CHANGING THE PROBABILITIES OF SELECTION WHEN TWO UNITS ARE SELECTED WITH PPS WITHOUT REPLACEMENT

Ivan P. Fellegi
Dominion Bureau of Statistics, Ottawa

Introduction

Let the population consist of N units and assume that a "measure of size", m_i (i=1, 2, ..., N), is associated with each unit. A general problem of selecting n units with probability proportional to size (pps) without replacement is to devise sampling schemes such that π_i, the probability that the i-th unit is in the sample is proportional to m_i, i.e.

$$\pi_i = nm_i/\Sigma m_i \tag{1}$$

Setting $$p_i = \frac{m_i}{\Sigma m_i} \tag{2}$$

(1) can be rewritten as $$\pi_i = np_i \tag{3}$$

The present author proposed a scheme [2] of successively selecting n units without replacement in such a way that the probability δ_i^k of selecting the i-th unit at the k-th selection equals p_i:

$$\delta_i^k = p_i \qquad i=1,2,\ldots,N;\ k=1,2,\ldots,n \tag{4}$$

Since the events of selecting unit i at any two draws are mutually exclusive, it follows that

$$\pi_i = \sum_{k=1}^{n} \delta_i^k = np_i$$

This general method of selection has some particularly useful features for n=2. A method has been outlined in [2] (and a FORTRAN programme is available) to solve the following system of equations for q_i(i=1,2,...,N) and a:

$$aq_i(1-q_i) = 2p_i \qquad i=1,2,\ldots,N \tag{5}$$

$$\sum_{i=1}^{N} q_i = 1 \tag{6}$$

$$a = 2/(1 - \sum_{i=1}^{N} q_i^2) \tag{7}$$

The selection of two units is then carried out* by selecting the first unit with probabilities $\{p_i;\ i=1,2,\ldots,N\}$ and having selected unit i at the first draw, selecting the second unit from among the remaining units with probabilities $\{q_j/(1-q_i);\ j=1,2,\ldots,N;\ j\neq i\}$. The probability of selecting units i and j in that order is, therefore, equal to

$$p_i \frac{q_j}{1-q_i} \tag{8}$$

while the probability of selecting the same units in reverse order is equal to

$$p_j \frac{q_i}{1-q_j} \tag{9}$$

It follows from (5) that the expressions under (8) and (9) are both equal to

$$\frac{1}{2} aq_iq_j = p_i \frac{q_j}{1-q_i} = p_j \frac{q_i}{1-q_j} \tag{10}$$

and hence π_{ij}, the probability of selecting units i and j in either order is equal to

$$\pi_{ij} = aq_iq_j \tag{11}$$

Consequently, given that units i and j are in the sample, the conditional probability of selecting them in either order equals 1/2. Formula (4) follows from formulae (5) to (11).

The problem posed in the present paper is the following. Suppose a sample of two units had been selected as described above, with probabilities proportional to the measures of size $\{m_i\}$. The two units might, for example, be two primary sampling units (areas) and the measures of size might be census counts. Some years later a new census is taken, resulting in new measures of size M_i. It is desirable to select a sample of two psu's with the following properties: I) The probabilities of selection are proportional to the new measures of size. II) Since, however, the original sample of two psu's often represents a capital investment (listings of households might have been prepared, enumerators trained, etc.), it is desirable to maximize the overlap between the old and new sample. More precisely, if X is a random variable (depending on the sampling scheme for the selection of the new sample), such that

X = 0 if the new sample coincides with the old

= 1 if the new and old samples have one unit in common

---

* Brewer [1] proved that equations (5) to (7) always have a unique solution if only $p_i < \frac{1}{2}$ (i=1,2,...,N).

= 2 if the new and old samples are disjoint

then it is desirable to have a sampling scheme for the selection of the new sample for which

E(X) = expected number of rejections

is minimized. III) it is also desirable to have the joint probabilities of selecting two units in the new sample satisfy equations (4) to (11) in terms of P_i, Q_i, A and Π_{ij} (defined in terms of the new size measures in analogy to p_i, q_i, a and π_{ij}) so that the process of revising the size measures in the future can be repeated again (actually property III implies I).

The procedure outlined below achieves I) and III) and achieves II) approximately.

The procedure

Keyfitz [3] provided a procedure to achieve the objectives outlined above when the sample consists of one unit. In this simple case sampling with and without replacement become indistinguishable and property I) automatically implies III). Since, however, the present procedure follows the methods of Keyfitz, it is instructive to recapitulate them here briefly.

