UNIVERSITY OF TARTU
Faculty of Mathematics and Computer Science
Institute of Mathematical Statistics

Mihhail Juhkam

POPULATIONS WITH LARGE NUMBER OF CLASSES: MODELS AND ESTIMATION OF SAMPLE COVERAGE AND SAMPLE SIZE

Master's thesis (40 CP)

Supervisor: prof. Kalev Pärna

Tartu 2006

Contents

1 Introduction **4**

2 Sample colority and coverage **7**

2.1 Multinomial and Poisson sampling scheme 7

2.2 Definitions 8

2.3 Mean and variance of sample colority and coverage in the case of known color probabilities 10

2.3.1 Mean of sample colority and coverage 10

2.3.2 Variance of sample colority and coverage 12

3 Two ways of defining distribution of color probabilities **17**

3.1 Direct definition of color probabilities 17

3.2 Defining color probabilities by density function 20

3.3 How to select density f that produces a given set of color probabilities 23

3.3.1 The case of approximately linearly decreasing color probabilities 23

3.3.2 Finding approximate density for arbitrary set of color probabilities 33

3.4 Density functions used to define color distribution 36

4 Modelling color probabilities by Gamma distribution **38**

4.1 Derivation of Engen's Extended Negative Binomial (ENB) model 38

4.1.1 Parametric Poisson-Gamma model definition 38
4.1.2 Mean number of colors with x representatives 40
4.1.3 Joint distribution of size indices T_x 44
4.2 Estimation of ENB model 45
4.2.1 Derivation of maximum likelihood function 45
4.2.2 Fitting the model by the ML estimation 46

5 Estimation of sample coverage 49
5.1 Review of literature on estimation of sample coverage 49
5.2 Estimation of sample coverage in the case of ENB model . . . 50
5.3 Inspection of goodness of ENB model: a Monte-Carlo experiment 53

6 Estimation of sample size required for achieving given coverage 56
6.1 Uniform color distribution 57
6.1.1 Method 1: Estimating of required sample size by the "two-point" method of moments 57
6.1.2 Method 2: Estimating of required sample size by non-linear regression 58
6.1.3 Method 3: "One-point" method of moments 59
6.1.4 Monte-Carlo comparison of Method1 and Method 3 . . 59
6.2 Linearly decreasing color distribution 61
6.2.1 Method 4 for estimation of required sample size 62
6.2.2 Monte-Carlo experiment: evaluation of Method 4 . . . 64

6.3 Exponentially decreasing color distribution 66

6.3.1 Method 5 for estimation of required sample size 68

6.3.2 Monte-Carlo experiment: evaluation of Method 5 . . . 70

7 Summary **74**

Resümee **78**

Appendix **80**

A1. SAS/IML functions for solving nonlinear optimization problems (NLP) 80

A2. Derivation of Turing estimator of sample coverage 82

1 Introduction

In many areas the following problem have been arisen. Each object in population belongs to some class, but the total number of classes s is unknown. We want to identify all the classes in population. In order to do this, we start to take objects into the sample. We should stop when all s classes are represented in the sample by at least one element. But since s itself is unknown, this stopping rule can not be applied.

In many cases the identification of membership of objects is costly. This is true, for example, when the researcher identifies all genotypes of a population. In this case we may limit ourselves to discovering only those classes, which represent the overwhelming part of the population, e. g. 99%. In this case the sample is said to have the *coverage* of 0.99. The following example explains why such limitation may be useful.

Consider two biological communities, both including 100 individuals belonging to 4 species. The frequencies of species in both communities are shown on Figure 1.

It is clear, that to disclose all the 4 species, in the first community it is sufficient to draw a smaller sample than in the second community. The reason is that in the second community, the probability of drawing the species D, which is represented by only one individual, is relatively small. It is quite probable, that we need to draw the most of individuals into the sample in order to disclose all the 4 species. Thus, it may be reasonable to draw individuals until the species A, B and C are represented in the sample. In this case, the coverage of the sample would be 99%. At the same time, the required sample size will be considerably smaller.

Thus, we may formulate the two main problems, which will be discussed in

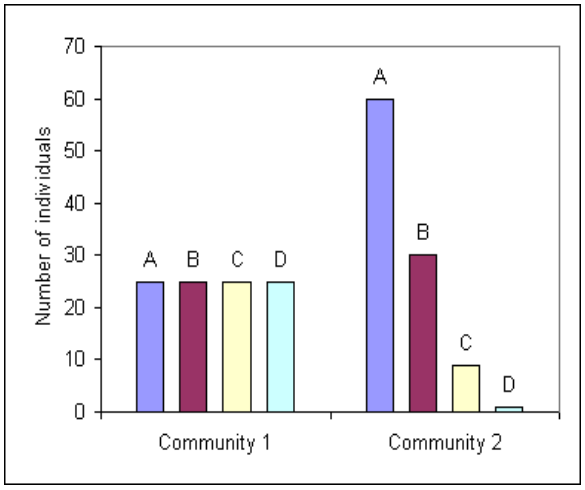

Figure 1: Frequencies of species in two communities

the present work:

(a) What is the coverage C of a given sample? In other words: what is the total percentage of the classes which are represented in the sample? If $C \geq 1 - \eta$, where η is small (say, $0 < \eta < 0.01$), then we should stop the sampling. Otherwise, the sample should be extended and the further question is:

(b) How many additional objects must we draw in order to achieve the sample coverage of $1 - \eta$?

The estimating of the sample coverage is first discussed by Good [7], who proposed the nonparametric estimator for the sample coverage. Another estimator has been derived by Engen [2], using parametric approach. Both estimators are discussed in the current work. The problem of estimating the sample size has not been discussed in the literature, but in article by Good and Toulmin [8], authors have discussed the close theme, estimating the increase of coverage if the sample is increased.

The thesis is organized as follows. In the Chapter 2 the terms of sample colority and coverage are defined. Also mean and variance of the sample colority and the coverage are evaluated in the case of known probabilities of classes. In the Chapter 3 there are proposed two ways of defining a set of color probabilities by a little number of parameters. The purpose of such defining is the further estimation of these parameters. The connection between the two ways of defining the probabilities of classes is also discussed. In the Chapter 4 we will discuss the Engen's Negative Binomial model (ENB). In this model the probabilities of classes are defined by the Gamma distribution, which has two parameters. The approximate distribution of size indices is also derived. Using this approximate distribution, the ML estimate of parameters of Gamma distribution is obtained. In the Chapter 5 there is proposed the review of the literature on the problem of the coverage estimation. Then the simulation experiment is conducted in order to inspect the reliability and validity of the Engen's ENB model in coverage estimation. In the Chapter 6 we discuss the estimation of the sample size, required for achieving the given coverage. We consider a simple case when the probabilities of classes are equal, and two more general cases when the sequence of probabilities of classes is either linearly or exponentially decreasing sequence.

2 Sample colority and coverage

2.1 Multinomial and Poisson sampling scheme

Consider the following urn model. From an urn containing balls of s different colors, n balls are drawn at random with replacement. Denote the number of balls of color i in the sample by N_i, $i = 1, \ldots, s$. Let the relative frequency of balls of color i in urn be equal to p_i, $i = 1, \ldots, s$. Under this model the joint distribution of N_i's is the multinomial distribution:

$$\mathbf{P}(\bigcap_{i=1}^{s}(N_i = n_i)) = n! \prod_{i=1}^{s} \frac{p_i^{n_i}}{n_i!}. \tag{1}$$

Further in the present work this sampling scheme will be referred to as ***multinomial scheme***. Note that under the multinomial scheme the sample size is fixed (nonrandom).

Besides the multinomial model, we will consider the following Poisson model. Suppose that the number of balls of ith color in the sample follows a homogeneous Poisson process ζ_i with intensity λ_i, $i = 1, \ldots, s$ and that processes $\zeta_1, \ldots, \zeta_s$ are independent. We will assume that observations have been made up to a fixed time ν. Then the counts N_i are independent Poisson random variables with expectations $\lambda_i \nu$, $i = 1, \ldots, s$. Hence, the joint distribution of N_i's is

$$\mathbf{P}(\bigcap_{i=1}^{s}(N_i = n_i)) = \prod_{i=1}^{s} \frac{(\lambda_i \nu)^{n_i} e^{-\lambda_i \nu}}{n_i!}.$$

Further this sampling scheme will be called the ***Poisson scheme***. The probability of drawing the ball of color i is λ_i/λ, where $\lambda = \sum_{i=1}^{s} \lambda_i$. Note that in the case of Poisson sampling scheme, the sample size is a Poisson random variable with mean $\lambda\nu$. The Poisson scheme is natural sampling model in ecology when the biologist counts species that he meets during fixed time interval $[0, \nu]$.

Considering the conditional joint distribution of N_i given $\sum_{i=1}^{s} N_i = n$ in the case of Poisson scheme, we get

$$\mathbf{P}(\bigcap_{i=1}^{s}(N_i = n_i)|N = n) = \frac{\prod_{i=1}^{s} \frac{(\lambda_i \nu)^{n_i} e^{-\lambda_i \nu}}{n_i!}}{\frac{(\lambda\nu)^n e^{-\lambda\nu}}{n!}} = n! \prod_{i=1}^{s} \frac{\left(\frac{\lambda_i}{\lambda}\right)^{n_i}}{n_i!}.$$

After substitution $p_i = \lambda_i/\lambda$, we obtain a multinomial distribution. Therefore, conditionally on the sample size, the color counts N_i are multinomially distributed.

