act, which has since been adopted by Colorado, Idaho, and North Dakota. The act provides in Section 5:

> The jury commissioner for each [county][district] shall compile and maintain a master list consisting of all [voter registration lists] [lists of actual voters] for the [county] [district] supplemented with names from other lists of persons resident therein, such as lists of utility customers, property [and income] taxpayers, motor vehicle registrations, and drivers licenses. . . . In compiling the master list the jury commission shall avoid duplication of names.[10]

The Uniform Act then goes on in Section 6 to specify a particular method for obtaining a random sample from the master list to comprise a "master jury wheel." It should be noticed that this act requires the use of supplemental lists and all states adopting this act will require methods of finding random samples from multiple name sources with overlaps.

The purpose of this paper is to present sampling methods that will allow jury commissioners to use multiple overlapping lists and to produce a random sample from those lists. The probability that any individual on at least one list is selected should be independent of on how many and what lists the individual appears. Although the methods are discussed in terms of juror selection, the results apply to sampling from overlapping lists in any context. For instance, if lists of equipment are kept according to possible uses, with versatile equipment listed many times, the methods of this paper can be used to draw a random sample of equipment to check for readiness.

1. FORMULATION OF THE SAMPLING PLANS

We assume that a population of potential jurors is composed of L distinct individuals. Each individual is listed on at least one and possibly as many as k different but overlapping frames. Regardless of any overlaps among the frames, we seek an equal probability sample of size n, which is defined to be any sample in which each individual has a probability of n/L of selection. We prefer a simple random element sample, defined as any sample in which each subset of d names has a probability $\binom{n}{d} \Big/ \binom{L}{d}$ of selection, for $d = 1, 2, \cdots, n$.

Statistical problems associated with multiple frame sampling have been discussed by several authors; however, their aims have been different. Kish, reference 7, pp. 394–395, outlines the problem of overlapping frames and describes the checking procedure used in this paper. Other work in this area has been centered on the estimation of the overlaps in the list, for example, Hartley,[5, 6] Cochran,[1] Goodman,[4] Deming and Glasser,[2] and Sirken and Levy.[11] We concentrate on the cost of obtaining random samples from multiple frames and its minimization.

Given that we can find several plans that yield equal probability samples or, simple random element samples, then the cost of a sampling plan will be one of the major considerations in choosing among plans. This cost arises primarily from the searching of a frame to determine if a certain name is present or absent from that frame. We assume a cost c_i, $i=1,\cdots,$ k, to verify if any particular name is on frame i. The cost of selecting at random a single name from any frame is positive but negligible compared to c_i, the searching cost. Plans 2, 3, and 4 are variations of a two-stage sampling scheme with blanks similar to that described by Kish, reference 7, p. 394. Plan 5 is a variation of a scheme suggested by Hartley (private communication), modified to reduce the cost of the plan. We assume that any name is mentioned on each list at most once, although our plans and analysis can be easily extended to the case where duplicates may appear on some or all lists.

Plan 1

The k frames are merged into a single master frame with one entry for each of the L individuals. A simple random element sample of size n is then chosen.

This plan is attractive when the sampling proportion n/L is large, which is seldom the case. For jury selection in Allegheny County in Pennsylvania, for example, L is about 1.5 million, and n is roughly 30,000, leading to a sampling proportion of 2 percent. Where n/L is small the effort to compile a single master frame is largely wasted, and plan 1 is very costly relative to plans 2, 3, 4, and 5. Additionally, the component lists become obsolete at different times, and each obsolescence may require recompilation of the master list.

Let n_i be the total number of individuals on frame i, and let l_i be the number of good names on frame i. We note $l_1=n_1$, $l_1+\cdots+l_k=L$, and the k frames are comprised of L good names and $N-L$ blanks where $N=n_1+\cdots+n_k$. The $\langle l_i\rangle_{i=1}^{i=k}$ depend upon the order of the frames and are not usually known.

Plan 2

Select a frame at random where frame i is selected with probability n_i/N, $i=1, \cdots, k$. Select a name at random from the frame and apply the checking procedure below. Repeat the process until a sample of size n is generated.