A sample of one unit had been selected with probabilities $\{p_j\}$. The desired new probabilities are $\{P_j\}$. The procedure of changing the probabilities of selection is summarized as follows: first the original unit is subjected to a test of retention; it is retained with probability

$$R_j$$

rejected with probability

$$1-R_j$$

In case of rejection, the second step consists of a subsequent selection. The new probability of selecting unit j is therefore

$$P_j = p_j R_j + S_j$$

S_j being the probability of selecting unit j at the second step. The objective is to maximize R_j (or minimize S_j) within the equation above. Clearly if

$$R_j < 1 \qquad \text{and} \qquad S_j > 0$$

the procedure is not optimal, since R_j could be further increased, S_j further decreased. In this case the probability of rejection is positive

$$1-R_j > 0$$

Hence a procedure is not optimal if it permits the rejection of a unit followed by its subsequent re-selection. The converse is also true: a procedure is optimal if the set of units with a positive probability of rejection and the set of units with a positive probability of subsequent selection are disjoint. The Keyfitz procedure is optimal in this sense. The set of units with positive probability of rejection is the set of decreasing units for which

$$p_j > P_j$$

None of these units can be selected subsequently. The set of units with positive probability of subsequent selection is the set of units which increased, i.e. for which

$$p_j \leq P_j$$

Formally

$$I = \{i:p_i \leq P_i\} \quad \text{and} \quad D = \{d:p_d > P_d\} \qquad (12)$$

The Keyfitz procedure consists of retaining the originally selected unit if it is in I. If the originally selected unit (d) is in D, then it is retained with probability $R_d = P_d/p_d$, rejected with probability $1-P_d/p_d$. Finally, if a unit in D is rejected, then a unit in I is selected with probabilities

$$S_i = \frac{P_i-p_i}{\sum_{i\epsilon I} (P_i-p_i)} \qquad i\epsilon I \qquad (13)$$

It is easy to show that the probability of unit j being in the sample after the adjustment is equal to $P_j (j=1,2,\ldots,N)$ and that the procedure is optimal in the sense of property II) above (it minimizes the expected number of rejections). These properties are based on the observation that since

$$\sum_{j=1}^{N} p_j = \sum_{j=1}^{N} P_j = 1$$

therefore

$$\sum_{d\epsilon D} (p_d-P_d) = \sum_{i\epsilon I} (P_i-p_i) \qquad (14)$$

The present procedure starts with the observation that the sampling procedure by which the original two units were selected is symmetric [2] i.e. the conditional probability of the two units having been selected in either order is equal to 1/2. Consequently if one does not know the order in which the units had been selected, one can "recreate" the order at random, e.g. by selecting a random number r $(0 < r \leq 1)$ and assuming that if $0 < r \leq 0.5$ than the order in which units i and j had been selected was higher subscript first, lower subscript second; if $0.5 < r \leq 1$ then the order in which the units had been selected was the reverse.

After this step we may assume that we have an ordered sample of two units (i first, j second) whose joint probability of selection was

$$i\pi_j = p_i \frac{q_j}{1-q_i} = \frac{1}{2} a q_i q_j \qquad (15)$$

Next we observe that the (unconditional) probability of unit i having been selected first was equal to p_i. Since (4) is to be maintained (in terms of the new measures) after the adjustment, the probability of unit i to be selected as the first should be equal to P_i. The Keyfitz procedure can be applied to this end without modification: the two sets I and D, defined under (12), are formed and the Keyfitz rules for retention, rejection and re-selection are applied.

After the adjustment of the probability of selection of the first unit the probability of selection of the second unit is adjusted. Here the Keyfitz principle has to be modified slightly. To illustrate this point, consider the sets

$$I_i = \{j: \ j\neq i \text{ and } \frac{q_j}{1-q_i} \leq \frac{Q_j}{1-Q_i}\} \qquad i=1,2,\ldots,N$$

$$D_i = \{j: \ j\neq i \text{ and } \frac{q_j}{1-q_i} > \frac{Q_j}{1-Q_i}\}$$

where $q_j/(1-q_i)$ and $Q_j/(1-Q_i)$ are the conditional probabilities of selecting unit j after i was selected.

Suppose in (15) $i\varepsilon D$ and $j\varepsilon D_i$.

According to the Keyfitz principle unit i is retained with probability

$$\frac{P_i}{p_i}$$

and unit j is retained with probability

$$\frac{Q_j}{1-Q_i} \Big/ \frac{q_j}{1-q_i}$$

This adjustment yields a probability of selection for units i and j in that order

$$i\Pi_j = i\pi_j \frac{P_i}{p_i} \cdot \frac{Q_j}{1-Q_i} \Big/ \frac{q_j}{1-q_i}$$

$$= P_i \frac{Q_j}{1-Q_i} = \frac{1}{2} A Q_i Q_j \qquad (16)$$

provided that neither unit has a chance of being re-selected once rejected (which is part of the Keyfitz principle). Formula (16) is the desirable equivalent of (15) after the adjustment. This reasonably simple procedure, however, has to be modified (as far as the adjustment of the probability of selection of the second unit is concerned!) if the first unit happens to be rejected. The reason for this is the fact that as long as the first unit is retained the Keyfitz procedure can be applied at the second step to the resulting conditional distribution (the condition being the selection of unit i at the first step). If, however, the first unit is rejected and another one is selected (e.g. i') then one has to compare

$$\frac{q_j}{1-q_i}$$

with

$$\frac{Q_j}{1-Q_{i'}},$$

i.e. one is compelled to consider two different conditional distributions.

The following modification of the Keyfitz principle yields the desirable result:

Theorem 1.