The difference between two schemes is that in the Poisson scheme counts N_i are independent. However, in the multinomial scheme, the covariance between N_i and N_j $(i \neq j)$ is negative, since the sum $\sum N_i$ is constrained to n. This covariance equals

$$\operatorname{cov}(N_i, N_j) = -np_ip_j. \tag{2}$$

Provided p_i's are small, the covariances (2) are close to zero and both schemes are approximately equivalent.

Both the multinomial and the Poisson schemes are discussed in articles dealing with coverage problems. The Poisson scheme is sometimes preferred for its mathematical simplicity.

2.2 Definitions

Let us define some terms that will be used further.

The set $\{p_i\}_{i=1}^{s}$ (or simply $\{p_i\}$) of relative frequencies of the classes in the population is called the ***color probabilities*** or the ***color distribution***.

The number of colors, which are represented in the sample by at least one ball, is called the ***sample colority***. In the case of multinomial sampling scheme the sample colority is denoted by V_n, where n is the sample size.

The sequence $\{V_n | n \in \{1, 2, \ldots\}\}$ of successive colorities may be regarded as the discrete-time random process (more precisely, a counting process). The colority of a sample may be written down as the sum of random indicators

$$V_n = \sum_{i=1}^{s} I_i^n, \tag{3}$$

where

$$I_i^n = \begin{cases} 1 & \text{if the color } i \text{ is represented in the sample of size } n, \\ 0 & \text{otherwise.} \end{cases}$$

In the case of Poisson scheme the sample colority at the fixed time ν is denoted by V_ν. Hence, we may consider a continuous-time counting process $\{V_\nu | \nu \in (0, \infty)\}$. For the case of Poisson scheme the colority can be expressed in a similar way

$$V_\nu = \sum_{i=1}^{s} I_i^\nu, \tag{4}$$

where

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

When adding an object to the sample, the sample colority either increases by 1 or stays the same. Therefore, every realization of the processes V_n and V_ν is a nondecreasing step-function, with step heights 1. These functions are called the ***colority curves***.

By the ***coverage of a sample*** we mean the sum of probabilities of colors, which are represented in the sample. Notation of the sample coverage depends on the sampling scheme available. In the case of multinomial scheme the coverage is denoted by C_n, and in the case of Poisson scheme by C_ν. According to the definition, the sample coverage can be expressed as

$$C_n = \sum_{i=1}^{s} p_i I_i^n \qquad \text{(the multinomial scheme)}, \tag{5}$$

$$C_\nu = \sum_{i=1}^{s} p_i I_i^\nu \qquad \text{(the Poisson scheme)}. \tag{6}$$

In the following section we will derive the means and the variances of the sample colority and the coverage.

2.3 Mean and variance of sample colority and coverage in the case of known color probabilities

2.3.1 Mean of sample colority and coverage

As it was seen in (3), (4), (5) and (6), both the colority and coverage are linear combinations of random indicators, either I_i^n or I_i^ν $(i = 1, \ldots, s)$. To find the mean of the colority and the coverage, we must first find the means of these indicators. The indicator I_i^n equals to 1 if there is at least one ball of color i in the sample. Thus,

$$\begin{aligned} &\mathbf{P}(I_i^n = 0) = \mathbf{P}(\text{no balls of color } i \text{ in the sample}) = (1 - p_i)^n, \\ &E(I_i^n) = \mathbf{P}(I_i^n = 1) = 1 - (1 - p_i)^n. \end{aligned} \tag{7}$$

The expression $(1 - p_i)^n$ is the probability that in n independent trials an event "the ball of color i is drawn" does not occur at any trial. If the probability p_i is close to zero and the number of trials n is large, then we may apply approximation by the Poisson distribution

$$\mathbf{P}(0 \text{ out of } n \text{ events occur}) = (1 - p_i)^n \approx \frac{np_i^0}{0!} e^{-np_i} = e^{-np_i} \tag{8}$$

and, hence

$$E(I_i^n) \approx 1 - e^{-np_i}. \tag{9}$$

In the Poisson scheme, the indicator I_i^ν equals to 1 if there is at least one occurrence of a Poisson process ζ_i up to the time ν. It follows that

$$\mathbf{P}(I_i^\nu = 0) = \mathbf{P}(\text{no balls of color } i \text{ up to the time } \nu) = \frac{(\lambda_i \nu)^0}{0!} e^{-\lambda_i \nu} = e^{-\lambda_i \nu},$$
$$E(I_i^\nu) = \mathbf{P}(I_i^\nu = 1) = 1 - e^{-\lambda_i \nu}. \tag{10}$$

Now, based on (7) and (10), we find the mean of sample colority:

(a) in the multinomial scheme

$$E(V_n) = \sum_{i=1}^{s} [1 - (1 - p_i)^n], \tag{11}$$

(b) in the Poisson scheme

$$E(V_\nu) = \sum_{i=1}^{s} (1 - e^{-\lambda_i \nu}). \tag{12}$$

Applying approximation (9) to (11), we get

$$E(V_n) \approx \sum_{i=1}^{s} (1 - e^{-np_i}). \tag{13}$$

Similarly, we find the expectation of the sample coverage

(a) in the multinomial scheme

$$E(C_n) = \sum_{i=1}^{s} p_i [1 - (1 - p_i)^n], \tag{14}$$

(b) in the Poisson scheme

$$E(C_\nu) = \sum_{i=1}^{s} p_i (1 - e^{-\lambda_i \nu}). \tag{15}$$

If p_i's are small, then approximation (9) may be applied to (14), giving us

$$E(C_n) \approx \sum_{i=1}^{s} p_i (1 - e^{-np_i}). \tag{16}$$

2.3.2 Variance of sample colority and coverage

In this paragraph we will find the variances of sample colority and coverage. Both the colority and the coverage are linear combinations of indicators I_i^n or I_i^ν. The variance of a linear combination of some random variables X_i $(i = 1, \ldots, s)$ expresses as

$$D\left(\sum_{i=1}^{s} a_i X_i\right) = \sum_{i=1}^{s} a_i^2 DX_i + 2\sum_{i=1}^{s-1}\sum_{j=i+1}^{s} a_i a_j \mathrm{cov}(X_i, X_j).$$

In order to find variances of colority and coverage, we need to obtain

1. the variances $D(I_i^n)$ and $D(I_i^\nu)$,
2. the covariances $\mathrm{cov}(I_i^n, I_j^n)$ and $\mathrm{cov}(I_i^\nu, I_j^\nu)$, $i \neq j$.

Multinomial scheme. First we find the variance DI_i^n:

$$\begin{aligned} D(I_i^n) &= E((I_i^n)^2) - (E(I_i^n))^2 = E(I_i^n) - (E(I_i^n))^2 \\ &= E(I_i^n)(1 - E(I_i^n)) = (1 - (1 - p_i)^n)(1 - p_i)^n. \end{aligned}$$

Next we find the covariance $\mathrm{cov}(I_i^n, I_j^n)$

$$\begin{aligned} \mathrm{cov}(I_i^n, I_j^n) &= E(I_i^n I_j^n) - (E(I_i^n))(E(I_j^n)) \qquad (17) \\ &= \mathbf{P}(I_i^n = 1 \cap I_j^n = 1) - (1 - (1 - p_i)^n)(1 - (1 - p_j)^n). \end{aligned}$$

We expand the probability $\mathbf{P}(I_i^n = 1 \cap I_j^n = 1)$ using the rule

$$\mathbf{P}(A \cap B) = 1 - \mathbf{P}(\bar{A} \cup \bar{B}) = 1 - \mathbf{P}(\bar{A}) - \mathbf{P}(\bar{B}) + \mathbf{P}(\bar{A} \cap \bar{B})$$

getting

$$\begin{aligned} &\mathbf{P}(I_i^n = 1 \cap I_j^n = 1) = \\ &1 - \mathbf{P}(I_i^n = 0) - \mathbf{P}(I_j^n = 0) + \mathbf{P}(I_i^n = 0 \cap I_j^n = 0). \qquad (18) \end{aligned}$$

The $\mathbf{P}(I_i^n = 0 \cap I_j^n = 0)$ is the probability that none of the n individuals of the sample belong to classes i or j. Therefore

$$\mathbf{P}(I_i^n = 0 \cap I_j^n = 0) = (1 - p_i - p_j)^n.$$

Finally, we obtain the expression of the covariance (17):

$$\begin{aligned}
\mathrm{cov}(I_i^n, I_j^n) &= 1 - \mathbf{P}(I_i^n = 0) - \mathbf{P}(I_j^n = 0) + \mathbf{P}(I_i^n = 0 \cap I_j^n = 0) \\
&- (1 - (1 - p_i)^n)(1 - (1 - p_j)^n) \\
&= 1 - (1 - p_i)^n - (1 - p_j)^n + (1 - p_i - p_j)^n \\
&- (1 - (1 - p_i)^n)(1 - (1 - p_j)^n),
\end{aligned}$$

which simplifies to

$$\mathrm{cov}(I_i^n, I_j^n) = (1 - p_i - p_j)^n - (1 - p_i)^n (1 - p_j)^n. \tag{19}$$

Note that the covariance (19) is always negative, because

$$(1 - p_i - p_j) < (1 - p_i)(1 - p_j)$$

and both p_i and p_j are nonzero.