Plans 2, 3, and 4 use the following checking routine. The frames are ordered, 1, 2, $\cdots, k$. Suppose a name is drawn from frame i. Frames $i-1, i-2, \cdots, 1$ are checked sequentially at costs $c_{i-1}, c_{i-2}, \cdots, c_1$ for the presence or absence of this name. If the name is present on any of the frames checked, the name is discarded as a 'blank' for that draw, and the name is flagged on list i as a blank. If the name is absent from $i-1, \cdots, 1$, the name is 'good' and is included in the sample of n names. All names selected from frame 1 are automatically 'good' names. Only those from frame $2, \cdots, k$ are subjected to checking. This checking procedure depends on the order of the lists, which is held fixed. This checking procedure ensures that each individual is a good name on exactly one frame and a blank on all other frames.

Administratively it might be convenient to implement plan 2 by choosing some group of size n' at random from all N names, and applying the checking procedure. This revision would have the same expected cost as plan 2, provided it yielded a number of good names not greater than the number sought.

Plan 3

Names are selected at random from frame i and are checked until nl_i/L good names have been chosen. The procedure is carried out for all k frames and yields exactly n good names.

Plan 4

Choose a number p, $0<p<1$. Names are selected at random from unflagged names on frame i until pn_i different names have been checked. This is repeated for each frame. Plan 4 results in a random number of good names whose distribution depends on the parameter p. The plan must be fully carried out and may yield extra names.

Plan 5

A name is chosen at random from the remaining entries on all frames. The remaining $k-1$ frames are searched for the presence of the chosen name.

Suppose the name appears on a total of f frames ($1 \leqq f \leqq k$). An auxiliary randomization is performed. With probability $1/f$ the name is good and is included in the sample. All f entries are flagged and cannot be chosen again. With probability $(f-1)/f$ the name is a blank and is not included. In this case $f-1$ of the entries are flagged and cannot be reselected. The remaining entry is given a special code. If this entry is subsequently chosen, it is included in the sample with probability 1 without any searching.

2. RANDOMNESS OF THE PLANS

Plans 1, 2, and 5 afford a simple random element sample of size n from the population with each subset of j names having a probability of $\binom{n}{j} \Big/ \binom{L}{j}$ of being selected for $j=1, \cdots, n$. This is obvious for plan 1. For plans 2 and 5 we divide the total number of trials needed into n 'good' trials and the rest 'blank' trials. Focus on a particular good name on list i (plan 2) or good name (plan 5). Let E_t be the event that this name is not selected on the tth good trial with $t=1, \cdots, n$; E_0 is the sure event; and $E=\bigcap_{i=1}^{i=n} E_t$, the event that the individual is not in the sample. $P(E)=\prod_{t=1}^{t=n} P(E_t \mid E_0, E_1, \cdots, E_{t-1})$. For plans 2 and 5 $P(E_t|E_0, E_1, \cdots, E_{t-1})=1-P(E_t^c|E_0, E_1, \cdots, E_{t-1})=1-1/(L-t+1)=(L-t)/(L-t+1)$. Thus $P(E)=\prod_{t=1}^{t=n}(L-t)/(L-t+1)=(L-n)/L$ independent of the name and frames on which it is listed. Consider any subset of n good names and order them $1, \cdots, n$. The probability that these names are selected in the required order is $\prod_{t=1}^{t=n} 1/(L-t+1)=(L-n)!/L!$, which holds for all $n!$ arrangements. Thus each subset of size n has probability $1 \Big/ \binom{L}{n}$ of selection. For subsets of size $1 \leqq j < n$ we note there are $\binom{L-j}{n-j}$ groups of size $n-j$ that can be chosen to complete the sample of size n. Thus there are $\binom{L-j}{n-j}$ ways to have the subset of size j be a part of the chosen n, and this gives a probability of $\binom{L-j}{n-j} \Big/ \binom{L}{n} = \binom{n}{j} \Big/ \binom{L}{j}$ of selection for all subsets of size j. This proves plans 2 and 5 yield a simple random element sample.

