UNBIASED ESTIMATION OF PROPORTIONS UNDER SEQUENTIAL SAMPLING

M.D. Bankier[1]

Under a sequential sampling plan, the proportion defective in the sample is generally a biased estimator of the population value. In this paper, an unbiased estimator is given. Also, an unbiased estimator of its variance is derived. These results are applied to an estimation problem from the 1976 Canadian Census.

1. INTRODUCTION

In a quality control (Q.C.) operation, a sample of m units from some lot is examined. The number x of defective units in the sample determines if the lot is accepted or rejected. Frequently, a sequential sampling plan is used. A definition of such plans is given in Section 2. For these sampling plans, the number of units sampled is a random variable and $\frac{x}{m}$ is usually a biased estimator of the proportion defective in the lot. In Section 3, an estimator $p(m,x)$ is presented which is unbiased under both sequential sampling with replacement (w.r.) and without replacement (w.o.r.). This estimator was first proposed by Girshick, Mosteller, and Savage [3] under sampling w.r. A result given by DeGroot [1] is also discussed. It states for which sampling plans $p(m,x)$ is the uniformly minimum variance unbiased (UMVU) estimator under sampling w.r. from an infinite population. In Section 4, unbiased estimators of $V(p(m,x))$ are derived under both sampling w.r. and w.o.r. These estimators are constructed using an approach taken under sampling w.r. in [3]. In Section 5, the results are applied to an estimation problem from the 1976 Canadian Census.

---

[1]M.D. Bankier, Census and Household Survey Methods Division, Statistics Canada.

2. DEFINITION OF A SAMPLING PLAN

Let (m,x) represent the event that m units have been sampled and of these, x are defective. A sampling plan is defined as a function S such that if

$$S((m,x)) = 1 \quad \text{then another unit is sampled and inspected.} \tag{2.1}$$

If

$$S((m,x)) = 0 \quad \text{then no more units are sampled.} \tag{2.2}$$

The event (m,x) is called a boundary point if $S((m,x)) = 0$ and $Pr((m,x)) > 0$. The set of all boundary points of a sampling plan will be labelled B. This set contains the possible outcomes of the sampling plan. A sampling plan is said to be closed if

$$\sum_{(m,x)\in B} Pr((m,x)) = 1 \tag{2.3}$$

A bounded sampling plan is one where there exists a positive integer b such that

$$Pr(m \leq b) = 1 \tag{2.4}$$

The size of a sampling plan is the smallest b for which (2.4) holds. Only closed and bounded sampling plans will be considered in this paper. Single, double, multiple and sequential bounded sampling plans used in quality control all satisfy the above criteria.

3. AN UNBIASED ESTIMATOR OF THE PROPORTION DEFECTIVE

Let

$$x_i = \begin{cases} 1 \text{ if the } i\text{th unit sampled is defective} \\ 0 \text{ otherwise.} \end{cases} \tag{3.1}$$

A path to (m,x) will be defined as a sequence of points

$$\left(j, \sum_{i=1}^{j} x_i\right) \qquad \begin{matrix} j = 1, \ldots, m \\ \sum_{i=1}^{m} x_i = x \end{matrix} \tag{3.2}$$

such that

$$s\left(\left(j, \sum_{i=1}^{j} x_i\right)\right) = 1 \text{ for } j = 1, \ldots, m-1. \tag{3.3}$$

Assume the units are sampled w.r. with P the proportion defective. Under these assumptions, it can be seen that

$$Pr((m,x)) = c(m,x)P^x(1-P)^{m-x} \tag{3.4}$$

where

$$c(m,x) = \text{the number of distinct paths to } (m,x). \tag{3.5}$$

Let

$$\bar{y}_i = \frac{\text{number of defective units in first } i \text{ sampled}}{i} \tag{3.6}$$