From (17) and (19) we can derive the variances $D(V_n)$ and $D(C_n)$:

$$\begin{aligned}
D(V_n) &= \sum_{i=1}^{s} ((1 - p_i)^n - (1 - p_i)^{2n}) \\
&+ 2 \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} ((1 - p_i - p_j)^n - (1 - p_i)^n (1 - p_j)^n),
\end{aligned} \tag{20}$$

$$\begin{aligned}
D(C_n) &= \sum_{i=1}^{s} p_i^2 ((1 - p_i)^n - (1 - p_i)^{2n}) \\
&+ 2 \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} p_i p_j ((1 - p_i - p_j)^n - (1 - p_i)^n (1 - p_j)^n).
\end{aligned} \tag{21}$$

Poisson scheme. Analogously to (17) we find the variance of indicator I_i^ν:

$$D(I_i^\nu) = E(I_i^\nu)(1 - E(I_i^\nu)) = (1 - e^{-\lambda_i \nu})e^{-\lambda_i \nu}.$$

In the Poisson scheme, the sample frequencies N_i of colors are independent. The indicators $I_i^n = I_{N_i>0}$ are also independent as functions of independent random variables. This means that the covariances $\mathrm{cov}(I_i^n, I_j^n)$ are equal to 0 and so the variances $D(V_\nu)$ and $D(C_\nu)$ take a simpler form as compared to the multinomial scheme:

$$D(V_\nu) = \sum_{i=1}^{s} e^{-\lambda_i \nu}(1 - e^{-\lambda_i \nu}),$$

$$D(C_\nu) = \sum_{i=1}^{s} p_i^2 e^{-\lambda_i \nu}(1 - e^{-\lambda_i \nu}).$$

Approximated multinomial scheme. If the probabilities p_i are small and the sample size is large then the binomial distribution is approximated well by the Poisson distribution and the formula (8) is accurate. When applying this formula to the expression (20), the variance of the sample colority becomes approximately

$$\begin{aligned} D(V_n) &\approx \sum_{i=1}^{s} e^{-np_i}(1 - e^{-np_i}) + 2\sum_{i=1}^{s-1}\sum_{j=i+1}^{s} (e^{-n(p_i+p_j)} - e^{-np_i}e^{-np_j}) \\ &= \sum_{i=1}^{s} e^{-np_i}(1 - e^{-np_i}), \end{aligned} \tag{22}$$

since the approximated covariances vanish to zero. Similarly, with the approximation (8), the variance (21) of the sample coverage is

$$\begin{aligned} D(C_n) &\approx \sum_{i=1}^{s} p_i^2 e^{-np_i}(1 - e^{-np_i}) + 2\sum_{i=1}^{s-1}\sum_{j=i+1}^{s} p_i p_j (e^{-n(p_i+p_j)} - e^{-np_i}e^{-np_j}) \\ &= \sum_{i=1}^{s} p_i^2 e^{-np_i}(1 - e^{-np_i}). \end{aligned} \tag{23}$$

The approximative expressions of $E(V_n)$, $E(C_n)$, $D(V_n)$ and $D(C_n)$ for the multinomial scheme are very similar to the corresponding expressions for the Poisson scheme. Based on this similarities, we conclude that the multinomial scheme can be approximated by the Poisson scheme with intensities λp_i where the sample is drawn until the time n/λ (here λ is an arbitrary positive number).

Example 1. (Case of equiprobable colors). Here we find expressions of mean and variance of sample colority and coverage in the case of one simple color distribution. This is the distribution, where all the colors are equiprobable, i. e. have equal probabilities ($p_i = 1/s$, $i = 1, \ldots, s$). We will assume the multinomial sampling scheme. According to (11), the mean colority in this case is

$$E(V_n) = s(1 - (1 - 1/s)^n).$$

According to (13), the approximated mean colority equals

$$E(V_n) \approx s(1 - e^{-n/s}). \tag{24}$$

By (14), the mean sample coverage equals

$$E(C_n) = (1 - (1 - 1/s)^n),$$

and the approximated value is

$$E(C_n) \approx 1 - e^{-n/s}. \tag{25}$$

By (20), the variance of the colority expresses as

$$D(\nu_n) \;=\; s\left((1 - \frac{1}{s})^n - (1 - \frac{1}{s})^{2n}\right) + s(s-1)\left((1 - \frac{2}{s})^n - (1 - \frac{1}{s})^{2n}\right)$$

This is an exact result. From the other side, the approximative formula (22) gives us

$$D(V_n) \approx se^{-n/s}(1 - e^{-n/s}). \tag{26}$$

By (21), the variance of coverage then equals

$$\begin{aligned} D(C_n) &= D(\nu_n/s) = D(\nu_n)/s^2 \\ &= \frac{1}{s}\left((1-\frac{1}{s})^n - (1-\frac{1}{s})^{2n}\right) + (1-\frac{1}{s})\left((1-\frac{2}{s})^n - (1-\frac{1}{s})^{2n}\right) \end{aligned}$$

or using the approximation (23):

$$D(C_n) \approx \frac{1}{s}e^{-n/s}(1-e^{-n/s}).$$

■

3 Two ways of defining distribution of color probabilities

We have seen in the previous chapter that in order to estimate the sample colority and coverage we need to know the probabilities of all colors in population. In the most of cases, however, the color distribution is unknown. The basic idea to overcome this difficulty is to assume that the set $\{p_i\}$ is defined by a small number of parameters and then to estimate these parameters. Two different approaches of defining the color probabilities are discussed in this chapter. One approach is to define the set $\{p_i\}$ by some function of i and the other is to define $\{p_i\}$ by some parametric density function.

3.1 Direct definition of color probabilities

Probabilities p_i of colors $i = 1, \ldots, s$ can be given by some function $\pi(i)$ of color number i, so that

$$p_i = \pi(i). \tag{27}$$

Function $\pi(\cdot)$ may also depend on some vector $\vec{\theta}$ of parameters. Between such parameters, one compulsory parameter is the number s of colors in the population. Without any loss of generality we will further assume that $\pi(\cdot)$ is a nondecreasing function. Next we provide some simple examples of different functions $\pi(\cdot)$.

Example 2. Uniform color probabilities is the simplest case of color probabilities:

$$\pi(i) = 1/s, \quad i = 1, \ldots, s.$$

On the Figure 2 this type of color probabilities is referred to as *CONST*.

■

Example 3. Piecewise constant probabilities. Suppose that the set of color numbers $\{1, \ldots, s\}$ can be divided into m classes $C_1, \ldots, C_m$, so that in each class $\pi(i)$ has constant value v_j, $j = 1, \ldots, m$:

$$\pi(i) = v_j, \quad i \in C_j, \quad j = 1, \ldots, m.$$

This is extremely wide class of functions. Many other function may be approximated by a piecewise constant function. On the Figure 2 one function of this type is referred to as *PIECE*.

■

Example 4. Linearly decreasing color probabilities are defined by the function

$$\pi(i) = p_0 - ai, \quad a > 0, \; i = 1, \ldots, s.$$

It suffices, when we fix only one parameter of p_0 and a, because the other is obtainable, when we account for constraint $\sum_{i=1}^{s} \pi(i) = 1$. One example of linearly decreasing function of probabilities is shown on the Figure 2 and referred to as *LINEAR*.

■

Example 5. Exponentially decreasing color probabilities are defined by the function

$$\pi(i, q) = p_0(q)q^i, \quad q < 1, \quad i = 1, \ldots, s.$$

Thus, the color probabilities compose a truncated geometric series with common ratio q and $p_0(q) = 1/\sum_{i=1}^{s} q^i$ is the coefficient, required to standardize

p_i's to add up to unity. Two special cases ($q = 0.95$ and 0.98) of this type of color probabilities are shown on the Figure 2. These functions are referred to as *EXP95* and *EXP98*.

■

Example 6. Inverse color probabilities are defined by the function

$$\pi(i) = p_0/i, \quad i = 1, \ldots, s.$$

Here, $p_0 = 1/\sum_{i=1}^{s} i^{-1}$ is the standardizing coefficient. This type of color probabilities is shown on the Figure 2, where it is referred to as *INV*.

■

Example 7. Quadratically decreasing probabilities are defined by

$$\pi(i) = p_0(s - i + 1)^2, \quad i = 1, \ldots, s,$$

where $p_0(s) = 1/\sum_{i=1}^{s} i^2$. The base number is $s - i + 1$ instead of i because we want the function to be monotonely decreasing. On the Figure 2 the plot of probabilities, defined by this function is referred to as *SQR*.

■

All the functions described in the examples above, are plotted on the Figure 2, provided that number of colors in population equals $s = 200$.

