On estimating entropy of sampling designs

Anton Grafström

Umeå University, Sweden
e-mail: anton.grafstrom@math.umu.se

Abstract

The focus of this paper is on finding suitable estimators of the entropy for different fixed size πps sampling designs. What estimator to use depends on the design and the situation. Some estimators are suitable only for small populations and samples. Other estimators can be used only for designs that have a known and practical probability function. Five different cases are covered. The entropy estimators presented are general and can be used for other sampling designs.

1 Introduction

Fixed size πps sampling is used to select a sample of fixed size n from a population of N units. The units are selected with unequal probabilities. Thus unit i in the population has a given probability π_i to be included in the sample. Some common designs for fixed size πps sampling are conditional Poisson, Sampford, and Pareto, cf. Bondesson *et al.* (2006). Several new and interesting designs have been discovered recently. Correlated Poisson sampling, cf. Bondesson & Thorburn (2008), is one of the new designs which is very general and flexible. The pivot method, cf. Tillé (2006 p. 106) or Bondesson (2008), is another new design which is easy to implement. It is necessary to ask how the different designs should be compared. The practical properties, such as how difficult the designs are to implement and how fast samples can be generated, are important. A sampling design should also have good mathematical properties. One such property is the entropy, which is a measure of randomness. Preferably a sampling design should be easy to implement, efficient and have a large entropy. As a first step, before comparisons can be made, it must be possible to calculate or estimate the entropy. That is why the focus here is on finding suitable estimators of the entropy for different fixed size πps sampling designs.

In section 2, the entropy is defined and different estimators are presented. A small simulation example can be found in section 3 and some comments are presented in section 4.

2 Estimating the entropy

A sampling design can be seen as a discrete distribution on the set of possible samples $\mathbf{x}_k$, $k = 1, 2, ..., M$. A sample $\mathbf{x}_k$ is seen as a vector of the inclusion indicators, thus we have $\mathbf{x}_k \in \{0, 1\}^N$, where N is the population size and a 1 indicates that the unit is included in the sample. The probability of getting sample $\mathbf{x}_k$ is here denoted by $p(\mathbf{x}_k)$. We have $\sum_{k=1}^{M} p(\mathbf{x}_k) = 1$. The entropy of a sampling design is defined as

$$H = -\sum_{k=1}^{M} p(\mathbf{x}_k) \log(p(\mathbf{x}_k)) = -E_p \left[\log(p(\mathbf{x}))\right], \tag{1}$$

where $\mathbf{x}$ is seen as a random variable. The entropy is a measure of randomness and it will be large when the probability mass is well distributed over the set of possible samples. The entropy will be small when a few samples have large probabilities. A sampling design with a high entropy is more robust than a design with low entropy.

It is well known that in the class of fixed size πps sampling designs, the adjusted conditional Poisson sampling design has maximum entropy cf. Hájek (1981) or Tillé (2006 p. 79). However, adjusted conditional Poisson sampling is a bit tricky to implement.

For many sampling designs it can be very hard to find the probabilities $p(\mathbf{x}_k)$. It is also a problem that the number of possible samples can be very large. When we have a fixed sample size n and a population of size N, there can be up to $\binom{N}{n}$ possible samples. Hence it can be extremely time consuming to calculate the entropy exactly and that is another reason why simulation might be needed to estimate the entropy. The objective is to find good estimates of the entropy for the different πps sampling designs so that they can be compared. The difficulty of calculating or estimating the entropy depends on the population and sample sizes and on which design we choose. There are at least five different cases, which will be described here.

2.1 Total count

If the probability function is known and the population and the sample sizes are small, then it is possible to calculate $p(\mathbf{x})$ for each possible sample $\mathbf{x}$. Thus, in this case we get an exact value of the entropy by using the definition (1) of entropy.

2.2 Infeasible probability function

In the situation described here it is assumed that it is possible to simulate samples via a sampling algorithm but the probability function is unknown or infeasible. This is the worst case, since it is time consuming to estimate the probability function without any information.

A large number, m, of samples need to be simulated. Let m_1 be the number of replicates of the first simulated sample and m_2 the number of replicates of the second simulated sample, etc. The entropy can then be estimated by the naive estimator

$$\hat{H}_{Naive} = -\frac{1}{m}\sum_{i=1}^{m}\log(m_i/m). \tag{2}$$

This estimator is biased, but it is consistent with respect to m and thus asymptotically unbiased.