Plan 3 yields a probability sample of size n with each individual having a probability of n/L of selection; however, all subsets of size $j (2 \leqq j \leqq n)$ are not equally likely. Plan 3 can only provide samples where each frame has at least one representative; thus samples with at least one frame missing are impossible.

Plan 4 yields a random sample size. Since exactly pn_1 names come from

frame 1, it cannot be a simple random element sample. It is, however, easily shown that each name has a probability p of selection; thus it does give a probability sample. Both plans 3 and 4 must be fully carried out to obtain a probability sample.

3. COST OF THE SAMPLING PLANS

In order to evaluate the cost of each of the plans, we introduce some notation. We partition frame i into i groups of names. The jth group $G(i,j)$, $j=1, \cdots, i-1$, consists of all names on frames i and $i-j$, but not frames $i-1, i-2, \cdots, i-j+1$. The ith group $G(i,i)$ consists of all good names on frame i. Suppose there are m_{ij} names in group $G(i,j)$ with $\sum_{j=1}^{j=i} m_{ij} = n_i$ and $m_{ii}=l_i$. The cost of checking a name from $G(i,j)$ is $c_{i-1}+\cdots+c_{i-j}$ with $c_0=0$.

Plan 2

Of the n good names chosen, suppose R_i are chosen from frame i, $R_1+\cdots+R_k=n$. The total cost T is the sum of the cost associated with checking good names, T_G, and the cost associated with checking blanks, T_B, with $T_G=R_1c_0+R_2c_1+\cdots+R_k(c_1+\cdots+c_{k-1})$ and $E(T_G)=\sum_{i=2}^{i=k} E(R_i)\cdot\sum_{j=1}^{i=1} c_j$.

To compute $E(R_i)$ we note that each of the l_i good names on frame i has a n/L probability of ultimate selection. Thus $E(R_i)=nl_i/L$ and $E(T_G)=(n/L)\sum_{i=2}^{i=k} l_i\sum_{j=1}^{i-1} c_j$.

To compute T_B we let s_{ij} be the number of names selected from $G(i,j)$. $E(T_B)=\sum_{i=2}^{i=k}\sum_{j=1}^{i-1} E(s_{ij})\sum_{l=1}^{l=j} c_{i-l}$.

To compute $E(s_{ij})$ we note that there is a total of L good names, of which n will be chosen, and m_{ij} names from group j of frame i, each of which is equally likely to be selected. The probability that a particular one of the m_{ij} is picked before the nth good name is $n/(L+1)$. Thus $E(s_{ij})=nm_{ij}/(L+1)$, and by substitution we find

$$E(T_B)=[n/(L+1)]\sum_{i=2}^{i=k}\sum_{j=1}^{i-1} m_{ij}\sum_{l=1}^{l=j} c_{i-l}$$
$$=[n/(L+1)]\sum_{l=1}^{k-1} c_l\sum_{i=l+1}^{i=k}\sum_{j=i-l}^{j=i-1} m_{ij}.$$

$$E(T)=E(T_G)+E(T_B)=[n/(L+1)]\sum_{l=1}^{k-1} c_l\sum_{i=l+1}^{i=k}\sum_{j=i-l}^{j=i-1} m_{ij}$$
$$+[n/(L+1)]\sum_{l=1}^{k-1} c_l\sum_{i=l+1}^{i=k} l_i[1+(1/L)]$$
$$=[n/(L+1)]\sum_{l=1}^{k-1} c_l\sum_{i=l+1}^{i=k}(l_i+\sum_{j=i-l}^{j=i-1} m_{ij})$$
$$+[n/L(L+1)]\sum_{l=1}^{k-1} c_l\sum_{i=l+1}^{i=k} l_i.$$

