ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 12, 2008

Estimation of the sample size required for obtaining given sample coverage

Mihhail Juhkam and Kalev Pärna

Abstract. We consider sampling from populations with large number of classes. The problem is to disclose a sufficiently big number of classes, which represent a dominating part of population (e.g. 99%). In many applications, e.g. in genetics, disclosure of all classes is not necessary, since it can require a very large sample and, hence, is too costly. In this paper we propose a method for estimation of the sample size, necessary to achieve a given sample coverage. We apply the method to populations where the class probabilities are the members of a geometric sequence. A Monte Carlo study demonstrates that the method we propose gives good results for values if the common ratio of the sequence is not too close to 1.

1. Introduction

Suppose that a population is divided into mutually exclusive classes, but the total number of classes is unknown. Sometimes it is necessary to draw a sample that contains at least one object from each class. For example, a biologist wishes to discover all species in an area, or a geneticist tries to identify all genotypes in a population. However, increasing the sample size and identification of the membership of objects is often costly. Therefore, we may limit ourselves with discovering a sufficiently big number of classes, which represent a dominating part of the population.

Let the unknown number of classes be s. Denote the probabilities of classes by $p_1 \geq p_2 \geq \ldots \geq p_s > 0$, $\sum_{i=1}^{s} p_i = 1$. This sequence $\{p_i\}_{i=1}^{s}$ will be called the *class distribution* of the population. Assume, for a moment, the *Poisson sampling scheme* where the number of objects from the class i in the sample follows the Poisson process with intensity p_i, $i = 1, \ldots, s$,

Received October 16, 2007.

2000 *Mathematics Subject Classification.* 62D05.

Key words and phrases. Class distribution, sample coverage, required sample size, size indices.

The research was partially supported by Estonian Science Foundation Grant 7044.

and all s processes are independent from each other. The Poisson scheme is natural sampling model in ecology where biologists count species that they observe during a fixed time interval $[0, \nu]$.

The *sample coverage* is defined by

$$C_\nu := \sum_{i=1}^{s} p_i I_i^\nu,$$

where

$$I_i^\nu = \begin{cases} 1, & \text{if the class } i \text{ is represented in the sample up to time } \nu, \\ 0, & \text{otherwise.} \end{cases}$$

Since $P\{I_i^\nu = 1\} = 1 - e^{-\nu p_i}$, $i = 1, \ldots, s$, the mean value of the sample coverage equals

$$EC_\nu = \sum_{i=1}^{s} p_i (1 - e^{-\nu p_i}).$$

Let the required coverage be $1-\eta$, where η is chosen close to zero. Suppose that a relatively small sample (of size n) is drawn successively and we wish to continue until the coverage $1 - \eta$ is achieved. One way to do this is to estimate the sample coverage each time a certain portion of new objects is drawn into the sample. The other possibility is to estimate the total sample size $n_{1-\eta}$, required to achieve the coverage $1-\eta$, and then increase the initial sample by $n_{1-\eta} - n$ units.

Estimation of the sample coverage was first discussed by Good [5] with application in studies of the literary vocabulary and accident proneness. The estimator proposed in the article was suggested by D. M. Turing and is called the Turing estimator. It is given by

$$\hat{C}_{Tur} = 1 - \frac{t_1}{n},$$

where t_1 is the number of classes in the sample, which are represented by one single object, and n is the sample size. The Turing estimator is derived using the Bayes' theorem. Normal limit law for this estimator has been proved by Mao and Lindsay [7]:

$$\exists \delta : \quad \frac{C_\nu - \hat{C}_{Tur}}{\delta} \sqrt{s} \to N(0, 1)$$

as $s \to \infty$. In contrast to the nonparametric approach, S. Engen [4] used a Gamma distribution to model class probabilities. Engen's idea was to estimate first the parameters of the Gamma distribution $f(p)$ and then use them to estimate the sample coverage. Boender and Rinnooy Kan [2] applied the Bayesian inference to estimate the sample coverage. Good and Toulmin [6] discussed the estimation of the increase of the coverage when the sample size was increased. A closely related problem of estimation of the number of undiscovered classes has been treated in [3], [8], and [9].