Suppose that a sample of two units had been selected without replacement with probabilities proportional to the measures of size $\{m_i\}$ satisfying formulae (1) to (11) (using the procedure described in [2]). Let $\{M_i\}$ be the new measures of size and let $\{P_i\}$, $\{Q_i\}$ and A be computed from formulae (2), (5), (6) and (7). Then the application of the rules below will result in a sample having the properties (1) to (11) in terms $\{M_i\}$, $\{P_i\}$, $\{Q_i\}$, and A.

Define the subsets of the set of integers between 1 and N as follows

$$I = \{c: p_c \leq P_c\} \qquad (17)$$

$$D = \{c: p_c > P_c\} \qquad (18)$$

$$I_j = \{c: c\neq j \text{ and } \frac{q_c}{1-q_j} \leq \frac{Q_c}{1-Q_j}\} \quad \text{for } j=1,2,\ldots,N \qquad (19)$$

$$D_j = \{c: c\neq j \text{ and } \frac{q_c}{1-q_j} > \frac{Q_c}{1-Q_j}\} \quad \text{for } j=1,2,\ldots,N \qquad (20)$$

$$I_{id} = \{c: c\neq i,\ c\neq d \text{ and } \frac{q_c}{1-q_i-q_d} \leq \frac{Q_c}{1-Q_i}\} \quad \text{for } i\varepsilon I \text{ and } d\varepsilon D \qquad (21)$$

$$D_{id} = \{c: c\neq i,\ c\neq d \text{ and } \frac{q_c}{1-q_i-q_d} > \frac{Q_c}{1-Q_i}\} \quad \text{for } i\varepsilon I \text{ and } d\varepsilon D \qquad (22)$$

Denote

$$S = \sum_{c\varepsilon I} (P_c - p_c) = \sum_{c\varepsilon D} (p_c - P_c) \tag{23}$$

$$S_j = \sum_{c\varepsilon I_j} \left(\frac{Q_c}{1-Q_j} - \frac{q_c}{1-q_j}\right)$$

$$= \sum_{c\varepsilon D_j} \left(\frac{q_c}{1-q_j} - \frac{Q_c}{1-Q_j}\right) \tag{24}$$

$$S_{id} = \sum_{c\varepsilon I_{id}} \left(\frac{Q_c}{1-Q_i} - \frac{q_c}{1-q_i-q_d}\right) \tag{25}$$

Rule 1: Given the original sample of two units determine the order in which they will be subjected to a test of rejection by selecting a random number r ($0 < r \leq 1$). If $0 < r \leq 0.5$ the unit with the lower subscript will be adjusted first (called "the first unit"); if $0.5 < r \leq 1$ then the unit with the higher subscript will be the "first unit".

Rule 2: Apply to the first unit the Keyfitz rule, i.e. a) if unit c is the first and $c\varepsilon I$ then retain it; b) if $c\varepsilon D$ retain it with probability P_c/p_c, reject with probability $1-P_c/p_c$ and in this case select a unit from I with probability $(P_c-p_c)/S$.

Rule 3: If the first unit was retained (with subscript j) and the second unit had subscript k:

(a) retain the second unit if $k\varepsilon I_j$

(b) if $k\varepsilon D_j$ retain it with probability

$$\frac{Q_k}{1-Q_j} \Big/ \frac{q_k}{1-q_j} = \frac{Q_k(1-q_j)}{q_k(1-Q_j)}$$

reject it otherwise; if it is rejected, select another one from I_j with probabilities

$$\frac{1}{S_j}\left(\frac{Q_c}{1-Q_j} - \frac{q_c}{1-q_j}\right); \qquad c\varepsilon I_j$$

Rule 4: If the first unit (d) was in D and if it was rejected and replaced by $i\varepsilon I$ and if the second unit had subscript k

(a) retain the second unit if $k\varepsilon I_{id}$

(b) if $k\varepsilon D_{id}$, retain it with probability

$$\frac{Q_k}{1-Q_i} \Big/ \frac{q_k}{1-q_i-q_d} = \frac{Q_k(1-q_i-q_d)}{q_k(1-Q_i)}$$

reject it otherwise; if it is rejected, select d with probability

$$\frac{Q_d}{Q_d+(1-Q_i)\,S_{id}}$$

and if d is not selected, select a unit from I_{id} with probabilities

$$\frac{1}{S_{id}}\left(\frac{Q_c}{1-Q_i} - \frac{q_c}{1-q_i-q_d}\right), \qquad c\varepsilon I_{id}$$

(c) if k=i, select from among all the units, excepting i, with probabilities

$$\frac{Q_c}{1-Q_i} \qquad c=1,2,\ldots,N; \quad c\neq i$$

Two lemmas are required for the proof of theorem 1:

Lemma 1. Let S_{id} be defined by formula (25). Then

$$S_{id} = \sum_{c\varepsilon I_{id}} \left(\frac{Q_c}{1-Q_i} - \frac{q_c}{1-q_i-q_d}\right)$$

$$= \sum_{c\varepsilon D_{id}} \left(\frac{q_c}{1-q_i-q_d} - \frac{Q_c}{1-Q_i}\right) - \frac{Q_d}{1-Q_i} \tag{26}$$

Lemma 2. Suppose that in the original sample the first unit (as obtained from Rule 1) was unit $d\varepsilon D$, the second unit was $k(\neq d)$ and that the application of Rule 2 resulted in the replacement of d by $i\varepsilon I$. Denote the conditional probability of obtaining (through retention or through rejection and a new selection) unit β as the second unit of the new sample by

$$P_\beta(d\to i,\ k)$$

Then

$$\sum_{\substack{k=1\\k\neq d}}^{N} q_k P_\beta(d\to i,\ k) = (1-q_d)\frac{Q_\beta}{1-Q_i} \tag{27}$$

for any $d\varepsilon D$, $i\varepsilon I$ and any $\beta(\neq i)$.