The expression $\sum_{i=l+1}^{i=k}(l_i+\sum_{j=i-l}^{j=i-1} m_{ij})$ gives the total number of distinct names on frames $l+1, l+2, \cdots, k$, and we symbolize this quantity by $n(\bigcup_{l+1}^{k})$. Thus $L=n(\bigcup)_{l+1}^{k}$. We have

$$E(T)=[n/(L+1)]\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})+[n/L(L+1)]\sum_{l=1}^{k-1} c_l[L-n(\bigcup_{l+1}^{k})]$$
$$=[n/(L+1)]\sum_{l=1}^{k-1} c_l[n(\bigcup_{l+1}^{k})-n(\bigcup_{1}^{l})/L]+[n/(L+1)]\sum_{l=1}^{k-1} c_l.$$

It is clear that $E(T)$ is well approximated by $[n/(L+1)]\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})$, the other terms being negligible in comparison. By noting that $n(\bigcup_{1}^{l})+n(\bigcup_{l+1}^{k})\geqq L$, we can easily show that

$$[n/(L+1)]\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})\leqq E(T)\leqq (n/L)\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})$$

or

$$L/(L+1)\leqq E(T)/n\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})\leqq 1,$$

and the expected cost per good name sampled is approximately $\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})/(L+1)$.

Plan 3

We again divide the total cost T into T_G and T_B. Here T_G is a constant given by

$$T_G=\sum_{i=2}^{i=k}(nl_i/L)\sum_{j=1}^{i-1} c_j=(n/L)\sum_{l=1}^{k-1} c_l\sum_{i=l+1}^{i=k} l_i.$$

Let D_i be the total cost arising from picking blanks on frame i, and s_{ij} be the number of names picked from $G(i, j)$, $j=1, \cdots, i-1$; $i=2, \cdots, k$. $D_i=\sum_{j=1}^{i-1} s_{ij}\sum_{l=1}^{l=j} c_{i-l}$ and $E(T_B)=\sum_{i=2}^{i=k}\sum_{j=1}^{i-1} E(s_{ij})\sum_{l=1}^{l=j} c_{i-l}$.

The exact distribution of s_{ij} can be calculated from simple combinatorial arguments; however, we merely find $E(s_{ij})$. Let $s_{ij}=I_1+\cdots+I_{m_{ij}}$, where

$$I_m=\begin{cases}1 \text{ if the } m\text{th name in } G(i, j) \text{ is chosen,}\\ 0 \text{ otherwise,}\end{cases}$$

$E(I_m)=nl_i/L(l_i+1)$ and $E(s_{ij})=nl_i m_{ij}/L(l_i+1)$. We have

$$E(T_B)=(n/L)\sum_{i=2}^{i=k}\sum_{j=1}^{i-1}[l_i/(l_i+1)]m_{ij}\sum_{l=1}^{l=j} c_{i-l}$$
$$=(n/L)\sum_{l=1}^{k-1} c_l\sum_{i=l+1}^{i=k}\sum_{j=i-l}^{j=i-1}[l_i/(l_i+1)]\, m_{ij}.$$

This quantity can be shown to be strictly less than the corresponding expected cost for plan 2. We simply note that $l_i/[(l_i+1)L]<1/(L+1)$ for every $i=2, \cdots, k$ since $l_i<L$. Therefore, $E(T_B)<n/(L+1)\sum_{l=1}^{k} c_l\cdot\sum_{i=l+1}^{i=k}\sum_{j=i-l}^{j=i-1} m_{ij}=E(T_B)$ for plan 2. Thus plan 3 gives smaller expected total cost than plan 2.

We have

$$E(T)=E(T_B)+E(T_G)$$
$$=(n/L)\sum_{l=1}^{k-1} c_l \sum_{i=l+1}^{i=k} \sum_{j=i-l}^{j=i-1} [l_i/(l_i+1)]l_i m_{ij}/(l_i+1)$$
$$+(n/L)\sum_{l=1}^{k-1} c_l \sum_{i=l+1}^{i=k} l_i$$
$$=(n/L)\sum_{l=1}^{k-1} c_l \sum_{i=l+1}^{i=k} [l_i/(l_i+1)](l_i+\sum_{j=i-l}^{j=i-1} m_{ij})$$
$$+(n/L)\sum_{l=1}^{k-1} c_l \sum_{i=l+1}^{i=k}(l_i/(l_i+1)).$$

We introduce a simple upper and lower bound for $E(T)$ as follows: let $\lambda=\min_{2\leqq i\leqq k} l_i$ and $\mu=\max_{2\leqq i\leqq k} l_i$. Now $1/L(1+1/\lambda)\leqq l_i/L(l_i+1)\leqq 1/L(1+1/\mu)$ for $2\leqq i\leqq k$. Thus $n/L(1+1/\lambda)\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})\leqq E(T)\leqq n/L(1+1/\mu)\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{l+k})$ and $n/L(1+1/\mu)\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})<n/(L+1)\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})\leqq$ expected total cost for plan 2.