The main problem, which will be discussed in the present work is the estimation of the sample size $n_{1-\eta}$ required to achieve a given coverage $1-\eta$. Sections 2 and 3 are preparatory — we present and analyze some useful models of class distributions. It is shown that for large values of s it is convenient to model the distribution of probabilities p_i by a density function. Also the procedures for deducing p_i from a density function $f(p)$ and vice versa, building the density $f(p)$ that produces given probabilities p_i, are presented. In Section 4 a general method for estimation of required sample size $n_{1-\eta}$ is proposed. The method is applied for a special class distribution in Section 5, where we assume that the class distribution $\{p_i\}_{i=1}^{s}$ is defined by a decreasing geometric sequence

$$p_i = ab^i, \quad 0 < b < 1, \quad i = 1, \dots, s. \tag{1}$$

Such a class distribution (1) will be called the *exponentially decreasing class distribution.* After showing an estimator $\hat{b}$ for b, we will deduce the following equation to find $n_{1-\eta}$:

$$\eta = \frac{\hat{b}^{n_{1-\eta}} - 1}{n_{1-\eta} \ln \hat{b}}. \tag{2}$$

The paper ends with a simulation experiment where the performance of our method is studied.

2. Specifying the class distribution

Following the parametric approach, we assume that the class distribution of the population is given by a parametric family of functions. We describe two ways of determining class distributions: (1) the direct method, (2) the method based on density functions.

2.1. Direct specification of the class distribution. Suppose we wish to define a class distribution for the population with s classes. It can be done as follows. Let ρ be a positive monotonically decreasing function on the interval $[1, s]$. Introduce the normalizing coefficient

$$\rho_0 = \sum_{i=1}^{s} \rho(i)$$

and let $\pi(x) = \rho(x)/\rho_0$. Then

$$p_i = \pi(i), \quad i = 1, \dots, s,$$

form a class distribution satisfying

$$p_1 \geq p_2 \geq \dots \geq p_s > 0, \quad \sum p_i = 1.$$

Examples are:

(1) A constant function $\rho(x) = c$, $c > 0$ defines the class distribution with equal class probabilities $p_i = 1/s$, $i = 1, \ldots, s$,
(2) A linearly decreasing function $\rho(x, a, b) = b - ax$, $a > 0$, $b > as$, defines the class probabilities, which form an arithmetic sequence,
(3) An exponentially decreasing function $\rho(x, b) = b^x$, $0 < b < 1$, defines the class probabilities, which are the terms of a geometric sequence.

2.2. Defining the class distribution by a density function. Except for some simple parametric cases as those described above, for a very large number of classes the direct method of specification of the class distribution $\{p_i\}$ becomes cumbersome. Then a natural idea is to model the distribution of probabilities p_i by a suitable density function. This approach has been earlier used in [4]. Let $f(p)$ be a density function satisfying

(a) $f(p) = 0$ for $p \leq 0$,
(b) $\int_0^\infty \frac{f(p)}{p} dp < \infty$.

An algorithm, which uniquely defines a class distribution, is:

(1) Start with density f satisfying (a) and (b). Denote $g(p) = f(p)/p$, $p > 0$.
(2) Find the points $0 = \xi_s < \xi_{s-1} < \ldots < \xi_1 < \xi_0 = \infty$ such that

$$\int_{\xi_i}^{\xi_{i-1}} g(p)dp = 1, \quad i = 1, \ldots, s-1, \quad 0 < \int_{\xi_s}^{\xi_{s-1}} g(p)dp \leq 1.$$

(3) Finally, define the class probabilities by

$$p_i = \int_{\xi_i}^{\xi_{i-1}} f(p)dp, \quad i = 1, \ldots, s.$$

Since

$$\sum_{i=1}^{s} p_i = \sum_{i=1}^{s} \int_{\xi_i}^{\xi_{i-1}} f(p)dp = \int_0^\infty f(p)dp = 1,$$

the obtained decreasing sequence $\{p_i\}_{i=1}^s$ is a valid class distribution, and is uniquely given by the density function f. Although $\xi_0 = \infty$, but $\xi_1 < 1$, as

$$1 = \int_{\xi_1}^{\xi_0} g(p)dp = \int_{\xi_1}^{\xi_0} \frac{f(p)}{p} dp < \int_{\xi_1}^{\xi_0} \frac{f(p)}{\xi_1} dp < \frac{1}{\xi_1}.$$

In the Figure 1 an example is shown where the class distribution is defined by a Gamma density.

