Journal of Propagations in Probability and Statistics
Vol. 2, No. 1, August 2001
International ed., pp. 117−122

An inductive proof for a closed form formula in truncated inverse sampling

Kuang-Chao Chang
Fu Jen Catholic University

Abstract

Inverse sampling is a sequential sampling procedure such that sampling is continued until a predetermined number of units possessing certain attribute are included in the sample. This sampling procedure does not have control over the total sample size, in particular, when the attribute under consideration prevails with rare frequency. A remedy for this shortcoming is to truncate the sampling procedure when the total sample size reaches a specified maximum number. We propose a closed form formula to compute the expected total sample size in this truncated version of inverse sampling, and we prove the formula by mathematical induction.

Keywords: Inverse sampling, Truncated inverse sampling, Mathematical induction.

□ Kuang-Chao Chang is Associate Professor, Department of Statistics and Information Science, Fu Jen Catholic University, Taipei, Taiwan, ROC. E-mail: stat1016@mails.fju.edu.tw.

1. Introduction

In sample survey theory and methodology, inverse sampling is often used to estimate the proportion of a certain rare item in the population. In so-called "standard" inverse sampling, samples are taken randomly and sequentially until a specified number of the rare item has been observed (see Haldane 1945 or Cochran 1977). A drawback in such sampling procedure is the lack of control on the final random sample size, which may be very large if the value of the proportion to be estimated is very small. To overcome this drawback, we may consider truncating the inverse sampling procedure when the random total sample size reaches a pre-determined positive integer. This modified version of inverse sampling will be called the truncated inverse sampling (TIS). In this paper, we propose a closed form formula to compute the expected final random total sample size in TIS under the assumption of infinite population, and we prove the formula by the method of mathematical induction. The contents of this article may be used as supplementary teaching material for teachers of probability, statistics, mathematics, and other related fields.

2. A closed form formula in TIS

We begin this section with the following lemma.

Lemma 2.1 Let X be distributed as Binomial(n, p), then

$$\sum_{i=0}^{n-1} \Pr(X \le i) = nq \quad \text{where } q = 1 - p.$$

Proof. Let Y be distributed as Binomial(n, q), then

$$\sum_{i=0}^{n-1} \Pr(X \le i) = \sum_{i=1}^{n} \Pr(Y \ge i) = E(Y) = nq$$

since Y is nonnegative and integer-valued (see Karr 1993, p.114, Corollary 4.24).

Q.E.D.

Next, in the following theorem, we give the closed form formula to compute the expected final random total sample size in TIS, assuming the population is infinite.

Theorem 2.2 Let Z have negative binomial distribution with p.m.f.

$$f(z) = \binom{z-1}{m-1} p^m q^{z-m}, \quad z = m, m+1, \cdots$$

and let $\tilde{Z}$ be the random variable defined by $\tilde{Z} = \min\{Z, M\}$ where M is a positive integer and $M \geqq m$. Then, the expected value of $\tilde{Z}$, denoted by $\tilde{E}(m, M)$, is

$$\tilde{E}(m,M) = \frac{m}{p} - \frac{1}{p}\sum_{i=1}^{m}\sum_{j=0}^{i-1}\binom{M}{j}p^j q^{M-j} \text{ , if } M > m$$

$$= m \text{ , if } M = m.$$

Proof. The case that $M = m$ is trivial. If $M > m$, we prove by induction on m. Let U be the random variable defined by

$$U = \begin{cases} 1, & \text{if the first trial results in a success,} \\ 0, & \text{otherwise.} \end{cases}$$