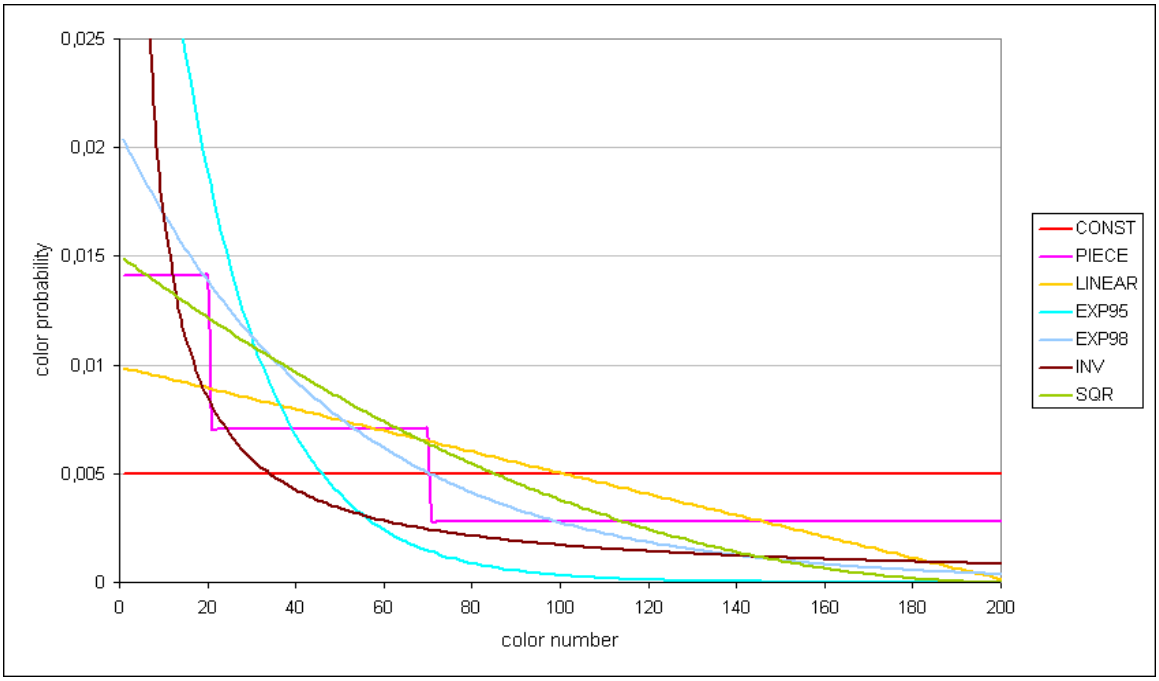

Figure 2: Plots of different types of color probabilities

3.2 Defining color probabilities by density function

Here we provide an alternative method of defining the color probabilities first described in [2]. In this method, the set of color probabilities is given by some density function $f(p)$ that satisfies the two following conditions:

(i) $s_1 := \int_{-\infty}^{\infty} \frac{f(p)}{p} dp < \infty$,

(ii) $\int_{-\infty}^{a} f(p)dp > 0$, where a satisfies the equation

$$\int_{-\infty}^{a} \frac{f(p)}{p} dp = \begin{cases} 1, & \tilde{s} = s_1 \\ \tilde{s} - s_1, & \tilde{s} > s_1 \end{cases}, \tag{28}$$

where $\tilde{s}$ is the smallest integer for which $\tilde{s} \geq s_1$.

The condition (i) guarantees that we get the finite number of probabilities. The condition (ii) ensures that all obtained probabilities are positive.

The set $\{p_i\}$ of color probabilities is obtained from $f(p)$ using the following Procedure 1.

Procedure 1. The procedure for defining a unique set of color probabilities by a density function

1. Start by giving a density function $f(p)$ satisfying conditions (i) and (ii).

2. Define the function $g(p) = f(p)/p$. By agreement, the value of $g(p)$ at the point $p = 0$ is replaced by the limit

$$\lim_{p \to 0} \frac{f(p)}{p}.$$

Let $\tilde{s}$ be the smallest integer satisfying $\tilde{s} \geq s_1$, where

$$s_1 := \int_{-\infty}^{\infty} g(p)dp.$$

Due to condition (i), s_1 is finite and, consequently, $\tilde{s}$ is also finite.

3. Let $m = \inf\{p|f(p) > 0\}$ and $M = \sup\{p|f(p) > 0\}$ Define the partition

$$m = \xi_{\tilde{s}} < \xi_{\tilde{s}-1} < \ldots < \xi_0 = M$$

of the interval $(-\infty, \infty)$ so that

$$\int_{\xi_i}^{\xi_{i-1}} g(p)dp = 1 \ (i = 1, \ldots, \tilde{s} - 1). \tag{29}$$

It means that the area under the curve $g(p)$ is divided into $\tilde{s}$ regions of area 1 (except for maybe region bounded by interval $[\xi_{\tilde{s}}, \xi_{\tilde{s}-1}]$). The integral over the leftmost interval $[\xi_{\tilde{s}}, \xi_{\tilde{s}-1}]$ equals to 1 when s_1 is integer and equals to $\tilde{s} - s_1$ (satisfying $0 < \tilde{s} - s_1 < 1$) otherwise. Hence,

$$0 < \int_{\xi_{\tilde{s}}}^{\xi_{\tilde{s}-1}} g(p)dp \leq 1.$$

Here $g(p)$ can be considered as a conditional density of p inside the intervals $[\xi_i, \xi_{i-1}]$ $(i = 1, \ldots, \tilde{s} - 1)$.

4. Define the color probabilities p_i $(i = 1, \ldots, \tilde{s})$ by the integral

$$p_i := \int_{\xi_i}^{\xi_{i-1}} pg(p)dp = \int_{\xi_i}^{\xi_{i-1}} f(p)dp. \tag{30}$$

The p_i can be considered a conditional expectation of p on the interval $[\xi_i, \xi_{i-1}]$ (except for maybe interval $[\xi_{\tilde{s}}, \xi_{\tilde{s}-1}]$, since the integral of $g(p)$ over this interval may be less than 1).

■

Obtained probabilities p_i form a decreasing sequence

$$p_1 > p_2 > \ldots > p_{\tilde{s}}.$$

Let us show that the condition (ii) guarantees that $p_{\tilde{s}} > 0$. Number $\xi_{\tilde{s}-1}$ satisfies the equation (28), since

$$\int_{-\infty}^{\xi_{\tilde{s}-1}} g(p)dp = \int_{\xi_{\tilde{s}}}^{\xi_{\tilde{s}-1}} g(p)dp = \begin{cases} 1, & \tilde{s} = s_1 \\ \tilde{s} - s_1, & \tilde{s} > s_1 \end{cases},$$

this means that the condition (ii) is equivalent to

$$\int_{\xi_{\tilde{s}}}^{\xi_{\tilde{s}-1}} f(p)dp = p_{\tilde{s}} > 0.$$

The condition (ii) is not needed to be checked if $f(p) = 0$ on the interval $(-\infty, 0)$, because in this case $\xi_{\tilde{s}} = m = \inf\{p | f(p) > 0\}$ is non-negative and $p_{\tilde{s}}$ is positive, being a conditional expectation of some random variable X given $X \in [\xi_{\tilde{s}}, \xi_{\tilde{s}-1}]$.

Furthermore, we see that $p_1 + p_2 + \ldots + p_{\tilde{s}} = 1$, since

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

and hence the color distribution of the population is uniquely given by the density function $f(p)$.

3.3 How to select density f that produces a given set of color probabilities

It was shown that the set $\{p_i\}$ of color probabilities can be defined either by a function of color number i, or by a density function. Suppose that we have defined the set $\{p_i\}$ directly by the function $\pi(i)$. Then the question is: "can we find a density f that produces the same set of color probabilities?". One form of such density f, which is simple from the computational point of view, is proposed in the following procedure. This type of density function, however, cannot be evaluated for all the sets of color probabilities. Furthermore, it will be seen, that such a function f is not unique.

3.3.1 The case of approximately linearly decreasing color probabilities

Procedure 2. The procedure for obtaining a density function that produces a given color distribution

1. The set $\{p_i\}_{i=1}^{s}$ of probabilities is given, where

$$p_1 > p_2 > \ldots > p_s.$$

2. Find the partition (assuming its existence at the moment)

$$\xi_s < \xi_{s-1} < \ldots < \xi_0 \tag{31}$$

of real axis, such that $p_i = \frac{\xi_i + \xi_{i-1}}{2}$, $i = 1, \ldots, s$, i.e. p_i is the midpoint of the interval $[\xi_i, \xi_{i-1}]$.

3. The density function, that generates the set $\{p_i\}_{i=1}^s$ is then

$$f(p) = \begin{cases} p/(\xi_0 - \xi_1), & p \in [\xi_1, \xi_0) \\ p/(\xi_1 - \xi_2), & p \in [\xi_2, \xi_1) \\ \vdots & \vdots \\ p/(\xi_{s-1} - \xi_s), & p \in [\xi_s, \xi_{s-1}] \\ 0, & p \in (-\infty, \xi_s) \cup (\xi_0, \infty) \end{cases}.$$

■

When applying the Procedure 1 to the function $f(p)$, we obtain exactly the same set $\{p_i\}_{i=1}^s$ of color probabilities, since

$$\int_{\xi_i}^{\xi_{i-1}} f(p)dp = \int_{\xi_i}^{\xi_{i-1}} \frac{p}{\xi_{i-1} - \xi_i} dp = \frac{\xi_{i-1}^2 - \xi_i^2}{2(\xi_{i-1} - \xi_i)} = \frac{\xi_{i-1} + \xi_i}{2} = p_i$$

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

The necessary condition for Procedure 2 to work is that the partition (31) in step 2 of the procedure exists. Necessary and sufficient condition for Procedure 2 to work is that the partition (31) in step 2 of the procedure exists, i. e. the system of equations and inequalities

$$\begin{cases} x_0 + x_1 = 2p_1 \\ x_1 + x_2 = 2p_2 \\ \vdots \\ x_{s-1} + x_s = 2p_s \\ x_{i-1} < x_i, \quad i = 1, \ldots, s \end{cases} \tag{32}$$

has at least one solution. In the following example the procedure works successfully.