The proof of lemma 1 follows immediately from (6) (i.e. the analogous result for Q_i) and from the observation that the union of D_{id} and I_{id} is the set of integers from 1 to N excepting i and d. The proof of lemma 2 will be presented after the proof of theorem 1.

Proof of theorem 1:

Denoting the first and second units of the new sample by α and β respectively, there are four possibilities:

(i) $\alpha\varepsilon I$, $\beta\varepsilon I_\alpha$

(ii) $\alpha\varepsilon I$, $\beta\varepsilon D_\alpha$

(iii) $\alpha\varepsilon D$, $\beta\varepsilon I_\alpha$

(iv) $\alpha\varepsilon D$, $\beta\varepsilon D_\alpha$

It will be shown that in each case the probability of obtaining units α and β in that order in the new sample is equal to

$$\frac{1}{2} AQ_\alpha Q_\beta \tag{28}$$

This will complete the proof of the theorem in that it will then follow from (5) (in terms of A, P_i, Q_i) that the probability of obtaining in the new sample unit α as the first unit is equal to

$$\sum_{\beta(\neq\alpha)} \frac{1}{2} AQ_\alpha Q_\beta = \frac{1}{2} AQ_\alpha(1-Q_\alpha) = P_\alpha$$

$$\alpha=1, 2, \ldots, N$$

Also the probability of obtaining unit β as the second unit is equal to

$$\sum_{\alpha(\neq\beta)} \frac{1}{2} AQ_\alpha Q_\beta = \frac{1}{2} A(1-Q_\beta)\, Q_\beta = P_\beta$$

$$\beta=1, 2, \ldots, N$$

and from the symmetry of (28) the probability of obtaining α and β in either order is equal to

$$\Pi_{\alpha\beta} = AQ_\alpha Q_\beta$$

In the following proofs we assume that rule 1 had already been applied and that consequently the original sample had already been ordered.

Proof of case (i): $\alpha\varepsilon I$, $\beta\varepsilon I_\alpha$

Such a sample can arise in the following three ways: the original sample was α and β in that order in which case they are both retained (rules 2 and 3a); the original sample was α and $k\varepsilon D_\alpha$, α was retained and $\beta\varepsilon I_\alpha$ selected (rules 2 and 3b); the first unit was $d\varepsilon D$, it was rejected, $\alpha\varepsilon I$ selected (rule 2) then β was obtained for the second unit as in lemma 2. The three probabilities corresponding to these three events are as follows:

$$\frac{1}{2} aq_\alpha q_\beta +$$

$$+ \sum_{k\varepsilon D_\alpha} \frac{1}{2} aq_\alpha q_k \left[1- \frac{Q_k(1-q_\alpha)}{q_k(1-Q_\alpha)}\right] \frac{1}{S_\alpha} \left(\frac{Q_\beta}{1-Q_\alpha} - \frac{q_\beta}{1-q_\alpha}\right) +$$

$$+ \sum_{d\varepsilon D} \sum_{k(\neq d)} \frac{1}{2} aq_d q_k \left(1- \frac{P_d}{p_d}\right) \frac{1}{S}(P_\alpha-p_\alpha)P_\beta(d\to\alpha, k) =$$

$$= \frac{1}{2} aq_\alpha q_\beta +$$

$$+ \frac{1}{2} aq_\alpha(1-q_\alpha) \frac{1}{S_\alpha} \left(\frac{Q_\beta}{1-Q_\alpha} - \frac{q_\beta}{1-q_\alpha}\right) \sum_{k\varepsilon D_\alpha} \left(\frac{q_k}{1-q_\alpha} - \frac{Q_k}{1-Q_\alpha}\right) +$$

$$+ \frac{1}{S}(P_\alpha-p_\alpha) \sum_{d\varepsilon D} \frac{1}{2} aq_d\left(1 - \frac{P_d}{p_d}\right)(1-q_d) \frac{Q_\beta}{1-Q_\alpha}$$

the last term following from lemma 2. Applying (5), (10), (23) and (24) we obtain

$$\frac{1}{2} aq_\alpha q_\beta +$$

$$+ p_\alpha \frac{Q_\beta}{1-Q_\alpha} - \frac{1}{2} aq_\alpha q_\beta$$

$$+ (P_\alpha-p_\alpha) \frac{Q_\beta}{1-Q_\alpha} = P_\alpha \frac{Q_\beta}{1-Q_\alpha} = \frac{1}{2} AQ_\alpha Q_\beta$$