2.3 Feasible probability function

If the probability function is known and practical in the sense that it can be calculated in an efficient way, then it is possible to get a good estimator of the entropy, even for large populations and sample sizes. Let $\mathbf{x}_k$, $k = 1, 2, ..., m$, be m simulated samples. The following estimator can then be used to estimate the entropy without bias

$$\hat{H}_{PF} = -\frac{1}{m}\sum_{k=1}^{m}\log(p(\mathbf{x}_k)). \tag{3}$$

This estimator can be used for correlated Poisson sampling, conditional Poisson sampling, Sampford sampling, systematic sampling and other designs with known and feasible probability functions.

2.4 Estimating the probability function

If it is possible to get good unbiased estimates of $p(\mathbf{x}_k)$ for simulated samples $\mathbf{x}_k$, $k = 1, 2, ..., m$, we can use that to estimate the entropy. If we estimate $p(\mathbf{x})$, we can approximate the entropy by

$$H = -E_p\left[\log(p(\mathbf{x}))\right] \approx -E_p\left[\log(\hat{p}(\mathbf{x})\right] - \frac{1}{2}E_p\left[\frac{Var_p(\hat{p}(\mathbf{x})|\mathbf{x})}{p^2(\mathbf{x})}\right]. \tag{4}$$

To get approximation (4) it is assumed that $\hat{p}(\mathbf{x})$ is close to $p(\mathbf{x})$, and then $\log(\hat{p}(\mathbf{x}))$ is approximated by a Taylor expansion around $p(\mathbf{x})$ with the first three terms. Then

$$\log(\hat{p}(\mathbf{x})) \approx \log(p(\mathbf{x})) + \frac{(\hat{p}(\mathbf{x}) - p(\mathbf{x}))}{p(\mathbf{x})} - \frac{1}{2}\frac{(\hat{p}(\mathbf{x}) - p(\mathbf{x}))^2}{p^2(\mathbf{x})}.$$

By then taking the expectation, and noting that

$$E_p\left[\frac{(\hat{p}(\mathbf{x}) - p(\mathbf{x}))^2}{p^2(\mathbf{x})}\right] = E_p\left[E_p\left(\frac{(\hat{p}(\mathbf{x}) - p(\mathbf{x}))^2}{p^2(\mathbf{x})}\middle|\mathbf{x}\right)\right] = E_p\left[\frac{Var_p(\hat{p}(\mathbf{x})|\mathbf{x})}{p^2(\mathbf{x})}\right],$$

we get approximation (4) of the entropy.

The entropy can then be estimated from the simulated samples by

$$\hat{H}_{EPF} = -\frac{1}{m}\sum_{k=1}^{m}\log(\hat{p}(\mathbf{x}_k)) - \frac{1}{2m}\sum_{k=1}^{m}\frac{\widehat{Var}(\hat{p}(\mathbf{x}_k))}{\hat{p}^2(\mathbf{x}_k)}. \tag{5}$$

This estimator will not be unbiased.

2.5 Approximating the probability function

If there exists a good approximation $p_0(\mathbf{x})$ of $p(\mathbf{x})$, then it is possible to construct a good estimator of the entropy. If $p_0(\mathbf{x})$ is close to $p(\mathbf{x})$, which we assume, then we can approximate $p(\mathbf{x})\log(p(\mathbf{x}))$, by a Taylor expansion around $p_0(\mathbf{x})$. Including the first three terms and then simplifying, we get

$$-p(\mathbf{x})\log(p(\mathbf{x})) \approx -p(\mathbf{x})\log(p_0(\mathbf{x})) - \frac{1}{2}\left(\frac{(p(\mathbf{x}))^2}{p_0(\mathbf{x})} - p_0(\mathbf{x})\right).$$

Summing over all possible samples on both sides, we then get

$$H \approx -E_p\left[\log(p_0(\mathbf{x})\right] - \frac{1}{2}\left(E_p\left[\frac{p(\mathbf{x})}{p_0(\mathbf{x})}\right] - 1\right).$$

Roughly, we have $H \approx -E_p\left[\log(p_0(\mathbf{x}))\right]$, which means that we can simulate m samples from $p(\cdot)$ and then use the estimator

$$\hat{H}_{APF1} = -\frac{1}{m}\sum_{k=1}^{m}\log(p_0(\mathbf{x}_k)). \tag{6}$$

This estimator will have some bias, but the estimator can be improved if $E_p\left[p(\mathbf{x})/p_0(\mathbf{x})\right]$ can be estimated. Since many samples are generated when estimating the entropy, a naive estimator of $p(\mathbf{x})$ can be used to estimate $E_p\left[p(\mathbf{x})/p_0(\mathbf{x})\right]$. If a total of m samples are simulated, then $p(\mathbf{x}_k)$ can be estimated, without bias, by

$$\hat{p}(\mathbf{x}_k) = \frac{(\#\text{ replicates of } \mathbf{x}_k) - 1}{m - 1}.$$

The improved estimator of the entropy would then be

$$\hat{H}_{APF2} = -\frac{1}{m}\sum_{k=1}^{m}\log(p_0(\mathbf{x}_k)) - \frac{1}{2}\left(\frac{1}{m}\sum_{k=1}^{m}\frac{\hat{p}(\mathbf{x}_k)}{p_0(\mathbf{x}_k)} - 1\right). \tag{7}$$

Estimator (7) will not be without bias, but it will have less bias than Estimator (6).