3. Finding a density that produces given class distribution

Here our aim is to show how to construct a density function f which produces the same class distribution $\{p_i\}_{i=1}^s$ as that produced directly by a function ρ (see Section 2.1). We will assume here that ρ is strictly decreasing

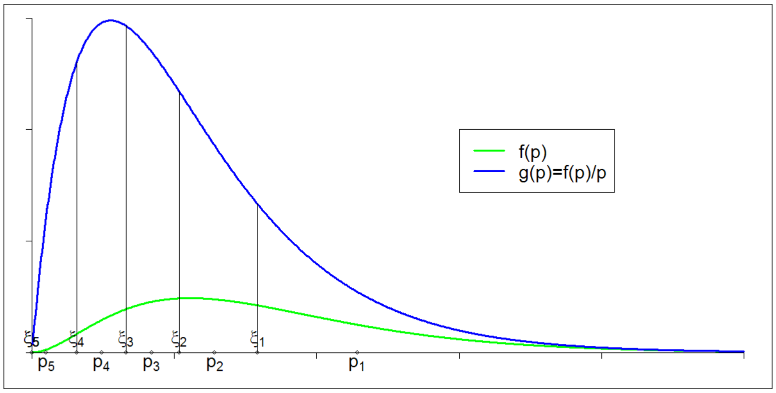

FIGURE 1. *Building the class distribution* $\{p_1, \ldots, p_s\}$ *by a Gamma density. The five areas under the curve of* $g(p)$ *are equal to* 1 *(except for the leftmost one).*

which also mean that strict inequalities $p_1 > p_2 > \ldots > p_s > 0$ hold. Hence, the inverse function ρ^{-1} exists on the interval $[\rho(s), \rho(1)]$.

3.1. General case. Let the class distribution $\{p_i\}_{i=1}^{s}$ be directly given by a function ρ, i.e.

$$p_i = \frac{\rho(i)}{\rho_0} = \pi(i), \quad i = 1, \ldots, s.$$

We use the notation and relationships given in Section 2.2. Let $f(p)$ be a density which produces the class distribution $\{p_i\}_{i=1}^{s}$ and let $g(p) = f(p)/p$. Then g must satisfy

$$\int_{\xi_i}^{\xi_{i-1}} g(p)dp = 1, \quad i = 1, \ldots, s-1$$

for a sequence $\{\xi_i\}_{i=0}^{s}$ where $\xi_i < p_i < \xi_{i-1}$. For simplicity, we also assume that

$$\int_{\xi_s}^{\xi_{s-1}} g(p)dp = 1.$$

Observe that for a large number of classes the difference between the three quantities, ξ_i, ξ_{i-1} and p_i, is negligible and we can take $\xi_i \approx p_i$. If we define

$$G(p) = \frac{\int_0^p g(t)dt}{s}, \tag{3}$$

then for $i = 1, \ldots, s$

$$G(p_i) \approx G(\xi_i) = \frac{s-i}{s} = 1 - \frac{i}{s} = 1 - \frac{\rho^{-1}(\rho_0 p_i)}{s}. \tag{4}$$

The approximation (4) can be extended to the whole interval $p \in [\xi_s, \xi_0]$

$$G(p) \approx 1 - \frac{\rho^{-1}(\rho_0 p)}{s}, \quad p \in [\xi_s, \xi_0].$$

From (3) we see that

$$g(p) = sG'(p),$$

providing

$$g(p) \approx -\left(\rho^{-1}(\rho_0 p)\right)', \quad p \in [\xi_s, \xi_0]. \tag{5}$$

Finally, by using relationship $f(p) = pg(p)$, we get an approximate density f that produces the class distribution $\{p_i\}_{i=1}^s$:

$$f(p) \approx -p\left(\rho^{-1}(\rho_0 p)\right)', \quad p \in [\xi_s, \xi_0]. \tag{6}$$

The bounds ξ_s and ξ_0 in (6) are calculated from the following system of equations:

$$\begin{cases} \int_{\xi_s}^{\xi_0} f(p)dp &= 1, \\ \\ \int_{\xi_s}^{\xi_0} g(p)dp &= s. \end{cases} \tag{7}$$

3.2. Producing exponentially decreasing classdistribution. Here we reveal the density which produces exponentially decreasing class distribution defined by

$$\rho(x) = b^x, \quad 0 < b < 1. \tag{8}$$

The inverse function ρ^{-1} is

$$\rho^{-1}(p) = \frac{\ln p}{\ln b}.$$

The functions g and f are obtained from (5) and (6):

$$\begin{aligned} g(p) &= -\left(\rho^{-1}(\rho_0 p)\right)' = -\left(\frac{\ln \rho_0 p}{\ln b}\right)' = -\frac{1}{p \ln b}, \quad p \in [\xi_s, \xi_0], \\ f(p) &= pg(p) = -\frac{1}{\ln b}, \quad p \in [\xi_s, \xi_0]. \end{aligned}$$