The sample mean $\bar{y}_m = \frac{x}{m}$ is generally a biased estimator of P except when the sample size m is fixed. However, because at least one unit is always sampled, $\bar{y}_1$ is unbiased. The estimator

$$p(m,x) = E(\bar{y}_1 | (m,x))$$

$$= \frac{Pr(\bar{y}_1 = 1 \text{ and } (m,x))}{Pr((m,x))}$$

$$= \frac{d(m,x)P^x(1-P)^{m-x}}{c(m,x)P^x(1-P)^{m-x}}$$

$$= \frac{d(m,x)}{c(m,x)} \tag{3.7}$$

is also an unbiased estimator of P with the same or a smaller variance than $\bar{y}_1$ by the Rao-Blackwell Theorem. In the above expression

$$d(m,x) = \text{the number of distinct paths to } (m,x) \text{ that pass through the point } (1,1) \tag{3.8}$$

Under sampling w.o.r. from a finite population, expression (3.7) still holds. The $P^x(1-P)^{m-x}$ is replaced by

$$G(m,x;M,X) = \begin{cases} \dfrac{\binom{X}{x}\binom{M-X}{m-x}}{\binom{m}{x}\binom{M}{m}} & \text{for } m \leq M,\ x \leq X,\ X \leq M,\ x \leq m \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

where M is the number of units in the population and X is the number of defectives. In this case, P = X/M.

A sampling plan is complete if the only estimator f such that E(f) = 0 for all Γ, is one defined by f(m,x) = 0 for all $(m,x) \in B$. In DeGroot [1], it is shown that under sampling w.r. a sampling plan of size b is complete if and only if the boundary B contains exactly b+1 points. If the sampling plan is complete and the population size is infinite, then by the Lehmann-Scheffé Uniqueness Theorem, (Roussas [4], p.216), p(m,x) is the UMVU estimator of P.

4. AN UNBIASED ESTIMATOR OF V(p(m,x))

For any closed, complete and bounded sampling plan with size $b \geq 2$, it is possible to construct an unbiased estimator $v(p(m,x))$ of $V(p(m,x))$ under both sampling w.r. and w.o.r. Expressions for $v(p(m,x))$ under sampling w.o.r. from a finite population are given below for three types of boundary point sets B. These expressions reduce to those for sampling w.r. if $\frac{M-1}{M}$ is replaced by 1 and $\frac{1}{M}$ is replaced by 0.

Case (1): It is assumed that $(2,2) \varepsilon B$. Then

$$v(p(m,x)) = \begin{cases} 0 \text{ for } (m,x) = (2,2) \\ p(m,x)(p(m,x) - \frac{1}{M}) \text{ otherwise for } (m,x) \varepsilon B. \end{cases} \tag{4.1}$$

Case (2): It is assumed that $(2,2) \notin B$ but $(1,1) \varepsilon B$. Then

$$v(p(m,x)) = \begin{cases} 0 \text{ for } (m,x) = (1,1) \\ \frac{M-1}{M} \frac{e(m,x)}{c(m,x)} \text{ otherwise for } (m,x) \varepsilon B \end{cases} \tag{4.2}$$

where $e(m,x)$ = the number of distinct paths going through (1,1) and (m,x) under the sampling plan with boundary point set B^*. B^* is identical to B except the point (1,1) has been removed and the point (2,2) has been added. (4.3)

Case (3): It is assumed that $(2,2) \notin B$ and $(1,1) \notin B$. Then

$$v(p(m,x)) = p(m,x)(p(m,x) - \frac{1}{M}) - \frac{M-1}{M} \frac{f(m,x)}{c(m,x)} \tag{4.4}$$

where $f(m,x)$ = the number of distinct paths going through (2,2) and (m,x) under the sampling plan with boundary point set B. (4.5)

Proof: This is given for case I only. The proofs for the other two cases are similar.

$$V(p(m,x)) = E(p^2(m,x))-P^2$$

$$= \sum_{(m,x)\in B} p^2(m,x)Pr((m,x))-P^2 \qquad (4.6)$$

It can be seen that $d(2,2) = 1$ and $c(2,2) = 1$. Thus, under sampling with replacement

$$p^2(2,2)Pr((2,2))-P^2 = P^2 - P^2 = 0. \qquad (4.7)$$

Thus

$$V(p(m,x)) = \sum_{\substack{(m,x)\in B \\ (m,x) \neq (2,2)}} p^2(m,x)Pr((m,x)) \ . \qquad (4.8)$$

