Comparisons of methods for generating conditional Poisson samples

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

Abstract

Methods for conditional Poisson sampling (CP-sampling) are compared and the focus is on the efficiency of the methods. The time it takes to generate samples is investigated by simulation in the R-programming language. A new method introduced by Bondesson, Traat & Lundqvist in 2004 is found to be efficient. The new method is an acceptance rejection method that uses the efficient Pareto sampling method.

1 Introduction

Both conditional Poisson sampling (CP-sampling) and Sampford sampling are fixed size πps sampling designs. Thus, the methods can be used to get a sample of fixed size n from a population of size N with unequal inclusion probabilities. In 2004, Bondesson, Traat & Lundqvist introduced new methods for both CP-sampling and Sampford sampling. The new methods use Pareto sampling, which was introduced by Rosén (1997a,b). The methods are acceptance rejection (A-R) methods and they use the fact that the Pareto sampling design is very close to the design of both CP-sampling and Sampford sampling. A Pareto sample, which is rapidly generated, can be adjusted to become a true CP-sample or a Sampford sample by the use of an A-R filter.

In Grafström (2005), methods for both CP-sampling and Sampford sampling were compared. The methods were compared by simulation in the Matlab programming language and the new methods were found to be efficient. The focus in this text is on the methods for CP-sampling and we present some simulation results using the R-programming language. It is more appealing to use R since it is a free software which is specialised on statistical computing and it is widely used. Four methods for CP-sampling are compared and we wonder which method is the most efficient one.

CP-sampling is a modification of Poisson sampling. Let p_i be the given target inclusion probability for unit i, $i = 1, ..., N$. Each unit i in the population is included with probability p_i but only samples of size n are accepted. Usually it is assumed that $\sum_{i=1}^{N} p_i = n$ since it will maximize the probability to get samples

of size n. The assumption $\sum_{i=1}^{N} p_i = n$ is not restrictive. If it is not satisfied, the p_is can be transformed to satisfy that condition (Hajek, 1981, p. 66, Broström and Nilsson, 2000). When using CP-sampling, the true inclusion probabilities will only be approximately p_i. However, there is a possibility to adjust the p_is to obtain desired inclusion probabilities (Dupacova, 1979, Chen *et al.*, 1994, Aires, 2000, Tillé, 2005).

In section 2 there is a description of each of the sampling methods. Then in section 3, the methods are tested by simulation in some different sampling situations. The conclusions are presented in section 4.

2 The methods

The different sampling methods are described in this section.

2.1 CP-reject

The CP-reject method for CP-sampling can be found in Hajek (1981). Let the target inclusion probability for unit i be p_i with $\sum_{i=1}^{N} p_i = n$. Also, let I_i be independent and $Bin(1, p_i)$ distributed inclusion variables. Then unit i is included in the sample if $I_i = 1$. Simulate I_i for $i = 1, ..., N$ and accept the sample as a CP-sample if $\sum_{i=1}^{N} I_i = n$. Repeat the procedure until a sample is accepted.

2.2 CP-with replacement

CP-with replacement (Hajek, 1981) is another method for CP-sampling. Let the target inclusion probability for unit i be p_i with $\sum_{i=1}^{N} p_i = n$. Draw n units with replacement where unit i is drawn with probability $p_i' \propto p_i/(1 - p_i)$ and $\sum_{i=1}^{N} p_i' = 1$. If all n units are distinct, the sample is accepted as a CP-sample. Otherwise the procedure is repeated from the beginning.