By solving the system (7) we obtain the values for ξ_s and ξ_0:

$$\xi_s = -\frac{b^s \ln b}{1 - b^s}, \quad \xi_0 = -\frac{\ln b}{1 - b^s}. \tag{9}$$

Therefore, the exponentially decreasing class distribution (8) is produced by the uniform density

$$f(p) = \begin{cases} -\frac{1}{\ln b}, & p \in [-\frac{b^s \ln b}{1-b^s}, -\frac{\ln b}{1-b^s}], \\ 0, & \text{otherwise}. \end{cases} \tag{10}$$

4. Estimating the required sample size

In this section a novel method for estimation of the sample size $n_{1-\eta}$, required to achieve the sample coverage $1-\eta$, is proposed. The method is applied to the exponentially decreasing class distribution and its goodness is studied by a simulation experiment.

4.1. A general method for estimation of the required sample size. Suppose that the class distribution is given by a parametric density $f(p,\vec{\theta})$. The first step is to estimate the vector of parameters $\vec{\theta}$. For that we use the sample values t_k of so called *size indices* T_k, $k=1,2,\ldots$. The size index T_k is the number of the classes which are represented in the sample by exactly k objects. Under the Poisson sampling scheme the mean value of the size index during the time ν is equal to

$$E(T_k)=\sum_{i=1}^{s}\frac{(\nu p_i)^k}{k!}e^{-\nu p_i}. \tag{11}$$

For large number of classes, each summand in (11) can be approximated by

$$\frac{(\nu p_i)^k}{k!}e^{-\nu p_i}\approx\int_{\xi_i}^{\xi_{i-1}}\frac{(\nu p)^k}{k!}e^{-\nu p}g(p,\vec{\theta})dp,$$

where $g(p,\vec{\theta})=f(p,\vec{\theta})/p$. Thus, the mean of the size index can be approximated by

$$E(T_k)\approx\int_{0}^{\infty}\frac{(\nu p)^k}{k!}e^{-\nu p}g(p,\vec{\theta})dp.$$

By replacing the first m mean size indices $E(T_k)$ by their realizations t_k, we obtain the system of equations

$$\int_{0}^{\infty}\frac{(\nu p)^k}{k!}e^{-\nu p}g(p,\vec{\theta})dp=t_k,\quad k=1,\ldots,m. \tag{12}$$

The system (12) can be solved for $\vec{\theta}$ using e.g. the method of least squares. Let $\hat{\theta}$ be the least squares estimator of the parameter vector $\vec{\theta}$. Further, we approximate the sample coverage by integral as follows:

$$\begin{aligned} EC_\nu &= \sum_{i=1}^{s}p_i(1-e^{-\nu p_i})\approx\sum_{i=1}^{s}\int_{\xi_i}^{\xi_{i-1}}p(1-e^{-\nu p})g(p,\vec{\theta})dp \\ &= \int_{0}^{\infty}p(1-e^{-\nu p})g(p,\vec{\theta})dp=1-\int_{0}^{\infty}e^{-\nu p}f(p,\vec{\theta})dp. \end{aligned}$$

Under the assumption of Poisson sampling scheme, the size n of the sample drawn during time ν is a random variable with mean ν. In practice, however, the sample size is recorded more often than the sampling time. Thus, we replace ν by n to obtain

$$E(C_n) \approx 1 - \int_0^\infty e^{-np} f(p, \vec{\theta}) dp.$$

For $n = n_{1-\eta}$, the required sample size, we have

$$E(C_{n_{1-\eta}}) \approx 1 - \int_0^\infty e^{-n_{1-\eta}p} f(p, \vec{\theta}) dp.$$

Here the mean coverage $E(C_{n_{1-\eta}})$ can be approximated by $1-\eta$, since $C_{n_{1-\eta}}$ is the coverage at the moment when the event "the coverage is greater than or equal to $1-\eta$" occurs. Now we replace the parameter $\vec{\theta}$ by its estimate $\hat{\theta}$. Therefore, an estimator of the required sample size is obtained by solving (for $n_{1-\eta}$) the equation

$$\eta = \int_0^\infty e^{-n_{1-\eta}p} f(p, \hat{\theta}) dp. \tag{13}$$

Next we apply the method to the exponentially decreasing class distribution.