Then, when $m = 1$, we have

$$\begin{aligned}
\tilde{E}(1,M) &= \tilde{E}(1, M \mid U=0)\Pr(U=0) + \tilde{E}(1, M \mid U=1)\Pr(U=1) \\
&= [\tilde{E}(1, M-1) + 1] \cdot q + 1 \cdot p \\
&= q\tilde{E}(1, M-1) + 1 \\
&= q[q\tilde{E}(1, M-2) + 1] + 1 \\
&= q^2\tilde{E}(1, M-2) + (q+1) \\
&= q^{M-1}\tilde{E}(1, 1) + (q^{M-2} + \cdots + 1) \\
&= q^{M-1} + q^{M-2} + \cdots + 1 \\
&= (1 - q^M)/(1-q) \\
&= \frac{1}{p} - \frac{1}{p}\sum_{i=1}^{1}\sum_{j=0}^{i-1}\binom{M}{j}p^j q^{M-j} .
\end{aligned}$$

Next, we assume that the theorem is true if the parameter value of the required number of successes for Z is $m-1$. Then, when the parameter value is m, we have

$$\begin{aligned}
\tilde{E}(m,M) &= \tilde{E}(m, M \mid U=0)\Pr(U=0) + \tilde{E}(m, M \mid U=1)\Pr(U=1) \\
&= [\tilde{E}(m, M-1) + 1] \cdot q + [\tilde{E}(m-1, M-1) + 1] \cdot p \\
&= q\tilde{E}(m, M-1) + p\left\{\frac{m-1}{p} - \frac{1}{p}\sum_{i=1}^{m-1}\sum_{j=0}^{i-1}\binom{M-1}{j}p^j q^{M-1-j}\right\} + 1 \\
&= q\tilde{E}(m, M-1) + m - \sum_{i=1}^{m-1}\sum_{j=0}^{i-1}\binom{M-1}{j}p^j q^{M-1-j} \\
&= q\left\{q\tilde{E}(m, M-2) + m - \sum_{i=1}^{m-1}\sum_{j=0}^{i-1}\binom{M-2}{j}p^j q^{M-2-j}\right\} \\
&\quad + m - \sum_{i=1}^{m-1}\sum_{j=0}^{i-1}\binom{M-1}{j}p^j q^{M-1-j}
\end{aligned}$$

$$= q^2\,\tilde{E}(m, M-2) + m(q+1) - \sum_{i=1}^{m-1}\sum_{j=0}^{i-1}\left[\binom{M-2}{j}+\binom{M-1}{j}\right]p^j q^{M-1-j}$$

$$= q^{M-m}\,\tilde{E}(m, m) + m\sum_{i=0}^{M-m-1} q^i \; - \sum_{i=1}^{m-1}\sum_{j=0}^{i-1}\sum_{k=m}^{M-1}\binom{k}{j}p^j q^{M-1-j}\,.$$

Now, $\tilde{E}(m, m) = m$ and

$$\sum_{k=m}^{M-1}\binom{k}{j} = \binom{M}{j+1} - \binom{m}{j+1}$$

(see Feller 1968, Vol. I, p.64, equation 12.8). Thus,

$$\tilde{E}(m, M) = m\sum_{i=0}^{M-m} q^i \; - \sum_{i=1}^{m-1}\sum_{j=0}^{i-1}\left[\binom{M}{j+1}+\binom{m}{j+1}\right]p^j q^{M-1-j}$$

$$= \frac{m}{p}(1-q^{M-m+1}) - \frac{1}{p}\sum_{i=2}^{m}\sum_{j=1}^{i-1}\left[\binom{M}{j}-\binom{m}{j}\right]p^j q^{M-j}$$

where

$$\sum_{i=2}^{m}\sum_{j=1}^{i-1}\left[\binom{M}{j}-\binom{m}{j}\right]p^j q^{M-j}$$

$$= \sum_{i=1}^{m}\sum_{j=0}^{i-1}\binom{M}{j}p^j q^{M-j} - \sum_{i=2}^{m}\binom{M}{0}q^M - \sum_{j=0}^{0}\binom{M}{j}p^j q^{M-j} - \sum_{i=2}^{m}\sum_{j=1}^{i-1}\binom{m}{j}p^j q^{M-j}$$

$$= \sum_{i=1}^{m}\sum_{j=0}^{i-1}\binom{M}{j}p^j q^{M-j} - \sum_{i=1}^{m}\sum_{j=0}^{i-1}\binom{m}{j}p^j q^{M-j}\,.$$