Proof of case (ii): $\alpha\varepsilon I$, $\beta\varepsilon D_\alpha$

Such a sample can arise in two ways: the original sample was α and β and β was retained (α is certainly retained by rule 2); the first unit was $d\varepsilon D$, it was rejected, $\alpha\varepsilon I$ selected (rule 2) then β was obtained for the second unit as in lemma 2. The probabilities corresponding to these events are as follow:

$$\frac{1}{2} aq_\alpha q_\beta \frac{Q_\beta(1-q_\alpha)}{q_\beta(1-Q_\alpha)} +$$

$$+ \sum_{d\varepsilon D} \sum_{k(\neq d)} \frac{1}{2} aq_d q_k \left(1- \frac{P_d}{p_d}\right) \frac{1}{S}(P_\alpha-p_\alpha)\, P_\beta\, (d\to\alpha, k)$$

$$= \frac{1}{2} aq_\alpha(1-q_\alpha) \frac{Q_\beta}{1-Q_\alpha} +$$

$$+ (P_\alpha-p_\alpha) \frac{Q_\beta}{1-Q_\alpha}$$

the second term on the right hand side following from lemma 2 in a way identical to the manipulation of the last term in the proof of (i). Applying (5) to the first term we obtain

$$p_\alpha \frac{Q_\beta}{1-Q_\alpha} + (P_\alpha-p_\alpha) \frac{Q_\beta}{1-Q_\alpha} = P_\alpha \frac{Q_\beta}{1-Q_\alpha} = \frac{1}{2} AQ_\alpha Q_\beta$$

Proof of case (iii): $\alpha\varepsilon D$, $\beta\varepsilon I_\alpha$

Such a sample can arise in two ways: the original sample was α and β and α was retained (rule 2; β is then automatically retained by rule 3a); the original sample was α and $k\varepsilon D_\alpha$, α retained (rule 2) k rejected and $\beta\varepsilon I_\alpha$ selected (rule 3b). The probabilities corresponding to these events are as follow:

$$\frac{1}{2} aq_\alpha q_\beta \frac{P_\alpha}{p_\alpha} +$$

$$+ \sum_{k\varepsilon D_\alpha} \frac{1}{2} aq_\alpha q_k \frac{P_\alpha}{p_\alpha} \left(1 - \frac{Q_k(1-q_\alpha)}{q_k(1-Q_\alpha)}\right) \frac{1}{S_j}$$

$$\left(\frac{Q_\beta}{1-Q_\alpha} - \frac{q_\beta}{1-q_\alpha}\right) =$$

$$= \frac{1}{2} aq_\alpha q_\beta \frac{P_\alpha}{p_\alpha} +$$

$$+ \frac{1}{2} aq_\alpha(1-q_\alpha) \frac{P_\alpha}{p_\alpha} \frac{1}{S_j} \left(\frac{Q_\beta}{1-Q_\alpha} - \frac{q_\beta}{1-q_\alpha}\right)$$

$$\sum_{k\varepsilon D_\alpha} \left(\frac{q_k}{1-q_\alpha} - \frac{Q_k}{1-Q_\alpha}\right)$$

$$= P_\alpha \frac{q_\beta}{1-q_\alpha} +$$

$$+ P_\alpha \left(\frac{Q_\beta}{1-Q_\alpha} - \frac{q_\beta}{1-q_\alpha}\right)$$

the first term following from (5), the second from (5) and (24). We obtain

$$P_\alpha \frac{Q_\beta}{1-Q_\alpha} = \frac{1}{2} AQ_\alpha Q_\beta$$

Proof of case (iv): $\alpha\varepsilon D$, $\beta\varepsilon D_\alpha$

Such a sample can only arise in one way: the original sample was α and β, α was retained (rule 2) and β was retained (rule 3b). The probability of this event is

$$\frac{1}{2} aq_\alpha q_\beta \frac{P_\alpha}{p_\alpha} \frac{Q_\beta(1-q_\alpha)}{q_\beta(1-Q_\alpha)} = P_\alpha \frac{Q_\beta}{1-Q_\alpha} = \frac{1}{2} AQ_\alpha Q_\beta$$

after two applications of (5).

This completes the proof of theorem 1.

Proof of lemma 2:

The original sample consisted of units $d\varepsilon D$ and $k(\neq d)$. The application of rule 2 resulted in the rejection of d and its replacement by $i\varepsilon I$. Lemma 2 is to be proved for all $\beta(\neq i)$, i.e. for the cases when $\beta\varepsilon D_{id}$, $\beta\varepsilon I_{id}$ and $\beta=d$.