4.2. Application to exponentially decreasing class distribution. We know that the exponentially decreasing class distribution can be produced by the uniform density (10). First simplify the expression of $E(T_k)$:

$$E(T_k) \approx \int_{\xi_s}^{\xi_0} \frac{(\nu p)^k}{k!} e^{-\nu p} g(p, b) dp = -\frac{\nu^k}{k! \ln b} \int_{\xi_s}^{\xi_0} p^{k-1} e^{-\nu p} dp, \tag{14}$$

where ξ_s and ξ_0 are defined by (9). Substituting $t = \nu p$ into the integral (14) we get

$$E(T_k) \approx -\frac{1}{k! \ln b} \int_{\nu\xi_s}^{\nu\xi_0} t^{k-1} e^{-t} dt. \tag{15}$$

The latter integral can be expressed as

$$\int_{\nu\xi_s}^{\nu\xi_0} t^{k-1} e^{-t} dt = \Gamma(k, \nu\xi_s) - \Gamma(k, \nu\xi_0),$$

where $\Gamma(k, c)$ is incomplete gamma function

$$\Gamma(k, c) = \int_c^\infty t^{k-1} e^{-t} dt,$$

which for integer values of k equals (see [1])

$$\Gamma(k, c) = (k-1)! e^{-c} \sum_{j=0}^{k-1} \frac{c^j}{j!}.$$

Hence

$$E(T_k) \approx -\frac{1}{k \ln b}\left(e^{-\nu\xi_s} \sum_{j=0}^{k-1} \frac{(\nu\xi_s)^j}{j!} - e^{-\nu\xi_0} \sum_{j=0}^{k-1} \frac{(\nu\xi_0)^j}{j!} \right).$$

Each term in parentheses is a sum of k first Poisson probabilities, one with a small expectation $(\nu\xi_s)$ and the other with a large expectation $(\nu\xi_0)$. Furthermore, it can be shown that if $s \to \infty$, $\nu \to \infty$ and $\nu b^s \to 0$, then[1]

$$E(T_k) \to -\frac{1}{k \ln b}.$$

If the realizations t_k of size indices T_k are available, then the following system of equations is obtained:

$$t_k = -\frac{1}{k \ln b}, \quad k = 1, 2 \ldots . \tag{16}$$

The system (16) provides us the least-squares (LS) estimate of $\ln b$, by minimizing the sum

$$\sum_{k=1}^{\infty} \left(t_k + \frac{1}{k\widehat{\ln b}} \right)^2 .$$

The required LS estimate is

$$\widehat{\ln b} = -\frac{\sum_{k=1}^{\infty} \frac{1}{k^2}}{\sum_{k=1}^{\infty} \frac{t_k}{k}} = -\frac{\pi^2}{6\sum_{k=1}^{\infty} \frac{t_k}{k}}.$$

Let m be the size of the largest class represented in the sample. Then $t_k = 0$, $k = m+1, m+2, \ldots$ and we can write

$$\widehat{\ln b} = -\frac{\pi^2}{6\sum_{k=1}^{m} \frac{t_k}{k}}.$$

The estimator of the required sample size is obtained by substituting the density

$$f(p) = -\frac{1}{\widehat{\ln b}}$$

into (13). This gives

$$\eta = -\frac{1}{\widehat{\ln b}} \int_{\xi_s}^{\xi_0} e^{-n_{1-\eta} p} dp.$$

After integration and replacing ξ_s by 0 the equation transforms to

$$\eta = \frac{\exp\left(-n_{1-\eta}\xi_0\right) - 1}{n_{1-\eta}\widehat{\ln b}}.$$

[1] The same result can be drawn formally from (15) by replacing $\nu\xi_s \approx 0$, and $\nu\xi_0 \approx \infty$, and using $\Gamma(k) = (k-1)!$.

Provided that $s \to \infty$, ξ_0 can be replaced by $-\widehat{\ln b}$. Thus the equation for calculation of the required sample size $n_{1-\eta}$ takes the form

$$\eta = \frac{\exp\left(n_{1-\eta}\widehat{\ln b}\right) - 1}{n_{1-\eta}\widehat{\ln b}}$$

which is equivalent to (2) where $\hat{b} = \exp(\widehat{\ln b})$.