Therefore

$$v(p(m,x)) = \begin{cases} 0 \text{ for } (m,x) = (2,2) \\ p^2(m,x) \text{ otherwise for } (m,x)\in B \end{cases} \qquad (4.9)$$

is an unbiased estimator of $V(p(m,x))$ under sampling w.r. Under sampling w.o.r., eq. (4.7) does not reduce to 0. In fact

$$p^2(2,2)Pr((2,2))-P^2$$

$$= G(2,2;M,X)-P^2$$

$$= -\frac{P(1-P)}{M-1} \ . \qquad (4.10)$$

This implies that the estimator (4.9) under sampling w.o.r. has a bias of $\frac{P(1-P)}{M-1}$. However if $(2,1)\varepsilon B$ then

$$q(m,x) = \begin{cases} 0 \text{ for } (m,x) = (2,2) \\ \\ \frac{d(m,x)}{c(m,x)} \quad \text{otherwise for } (m,x)\varepsilon B \end{cases} \tag{4.11}$$

is an unbiased estimator of $P(1-P)\ \frac{M}{M-1}$ since

$$d(m,x) = \begin{cases} 1 \text{ for } (m,x) = (2,1) \text{ or } (2,2) \\ \\ 0 \text{ otherwise for } (m,x)\varepsilon B \end{cases} \tag{4.12}$$

and

$$G(2,1;M,X) = P(1-P)\ \frac{M}{M-1}\ . \tag{4.13}$$

If $(2,1)\notin B$, then let B' be the set of boundary points B with (2,1) added. Also let

$$k(m,x) = \text{the number of distinct paths to } (m,x) \text{ under sampling plan with boundary point set } B' \tag{4.14}$$

and

$$g(m,x) = \begin{cases} 0 \text{ if } (m,x) = (2,1) \\ \\ \text{the number of paths going through } (2,1) \text{ and } (m,x) \text{ under the sampling plan with boundary point set B if } (m,x) \neq (2,1). \end{cases} \tag{4.15}$$

It is obvious that

$$k(m,x) = c(m,x) - g(m,x). \tag{4.16}$$

Also,

$$\sum_{(m,x)\in B'} k(m,x)\, G(m,x;M,X) = 1 \quad , \tag{4.17}$$

$$k(2,1)G(2,1;M,X) + \sum_{(m,x)\in B} k(m,x)\, G(m,x;M,X) = 1 \ , \tag{4.18}$$

$$k(2,1)G(2,1;M,X) + \sum_{(m,x)\in B} c(m,x)\, G(m,x;M,X)$$

$$- \sum_{(m,x)\in B} g(m,x)\, G(m,x;M,X) = 1 \tag{4.19}$$

and

$$\sum_{(m,x)\in B} g(m,x)\, G(m,x;M,X) = k(2,1)\, G(2,1;M,X) \ . \tag{4.20}$$

Now if $(1,0)\in B$, then $k(2,1) = 1$ and

$$g(m,x) = \begin{cases} 0 \text{ for } (m,x) = (2,2) \\ \\ d(m,x) \text{ otherwise for } (m,x)\in B. \end{cases} \tag{4.21}$$

If $(1,0)\notin B$ then $k(2,1) = 2$ and

$$g(m,x) = \begin{cases} 0 \text{ for } (m,x) = (2,2) \\ \\ 2d(m,x) \text{ otherwise for } (m,x)\in B. \end{cases} \tag{4.22}$$

Thus from eq. (4.21) and (4.22), eq. (4.20) can be rewritten

$$\sum_{\substack{(m,x)\in B \\ (m,x)\neq(2,2)}} d(m,x)\, G(m,x;M,X) = G(2,1;M,X) \ . \tag{4.23}$$

Thus $q(m,x)$ given in eq. (4.11) is also an unbiased estimator of $G(2,1;M,X)$ if $(2,1)\notin B$. Therefore, an unbiased estimator of $v(p(m,x))$ under sampling w.o.r. from a finite population is given by eq. (4.1).