Let $X \sim$Binomial(m, p), then by Lemma 2.1

$$\sum_{i=1}^{m}\sum_{j=0}^{i-1}\binom{m}{j}p^j q^{M-j} = q^{M-m}\sum_{i=1}^{m}\sum_{j=0}^{i-1}\binom{m}{j}p^j q^{m-j} = q^{M-m}\sum_{i=0}^{m-1}\Pr(X \le i) = q^{M-m}(mq)\,.$$

Combining all the above results, we obtain

$$\tilde{E}(m, M) = \frac{m}{p}(1-q^{M-m+1}) - \frac{1}{p}\left\{\sum_{i=1}^{m}\sum_{j=0}^{i-1}\binom{M}{j}p^j q^{M-j} - q^{M-m}(mq)\right\}$$

$$= \frac{m}{p} - \frac{1}{p}\sum_{i=1}^{m}\sum_{j=0}^{i-1}\binom{M}{j}p^{j}q^{M-j} \ . \qquad \text{Q.E.D.}$$

The distribution of $\tilde{Z}$ in Theorem 2.2 may be considered as a right truncated negative binomial distribution, differing from those left truncated ones largely discussed in literature (see Sampford 1955, Rider 1955, Cacoullos and Charalambides 1975, Johnson, Kotz, and Kemp 1992 pp. 225-227, etc.).

3. Conclusion

In this paper, we proposed a closed form formula to compute the expected final random total sample size in TIS, and we proved the formula by mathematical induction. The method of mathematical induction may be used to prove many formulas in sampling theory. We conclude this paper by introducing the following Lemma 3.1 in which a well-known formula in capture-recapture sampling is given (see Singh and Chaudhary 1986, p.314, or Chapman 1952). The inductive proof of the formula is left as an exercise for readers [*] .

Lemma 3.1 Let N be the size of a finite population consisting of two strata and let N_h be the h^{th} stratum size, h = 1, 2. Samples are taken sequentially without replacement until m observations are obtained from the first stratum. Then, the expected value of the final random sample size, denoted by $E_m(N, N_1, N_2)$, is

$$E_m(N, N_1, N_2) = \frac{m(N+1)}{N_1 + 1}$$

where $1 \leqq m \leqq N_1$.

[*] The proof will be given in the next issue of this journal.

Acknowledgements

The author would like to thank Professor Chien-Pai Han and Professor Shaw-Hwa Lo for their careful reading and helpful comments.

References

Cacoullos, T. and Charalambides, C. A. (1975). On MVUE for truncated binomial and negative binomial distributions, *Annals of the Institute of Statistical Mathematics*, **27**, 235-244.

Chapman, D. G. (1952). Inverse, multiple and sequential sample censuses, *Biometrics*, **8**, 286-306.

Cochran, W. G. (1977). Sampling Techniques, 3rd ed., Wiley.

Feller, W. (1968). An Introduction to Probability Theory and Its Applications, Vol. I, 3rd ed., Wiley.

Haldane, J. B. S. (1945). On a method of estimating frequencies, *Biometrika*, **33**, 222-225.

Johnson, N. L., Kotz, S., and Kemp, A. W. (1992). Univariate Discrete Distributions, 2nd ed., Wiley.

Karr, A. F. (1993). Probability., Springer-Verlag.

Rider, Paul R.(1955). Truncated binomial and negative binomial distributions, *Journal of the American Statistical Association*, **50**, 877-883.

Sampford, M. R. (1955). The truncated negative binomial distribution, *Biometrika*, **42**, 58-69.

Singh, D. and Chaudhary, F. S. (1986). Theory and Analysis of Sample Survey Designs, Wiley.