(a) If $\beta\varepsilon D_{id}$:

According to rule 4a) and rule 4b) if $k\varepsilon I_{id}$ or if $k\varepsilon D_{id}$ but $k\neq\beta$

$$P_\beta(d\to i, k) = 0$$

According to rule 4b) if $k=\beta\varepsilon D_{id}$

$$P_\beta(d\to i, \beta) = \frac{Q_\beta(1-q_i-q_d)}{q_\beta(1-Q_i)}$$

According to rule 4c)

$$P_\beta(d\to i, i) = \frac{Q_\beta}{1-Q_i}$$

Consequently

$$\sum_{\substack{k=1\\(k\neq d)}}^{N} q_k P_\beta(d\to i, k) = q_\beta \frac{Q_\beta(1-q_i-q_d)}{q_\beta(1-Q_i)} + q_i \frac{Q_\beta}{1-Q_i}$$

$$= (1-q_d) \frac{Q_\beta}{1-Q_i}$$

(b) If $\beta\varepsilon I_{id}$

According to rule 4a) if $k\varepsilon I_{id}$ but $k\neq\beta$ then

$$P_\beta(d\to i, k) = 0$$

and

$$P_\beta(d\to i, \beta) = 1$$

According to rule 4c)

$$P_\beta(d\to i, i) = \frac{Q_\beta}{1-Q_i}$$

According to 4b) if $k\varepsilon D_{id}$

$$P_\beta(d\to i, k) = \left(1 - \frac{Q_k(1-q_i-q_d)}{q_k(1-Q_i)}\right)\left(1 - \frac{Q_d}{Q_d+(1-Q_i)S_{id}}\right)$$

$$\frac{1}{S_{id}} \left(\frac{Q_\beta}{1-Q_i} - \frac{q_\beta}{1-q_i-q_d}\right)$$

Hence

$$\sum_{\substack{k=1\\(k\neq d)}}^{N} q_k P_\beta(d\to i, k) = q_\beta + q_i \frac{Q_\beta}{1-Q_i} +$$

$$+ \left(1 - \frac{Q_d}{Q_d+(1-Q_i)S_{id}}\right) \frac{1}{S_{id}} \left(\frac{Q_\beta}{1-Q_i} - \frac{q_\beta}{1-q_i-q_d}\right)$$

$$(1-q_i-q_d) \sum_{k\varepsilon D_{id}} \left(\frac{q_k}{1-q_i-q_d} - \frac{Q_k}{1-Q_i}\right)$$

Applying lemma 1 to the last term we obtain

$$q_\beta + q_i \frac{Q_\beta}{1-Q_i} + \frac{(1-Q_i)S_{id}}{Q_d+(1-Q_i)S_{id}} \frac{1}{S_{id}} \left(\frac{Q_\beta}{1-Q_i} - \frac{q_\beta}{1-q_i-q_d}\right)$$

$$(1-q_i-q_d)\left(S_{id} + \frac{Q_d}{1-Q_i}\right)$$

$$= (1-q_d) \frac{Q_\beta}{1-Q_i}$$

(c) If $\beta=d$

According to rule 4b) if $k\varepsilon I_{i\beta}$

$$P_\beta(\beta\to i, k) = 0$$

According to rule 4c)

$$P_\beta\ (\beta\to i,\ i) = \frac{Q_\beta}{1-Q_i}$$

According to rule 4b) if $k\epsilon D_{i\beta}$

$$P_\beta(\beta\to i,\ k) = \left(1-\frac{Q_k(1-q_i-q_\beta)}{q_k(1-Q_i)}\right)\frac{Q_\beta}{Q_\beta+(1-Q_i)S_{i\beta}}$$

Consequently

$$\sum_{\substack{k=1\\(k\neq\beta)}}^{N} q_k P_\beta(\beta\to i,k) = q_i\frac{Q_\beta}{1-Q_i} + \frac{Q_\beta}{Q_\beta+(1-Q_i)S_{i\beta}}(1-q_i-q_\beta)$$

$$\sum_{k\epsilon D_{i\beta}}\left(\frac{q_k}{1-q_i-q_\beta}-\frac{Q_k}{1-Q_i}\right)$$

Applying lemma 1 to the last term, we obtain

$$q_i\frac{Q_\beta}{1-Q_i} + \frac{Q_\beta}{Q_\beta+(1-Q_i)S_{i\beta}}(1-q_i-q_\beta)\left(S_{i\beta}+\frac{Q_\beta}{1-Q_i}\right)$$

$$= (1-q_\beta)\frac{Q_\beta}{1-Q_i}$$

This completes the proof of Lemma 2.

The expected number of rejections

It has been proved in the previous section that the procedure as described by the rule of theorem 1 satisfies objectives I) and III) as set out in the introduction. With respect to the objective of minimizing the expected rejections the following may be said. The procedure would be optimal if it rejected originally selected units only to the extent necessary, i.e. if it rejected a unit only if its probability of selection for the new sample had to be diminished, and if in this case the original probability of selection times the probability of retention were equal the desired new probability of rejection. A procedure is not optimum if a rejected unit can be re-selected since this means that the rejection procedure renders its probability for the new sample too small (it is rejected with an unnecessarily large probability).

The present procedure is such that a first unit (or a second unit), once rejected, can never be re-selected as a first unit (or a second unit). There is, however, one (and only one) condition, under which a unit rejected as a first unit can be re-selected as a second unit. This is embodied in rules 4b) and 4c) when a first unit $d\epsilon D$ is rejected, $i\epsilon I$ is selected, the second unit k is in D_{id} or k=i, it is rejected and d is re-selected as a second unit. It is not too difficult to show that if one did not permit the re-selection of d as a second unit at all, then the procedure would not yield the required probabilities for the new sample. Whether the present procedure is actually optimal, however, is not known. Neither is it known whether there is an optimal procedure at all (subject to objectives I) and III)). Yet the probability of the event described above which may result in a departure from optimality is sufficiently small to render the statement concerning the "near optimality" of the procedure more or less justified.

