QUOTA FULFILMENT AND ERRORS IN STRATIFICATION

Norman L. Johnson
Department of Statistics
University of North Carolina
at Chapel Hill
Chapel Hill, NC 27599-3260, USA

Samuel Kotz
Department of Management
Science & Statistics
University of Maryland
College Park, MD 20742, USA

Xizhi Wu
Department of Statistics
University of North Carolina
at Chapel Hill

SUMMARY

It is desired to draw a random sample containing specified numbers of individuals from each stratum of a population. First a random sample of size N is chosen from the whole populations and the stratum of each individual ascertained; then any shortfall is made up by selecting individuals with known stratum affiliation. Optimal values of N are sought, allowing for cost structure and also the possibility of error in ascertaining the strata to which individuals in the first sample belong.

Some Key Words: Optimization; Quota; Stratified Sampling

QUOTA FULFILMENT AND ERRORS IN STRATIFICATION

Norman L. Johnson
Univ. of North Carolina
at Chapel Hill

Samuel Kotz
Univ. of Maryland,
College Park

Xizhi Wu
Univ. of North Carolina
at Chapel Hill

1. Introduction.

Johnson (1957, 1963) considered a problem arising when it is desired to obtain a random sample containing specified numbers $n_1, n_2, \ldots, n_k$ of individuals from strata $\pi_1, \pi_2, \ldots, \pi_k$ respectively by

(i) first taking a random sample of size N from the whole population, and subsequently determining to which each chosen individual belongs, and

(ii) if the number - M_i, say - of individuals chosen from π_i is less than n_i, making good the shortfall by random selection of a further $(n_i - M_i)$ individuals from a set known to belong to π_i (for $i=1,2,\ldots,k$).

The problem was to find an optimal value for N, the size of the first sample. It was supposed that the cost of obtaining this sample is $(a + cN)$, and that the cost of obtaining an individual from π_i in the second step is c_i'. It is to be expected that c_i' exceeds c, at least for some values of i; otherwise the optimum value of N would be zero. Any excess individuals from π_i (if $M_i > n_i$) were supposed to have value v_i' each, though this value might be zero. Clearly,

realistic values of v_i' must be less than c, at least for some i, otherwise one could gain by taking N as large as possible. In Johnson (1957), sampling from an effectively infinite population was discussed; in Johnson (1963), finite population size was allowed for.

2. The Problem.

In the present paper, we will consider only the case of an effectively infinite population, with proportions $p_1, p_2, \ldots, p_k$ of its members in $\pi_1, \pi_2, \ldots, \pi_k$ respectively ($p_1 + p_2 + \ldots + p_k = 1$). The essential new feature is that in step (i) it will not be assumed that determination of the stratum to which an individual belongs is achieved without error.

This extension is closely related to problems in faulty inspection sampling which we have studied over the last few years (Johnson & Kotz (1985)). It also reflects aspects of currently prevailing models of stratified sampling, which attempt to be more realistic then earlier ones.

We will denote the probability that an individual, actually belonging to π_i, is assigned to π_j by P_{ji}. The number of individuals, in the first sample of size N, actually belonging to π_i will be denoted by Y_i; the number ascribed to π_i will be denoted by M_i (as above). We have

$$\sum_{i=1}^{k} M_i = \sum_{i=1}^{k} Y_i = N.$$

Assuming that determinations of strata of different individuals are mutally independent,

Y_i is distributed binomially with parameters (N, p_i); M_i is distributed binomally with parameters (N, ω_i) where $\omega_i = \sum_{j=1}^{k} p_j P_{i|j}$.

Given Y_i, M_i is distributed as the convolution of binomial distributions with parameters $(Y_i, P_{i|i})$ and $(N-Y_i, \omega_i')$, where

$$\omega_i' = (1-p_i)^{-1} \sum_{j\neq i}^{k} p_j P_{i|j} = (1-p_i)^{-1}(\omega_i - p_i P_{i|i}).$$

It will be assumed, here, that no error occurs in the selection of individuals in step (ii) - that is, when (n_i-M_i) individuals are chosen 'from π_i', they really do belong to π_i. (Allowance for the possibility of errors in selection can be made in a straightforward manner, though it leads to greater complexity in the formulae.)

