Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

An interesting property of the entropy of some sampling designs

by Yves Tillé and David Haziza

December 2010

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 229-231
Statistics Canada, Catalogue No. 12-001-X

An interesting property of the entropy of some sampling designs

Yves Tillé and David Haziza [1]

Abstract

In this short note, we show that simple random sampling without replacement and Bernoulli sampling have approximately the same entropy when the population size is large. An empirical example is given as an illustration.

Key Words: Conditional Poisson sampling; Entropy; Simple random sampling; Poisson sampling.

1. Introduction

Consider a finite population of size N and let $U = \{1, ..., k, ..., N\}$ be the set of labels of this population. A sample s is a subset of U and a sampling design is a probability law $p(.)$ on the subsets of U such that $p(s) \geq 0$ for all $s \subset U$, and

$$\sum_{s \subset U} p(s) = 1.$$

Let $\pi_k = P(k \in s)$ be the first-order inclusion probability of unit k in the sample:

$$\pi_k = \sum_{\substack{s \subset U \\ s \ni k}} p(s).$$

Similarly, let $\pi_{k\ell} = P(k \in s \text{ and } \ell \in s)$ be the second-order inclusion probability of unit k and ℓ in the sample:

$$\pi_{k\ell} = \sum_{\substack{s \subset U \\ s \ni k,\, \ell}} p(s).$$

The entropy of a sampling design $p(.)$, denoted by $I(p)$, is defined as

$$I(p) = -\sum_{s \in Q} p(s) \log p(s), \qquad (1)$$

where $Q = \{s \mid p(s) > 0\}$ is the support of the sampling design $p(.)$. A sampling design has high entropy when there is a high amount of uncertainty or high amount of surprise in the sample which will be selected. In other words, when a sampling design has high entropy, it is very difficult to predict the type of sample we would obtain. Many sampling designs used in practice are high entropy designs. One notable exception is systematic sampling that has a very low entropy. The concept of entropy is useful in the context of variance estimation. When a sampling design has a high entropy, it is possible to obtain approximation of the second-order inclusion probabilities, $\pi_{k\ell}$, in terms of the first-order inclusion probabilities, which simplifies considerably the problem of variance estimation in the context of unequal probability sampling; *e.g.*, Brewer and Donadio (2003), Matei and Tillé (2005), Henderson (2006) and Haziza, Mecatti and Rao (2008).

It is well known that the sampling design with maximum entropy is Poisson sampling:

$$p_{\text{poiss}}(s) = \left(\prod_{k \in s} \pi_k \right) \left(\prod_{k \in U \setminus s} (1 - \pi_k) \right) \qquad (2)$$

for all $s \in Q$; *e.g.*, Tillé (2006). A special case of Poisson sampling is Bernoulli sampling, which is obtained from (2) by setting $\pi_k = \pi \in (0, 1)$, which leads to

$$p_{\text{bern}}(s) = \pi^{n_s} (1 - \pi)^{N - n_s}, \text{ for all } s \subset U,$$

where n_s is the random size of s. Using (1) and noting that $\sum_{s \in Q} n_s p(s) = N\pi$, the entropy of Bernoulli sampling is given by

$$I(p_{\text{bern}}) = -N(1 - \pi) \log(1 - \pi) - N\pi \log \pi, \qquad (3)$$

which is maximum when $\pi = 1/2$. In this case, we have $I(p_{\text{bern}}) = N \log 2$.

If we restrict to the class of fixed size sampling designs with first-order inclusion probabilities π_k, $k \in U$, the maximum entropy design is the so-called Conditional Poisson Sampling (CPS); (see Chen, Dempster and Liu 1994; Deville 2000; Tillé 2006). The CPS design can be implemented by repeatedly selecting samples according to Poisson sampling until the desired sample size, n (say), has been obtained. When $\pi_k = n/N$ for all $k \in U$, the CPS design reduces to simple random sampling without replacement:

$$p_{\text{srs}}(s) = \binom{N}{n}^{-1}$$

for all $s \in Q$. From (1), it follows that the entropy of simple random sampling is given by

$$I(p_{\text{srs}}) = \log N! - \log n! - \log(N - n)!. \qquad (4)$$

In other words, simple random sampling without replacement is the maximum entropy design in the class of equal probability fixed size sampling designs.

1. Yves Tillé, Institut de Statistique, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland; David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, QC, Canada, H3C 3J7. E-mail: haziza@dms.umontreal.ca.