The actual formula for the expected number of rejections can be derived without difficulty.

Theorem 2.

The expected number of rejections when applying the rules of theorem 1 is given by

$$\sum_{i\epsilon I} p_i S_i + \sum_{d\epsilon D} P_d S_d + S + \sum_{i\epsilon I}\sum_{d\epsilon D}(P_i-p_i)(p_d-P_d)\frac{S_{id}}{S}$$

$$-\sum_{d\epsilon D}\sum_{i\epsilon I}\frac{1}{2}aq_dq_i\left(1-\frac{P_d}{p_d}\right)\frac{P_i-p_i}{S}\frac{Q_d}{1-Q_i}$$

$$-\sum_{d\epsilon D}\sum_{i\epsilon I}\frac{1}{2}aq_dq_i\left(1-\frac{P_d}{p_d}\right)\frac{P_i-p_i}{S}S_{id} \qquad (29)$$

and an upper bound of (29) is given by

$$\sum_{i=1}^{N} P_i S_i + S \qquad (30)$$

Proof:

The event "at least one rejection" is composed of the following mutually exclusive events:

(i) the first unit is rejected

(ii) the first unit is retained and the second unit rejected.

It is easy to show that the probability of event (i), as a result of applying rule 2, is given by

$$S = \sum_{i\epsilon I}(P_i-p_i) = \sum_{d\epsilon D}(p_d-P_d) \qquad (31)$$

The probability of event (ii) is given by (rule 3)

$$\sum_{i\epsilon I}\sum_{k\epsilon D_i}\frac{1}{2}aq_iq_k\left(1-\frac{Q_k(1-q_i)}{q_k(1-Q_i)}\right) + \sum_{d\epsilon D}\sum_{k\epsilon D_i}\frac{1}{2}aq_dq_k\frac{P_d}{p_d}\left(1-\frac{Q_k(1-q_d)}{q_k(1-Q_d)}\right)$$

$$= \sum_{i\epsilon I} p_i\sum_{k\epsilon D_i}\left(\frac{q_k}{1-q_i}-\frac{Q_k}{1-Q_i}\right) + \sum_{d\epsilon D} P_d\sum_{k\epsilon D_i}\left(\frac{q_k}{1-q_d}-\frac{Q_k}{1-Q_d}\right)$$

$$= \sum_{i\epsilon I} p_i S_i + \sum_{d\epsilon D} P_d S_d \qquad (32)$$

the middle line following from (5).

(31) plus (32) represent the probability of at least one rejection in the course of selecting the new sample. It is possible to arrive at the new sample via rejections and obtain the old sample. This happens according to rule 4c) with the probability (dεD and iεI constituted the first sample in that order, d was rejected and i selected for the first unit in the new sample and d was then selected for the second unit in the new sample):

$$\sum_{d\epsilon D} \sum_{i\epsilon I} \frac{1}{2} a q_d q_i \left(1 - \frac{P_d}{p_d}\right) \frac{P_i - p_i}{S} \frac{Q_d}{1-Q_i} \qquad (33)$$

Consequently (31) plus (32) minus (33) represents the probability that the new sample is not identical with the old. Adding to this the probability of the event that the new sample will be entirely different from the old we shall obtain the expected number of rejections. The probability of the latter event (rule 4b) is given by

$$\sum_{d\epsilon D} \sum_{i\epsilon I} \sum_{k\epsilon D_{id}} \frac{1}{2} a q_d q_k \left(1 - \frac{P_d}{p_d}\right) \frac{P_i - p_i}{S}$$

$$\left(1 - \frac{Q_k(1-q_i-q_d)}{q_k(1-Q_i)}\right) \left(1 - \frac{Q_d}{Q_d + (1-Q_i)S_{id}}\right)$$

$$= \sum_{d\epsilon D} \sum_{i\epsilon I} \frac{1}{2} a q_d \left(1 - \frac{P_d}{p_d}\right) \frac{P_i - p_i}{S} (1-q_i-q_d) S_{id}$$

$$= \sum_{d\epsilon D} \sum_{i\epsilon I} (p_d - P_d)(P_i - p_i) \frac{S_{id}}{S} - \sum_{d\epsilon D} \sum_{i\epsilon I} \frac{1}{2} a q_d q_i$$

$$\left(1 - \frac{P_d}{p_d}\right) \frac{P_i - p_i}{S} S_{id} \qquad (34)$$

the second line following from lemma 1.

Since the expected number of rejections is given by

$$(31)+(32)-(33)+(34)$$

this completes the proof of (29).

In order to prove (30), we note that I_{id} is a subset of I_i, consequently

$$S_{id} \leq S_i$$

hence

$$\sum_{i\epsilon I} \sum_{d\epsilon D} (P_i - p_i)(p_d - P_d) \frac{S_{id}}{S} \leq$$

$$\leq \sum_{i\epsilon I} \sum_{d\epsilon D} (P_i - p_i) \frac{p_d - P_d}{S} S_i$$

$$= \sum_{i\epsilon I} (P_i - p_i) S_i \qquad (35)$$

Omitting the last two terms in (29) and substituting the right hand side of (35) for the fourth term in (29) proves (30).