3. Costs.

In general, $M_i \neq Y_i$, so the final sample may be deficient in numbers for some strata and in excess for others. We introduce the following symbols to represent the cost of the sample:

v_i - a 'penalty' for each individual lacking from π_i.

v_i' - the value (if any) of each individual from π_i in excess of the required number, n_i.

So we have, with $z^+ = \begin{cases} z & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$:

COMPONENT	COST
Obtaining First Sample	$a + cN$
Obtaining Second Sample	$\sum_{i=1}^{k} c_i'(n_i-M_i)^+$
Penalty	$\sum_{i=1}^{k} v_i\{n_i-Y_i-(n_i-M_i)^+\}^+$

Value of excess $$-\sum_{i=1}^{k} v_i'\{Y_i - n_i + (n_i - M_i)^+\}^+$$

The expected total cost is

$$C_N = a + cN + \sum_{i=1}^{k} \{v_i E[n_i - Y_i \mid Y_i < n_i] \Pr[Y_i < n_i] - v_i' E[Y_i - n_i \mid Y_i \geq n_i] \Pr[Y_i \geq n_i]\}$$
$$+ \sum_{i=1}^{k} c_i' \, E[n_i - M_i \mid M_i < n_i] \, \Pr[M_i < n_i]$$
$$- \sum_{i=1}^{k} [\{v_i \, E[n_i - Y_i \mid M_i \leq Y_i \leq n_i] + v_i' \, E[Y_i - M_i \mid M_i \leq Y_i \leq n_i]\} \, \Pr[M_i \leq Y_i \leq n_i]$$
$$+ v_i E[n_i - M_i \mid Y_i < M_i < n_i] \Pr[Y_i < M_i < n_i] + v_i' E[n_i - M_i \mid M_i < n_i \leq Y_i] \Pr[M_i < n_i \leq Y_i]]. \tag{1}$$

On the right-hand side of (1), the first line is the expected cost associated with the first sample of size N, allowing for shortfall penalties and value of excess individuals; the second line is the expected cost of sampling in step (ii); the third and fourth lines represent the savings from the expected value of the extra individuals chosen in step (ii) to make up shortfalls.

Direct minimization of C_N with respect to N is a formidable task, even with the introduction of some approximations. As in Johnson (1957), we will approach the problem by considering the change in cost,

$$\Delta C_N = C_{N+1} - C_N,$$

if the size of the first sample is increased from N to (N+1).

The immediate increase in sampling cost is c. If the additional individual is from π_i <u>and</u> is classified as belonging to π_j (probability, $p_i \, P_{j|i}$), this extra cost is offset by (a) the value of the extra individual, which is v_i if $Y_i < n_i$, v_i' if $Y_i \geq n_i$, and (b) if $M_j < n_j$, the saving (c_j') arising from reduction (by 1) in the number of

individuals to be chosen from π_j in step (ii), less the value of this individual if it had been chosen (immediately after the first sample), which would be v_j if $Y_j < n_j$, v_j' if $Y_j \geq n_j$. Hence

$$\begin{aligned}\Delta C_N &= c - \sum_{i=1}^{k} p_i\{v_i \Pr[Y_i < n_i] + v_i' \Pr[Y_i \geq n_i]\} \\ &\quad - \sum_{i=1}^{k} p_i[P_{j|i}\{c_j' \Pr[M_j < n_j] - v_j \Pr[(M_j < n_j) \cap (Y_j < n_j)] \\ &\quad - v_j' \Pr[(M_j < n_j) \cap (Y_i \geq n_j)]\}] \qquad (2) \\ &= c - \sum_{i=1}^{k} p_i\{v_i' + (v_i - v_i') \Pr[Y_i < n_i]\} \\ &\quad - \sum_{i=1}^{k} p_i \left[\sum_{j=1}^{k} P_{j|i}\{(c_j' - v_j') \Pr[M_j < n_j] \right. \\ &\quad \left. -(v_j - v_j') \Pr[(M_j < n_j) \cap (Y_j < n_j]\} \right], \qquad (2)'\end{aligned}$$