2.6 Other possibilities

When the probability function is unknown it is possible to use inverse sampling to estimate the entropy without bias. Inverse sampling means that samples must be generated until we get a specific sample a fixed number of times. This will probably not be an efficient way since generally there are a lot of possible samples and the probability of getting a specific sample is very small.

Another possibility is to use a Bayesian method and put a prior distribution on each $p(\mathbf{x}_k)$. A Dirichlet prior can be used for the vector of such probabilities. From the simulation result we get a posterior distribution that can be used to estimate the entropy.

3 Simulation example

In this example two estimators of the entropy are compared, the naive Estimator (2) and Estimator (3). The naive Estimator (2) uses only the number of replicates of each sample in the simulation, whereas Estimator (3) uses the known probability function. A small population of size $N = 10$ is used and samples of size $n = 5$ are drawn. The inclusion probability vector is

$$\boldsymbol{\pi} = (0.2, 0.25, 0.35, 0.4, 0.5, 0.5, 0.55, 0.65, 0.7, 0.9).$$

This small population was used by Sampford (1967) and in this example, the Sampford design has been used to generate samples. The probability function for Sampford sampling can be written as

$$p_{Sampf}(\mathbf{x}) = C_S \prod_{i=1}^{N} \pi_i^{x_i}(1-\pi_i)^{1-x_i} \times \sum_{k=1}^{N}(1-\pi_k)x_k, \tag{8}$$

where C_S is found from the normalizing condition $\sum_{\mathbf{x}:|\mathbf{x}|=n} p(\mathbf{x}) = 1$. The exact value of the entropy for the Sampford design is 4.727 for this population, which can be confirmed by summing over all possible samples, i.e. by using the definition of entropy.

Table 1: Comparisons of estimators $\hat{H}_{Naive}$ and $\hat{H}_{PF}$

Estimator	# Samples	$\hat{H}$	$SE(\hat{H})$	95% CI
Naive (2)	500	4.489	0.039	(4.413, 4.565)
-	1000	4.600	0.031	(4.539, 4.661)
-	5000	4.689	0.015	(4.659, 4.719)
-	10000	4.714	0.011	(4.692, 4.735)
PF (3)	500	4.758	0.049	(4.661, 4.856)
-	1000	4.751	0.034	(4.683, 4.819)
-	5000	4.735	0.016	(4.705, 4.766)
-	10000	4.719	0.011	(4.698, 4.741)

When studying the simulation results in Table 1, we clearly see that the naive Estimator (2) is negatively biased, but the bias reduces when more samples are used. The 95% confidence intervals for this estimator did not cover the true value for $m = 500$, 1000 and 5000 samples. Estimator (3), when the known probability function is used is however unbiased. The standard error of the two estimators are approximately equal for this population. For larger populations and samples, it is even more preferable to know the probability function and be able to use Estimator (3) since Estimator (2) would be more biased.

4 Comments

The simulation example indicates that the naive Estimator (2) may require a lot of simulated samples in order to estimate the entropy with small bias.

Estimation and approximation of the probability function might be useful when the probability function is known but difficult or time consuming to calculate. One such design is Brewer's method, cf. Tillé (2006 p. 112), for πps sampling, for which it is easy to find the probability of an ordered sample. To find the probability of an unordered sample it is nessessary to sum over all $n!$ ordered samples. Further work on entropy of πps sampling designs will be presented in an extended version of this paper.

References

Bondesson, L. & Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scand. J. Statist.* To appear, doi: 10.1111/j.1467-9469.2008.00596.x.

Bondesson, L. (2008). Variants of the splitting method for unequal probability sampling. In this volume.

Bondesson, L., Traat, I., Lundqvist, A. (2006). Pareto Sampling versus Sampford and Conditional Poisson Sampling *Scand. J. Statist.* **33**, 699-720.

Hájek, J. (1981). *Sampling from a Finite Population.* Marcel Dekker, New York.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika* **54**, 499-513.

Tillé, Y. (2006). *Sampling Algorithms.* Springer series in statistics, Springer science + Business media, Inc., New York.