2.3 CP-list sequential

The CP-list sequential method uses the definition of conditional probability and it was found to be efficient by Öhlund (1999). The method can also be found in Chen & Liu (1997), Traat *et al.* (2004) and in Tillé (2005). Let the target inclusion probability for unit i be p_i and $\sum_{i=1}^{N} p_i = n$. Also, let I_i be independent

$Bin(1, p_i)$ distributed random inclusion variables. Then the inclusion variables I_i can be successively generated from the conditional distributions

$$P\left(I_i = x \left| \sum_{j=i}^{N} I_j = n - n_{i-1} \right.\right), \quad x = 0, 1,$$

where $n_{i-1} = \sum_{j=0}^{i-1} I_j$ and $I_0 = 0$. We will always get a sample of size n. The conditional probabilities can be written as

$$P\left(I_i = 1 \left| \sum_{j=i}^{N} I_j = n - n_{i-1} \right.\right) = \frac{P\left(I_i = 1\right) P\left(\sum_{j=i+1}^{N} I_j = n - n_{i-1} - 1\right)}{P\left(\sum_{j=i}^{N} I_j = n - n_{i-1}\right)}.$$

To use this formula, one first has to calculate the probabilities $P\left(\sum_{j=i}^{N} I_j = k\right)$ for all i and k. That can be done recursively. Fortunately these probabilities need only to be calculated once. Then they can be used to generate as many samples as desired. The calculation may still be too time-consuming if N and n are large. Then it is possible to calculate only some of the probabilities exactly and use normal approximations for the rest of them.

2.4 Pareto sampling

Pareto sampling (Rosén, 1997a,b) is used to select a sample of fixed size n from a population of size N. Let λ_i be the given target inclusion probability for unit i and $\sum_{i=1}^{N} \lambda_i = n$. The method works as follows.

Generate $U_1, U_2, ..., U_N$, where the U_is are independent $U(0, 1)$ variables. Then calculate the Pareto ranking variables

$$Q_i = \frac{U_i/(1 - U_i)}{\lambda_i/(1 - \lambda_i)}$$

for each unit. Select the n units with the smallest Q-values as a Pareto sample of fixed size n. The true inclusion probabilities will be approximately λ_i.

2.5 CP-sampling via Pareto sampling

CP-sampling via Pareto sampling is the new method that was introduced by Bondesson, Traat & Lundqvist (2004). Let the target inclusion probability for unit i be p_i and $\sum_{i=1}^{N} p_i = n$. First a Pareto sample is generated with $\lambda_i = p_i,\ i = 1, ..., N$. Then the Pareto sample is either rejected or accepted as a CP-sample using the probability functions for the Pareto and CP designs. Let

$\mathbf{I} = (I_1, I_2, ...I_N)$ be the vector of random inclusion variables, i.e. $I_i \in \{0, 1\}$ and if $I_i = 1$ then unit i is sampled. Also, let $|\mathbf{I}| = \sum_{i=1}^{N} I_i = n$ be the sample size. The probability functions $p(\mathbf{x}) = P(\mathbf{I} = \mathbf{x})$ for the designs can then be written as

$$p_{CP}(\mathbf{x}) = C_{CP} \prod p_i^{x_i} (1 - p_i)^{1-x_i}, \quad |\mathbf{x}| = n,$$

and, for $\lambda_i = p_i$,

$$p_{Par}(\mathbf{x}) = \prod p_i^{x_i} (1 - p_i)^{1-x_i} \times \sum c_k x_k, \quad |\mathbf{x}| = n,$$

where

$$c_k = \int_0^\infty x^{n-1} \prod \frac{1 + \tau_i}{1 + \tau_i x} \cdot \frac{1}{1 + \tau_k x} dx \quad \text{and} \quad \tau_i = \frac{p_i}{1 - p_i}.$$

The sums and products are taken over the integers $1, 2, ..., N$. The constant C_{CP} is found from the normalizing condition $\sum_{\mathbf{x}:|\mathbf{x}|=n} p(\mathbf{x}) = 1$. We also have $C_{CP} \approx \sqrt{2\pi d}$ for large values of $d = \sum p_i(1 - p_i)$. The c_ks can be calculated exactly or approximated by Laplace approximations. One approximation is

$$c_k \approx c_k^* = (1 - p_k)\sqrt{2\pi}\sigma_k \exp\{\sigma_k^2 p_k^2/2\}, \text{ where } \sigma_k^2 = \frac{1}{d + p_k(1 - p_k)}.$$

This approximation can be improved by the following calibration

$$c_k^{*(cal)} = \frac{(N - n)c_k^*}{\sum_i c_i^*} c_0, \text{ where } c_0 = \int_0^\infty x^{n-1} \prod \frac{1 + \tau_i}{1 + \tau_i x} dx.$$

The constant c_0 can be calculated exactly or approximated by $c_0^* = \sqrt{2\pi/d}$. See Bondesson, Traat & Lundqvist (2004) for a full description of these approximations.