Example 8. Let us find the function f that produces the exponentially decreasing set of probabilities where $q = 0.5$ and $s = 5$. In this case color

probabilities express as

$$p_i = \frac{0.5^i}{\sum_{i=1}^{5} 0.5^i} = \frac{0.5^i}{1 - 0.5^5}, \quad i = 1, \ldots, 5.$$

Numerically these probabilities equal

$$p_1 = \frac{31}{64},\ p_2 = \frac{31}{128},\ p_3 = \frac{31}{256},\ p_4 = \frac{31}{512},\ p_5 = \frac{31}{1024}.$$

First we solve the system (32) for general p_i's. The system then takes the following form

$$\begin{cases} x_{i-1} + x_i = 2p_i \quad i = 1, \ldots, 5 \\ x_{i-1} < x_i \quad i = 1, \ldots, 5 \end{cases} \tag{33}$$

The system $x_{i-1} + x_i = 2p_i$ $(i = 1, \ldots, 5)$ of linear equations has infinitely many solutions, since the number of unknowns exceeds the number of equations by one. It means that four unknowns can be expressed through the remaining one. Express for example $x_0, \ldots, x_4$ through x_5:

$$\begin{cases} x_4 = -x_5 + 2p_5 \\ x_3 = x_5 + 2(p_4 - p_5) \\ x_2 = -x_5 + 2(p_3 - p_4 + p_5) \\ x_1 = x_5 + 2(p_2 - p_3 + p_4 - p_5) \\ x_0 = -x_5 + 2(p_1 - p_2 + p_3 - p_4 + p_5) \end{cases}.$$

After substitution of p_i's we get

$$\begin{cases} x_4 = -x_5 + \frac{31}{512} \\ x_3 = x_5 + \frac{31}{512} \\ x_2 = -x_5 + \frac{93}{512} \\ x_1 = x_5 + \frac{155}{512} \\ x_0 = -x_5 + \frac{341}{512} \end{cases}. \tag{34}$$

Besides that inequalities in system (33) must be satisfied. If we account for equations (34) then inequalities transform to the following system

$$\begin{cases} -2x_5 + \frac{31}{512} > 0 \\ 2x_5 > 0 \\ -2x_5 + \frac{62}{512} > 0 \\ 2x_5 + \frac{62}{512} > 0 \\ -2x_5 + \frac{186}{512} > 0 \end{cases}, \tag{35}$$

which has the solution $x_5 \in (0, \frac{31}{1024})$. Hence the system (33) is equivalent to

$$\begin{cases} x_4 = -x_5 + \frac{31}{512} \\ x_3 = x_5 + \frac{31}{512} \\ x_2 = -x_5 + \frac{93}{512} \\ x_1 = x_5 + \frac{155}{512} \\ x_0 = -x_5 + \frac{341}{512} \\ 0 < x_5 < \frac{31}{1024} \end{cases}. \tag{36}$$

The system (36) has infinitely many solutions. One of the solutions is, for example (all fractions are rounded up to 3 decimal places),

$$x_5 = 0.025,\ x_4 = 0.036,\ x_3 = 0.086,\ x_2 = 0.157,\ x_1 = 0.328,\ x_0 = 0.641.$$

Now we can construct the density function f, that produces the set $\{p_i\}$. Required f expresses as

$$f(p) = \begin{cases} 3.19p, & p \in [0.328, 0.641] \\ 5.84p, & p \in [0.157, 0.328) \\ 14.1p, & p \in [0.086, 0.157) \\ 20.0p, & p \in [0.036, 0.086) \\ 95.2p, & p \in [0.025, 0.036) \\ 0, & p \in (-\infty, 0.025) \cup (0.641, \infty) \end{cases} \tag{37}$$

The plot of the function f is presented in Figure 3.

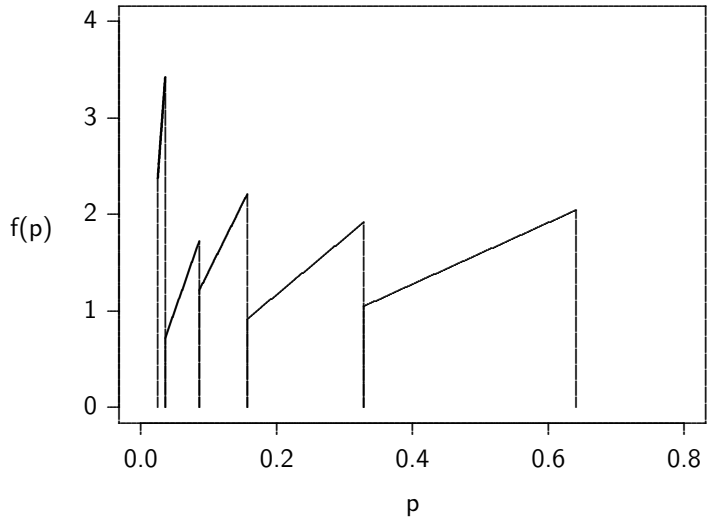

Figure 3: The function $f(p)$ that generates the given set of color probabilities

■

In the following lemma we provide the condition that guarantees a successful work of Procedure 2. Before the formulating the lemma we will introduce the notation

$$\Delta_j := p_j - p_{j+1}, \quad j = 1, \ldots, s-1.$$

In the lemma we will also use the following definition.

Definition 1. The set of color probabilities, that satisfies the condition

$$\sum_{i=0}^{2l} (-1)^i \Delta_{k+i} > 0, \quad k, l \in \{1, 2, \ldots\}, \quad 2l + k \leq s - 1 \tag{38}$$

is called an ***nearly linearly decreasing*** set of color probabilities.

■

The intuitive meaning of the condition (38) is that the sequence $\{\Delta_j\}_{j=1}^{s-1}$ is not allowed to increase or decrease "too quickly". The "ideal" sequence in this sense, is the constant sequence

$$\Delta_j \equiv \Delta, \quad j = 1, \ldots, s-1.$$

In this case, the sequence $\{p_i\}_{i=1}^{s}$ is a linearly decreasing sequence.

Lemma 1. *Suppose we are given the set $\{p_i\}_{i=1}^{s}$ of probabilities. Then Procedure 2, applied to the set $\{p_i\}_{i=1}^{s}$, produces a density function f if and only if the set $\{p_i\}_{i=1}^{s}$ is nearly linearly decreasing.*

Proof: The Procedure 2 produces some density function if and only if the system (32) has a solution. If we express $x_0, \ldots, x_{s-1}$ through x_s from the system (32) then we get

$$\begin{cases} x_{s-1} = 2p_s - x_s \\ x_{s-2} = 2p_{s-1} - 2p_s + x_s \\ x_{s-3} = 2p_{s-2} - 2p_{s-1} + 2p_s - x_s \\ \vdots \\ x_{i-1} = 2\sum_{j=i}^{s}(-1)^{j-i}p_j - (-1)^{s-i+1}x_s \\ \vdots \\ x_0 = 2\sum_{j=1}^{s}(-1)^{j-1}p_j - (-1)^s x_s \end{cases} . \tag{39}$$

Hence the difference $x_{i-1} - x_i$ expresses as follows

$$\begin{aligned} x_{i-1} - x_i &= 2\sum_{j=i}^{s}(-1)^{j-i}p_j - (-1)^{s-i+1}x_s - 2\sum_{j=i+1}^{s}(-1)^{j-i-1}p_{j+i+1} + (-1)^{s-i}x_s \\ &= 2(\sum_{j=i}^{s-1}(-1)^{j-i}(p_i - p_{i+1}) + (-1)^{s-i}(p_s - x_s)) \\ &= 2(\sum_{j=i}^{s-1}(-1)^{j-i}\Delta_j + (-1)^{s-i}(p_s - x_s)), \quad i = 1, \ldots, s-1, \\ x_{s-1} - x_s &= 2(p_s - x_s). \end{aligned}$$

For further convenience we introduce the notation

$$S(i_1, i_2) := \sum_{j=i_1}^{i_2}(-1)^{j-i_1}\Delta_j$$

Therefore, the system (32) has a solution if and only if the following system of inequalities has a solution

$$\begin{cases} S(i, s-1) + (-1)^{s-i}(p_s - x_s) > 0, \quad i = 1, \ldots, s-1 \\ p_s - x_s > 0 \end{cases} . \tag{40}$$

Further we will consider two cases: where s is odd or even.