Again as in plan 2 the expression $\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})$ is the important part of the expected cost.

Plan 4

This plan is not easy to compare directly with plans 2, 3, and 5, since it yields a random number of good names. One way to compare them is to use plan 4 with a sufficiently small p so that $pL<n$. The rest of the needed names are produced by plan i, where $i=2$, 3, or 5. The cost of this composite plan can be compared with the cost of plan i as follows:

E(Cost for plan 4, then plan i) $-E$(Cost for plan i)

$=E$(number from plan 4) [E(cost for plan 4)/E(number from plan 4)
$-E$(cost per good name from plan i)].

Thus the relevant quantity to compute is the ratio of the two marginal expectations.

It is easy to compute the two marginal expectations. A total of pn_i names is selected from frame i. Consider a name in $G(i, j)$ $j=1, \cdots, i$. The probability this name is chosen is p. The cost of checking this name is $C_{i-1}+\cdots+C_{i-j}$ with $C_0=0$. We have E(number of good names chosen from frame i) $=pl_i$ and E(cost from names in frame i) $=\sum_{j=1}^{j=i} pm_{ij}\sum_{l=1}^{l=j} c_{i-l}$.

Thus E(total number of good names selected) $=\sum_{i=1}^{i=k} pl_i=pL$ and E(total cost) $=\sum_{i=2}^{i=k}\sum_{j=1}^{j=i} pm_{ij}\sum_{l=1}^{l=j} c_{i-l}=p\sum_{l=1}^{k-1} c_l\sum_{i=l+1}^{i=k}\sum_{j=i-l}^{j=i} m_{ij}$ $=p\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})$. Then E(total cost for plan 4)/E(good names from plan 4) $=\sum_{l=1}^{k-1} c_l n(\bigcup_{l+1}^{k})/L$.

Plan 5

It is difficult to calculate exactly the expected cost for plan 5; however, we can show that this cost is larger than the cost associated with plan 2

provided the lists are ordered with $c_k \geqq c_i$ $i=1, 2, \cdots, k-1$. To simplify the calculation we let $n=L$, that is, we require that every name be selected. The expected cost per name for plan 2 is independent of n, and the expected cost per name for plan 5 is a decreasing function of n. In the case of $c_i=c_j=c$ $1 \leqq i, j \leqq k$, for $n=1$ the expected cost per name for plan 5 is $c(k-1)N/L$ and decreases to $c(k-1)$ with $n=L$.

In the unequal-cost case we consider each distinct name. Suppose a name appears on lists $i_1, i_2, \cdots, i_j$ with $1 \leqq j \leqq k$. For plan 5 every frame must be searched except for the frame from which the name is chosen. Thus the expected cost to select this name is $\sum_{i=1}^{i=k} c_i - j^{-1} \sum_{l=1}^{l=j} c_{i_l} \geqq \sum_{i=1}^{k-1} c_i$, if $c_k \geqq c_i$ $i=1, \cdots, k-1$. For plan 2 the cost to select this name will depend upon whether this name is selected on any of frames $i_2, i_3, \cdots, i_j$ before all L good names have been selected. With probability $L/(L+1)$ the name on i_l, $l=2, \cdots, j$ will have been encountered and a cost of $c_{i_{l-1}} + c_{i_{l-1}+1} + \cdots + c_{i_l-1}$ charged. Thus the expected cost from plan 2 for this name is $\sum_{h=1}^{i_1-1} c_h + \sum_{h=i_1}^{i_j-1} (L/L+1)c_h \leqq \sum_{h=1}^{i_j-1} c_h \leqq \sum_{h=1}^{k-1} c_h$, showing that plan 2 always gives a smaller expected cost than plan 5 if $c_k \geqq c_i$, $i=1, \cdots, k-1$. In fact, in most examples with n small, plan 5 will yield a cost that is large compared to plan 2 because of the necessity for large amounts of checking.