An alternative procedure

An alternative procedure, also satisfying objectives I) and III) of the introduction is a direct application of the Keyfitz procedure to the entire sample (instead of the individual units). The procedure can be briefly described as follows:

Let π_{ij} be the probabilities of the original samples, Π_{ij} the desired new probabilities given by (11) in terms of A, Q_i, Q_j. Define the sets of distinct samples (ordered pairs of integers) as follows:

$$K = \{(i,j):\ i<j \text{ and } \pi_{ij} \leq \Pi_{ij}\}$$

$$L = \{(i,j):\ i<j \text{ and } \pi_{ij} > \Pi_{ij}\}$$

Let

$$T = \sum_{(i,j)\epsilon K} (\Pi_{ij} - \pi_{ij}) = \sum_{(i,j)\epsilon L} (\pi_{ij} - \Pi_{ij})$$

If the units of the old sample are in K, retain them. If they are in L retain them with probability Π_{ij}/π_{ij}. If the old sample is rejected select a new sample from K with probabilities

$$\frac{\Pi_{ij} - \pi_{ij}}{T}$$

Keyfitz's proof concerning his procedure applies here without change to show that this procedure satisfies objective III) and hence also I). It also follows from the Keyfitz proof that this procedure maximizes the probability of maintaining the entire old sample (i.e. it minimizes the probability of at least one rejection). In case of a rejection, however, the entire old sample is rejected. The procedure is not optimal, since it permits the re-selection of a rejected unit (if a sample is rejected and a new sample is selected, one of the units of the new sample may be identical to one of the units of the old).

Numerical examples show the relative performance of the two procedures: the alternative procedure results in 25 to 40 per cent (in one example 100 per cent) more rejections than the procedure of theorem 1. It is also easier to apply (although the rules appear to be more complicated) since under that procedure in

any given concrete case at most 2N comparisons need to be made between old and new probabilities while the alternative procedure requires N(N-1)/2 such comparisons.

Numerical examples

In examples 1 to 4 the same "original measures of size" are used.

In example 1 the "new measures of size" are obtained by reversing the ordering of the original measures of size: the unit with the smallest original size measure becomes the largest, etc. This represents a drastic change in the original size distribution. In example 2 the new size measures are obtained by interchanging consecutive pairs of the original size measures. This represents "intermittent moderate growth". In example 3 the fourth and fifth original measures of size are interchanged to obtain the new measures of size. This represents "some small growth and some small decline among the larger units". In example 4 the new measures are identical to the old except that the largest measure of size was doubled. This represents "large growth of a few units".

Examples 5 to 8 are analogous to examples 1 to 4 except that the population is larger (N=20) and there is a substantially higher variation in the original size measures.

The expected number of rejections, as in (29), and its upper bound, as in (30), is shown. Also the expected number of rejections under the alternative procedure of the previous section is shown.

Examples 1 to 4

i	m_i	M_i Ex. 1	Ex. 2	Ex. 3	Ex. 4
1	10	22	14	10	10
2	14	19	10	14	14
3	17	18	18	17	17
4	18	17	17	19	18
5	19	14	22	18	19
6	22	10	19	22	44
Expected No. of rejections		0.3759	0.1671	0.0239	0.3143
Upper bound of Expected No. of rejections		0.4000	0.1752	0.0245	0.3237
Alternative Expected No. of rejections		0.4998	0.1999	0.0342	0.4455

Examples 5 to 8

i	m_i	M_i Ex. 5	Ex. 6	Ex. 7	Ex. 8
1	5	95	10	5	5
2	10	90	5	10	10
3	15	85	20	15	15
4	20	80	15	20	20
5	25	75	30	25	25
6	30	70	25	30	30
7	35	65	40	35	35
8	40	60	35	40	40
9	45	55	50	45	45
10	50	50	45	50	50
11	50	50	55	50	50
12	55	45	50	55	55
13	60	40	65	60	60
14	65	35	60	65	65
15	70	30	75	70	70
16	75	25	70	80	170
17	80	20	85	75	80
18	85	15	80	85	135
19	90	10	95	90	90
20	95	5	90	95	200
Expected No. of rejections		0.9039	0.1012	0.0108	0.3066
Upper bound of Expected No. of rejections		0.9194	0.1027	0.0109	0.3109
Alternative Expected No. of rejections		1.2170	0.1252	0.0193	0.4051

In seven of the eight examples the alternative procedure yields between 20 to 45 per cent more rejections than the procedure of theorem 1. In example 7 the alternative procedure yields almost 75 per cent more rejections. In every case the upper bound for the expected number of rejections provided satisfactory approximations.

References

[1] Brewer, K.R.W.: "A note on Fellegi's method of sampling without replacement with probability proportional to size". To be published in the Journal of the American Statistical Association.

[2] Fellegi, I.P.: "Sampling with varying probabilities without replacement: rotating and non-rotating samples". Journal of the American Statistical Association, 58 (1963), 183-201.

[3] Keyfitz, N.: "Sampling with probabilities proportional to size: adjustment for changes in probabilities". Journal of the American Statistical Association, 46 (1951), 105-109.