with $$\Pr[Y_i < n_i] = \sum_{y=0}^{n_i-1} \binom{N}{y} p_i^y (1-p_i)^{N-y}$$

$$\Pr[M_i < n_i] = \sum_{m=0}^{n_i-1} \binom{N}{m} \omega_i^m (1-\omega_i)^{N-m}$$

and $$\Pr[(M_i < n_i) \cap (Y_i < n_i)] = \sum_{y=0}^{n_i-1} \Pr[(M_i < n_i) \cap (Y_i = y)]$$

$$= \sum_{y=0}^{n_i-1} \binom{N}{y} p_i^y (1-p_i)^{N-y} \sum\sum_{u+v<n_i} \binom{y}{u}$$

$$p_{i|i}^u (1-P_{i|i})^{y-u} \binom{N-y}{v} \omega_i'^v (1-\omega_i')^{N-y-v}.$$

Under the conditions $c_i' > (v_i, c) > v_i'$ (for all i), ΔC_N increases as N increases, and $\Delta C_0 < 0$. The optimal N is then the integer part of the solution of the equation $\Delta C_N = 0$. This solution depends on the values of c, $\{c_i'\}$, $\{v_i\}$ and $\{v_i'\}$ only through the ratios

$$\{c-v_i'\} : \{c_i'-v_i'\} : \{v_i-v_i'\}.$$

4. Special Cases

We now consider a succession of special cases. If $v_i = v$, $v_i' = v'$ and $c_i' = c'$ for all $i=1,2,\ldots,k$, then

$$\Delta C_N = c-v'-(c'-v') \sum_{i=1}^{k} p_i \sum_{j=1}^{k} P_{j|i} \Pr[M_j < n_j]$$
$$- (v-v')\left[\sum_{i=1}^{k} p_i\{\Pr[Y_i < n_i] - \sum_{j=1}^{k} P_{j|i} \Pr[(M_j<n_j) \cap (Y_j<n_j)]\}\right]. \tag{3}$$

The solution of $\Delta C_N = 0$ now depends only on the ratios

$$c-v' : c'-v' : v-v'$$

If we suppose, further, that $P_{i|i} = P$ and $P_{j|i} = (1-p_i)^{-1} p_j(1-P)$ for all i and all $j \neq i$ $(i=1,2,\ldots,k)$, so that P is the probability of correct classification, and incorrectly assigned individuals are ascribed to other strata with probabilities proportional to stratum sizes, then

$$\omega_i = p_i\{P + (1-P) \sum_{j\neq i} p_j(1-p_j)^{-1}\} \tag{4.1}$$

and

$$\omega_i' = (1-p_i)^{-1} (1-P)\, p_i \sum_{j\neq i} p_j(1-p_j)^{-1}. \tag{4.2}$$

In the completely symmetric case when, in addition to all the above-specific conditions, we have $p_i = k^{-1}$ $(i=1,2,\ldots,k)$ and $n_1 = n_2 = \ldots = n_k = N/k = n$, say, the variables M_i and Y_i have the same binomial distribution, with parameters (N,k^{-1}). They are not, of course, independent. Generally, the correlation between Y_i and M_j is

$$p_i(P_{j|i} - \sum_{h=1}^{k} p_h P_{j|h}) \left[p_i(1-p_i)(\sum_{h=1}^{k} p_h P_{j|h})(1- \sum_{h=1}^{k} p_h P_{j|h}) \right]^{-\frac{1}{2}}.$$