4. SELECTION OF A SAMPLING PLAN

Based on cost alone, plans 2, 3, 4, and 5 are far superior to plan 1, the creation of a master file. Furthermore, plan 3 gives the lowest cost per name and plans 4 and/or 5 the highest. There are two other important considerations. First, plans 2 and 5 allow greater flexibility than plans 3 and 4. They both provide quick methods to pick a very small number of names, and they can be terminated before the sample size of n is achieved without altering the randomness of the current sample. This is not true of plans 3 and 4, which must be fully carried out to provide the required random sample. Second, plan 3 requires nl_i/L good names to be selected from frame i. In the jury selection problem, as in many other applications, l_i is unknown and would have to be estimated for $i=2, \cdots, k$ using, for example, the methods of Hartley[5,6] or Cochran.[1] A preliminary sampling experiment would have to be undertaken to estimate $l_2, \cdots, l_k$. Not only is such a procedure costly, but the sampling plans no longer provide random samples. If the $\hat{l}_i$ are not exactly correct, then some names will have higher probability of selection than others. Plans 2, 4, and 5 require only the n_i's and L for implementation.

As a result, plan 2 seems best if the l_i's are not known, and plan 3 seems

TABLE I

Plan	Description	Randomness	Cost	Comment
1	Creation of a master file	Simple random element sample	1, Most costly	Costly and obsolete with new lists
2	Random selection from entire list	Simple random element sample	4	Flexible
3	Frame-by-frame selection using nl_i/L	Probability sample	5, Least costly	Must know l_i and carry out fully
4	Frame-by-frame selection using pn_i	Probability sample	3	Random number of good names, must carry out fully
5	Selection from entire list and randomization based on number of lists	Simple random element sample	2	Costly

best if they are known. Generally, the l_i's are not known. Table I summarizes the attributes of each plan.

5. ORDERING OF FRAMES

The cost per good name for methods 2, 3, and 4 is approximately $f_S(L)$ $\{\sum_{l=1}^{k-1} c_l n(\cup_{l+1}^{k})\}$, where $f_4(L)=1/L>f_2(L)>1/(L+1)>f_3(L)$.

The cost per name for methods 1 and 5 is substantially larger, and these plans do not change in cost when the frames are reordered. Thus we choose the quantity in brackets above to minimize by choice of ordering of the frames.

We use the notation n_{i+j} to denote the number of individuals on frame i or frame j or both and $n_{i\bar{j}}$ to denote the number of individuals on frame i but not on frame j. We let $C(i_1, i_2, \cdots, i_k)=\sum_{l=1}^{k-1} c_{i_l} n(\cup_{l+1}^{i_k})$.

The case $k=2$ is simple since $C(1,2)-C(2,1)=n_2c_1-n_1c_2$. Thus the optimal order is $(1,2)$ if and only if $c_1/n_1 \leqq c_2/n_2$.

We consider the case $k=3$ in detail to gain some insight into the nature of the ordering problem. We have $C(i_1, i_2, i_3)=n_{i_3}c_{i_2}+n_{i_2+i_3}c_{i_1}$. Suppose that the order $(1,2,3)$ minimizes C. Then the effect of switching the order of frames 2 and 1 is $0 \leqq C(2,1,3)-C(1,2,3)=n_3c_1+n_{1+3}c_2-n_3c_2-n_{3+2}c_1=n_{23}n_{13}[c_2/n_{23}-c_1/n_{13}]$; hence $c_1/n_{13} \leqq c_2/n_{23}$. Similarly, $0 \leqq C(1,3,2)-C(1,2,3)=n_2c_3-n_3c_2$, which shows $c_2/n_2 \leqq c_3/n_3$.