5. ESTIMATING THE PROPORTION OF DEFECTIVE QUESTIONNAIRES IN THE 1976 CENSUS

In the 1976 Canadian Census, Quality Control Technicians were used to examine and reject enumeration areas (EAs) where the data, on a sample basis, was of poor quality. A sample of population and housing questionnaires (Forms 2B) were sampled w.o.r. one at a time. If a sampled Form 2B did not meet certain quality standards, it was rejected. After examining a Form 2B, the technician looked at Table 1 to determine whether to accept the EA, reject the EA or continue sampling Forms 2B. For example, after 17 households had been examined, the EA was accepted, rejected or sampling continued if 2, 4 or 3 households respectively had been rejected. A minimum of two Forms 2B and a maximum of 24 Forms 2B were sampled. A sequential sampling plan requires a smaller sample on the average than a fixed sample size plan to distinguish between good and bad quality EAs with reasonable accuracy (see Duncan [2], p. 178).

For future planning purposes, an estimate for Canada was needed of the proportion of defective Forms 2B (those that did not meet the quality standards) at the beginning of the Q.C. operation. A stratified sample of EAs was picked with proportional allocation. The EAs were stratified by regional office and EA methodology. Information recorded from the QC forms included the number of Forms 2B sampled and the number rejected. The proportion of defective Forms 2B on the Canadian level was estimated by

$$\hat{p} = \frac{\sum_i \frac{N_i}{n_i} \sum_{j=1}^{n_i} M_{ij} p_{ij}}{\hat{M}} \tag{5.1}$$

where

N_i = the number of EAs in the ith stratum, (5.2)

n_i = the number of EAs in the ith stratum sample, (5.3)

M_{ij} = the number of Forms 2B in the jth EA sampled from the ith stratum, (5.4)

p_{ij} = the estimator for the proportion of defective Forms 2B in the jth EA sampled from the ith stratum . (5.5)

Table 1: Decision Table For the 1976 Census Form 2B Sequential Sampling Plan

Number of Forms 2B Sampled	Accept EA if the Following Number of Forms 2B are Rejected.	Reject EA if the Following Number of Forms 2B are Rejected
2	*	2
3	*	2
4	*	2
5	*	2
6	*	2
7	0	3
8	0	3
9	0	3
10	0	3
11	0	3
12	1	4
13	1	4
14	1	4
15	1	4
16	1	4
17	2	4
18	2	5
19	2	5
20	2	5
21	2	5
22	2	5
23	3	5
24	4	5

* indicates that the EA cannot be accepted with this sample size

and

$$\hat{M} = \sum_i \frac{N_i}{n_i} \sum_{j=1}^{n_i} M_{ij}. \qquad (5.6)$$

Estimators of the variance and bias of p are given by

$$v(\hat{p}) = \frac{1}{\hat{M}^2} \sum_i \left(s(r_{ij}, r_{ij}) + \left(\frac{N_i}{n_i}\right)^2 \sum_{j=1}^{n_i} M_{ij}\, v(p_{ij})\right) \qquad (5.7)$$

and

$$b(\hat{p}) = -\frac{1}{\hat{M}^2} \sum_i s(M_{ij}, r_{ij}) \qquad (5.8)$$

where

$$r_{ij} = M_{ij}(p_{ij} - \hat{p}) \qquad (5.9)$$

and

$$s(w_{ij}, v_{ij}) = N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) \frac{1}{n_i - 1} \left(\sum_{j=1}^{n_i} w_{ij} v_{ij} - \frac{1}{n_i} \left(\sum_{j=1}^{n_i} w_{ij}\right)\left(\sum_{j=1}^{n_i} v_{ij}\right)\right). \qquad (5.10)$$

The standard formulae were used to linearize the ratio estimator $\hat{p}$ in $v(p)$ and $b(\hat{p})$. Conditional expectations were applied in the derivations because of the two-stage sampling. Sample quantities were substituted for population values where necessary.