(a) If s is odd ($s = 2r-1$, $r = 1, 2, \ldots$), the system (40) transforms to

$$\begin{cases} x_s < p_s \\ x_s > p_s - S(s-1, s-1) \\ x_s < p_s + S(s-2, s-1) \\ x_s > p_s - S(s-3, s-1) \\ \vdots \\ x_s < p_s + S(3, s-1) \\ x_s > p_s - S(2, s-1) \\ x_s < p_s + S(1, s-1) \end{cases} . \tag{41}$$

Denote

$$M_0 := p_s, \quad M_x := p_s + S(s-2x, s-1), \quad x = 1, \ldots, r-1$$

and

$$m_y := p_s - S(s-2y+1, s-1), \quad y = 1, \ldots, r-1.$$

The system (41) is equivalent to

$$\begin{cases} x_s < M_x, & x = 0, 1, \ldots, r-1 \\ x_s > m_y, & y = 1, 2, \ldots, r-1 \end{cases}. \tag{42}$$

The system (42) has a solution if and only if

$$\max\{m_y | y = 1, \ldots, r-1\} \leq \min\{M_x | x = 0, \ldots, r-1\}. \tag{43}$$

It is clear, that (43) is equivalent to the condition

$$M_x - m_y > 0, \quad x = 0, \ldots, r-1, \quad y = 1, \ldots, r-1. \tag{44}$$

The difference $M_x - m_y$ expresses as follows

$$M_x - m_y = \begin{cases} S(s-2x, s-2y), & x \geq y \\ S(s-(2y-1), s-(2x+1)), & x < y \end{cases}. \tag{45}$$

Hence, (44) is equivalent to the system

$$\begin{cases} \sum_{j=s-2x}^{s-2y} (-1)^{j-(s-2x)} \Delta_j > 0, & x \geq y, \\ \sum_{j=s-(2y-1)}^{s-(2x+1)} (-1)^{j-(s-2y+1)} \Delta_j > 0, & x < y \end{cases}, \tag{46}$$

which is equivalent, in turn, to

$$\begin{cases} \sum_{j=0}^{2(x-y)} (-1)^j \Delta_{j+s-2x} > 0, & x \geq y, \\ \sum_{j=0}^{2(y-x-1)} (-1)^j \Delta_{j+s-2y+1} > 0, & x < y \end{cases}. \tag{47}$$

Substitute the upper limit of summation in each sum of (47) by $2l$ and the first summand by Δ_k. Then the system of inequalities (47) is equivalent to the following system

$$\begin{cases} \sum_{j=0}^{2l}(-1)^j\Delta_{j+k} > 0, \quad l \in \{0,1,\ldots\},\ k \in \{1,3,5,\ldots\},\ 2l+k \le s-2 \\ \sum_{j=0}^{2l}(-1)^j\Delta_{j+k} > 0, \quad l \in \{0,1,\ldots\},\ k \in \{2,4,6,\ldots\},\ 2l+k \le s-1 \end{cases}. \tag{48}$$

The system (48) can be summarized like follows

$$\sum_{j=0}^{2l}(-1)^j\Delta_{j+k} > 0,\ l \in \{0,1,\ldots\},\ k \in \{1,2,3,\ldots\},\ 2l+k \le s-1. \tag{49}$$

Inequalities $\Delta_k > 0$ of the system (49), which correspond to the case $l = 0$ are always satisfied and may be left out from the system. Hence, the (49) is equivalent to the condition (38).

(b) if s is even ($s = 2r$, $r = 1, 2, \ldots$), the system (40) transforms to

$$\begin{cases} x_s < p_s \\ x_s > p_s - S(s-1, s-1) \\ x_s < p_s + S(s-2, s-1) \\ x_s > p_s - S(s-3, s-1) \\ \vdots \\ x_s > p_s - S(3, s-1) \\ x_s < p_s + S(2, s-1) \\ x_s > p_s - S(1, s-1) \end{cases}. \tag{50}$$

Denote

$$M_0 := p_s, \quad M_x := p_s + S(s-2x, s-1), \quad x = 1, \ldots, r-1$$

and

$$m_y := p_s - S(s-2y+1, s-1), \quad y = 1, \ldots, r.$$

The system (50) is equivalent to

$$\begin{cases} x_s < M_x, & x = 0, 1, \ldots, r-1 \\ x_s > m_y, & y = 1, 2, \ldots, r \end{cases} . \tag{51}$$

The latter system has a solution if and only if

$$M_x - m_y > 0, \quad x = 0, \ldots, r-1, \quad y = 1, \ldots, r. \tag{52}$$

The difference $M_x - m_y$ expresses as (45). As in the case of odd s, we conclude, that (52) is equivalent to the system (47). After making the same substitution in (47), like in the case of odd s we see that the system of inequalities (47) is equivalent to the following system

$$\begin{cases} \sum_{j=0}^{2l}(-1)^j \Delta_{j+k} > 0, & l \in \{0, 1, \ldots\}, \ k \in \{1, 3, 5, \ldots\}, \ 2l + k \leq s - 1 \\ \sum_{j=0}^{2l}(-1)^j \Delta_{j+k} > 0, & l \in \{0, 1, \ldots\}, \ k \in \{2, 4, 6, \ldots\}, \ 2l + k \leq s - 2 \end{cases} . \tag{53}$$

The system (53) can be summarized as follows

$$\sum_{j=0}^{2l}(-1)^j \Delta_{j+k} > 0, \ l \in \{0, 1, \ldots\}, \ k \in \{1, 2, 3, \ldots\}, \ 2l+k \leq s-1. \tag{54}$$

Inequalities $\Delta_k > 0$ of the system (54), which correspond to the case $l = 0$ are always satisfied and may be left out from the system. Hence, the (54) is equivalent to the condition (38).

Now we have shown that in the case of both even and odd s the system (40) has a solution if and only if the condition (38) is fulfilled. So we have proved that (38) is the necessary and sufficient condition for the Procedure 2 to produces a density function.

■

3.3.2 Finding approximate density for arbitrary set of color probabilities

In this paragraph we propose another form $\tilde{f}$ of density function, that produces a given set of color probabilities. However, this type of density is approximative, since it produces only approximately the same set of probabilities. In Definition 2 we will specify what is meant by "approximately equal sets of probabilities". The advantage of the density $\tilde{f}$, as compared to the density described by the Procedure 2, is that $\tilde{f}$ can be calculated for arbitrary set $\{p_i\}_{i=1}^s$ of probabilities, whenever all the color probabilities are distinct, i.e.

$$p_1 > p_2 > \ldots > p_{s-1} > p_s. \tag{55}$$

Definition 2. Consider two sets $\{p_i^{'}\}_{i=1}^s$ and $\{p_i^{''}\}_{i=1}^s$ of color probabilities. Define the functions

$$G^{'}(x) = \#\{p_i^{'} \leq x\} \quad \text{and} \quad G^{''}(x) = \#\{p_i^{''} \leq x\}.$$

Then the sets $\{p_i^{'}\}$ and $\{p_i^{''}\}$ are called ***approximately equal sets*** of color probabilities if

$$\max_{0 \leq x \leq 1} |G^{'}(x) - G^{''}(x)|$$

is small as compared to s.

■

Lemma 2. *Suppose that the set $\{p_i\}_{i=1}^s$ of color probabilities is given and that (55) is satisfied. Then the density function that produces (via Procedure 1) approximately the color distribution $\{p_i\}$ is given by*

$$\tilde{f}(p) = \begin{cases} p/(\xi_{i-1} - \xi_i), & p \in [\xi_i, \xi_{i-1}),\ i = 1, \ldots, s \\ 0, & p \in (-\infty, \xi_s) \cup [\xi_0, \infty) \end{cases},$$

where

$$\xi_i = \frac{1}{2}(p_i + p_{i+1}),\ i = 1, \ldots, s-1, \quad \xi_s = 0, \quad \xi_0 = p_1 + p_s.$$

Proof: First we show that $\tilde{f}(p)$ is a density function:

$$\begin{aligned}
\int_0^\infty \tilde{f}(p)dp &= \sum_{i=1}^{s} \int_{\xi_i}^{\xi_{i-1}} \tilde{f}(p)dp = \sum_{i=1}^{s} \left[\frac{1}{\xi_{i-1} - \xi_i} \int_{\xi_i}^{\xi_{i-1}} p dp \right] \\
&= \frac{1}{2} \sum_{i=1}^{s} \frac{\xi_{i-1}^2 - \xi_i^2}{\xi_{i-1} - \xi_i} = \frac{1}{2} \sum_{i=1}^{s} (\xi_{i-1} + \xi_i) = \frac{\xi_0 + \xi_s}{2} + \sum_{i=1}^{s-1} \xi_i \\
&= \frac{p_1 + p_s}{2} + \sum_{i=1}^{s-1} \frac{p_i + p_{i+1}}{2} = 1.
\end{aligned}$$

Therefore, the function $\tilde{f}(p)$ is indeed a density function.

Next we have to check the conditions (i) and (ii), that must be satisfied before applying the Procedure 1. Condition (i) is satisfied because integral of piecewise constant function

$$\tilde{g}(p) = \tilde{f}(p)/p = \begin{cases} 1/(\xi_{i-1} - \xi_i), & p \in [\xi_i, \xi_{i-1}),\ i = 1, \ldots, s \\ 0, & p \in (-\infty, \xi_s) \cup [\xi_0, \infty) \end{cases},$$

is finite. The condition (ii) is also satisfied, since all the probabilities produced are greater than $\xi_s = 0$.

Suppose that after applying the Procedure 1 to the function $\tilde{f}(p)$ we get the new set $\{\tilde{p}_i\}_{i=1}^{\tilde{s}}$. Integral of function $\tilde{g}(p)$ equals

$$\int_0^\infty \tilde{g}(p)dp = \sum_{i=1}^{s} \int_{\xi_i}^{\xi_{i-1}} \tilde{g}(p)dp = \sum_{i=1}^{s} \left[\frac{1}{(\xi_{i-1} - \xi_i)} \int_{\xi_i}^{\xi_{i-1}} dp \right] = \sum_{i=1}^{s} 1 = s,$$

which means that $\tilde{s} = s$ and, hence, there are s elements in the set $\{\tilde{p}_i\}$, as much as in the set $\{p_i\}$. Elements of the set $\{\tilde{p}_i\}$ equal

$$\tilde{p}_i = \int_{\xi_i}^{\xi_{i-1}} \tilde{f}(p)dp = (\xi_{i-1} + \xi_i)/2 = \begin{cases} \frac{1}{4}p_{i-1} + \frac{1}{2}p_i + \frac{1}{4}p_{i+1}, & i = 2, \ldots, s-1 \\ \frac{1}{2}p_s + \frac{3}{4}p_1 + \frac{1}{4}p_2, & i = 1 \\ \frac{1}{4}p_{s-1} + \frac{1}{4}p_s, & i = s \end{cases}.$$

Let us show that the new set $\{\tilde{p}_i\}$ is approximately the same as the set $\{p_i\}$. Define the functions

$$G(x) = \#\{p_i \le x\}, \quad \tilde{G}(x) = \#\{\tilde{p}_i \le x\} \quad \text{and} \quad \Delta G(x) = |G(x) - \tilde{G}(x)|.$$

Our purpose is to find the maximum

$$\max_{0 \le x \le 1} \Delta G(x)$$

and to show that this maximum is small compared to s.