These two results indicate that the problem of ordering frames does not have a simple structure. In particular, there is no function $\phi(i,j)$

of two frames alone such that if $\phi(i, j)>0$, then it is optimal to order frame i ahead of frame j (see Smith[12] for problems that do have this property).

Minimization by direct enumeration of all the possibilities is possible, of course, for small values of k, say up to 8. Beyond that range we can provide only heuristic approaches that might yield a good but not necessarily optimal ordering. We hope that other researchers will be able to provide improved algorithms.

The heuristic algorithm has two parts, finding a good starting order and then making switches in adjacent frames when it decreases the cost. We present the switching algorithm first.

Suppose the first $u(0\leqq u\leqq k-2)$ frames and the last $k-u-2$ frames are fixed. Two frames remain, call them i and j, to be inserted into locations $u+1$ and $u+2$. We wish to find a criterion that will decide between (i, j) or (j, i). Let H be the set of names on the last $k-u-2$ frames that must be checked against frame $u+2$; that is, H consists of the names on exactly one of the last $k-u-2$ frames. If $u=k-2$, then H is empty. Let n_H be the number of elements in H, n_{Hi} be the number of elements in H or frame i, and $n_{H\bar{i}}$ be the number of elements in H not in frame i. Now the difference in costs from exchanging the positions of i and j is given by $C(\cdots, i, j, \cdots)-C(\cdots, j, i, \cdots)=n_Hc_j+n_{Hj}c_i+n_jc_i-(n_Hc_i+n_{Hi}c_j+n_ic_j)=c_i(n_{Hj}-n_H)-c_j(n_{Hi}-n_H)$. This shows that, given that all but two adjacent locations have been filled, (i, j) is optimal if and only if $c_i/n_{i\bar{H}}\leqq c_j/n_{j\bar{H}}$. In the particular case $u=k-2$, H is empty and $n_{i\bar{H}}=n_i$. Thus, given that the first $k-2$ frames have been specified, the $k-1st$ should be the remaining frame with smallest c_i/n_i. By comparing $c_i/n_{i\bar{H}}$ to $c_j/n_{j\bar{H}}$, we can decide whether any adjacent pairs should be switched.

We now offer a heuristic for obtaining a starting permutation. Let

$$x_{j,l}=\begin{cases}1 \text{ if list } j \text{ is put in position } l,\\ 0 \text{ otherwise;}\end{cases} \qquad y_{ij}=\begin{cases}0 \text{ if person } i \text{ is in list } j,\\ 1 \text{ otherwise.}\end{cases}$$

Then

$$C=\sum_{j,j'} x_{j,k-1}c_jx_{j',k}\sum_{i=1}^{i=L}(1-y_{ij'}) + \sum_{j,j'j''} x_{j,k-2}c_jx_{j',k-1}x_{j'',k}\sum_{i=1}^{i=L}(1-y_{ij'}y_{ij''})+\cdots$$

where C is to be minimized over the $x_{j,l}$'s subject to the constraints: (i) $x_{j,l}=0$ or 1, and (ii) $\sum_j x_{jl}=\sum_l x_{jl}=1$.

The constraint set (i) and (ii) implies that the minimization of C is an assignment problem, in the literature of integer programming. However, the first term of C is quadratic, the second term cubic, etc. For quadratic assignment problems alone only heuristics are applicable in large problems (see Koopmans and Beckman,[10] Gilmore,[3] and Lawler[11]; hence the direct minimization of C here seems beyond present integer programming methods.