To calculate p_{ij} required finding $c(m,x)$ and $d(m,x)$ in expression (3.7). This was done in Figures 1 and 2. The number in a cell in Figure 1, for example, is the value of $c(m,x)$ for that m and x. Cells with solid lines drawn around them are points where an EA could be accepted or rejected in Table 1. The calculations began in Figure 1 by placing the number 1 in the first column. The number in a cell to the right was calculated by adding together the numbers in any cells to the left with arrows pointing into that

Figure 1: Calculating c(m,x) for the Table 1 Sampling Plan

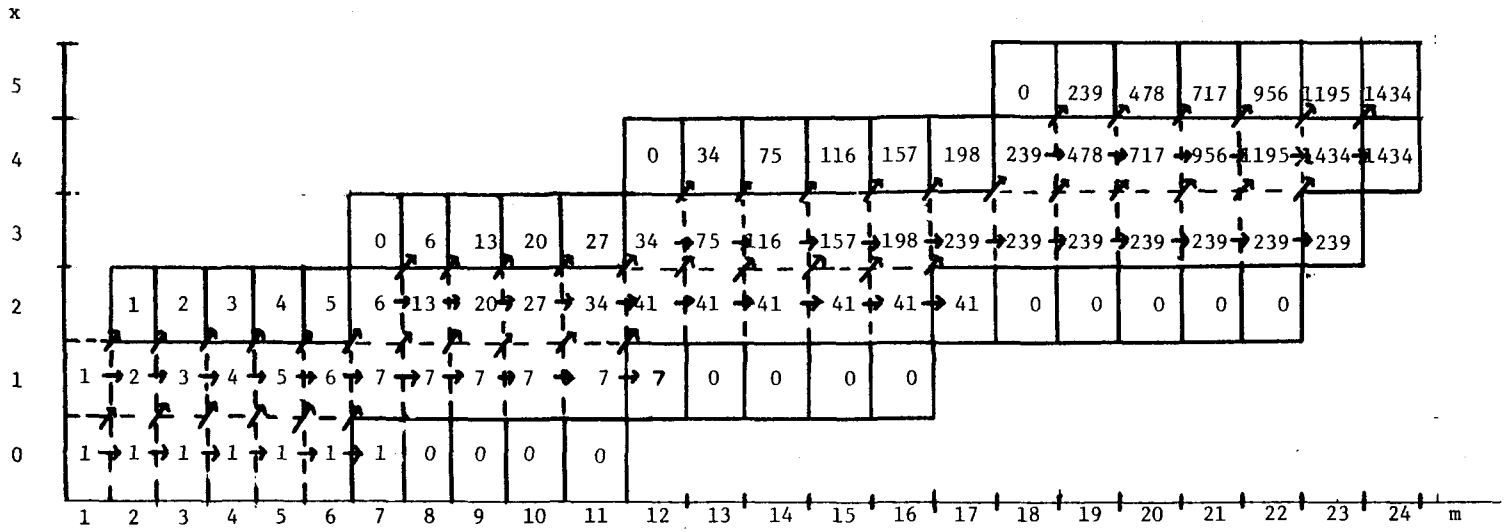
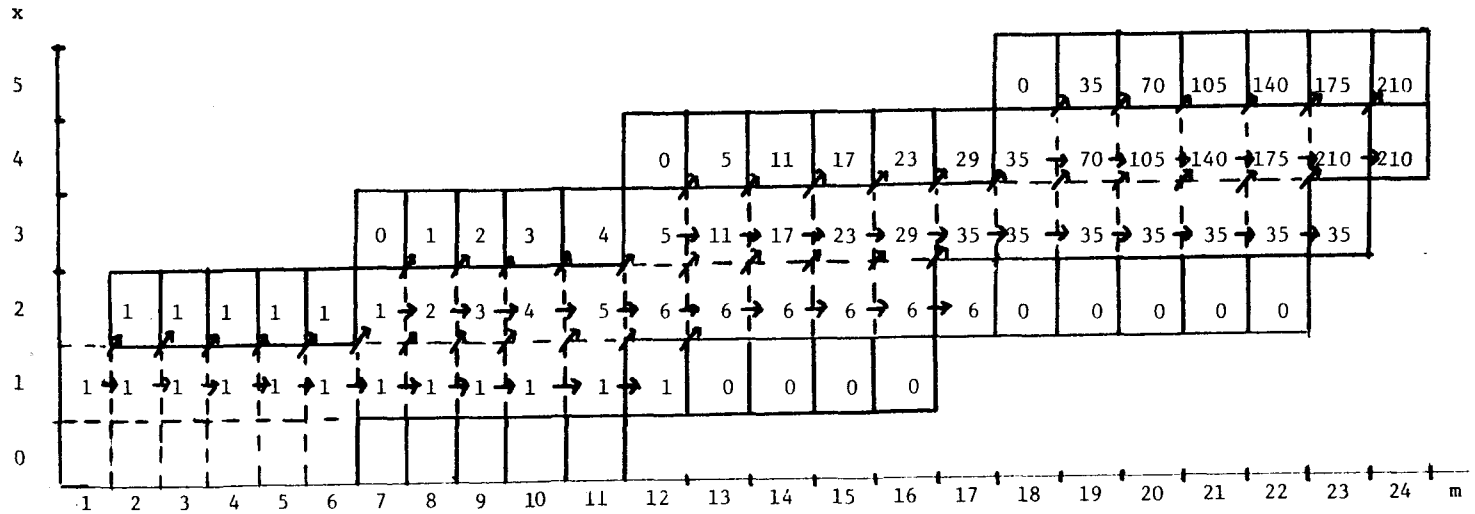