4.3. A simulation study. To study the performance of the proposed method of estimation of the required sample size, a simulation experiment was carried out. We simulated 100 multinomial samples, each of size $n = 500$, from a population with $s = 500$ classes and exponentially decreasing class distribution with $b = 0.95$, 0.97, 0.98 and 0.99. The required sample sizes $\hat{n}_{0.99}$, $\hat{n}_{0.995}$ and $\hat{n}_{0.999}$ were estimated using the method proposed above. The actual required sample sizes $n_{0.99}$, $n_{0.995}$ and $n_{0.999}$ were also obtained by continuing simulation until the required coverage was achieved. The mean relative error E was used to measure the error of estimation. To calculate E, the absolute values of relative errors of estimates were obtained and averaged over 100 samples as follows

$$E = AVG\left(\left|\frac{\hat{n}_{1-\eta} - n_{1-\eta}}{n_{1-\eta}}\right|\right).$$

In the Figure 2, the relative errors of required sample size are shown.

We can see that E is in range 0.1– 0.2 for values of b in $[0.94, 0.985]$. However, for $b = 0.99$ relative errors become unacceptably large, especially for the smallest values of η. As the limiting class distribution in the case $b \to 1$ is the uniform class distribution $p_i = 1/s$, $i = 1, \dots, s$, we can conclude that the method proposed does not perform well for class distributions which are close to the uniform class distribution. Nevertheless, the method gives satisfactory results when the class distribution differs significantly from the uniform one.

5. Summary

A method for estimation of the sample size which is necessary to achieve a given sample coverage was proposed. The method works well under the assumption that the population consists of a large number of classes and for at least exponentially decreasing class probabilities if the common ratio is not too close to 1. A further research is needed to elaborate (and test) similar estimation methods for other families of class distributions.

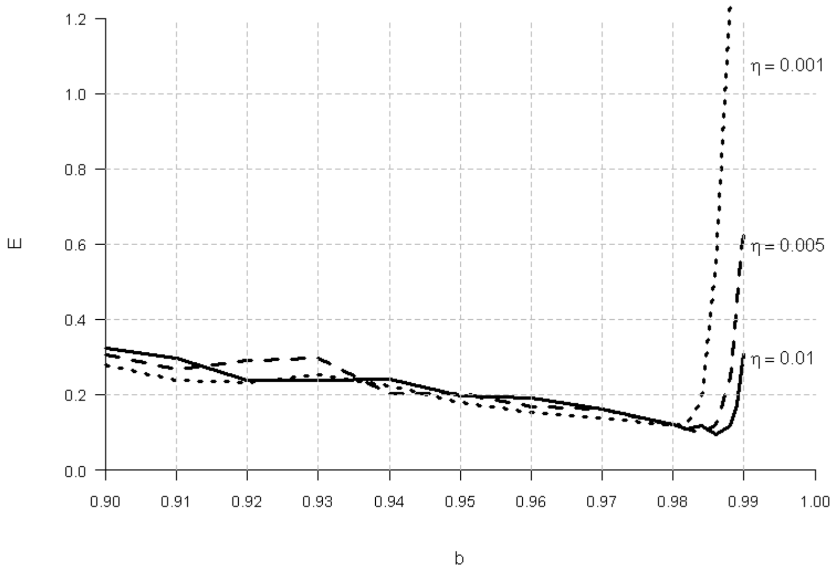

FIGURE 2. *Results of the simulation experiment: average relative error of the estimated sample size.*

References

[1] M. Abramowitz and I. A. Stegun (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York.

[2] C. G. E. Boender and A. H. G. Rinnooy Kan (1987), *A multinomial Bayesian approach to the estimation of population and vocabulary size*, Biometrika **74**(4), 849–856.

[3] A. Chao (1984), *Nonparametric estimation of the number of classes in a population*, Scand. J. Statist. **11**, 265–270.

[4] S. Engen (1974), *On species frequency models*, Biometrika **61**(2), 263–270.

[5] I. J. Good (1953), *The population frequencies of species and the estimation of population parameters*, Biometrika **40**(3), 237–264.

[6] I. J. Good and G. H. Toulmin (1956), *The number of new species and the increase in population coverage when a sample is increased*, Biometrika **43**(1), 45–63.

[7] C. X. Mao and B. G. Lindsay (2002), *A Poisson model for the coverage problem with a genomic application*, Biometrika **89**(3), 669–681.

[8] T.-J. Shen, A. Chao and C.-F. Lin (2003), *Predicting the number of new species in further taxonomic sampling*, Ecology **84**(3), 798–804.

[9] A. R. Solow and S. Polasky (1999), *A quick estimator for taxonomic surveys*, Ecology **80**(8), 2799–2803.

INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITY OF TARTU, LIIVI 2, 50409 TARTU, ESTONIA

E-mail address: mihhail.juhkam@ut.ee

E-mail address: kalev.parna@ut.ee