In the completely symmetric case, the correlation between Y_i and M_i is $(k-1)^{-1}(k\,P-1)$.) In this case

$$\omega'_j = P_{j|i} = (k-1)^{-1}(1-P) \quad \text{for } j \neq i; \tag{5.1}$$

$$\Pr[Y_i < n] = \Pr[M_i < n] \equiv P(n) = k^{-N} \sum_{y=0}^{n-1} \binom{N}{y}(k-1)^{N-y}$$

$$\text{for all } i \text{ and } j; \tag{5.2}$$

$$\Pr[(Y_i<n) \cap (M_i<n)] \equiv P^*(n)$$

$$= k^{-N} \sum_{y=0}^{n-1} \binom{N}{y} \sum\sum_{u+v<n} \binom{y}{u}\binom{N-y}{v} P^u (1-P)^{y-u+v} (k-2+P)^{N-y-v}$$

$$\text{for all } i; \tag{5.3}$$

and $$\Delta C_N = c-v' - (c'-v')\,P(n) - (v-v')\{P(n)-P^*(n)\}. \tag{5.4}$$

In this completely symmetric case, P appears only in $P^*(n)$ and so one might expect that the optimal value of N would not depend much on P, unless v-v' is large, relative to c-v' and c'-v'. Table 1 supports this conjecture to a remarkable extent. Indeed, so weak is the dependence on P that it would seem reasonable to use the optimal values corresponding to P=1 (errorless inspection) except, perhaps for values of P so small as to be very unlikely.

Of course the minimized value of <u>expected cost</u> will depend substantially on P, even though the optimal value of N does not.

(Note: Two FORTRAN programs were prepared for calculations of optimal values of N. One is for a personal computer, and the other - a faster one - is suitable for a main frame. These programs are available on request from the third author. Computation time for $n < 20$ is negligible, but for $n \geq 50$ it is quite substantial. A bivariate normal approximation to the joint distribution of Y_i and M_i might be used to evaluate $P^*(n)$, but it should be noted that although the regressions are linear, variation about the regression line is not homoscedastic.)

REFERENCES

Johnson, N.L. (1957) Optimal sampling for quota fulfilment, Biometrika, 44, 518-523.

Johnson, N.L. (1963) Quota fulfilment in finite populations, In Classical and Contagious Discrete Distributions, (G.P. Patil, ed.) Statistical Publ. Soc., Calcutta, India and Pergamon Press, pp. 419-426.

Johnson, N.L. and Kotz, S. (1985) Some distributions arising as a consequence of errors in inspection, Naval Res. Logist. Qtly., 32, 35-43.

TABLE 1: OPTIMAL VALUES OF N

		$c \cdot v' : c' - v' : v - v' = 1 : 3 : \frac{1}{2}$		$1 : 3 : 2$		$1 : 5 : \frac{1}{2}$		$1 : 5 : 2$	
k	n	P=0.85	0.70	0.85	0.70	0.85	0.70	0.85	0.70
2	25	53	53	54	55	56	56	57	57
2	50	104	105	106	107	109	109	110	110
2	100	206	207	209	210	212	213	213	214
3	17	55	55	56	57	59	60	60	61
3	33	105	105	107	108	111	111	112	113
3	67	210	210	213	214	218	218	220	221
4	13	57	57	59	59	62	63	63	64
4	25	107	108	109	111	115	115	116	117
4	50	211	211	214	215	221	221	223	224
5	10	55	56	57	58	62	62	63	63
5	20	108	109	111	112	117	117	118	119
5	40	212	213	216	217	224	224	226	227
5	100	520	521	525	528	538	539	542	543
6	8	54	54	56	57	61	61	62	63
6	17	111	111	114	115	121	121	123	124
6	33	211	212	215	217	225	225	227	228
6	83	520	521	526	529	541	541	544	546
7	7	55	55	57	58	63	63	64	65
7	14	108	108	110	112	118	118	120	121
7	29	218	218	222	224	232	233	235	236
7	71	521	522	527	531	544	544	548	550
8	6	54	55	57	58	63	63	64	65
8	13	115	115	118	120	126	127	128	130
8	25	215	216	220	222	232	232	234	236
8	63	530	531	537	541	555	555	559	561
9	6	61	62	64	65	71	71	72	73
9	11	110	110	113	115	122	123	124	126
9	22	214	215	219	222	231	232	234	236
9	56	531	533	539	543	558	559	563	565
10	5	57	58	60	61	67	67	69	70
10	10	111	112	115	117	125	125	127	128
10	20	217	218	222	225	236	236	239	240
10	50	529	530	537	541	557	558	562	564

$$N = \min\{n : \Delta C_n > 0\}$$