Figure 2: Calculating d(m,x) for the Table 1 Sampling Plan



cell. Figure 1 shows that this sampling plan has 25 boundary points. The point (8,0), for example, is not a boundary point since $Pr((8,0)) = 0$. This indicates (see section 3) that if the questionnaires had been sampled w.r. from an infinite population, then $p(m,x)$ would have been the UMVU estimator of P. The unbiased estimator of $V(p_{ij})$ is given by eq. (4.1).

The results below are based on a sample of 1199 EAs:

$$\hat{p} = 2.63 \times 10^{-2}, \tag{5.11}$$

$$b(\hat{p}) = 2.82 \times 10^{-8}, \tag{5.12}$$

$$\sqrt{v(\hat{p})} = 3.22 \times 10^{-3}, \tag{5.13}$$

$$\frac{\sqrt{v(\hat{p})}}{\hat{p}} \times 100\% = 12.2\%. \tag{5.14}$$

It can be seen that the estimate of the Canadian proportion defective has a reasonably small coefficient of variation and a very small bias.

ACKNOWLEDGEMENTS

I would like to thank my supervisor R. Burgess for his advice and also G. Brackstone for his suggestion that $p(m,x)$ be investigated.

RESUME

Dans un plan de sondage séquentiel, la proportion défectueuse de l'échantillon est en général un estimateur biaisé de la valeur de la population. L'auteur de l'article propose un estimateur sans biais, dont un estimateur sans biais de la variance est également défini. Les résultats sont appliqués à un problème d'estimation tiré du recensement de 1976.

REFERENCES

[1] DeGroot, M.H. (1959), "Unbiased Sequential Estimation for Binomial Populations", Ann. Math. Stat., 30, 80-101.

[2] Duncan, A.J. (1965), "Quality Control and Industrial Statistics", Third Edition, Richard D. Irwin, Inc.

[3] Girshick, M.A., Mosteller, F. and Savage, L.J. (1946), "Unbiased Estimates for Certain Binomial Sampling Problems With Applications", Ann. Math. Stat., 17, 13-23.

[4] Roussas, G.G. (1973), "A First Course in Mathematical Statistics", Addison-Wesley Publishing Company.