Now C can be rewritten as $C = \sum_{j_1} x_{j_1,1} c_{j_1} \sum_{i=1}^{i=L} (1 - \prod_{j_1 \neq j'} y_{ij'}) + \cdots$ so j_1 can be chosen to minimize $c_j \sum_{i=1}^{i=L} (1 - \prod_{j \neq j'} y_{ij'})$. The second term in the above sum is $\sum_{j_2, j_1} x_{j_1,1} x_{j_2,2} c_{j,2} \sum_{i=1}^{i=L} (1 - \prod_{j' \neq j_1, j_2} y_{i,j'})$ so j_2 can be chosen to minimize over $j \neq j_1$, $c_j \sum_{i=1}^{i=L} (1 - \prod_{j' \neq j_1, j_2} y_{i,j'})$; and generally j_r can be chosen to achieve

$$\text{Min}_{j \neq j_1, j_2, \cdots, j_{r-1}} \{c_j \sum_{i=1}^{i=L} (1 - \prod_{j \neq j_1, j_2, \cdots, j_r} y_{ij'})\}.$$

The resulting order is then $(j_1, j_2, \cdots, j_k)$. This will not be an optimal order in general, but hopefully will be close to optimal. Then we can search the initial order $(j_i, j_2, \cdots, j_k)$ to see if any adjacent switches can be profitably made, using the results above.

Another, simpler heuristic would be to order the lists according to c_j/n_j, smallest first.

All these methods, except the last, require detailed knowledge of the overlap structure of the list. Although this overlap might be estimated at first by sampling, the operation of any of the 5 plans would supply additional information, which should be properly recorded as the system is implemented.

6. CONCLUSION AND LIMITATIONS

This paper has provided jury commissioners with a number of methods for selecting jurors from multiple list sources and given an analysis of the cost and randomness characteristics of each. With the adoption of the Uniform Act by a number of states these results have practical significance. We wish to stress certain points this paper does not address. It does not give advice on how to decide whether a given method of jury selection is adequate, nor does it provide guidance on how many and which lists to include in the multiple list system. The latter decision will have to take into account the legal mandate to consider as many citizens as possible, the legal mandate to achieve a cross-section of the community, and cost. Finally, we have not given practical advice on how to decide whether a given name is matched by another where different conventions about first names, addresses, and titles, as well as possible errors, cloud the issue.

ACKNOWLEDGMENT

The authors are grateful to H. O. Hartley, T. Kerr, G. T. Munsterman, and W. Pabst, Jr., for helpful comments. J. B. Kadane's research was supported in part by the Office of Naval Research under contract N00014-

67-A-0314-0022, Task NR 042-309 and J. P. Lehoczky's by the Air Force Office of Scientific research under Grant AFOSR 74-2642.

REFERENCES

1. R. S. Cochran. "The Estimation of Domain Sizes When Sampling Frames Are Interlocked," *Proc. Social Statistics Section American Statistical Assoc. Meetings, 1967*, 332–335, American Statistical Association, Washington, D.C.
2. W. E. Deming and G. J. Glasser, "On the Problem of Matching Lists by Samples," *JASA* **54,** 403–415 (1959).
3. P. C. Gilmore, "Optimal and Suboptimal Algorithms for the Quadratic Assignment Problem," *J. Soc. Indust. Appl. Math.* **10,** 305–313 (1962).
4. L. Goodman, "On the Analysis of Samples from k Lists," *Ann. Math. Stat.* **23,** 632–634 (1952).
5. H. O. Hartley, "Multiple Frame Surveys," *Proc. Social Statistics Section of the American Statistical Assoc. Meetings*, 203–206 (1962).
6. H. O. Hartley, "Multiple Frame Methodology and Selected Applications," *Sankhya*, Series C, **36,** 99–118 (1974).
7. L. Kish, *Survey Sampling*, Wiley, New York, 1965.
8. T. C. Koopmans and M. Beckman, "Assignment Problems and the Location of Economic Activities," *Econometrica* **25,** 53–76 (1957).
9. E. Lawler, "The Quadratic Assignment Problem," *Management Sci.* **9,** 586–599 (1963).
10. V. L. McKusick and D. E. Boxer "Uniform Jury Selection and Service Act," *Harv. J. Legislation* **8,** 280 (1971).
11. M. G. Sirken and P. S. Levy, "Multiplicity Estimation of Proportions Based on Ratios of Random Variables," *JASA* **69,** 69–73 (1974).
12. W. E. Smith, "Various Optimizers for Single-Stage Production," *Naval Res. Log. Quart.* **3,** 59–66 (1956).
13. *United States Congressional Record*, pp. S5985–S5987. April 15, 1975.