Let us find the value of $\Delta G(x)$ in some point $x_0 \in [0, 1]$. If $x_0 > \xi_0$ then $\Delta G(x_0) = s - s = 0$, which is of course small compared to s. If $x_0 \le \xi_0$ then there exists such r that $x_0 \in [\xi_r, \xi_{r-1}]$. Introduce the following notations

$$p_{min} = \min\{p_r, \tilde{p}_r\}, \quad p_{max} = \max\{p_r, \tilde{p}_r\}.$$

Let us find all the possible values of the $\Delta G(x_0)$ depending on the mutual location of points x_0, p_{min} and p_{max} inside the interval $[\xi_r, \xi_{r-1}]$.

1. If $p_{min} = p_{max}$ then $\Delta G(x_0) = 0$ independently of location of x_0.
2. If $p_{min} < p_{max}$ then
   (a) If $x_0 \in [\xi_r, p_{min})$ then $\Delta G = |(s - r) - (s - r)| = 0$.
   (b) If $x_0 \in [p_{min}, p_{max})$ then $\Delta G = |(s - r + 1) - (s - r)| = 1$.
   (c) If $x_0 \in [p_{max}, \xi_{r-1}]$ then $\Delta G = |(s - r + 1) - (s - r + 1)| = 0$.

Therefore

$$\max_{0 \le x \le 1} \Delta G(x) = 1,$$

which is small relative to s, given that s is large. Moreover, in the process $s \to \infty$ the maximum value of $|\Delta G|$ is infinitely small. This may be written as follows

$$\max_x \Delta G(x) = o(s).$$

■

3.4 Density functions used to define color distribution

In the section 3.2 we have seen that the set $\{p_i\}$ of color probabilities can be defined by some density function f. This allows us to deal with the density f instead of the set $\{p_i\}$. Suppose that we want to estimate the coverage of a sample from formulae (3) or (4). Then it can be assumed that the density f is a member of some well-known family $f(\vec{\theta})$ of distributions, where $\vec{\theta} = (\theta_1, \ldots, \theta_m)$ is a vector of parameters. So the problem reduces to the estimating of a small number of parameters. In the Chapter 5 it will be shown how to estimate the coverage when the parameters of the density function f are already estimated. Now we introduce some types of densities used elsewhere in the literature for defining distribution of color probabilities.

1. In 1943 R. Fisher [6] proposed the Gamma distribution

$$f(p) = \frac{\alpha^{k+1}}{\Gamma(k+1)} p^k e^{-\alpha p}, \ (k > -1, \alpha > 0), \tag{56}$$

   with $k = 0$ as a density of color probabilities. The author asserted that the parameter k measures the variability, or heterogenity, of color probabilities. If the population is very heterogenous, then k must be close to zero, which corresponds to the exponential distribution.

   Later, Engen [2] used the Gamma distribution where the possible values of k were in the interval $(-1, \infty)$.

2. MacArthur [13] suggested, that, if there are s species in the population, then their relative frequencies p_i might be proportional to the lengths of the segments of a line (or stick) broken at random into s

pieces. This model is equivalent to supposing that p_i are independent and identically distributed variates having approximately exponential distribution. The exponential distribution is the special case of Gamma distribution with $k = 0$ and thus, has density

$$f(p) = \alpha e^{-\alpha p}, \ (\alpha > 0).$$

3. Bulmer [1] extended MacArthur's "broken stick" model by supposing that the stick is not broken into s pieces simultaneously, but the breakage occurs sequentially in a series of stages. If, at each stage, the law of breakage is independent of the size of the stick, then the distribution of lengths p_i of s pieces is lognormal with density

$$f(p) = \frac{1}{pd\sqrt{2\pi}} \exp\left(-\frac{(\ln(p) - \mu)^2}{2d^2}\right), \ (d > 0).$$

Still, the models, using Gamma distribution for defining the color probabilities, are mathematically more simple than lognormal model. At the same time, Gamma models are general enough to cover the most patterns of population structures. Therefore, in the next chapters we will consider the Gamma model of color probabilities.

4 Modelling color probabilities by Gamma distribution

In the previous section we have discussed how population probabilities of colors can be defined by a density function $f(p)$. Now we will consider a particular density, namely, the Gamma density as a model of color probabilities.

4.1 Derivation of Engen's Extended Negative Binomial (ENB) model

4.1.1 Parametric Poisson-Gamma model definition

Here we will demonstrate how the set of color probabilities $\{p_i\}, i = 1, 2, \ldots, s$ (s may be infinite) can be defined by a Gamma density. We use the following notation for Gamma density

$$f(x) = \frac{\alpha^\gamma}{\Gamma(\gamma)} x^{\gamma-1} e^{-\alpha x}, \ (x > 0, \ \gamma > 0, \ \alpha > 0). \tag{57}$$

In the previous chapter we have used the Procedure 1 for the defining of color probabilities. However, before applying the procedure, we must check if the conditions (i) and (ii) are satisfied. The condition (ii) is satisfied, since $f(x) > 0$ if $x \geq 0$, and, hence, the non-positive probabilities cannot be produced.

If we define the function $g(x) = f(x)/x$, which is (up to the constant multiplier α/k) another Gamma-density, we have

$$\int_0^\infty g(x)dx = \frac{\alpha^{k+1}}{\Gamma(k+1)} \int_0^\infty x^{k-1} e^{-\alpha x} dx = \begin{cases} \alpha/k & \text{if } k > 0, \\ \infty & \text{if } k \leq 0. \end{cases} \tag{58}$$

If $k > 0$ then the condition (i) is satisfied and the function $g(x)$ may be used for defining color probabilities as in Procedure 1. If $k \leq 0$ then we use the procedure similar to the Procedure 1:

1. Divide area under the curve $g(x)$ into infinitely many regions by points

$$0 < \ldots < \xi_2 < \xi_1 < \xi_0 = \infty, \tag{59}$$

so that the area of each region is 1 (except for maybe the region in the interval, i.e.

$$\int_{\xi_i}^{\xi_{i-1}} g(x)dx = 1, \quad i \geq 1. \tag{60}$$

2. Find the mean of the density $g(x)$ in each interval $[\xi_i, \xi_{i-1}]$:

$$p_i = \int_{\xi_i}^{\xi_{i-1}} xg(x)dx. \tag{61}$$

Then

$$\sum_i p_i = \int_0^\infty xg(p)dx = \int_0^\infty f(x)dx = 1 \tag{62}$$

and hence we have obtained the necessary set $\{p_i\}$.

To avoid infinite number of colors in the case of $k \leq 0$, we can consider the set of color probabilities

$$\left\{ p_1, p_2, \ldots, p_{j-1}, \sum_{i=j}^{\infty} p_i \right\}$$

with a suitable choice of j. In this case all colors with indices $j, j+1, \ldots$ (these colors have the smallest probabilities) are considered as one single color.

Together with the Gamma density of color probabilities, Engen considered the Poisson sampling scheme. This means, that the balls of color i are drawn during time ν according to independent Poisson processes having standardized intensities p_i, $\sum_i p_i = 1$. Then the mean number of balls of color i in

the sample is νp_i and the mean size of the total sample is ν. Since we have Gamma distribution of color probabilities and Poisson sampling scheme, the population model is called the ***Poisson-Gamma model***.

4.1.2 Mean number of colors with x representatives

Let T_x denote the number of colors which are represented by exactly x balls in the sample. Quantities T_x are called ***size indices*** or, alternatively, ***frequencies of frequencies*** (Good [7]). In this section we will find the expectation of T_x, $x \in \{0, 1, 2, \ldots\}$.

Let the random variables F_i denote the sample frequency of color i, i. e. number of balls of color i in the sample. Then T_x can be expressed as

$$T_x = \sum_{i=1}^{s} I(F_i = x), \tag{63}$$

where

$$I(F_i = x) = \begin{cases} 1, & F_i = x, \\ 0, & F_i \neq x. \end{cases} \tag{64}$$

Note that the sum

$$\sum_{x=1}^{s} T_x = s - T_0$$

equals to the number of colors represented in the sample. Next we are interested in finding the mean value of T_x

$$E(T_x) = E(\sum_{i=1}^{s} I(F_i = x)) = \sum_{i=1}^{s} \mathbf{P}(F_i = x).$$

Since the frequency F_i of color i have Poisson distribution with mean νp_i, $i = 1, \ldots, s$, then

$$E(T_x) = \sum_{i=1}^{s} \frac{(\nu p_i)^x}{x!} e^{-\nu p_i}. \tag{65}$$

Next we use approximation in each interval $[\xi_i, \xi_{i-1}]$

$$\frac{(\nu p_i)^x}{x!}e^{-\nu p_i} \approx \int_{\xi_i}^{\xi_{i-1}} \frac{(\nu p)^x}{x!}e^{-\nu p}g(p)dp. \tag{66}$$

Comment 1. The approximation (66) is a special case of the approximation

$$E[h(X)] \approx h(EX),$$

where the random variable X has density $g(p)$ in the interval $[\xi_i, \xi_{i-1}]$ and mean p_i. The function $h(\cdot)$ is given by

$$h(p) = \frac{(\nu p)^x}{x!}e^{-\nu p}.$$

The shorter is the interval $[\xi_i, \xi_{i-1}]$, the more accurate is the approximation.