Not all sampling designs possess a high entropy. For example, the 1-in-G systematic sampling design has a very low entropy. Here, the number of samples, $G = N/n$, is assumed to be an integer value. Since $p_{\text{syst}}(s) = 1/G$ for all $s \in Q$, the entropy of systematic sampling is given by

$$I(p_{\text{syst}}) = \log N - \log n,$$

which is much smaller than (4), especially for large values of N.

2. Main result

In this section, we compare the entropy of Bernoulli sampling with that of simple random sampling without replacement. Since the support of the Bernoulli sampling designs is much larger than that of simple random sampling without replacement, we expected the entropy of Bernoulli sampling to be much larger than that of simple random sampling without replacement. Table 1 shows the entropy for simple random sampling and Bernoulli sampling for different values of N and π. Surprisingly, we found the entropy of both sampling designs for the same inclusion probabilities and the same sample size to be approximately equal. From Table 1, it is clear that both sampling designs have similar entropies, even for moderate population sizes (*e.g.*, $N = 100$), independently of the value of π. This result is somehow curious considering the strong reduction of possible samples by fixing the sample size. Indeed, recall that the size of the support is $\binom{N}{n}$ for simple random sampling without replacement, whereas it is 2^N for Bernoulli sampling. For example, for $N = 100$ and $n = 20$, the size of the support for simple random sampling without replacement is equal to $\binom{100}{20} \approx 5.36 \times 10^{20}$, whereas it is equal to $2^{100} \approx 1.26 \times 10^{30}$ for Bernoulli sampling. In other words, the size of the support of Bernoulli sampling is approximately 2.36×10^9 larger than that of simple random sampling without replacement.

Result 1. *Let* $I(p_{\text{bern}})$ *and* $I(p_{\text{srs}})$ *be the entropy for Bernoulli sampling and simple random sampling without replacement*, *respectively given by* (3) *and* (4). *Then*,

$$\lim_{N\to\infty} \frac{I(p_{\text{srs}})}{I(p_{\text{bern}})} = 1.$$

Proof. By considering Stirling's formula (see Abramowitz and Stegun 1964, page 257)

$$\lim_{n\to\infty} \frac{n\log n - n}{\log n!} = 1,$$

we get

$$\lim_{\substack{N\to\infty \\ n\to\infty \\ N-n\to\infty}} \frac{N\log N - n\log n - (N-n)\log(N-n)}{\log\binom{N}{n}} = 1,$$

from which we obtain

$$\lim_{N\to\infty} = \frac{\log\binom{N}{N\pi}}{-N(1-\pi)\log(1-\pi) - N\pi\log\pi} = 1.$$

3. Conclusion

In this note, we showed that Bernoulli sampling and simple random sampling without replacement have very similar entropies, even for moderate population sizes. We conjecture that the same should be observed when comparing the Poisson sampling design and the CPS design for a given set on first-order inclusion probabilities. However, the proof of this result seems to be considerably more complex.

Table 1
Entropy of (Bernoulli sampling, simple random sampling) designs

N	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$	$\pi = 0.4$	$\pi = 0.5$
10	(3.3, 2.3)	(5, 3.8)	(6.1, 4.8)	(6.7, 5.3)	(6.9, 5.5)
100	(32.5, 30.5)	(50, 47.7)	(61.1, 58.6)	(67.3, 64.8)	(69.3, 66.8)
1,000	(325.1, 321.9)	(500.4, 496.9)	(610.9, 607.3)	(673, 669.4)	(693.1, 689.5)
10,000	(3,250.8, 3,246.5)	(5,004, 4,999.4)	(6,108.6, 6,103.9)	(6,730.1, 6,725.3)	(6,931.5, 6,926.6)
100,000	(32,508.3, 32,502.8)	(50,040.2, 50,034.5)	(61,086.4, 61,080.5)	(67,301.2, 67,295.2)	(69,314.7, 69,308.7)
1,000,000	(325,083, 325,076)	(500,402, 500,396)	(610,864, 610,857)	(673,012, 673,005)	(693,147, 693,140)

Acknowledgements

We thank an Associate Editor and a referee for constructive comments. Work of David Haziza was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Abramowitz, M., and Stegun, I.A. (1964). *Handbook of Mathematical Functions*. New York: Dover.

Brewer, K.R.W., and Donadio, M.E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29, 189-196.

Chen, S.X., Dempster, A.P. and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457-469.

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Technical report, CREST-ENSAI, Rennes.

Haziza, D., Mecatti, F. and Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron*, 66, 91-108.

Henderson, T. (2006). Estimating the variance of the Horvitz-Thompson estimator. Master's thesis, School of Finance and Applied Statistics, The Australian National University.

Matei, A., and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21, 4, 543-570.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.