Now let us consider when we can accept a Pareto sample as a CP-sample. Let $p_1(\cdot)$ and $p_2(\cdot)$ be two probability functions. If there exists a constant B such that $p_1(\mathbf{x}) \leq Bp_2(\mathbf{x})$ for all $\mathbf{x}$, then a sample from $p_2(\cdot)$ can be generated and accepted as a sample from $p_1(\cdot)$ if $U \leq p_1(\mathbf{x})/(Bp_2(\mathbf{x}))$, where U is a random number from $U(0, 1)$. The procedure is repeated from the beginning until a sample is accepted.

If $p_1(\cdot) = p_{CP}(\cdot)$ without C_{CP} and $p_2(\cdot) = p_{Par}(\cdot)$, then the constant B must be chosen so that $1 \leq B \sum c_k x_k$ for all $\mathbf{x}$. If the probabilities $p_i,\ i = 1, ..., N$, are given in increasing order, then the $c_k s$ will decrease. The best choice of B will be $B^{-1} = \sum_{k=m}^{N} c_k$ where $m = N - n + 1$.

The conditional acceptance rate for accepting a Pareto sample as a CP-sample is

$$CAR(\mathbf{x}) = \frac{1}{B \sum c_k x_k}.$$

Thus a generated Pareto sample with $\lambda_i = p_i$ will be accepted as a CP-sample if $U \leq CAR(\mathbf{x})$, where $U \sim U(0, 1)$. See Bondesson, Traat & Lundqvist (2004) for more details.

3 Simulation and results

The sampling methods have been implemented in the R-programming language. For CP-sampling via Pareto we have used the calibrated Laplace approximation for calculation of the c_ks. In the CP-list sequential method all necessary probabilities for sums are calculated exactly.

The methods are first tested on a relatively small population and then a larger population is used, where the differences are more apparent.

Example 1. Sampling from the MU284 population. The population that consists of the 284 municipalities of Sweden is called the MU284 population and can be found in Särndal, Swensson & Wretman (1992, pp. 652-659). We use the variable P85, which is the population size in a municipal in the year 1985. Sampling is performed proportional to the size of the population (P85) in each municipal. We generated 1000 samples of size 50 and the results can be found in Table 1. The acceptance rate for CP-with replacement was too low for that method to be used in this example.

Table 1: Results for the MU284 population. We generated 1000 samples of size 50. The times are in seconds and $\hat{AR}_{Sim}$ is the acceptance rate for this simulation.

Method	n	Prel. calc.	Mean time	Total time	$\hat{AR}_{Sim}$
CP-reject	50	0	0.00232	2.32	0.069
CP-list sequential	50	0.66	0.01182	12.82	1
CP via Pareto	50	0	0.00336	3.36	0.791

We see from Table 1 that CP-reject has the lowest mean time. The simplicity of that method makes it efficient as long as the acceptance rate is not too low. CP-sampling via Pareto is also quite efficient and the high acceptance rate (0.791) implies that the probability functions for CP and Pareto are close. The CP-list sequential method is not as efficient as the other methods.

Example 2. Sampling from a large population. Let $N = 10000$ be the population size and $n = 2000$ be the sample size. Also let the target inclusion probabilities be

$$p_1 = 0.1,\ p_2 = 0.15,\ p_3 = 0.2,\ p_4 = 0.25,\ p_5 = 0.3,$$

where each p-value is used for 2000 units (thus we have $\sum_{i=1}^{N} p_i = n$). The acceptance rate for CP-with replacement was too low for that method to be used in this example. We generated 100 samples of size 2000 from this population and the results can be found in Table 2.

Table 2: Results for the large population. We generated 100 samples of size 2000. The times are in seconds and $\hat{AR}_{Sim}$ is the acceptance rate for this simulation.

Method	Prel. calc.	Mean time	Total time	$\hat{AR}_{Sim}$
CP-reject	0	0.373	37.31	0.009
CP-list sequential	616	0.3422	650.22	1
CP via Pareto	0	0.0501	5.01	0.901

In Table 2, we see that CP via Pareto has the lowest mean time. We also see that the acceptance rate (0.901) is even higher than in Example 1. If we look at the acceptance rate for CP-reject, we see that it is much lower now than in Example 1. The time for preliminary calculations in the list sequential method has increased a lot. After the preliminary calculations have been performed, the method is a little bit more efficient than CP-reject.

4 Conclusions

We found that the method CP-reject is efficient for sampling from a small population, but the acceptance rate decreases when the population size increases. We found that CP-with replacement is efficient only when n is much smaller than N. The method becomes inefficient very fast when the sample size n increases. The CP-list sequential method has preliminary calculations and the time for these calculations increases rapidly when the sample size n and the population size N increases. However, after the preliminary calculations have been performed the method is quite efficient. CP-sampling via Pareto seems to be very efficient in all situations. We have used Laplace approximation of the c_ks, but the approximation is very good and it makes this method faster than if the c_ks are calculated exactly. It is also easy to implement. The time it takes to generate a sample with this method is rather independent of the sample size n. The new method is the most efficient one in general, but not always. If the population and the sample size are not too big, then the list sequential method can be efficient and useful (Öhlund, 1999). The list sequential method might even be the most efficient one if many samples are to be generated, since the samples always are accepted.

References

Aires, N. (2000). *Techniques to calculate exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs.* Doctoral thesis, Chalmers University of technology and Göteborg Univerity, Göteborg, Sweden.

Bondesson, L., Traat, I., Lundqvist, A. (2004). Pareto Sampling versus Sampford and Conditional Poisson Sampling. Research Report No. 6 2004, Department of Mathematical Statistics, Umeå University. *To appear in Scand. J. Statist.*

Broström, G. & Nilsson, L. (2000). Acceptance-Rejection sampling from the conditional distribution of independent discrete random variables, given their sum. *Statistics* **34**, 247-257.

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

Chen, S.X. and Liu, J.S. (1997). Statistical applications of the Poissonbinomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875-892.

Dupacova, J. (1979). A note on Rejective Sampling, *Contributions to Statistics, Jaroslav Hajek Memorial Volume.* Reidel, Holland and Academia, Prague, 71-78.

Grafström, A. (2005). Comparisons of methods for generating conditional Poisson samples and Sampford samples. Master's thesis, Department of Mathematics and Mathematical Statistics, University of Umeå, Sweden.

Hajek, J. (1981). *Sampling from a finite population.* Marcel Dekker, New York.

Öhlund, A. (1999). Jämförelse av olika metoder att generera Bernoullifördelade slumptal givet deras summa (Comparisons of different methods to generate Bernoulli distributed random numbers given their sum). Master's thesis, Department of Mathematical Statistics, University of Umeå, Sweden.

Rosén, B. (1997a). Asymptotic theory for order sampling. *J. Statist. Plann. Inference* **62**, 135-158.

Rosén, B. (1997b). On sampling with probability proportional to size. *J. Statist. Plann. Inference* **62**, 159-191.

Särndal, C-E, Swensson, B. & Wretman, J. (1992). *Model assisted survey sampling.* Springer-Verlag, New York.

Tillé, Y. (2005). *Sampling algorithms.* Technical Report, Neuchâtel, Switzerland.

Traat, I., Bondesson, L. & Meister, K. (2004). Sampling design and sample selection through distribution theory. *J. Statist. Plann. Inference* **123**, 395-413.