■

Thus, the mean size index $E(T_x)$ may be approximated by

$$E(T_x) \approx \int_0^\infty \frac{(\nu p)^x}{x!}e^{-\nu p}g(p)dp. \tag{67}$$

Substituting Gamma distribution function (57) divided by p into the place of $g(p)$ we get

$$\begin{aligned} E(T_x) &\approx \int_0^\infty \frac{(\nu p)^x}{x!}e^{-\nu p}\left[\frac{\alpha^{k+1}}{\Gamma(k+1)}p^{k-1}e^{-\alpha p}\right]dp \\ &= \frac{\nu^x\alpha^{k+1}}{x!\Gamma(k+1)}\int_0^\infty p^{x+k-1}e^{-(\alpha+\nu)p}dp. \end{aligned} \tag{68}$$

Using the definition of gamma function

$$\Gamma(t) = \int_0^\infty p^{t-1}e^{-p}dp \tag{69}$$

(68) simplifies to

$$E(T_x) \approx \frac{\nu^x\alpha^{k+1}\Gamma(x+k)}{x!\Gamma(k+1)(\nu+\alpha)^{x+k}} = \alpha\frac{\Gamma(x+k)}{x!\Gamma(k+1)}\omega^k(1-\omega)^x, \tag{70}$$

where

$$\omega = \frac{\alpha}{\nu + \alpha}.$$

Let us show that the right side of (70) is proportional to the probability of negative binomial distribution. Recall that the probability mass function of the negative binomial distribution is

$$Pr(x|k,\omega) = \binom{k+x-1}{k-1}\omega^k(1-\omega)^x, \quad x = 0,1,2,\ldots, \quad 0 < \omega < 1. \quad (71)$$

Usually the negative binomial distribution is defined only for non-negative integer values of k. In this case the probability $Pr(x|k,\omega)$ has a simple interpretation. Namely, $Pr(x|k,\omega)$ is the probability of getting x failures before kth success occurs in series of $x+k$ independent identical trials with probability of success ω.

If we recall that

$$(x-1)! = \Gamma(x), \quad x \notin \{0,-1,-2,\ldots\}$$

and

$$\Gamma(x+1) = x\Gamma(x), \quad x \notin \{0,-1,-2,\ldots\}$$

then probability (71) becomes

$$Pr(x|k,\omega) = \frac{\Gamma(k+x)}{\Gamma(k)x!}\omega^k(1-\omega)^x, \quad x = 0,1,2,\ldots, \quad 0 < \omega < 1. \quad (72)$$

The latter expression is defined for arbitrary $k > 0$. It can be shown that (72) gives a valid probability mass function for all $k > 0$. The expression (72) is defined also for $k \in (-1,0)$ and

$$\sum_{x=0}^{\infty} Pr(x|k,\omega) = 1,$$

but in this case we cannot talk about the probability distribution, since in the case $k \in (-1,0)$

$$Pr(0|k,\omega) > 1$$

and

$$Pr(i|k,\omega) < 0, \quad i \in \{1,2,\ldots\}.$$

In the case $k = 0$, the probabilities (72) are not defined, since the Γ function has no value at the point 0. But in the process $k \to 0$ the probability $Pr(0|k,\omega)$ approaches value 1 and the other probabilities tend to 0.

It follows that

$$E(T_x) \approx \frac{\alpha}{k} Pr(x|k,\omega). \tag{73}$$

We now see that the total number of colors s in population approximately equals to α/k, since

$$s = \sum_{x=0}^{\infty} E(T_x) \approx \frac{\alpha}{k} \sum_{x=0}^{\infty} Pr(x|k,\omega) = \frac{\alpha}{k}. \tag{74}$$

The mean value of number S of colors represented in the sample is then

$$E(S) = \frac{\alpha}{k} - E(T_0) = \frac{\alpha}{k}(1 - Pr(0|k,\omega)) = \frac{\alpha}{k}(1 - \omega^k). \tag{75}$$

Note 1. *Note that if $-1 < k < 0$, then $E(T_0)$ cannot be approximated from (70), because the estimate becomes negative. However, estimates of other size indices $E(T_x)$ $x = 1, 2, \ldots$ are available.*

Because of the property that size indices $E(T_x)$ are approximately proportional to the probabilities $Pr(x|k,\omega)$, the Gamma-Poisson model is called ***Engen's Extended Negative Binomial (ENB) model***. The "extension" of this model implies the new region $-1 < k < 0$, that was introduced by Engen. Before that, only values $k \geq 0$ have been used. Because of this extension Engen's ENB model describes populations of very different structures.

We have seen that in the case of the Gamma-Poisson model the sample size N is a random variable with Poisson distribution with mean ν. But in practice,

the sample size n is usually given. Due to (70),

$$\begin{aligned} E(N) &= \sum_{x=1}^{\infty} xE(T_x) \approx \alpha \sum_{x=1}^{\infty} \frac{\Gamma(x+k)}{(x-1)!\Gamma(k+1)} \omega^k (1-\omega)^x \\ &= \frac{\alpha(1-\omega)}{\omega} \sum_{x=1}^{\infty} Pr(x-1|k+1,\omega) = \frac{\alpha(1-\omega)}{\omega}. \end{aligned}$$

If we approximate the expectation $E(N)$ of sample size by its realization n then we get

$$\frac{\alpha(1-\omega)}{\omega} \approx n. \tag{76}$$

Thus, the expectation of size indice $E(T_x)$ can be reformulated as

$$E(T_x) \approx n \frac{\Gamma(x+k)}{x!\Gamma(k+1)} \omega^{k+1} (1-\omega)^{x-1}. \tag{77}$$

The latter formula is simpler to use for estimation purpose than approximative formula (70).

4.1.3 Joint distribution of size indices T_x

Consider a Gamma-Poisson model with parameters α and k. Recall that α and k approximately define s by $s \approx \alpha/k$, where $0 \leq s - \alpha/k < 1$. Hoshino [10] has shown, that in the process

$$s \to \infty \text{ and } k \to 0 \text{ with } sk = \alpha \text{ fixed} \tag{78}$$

the joint distribution of size indices $T_1, T_2, \ldots$ converges to the logarithmic series model. This model is given by

$$P(T_1 = t_1, T_2 = t_2, \ldots) = \prod_{x=1}^{\infty} \frac{\mu_x^{t_x} e^{-\mu_x}}{t_x!}, \tag{79}$$

where

$$\mu_x = E(T_x).$$

Note that number T_0 of undiscovered colors is not included into the model (79). The model (79) is equivalent to the assumption that the size indices T_x are independent random variables having Poisson distribution with mean μ_x. Further, the model (79) will be used to derive the maximum likelihood estimate of parameters of Gamma distribution.

4.2 Estimation of ENB model

4.2.1 Derivation of maximum likelihood function

In this section we derive a log-likelihood function required to estimate population parameters ω and k based on the distribution (79).

If we substitute expression (77) of $E(T_x)$ into joint distribution of size indices (79) then we obtain the maximum likelihood (ML) function $L(\omega, k)$ in terms of parameters ω and k. To avoid working with products, we simplify the ML function $L(\omega, k)$ by taking logarithm of it and finding the log likelihood function $l(\omega, k) = \ln L(\omega, k)$

$$l(\omega, k) = \sum_{x=1}^{\infty} t_x \ln E(T_x) - \sum_{x=1}^{\infty} E(T_x) - \sum_{x=1}^{\infty} \ln(t_x!) \tag{80}$$

We omit $\ln(t_x!)$, since it doesn't depend on parameters ω and k and thus doesn't affect the maximum of $l(\omega, k)$. Let us find the first sum of (80)

$$\begin{aligned}
&\sum_{x=1}^{\infty} t_x \ln E(T_x) = \sum_{x=1}^{\infty} t_x \ln \left[n \frac{\Gamma(x+k)}{x!\Gamma(k+1)} \omega^{k+1} (1-\omega)^{x-1} \right] \qquad (81)\\
&= \sum_{x=1}^{\infty} t_x ((k+1)\ln\omega + (x-1)\ln(1-\omega) + \ln\Gamma(x+k) - \ln\Gamma(k+1)) + c_1\\
&= S(k+1)\ln\omega + (n-S)\ln(1-\omega) + \sum_{x=1}^{\infty} t_x (\ln\Gamma(x+k) - \ln\Gamma(k+1)) + c_1,
\end{aligned}$$

since

$$\sum_{x=1}^{\infty} t_x = s, \quad \sum_{x=1}^{\infty} x t_x = n,$$

where c_1 denotes an expression which doesn't depend on ω and k. The second sum of (80) simplifies to

$$\begin{aligned} \sum_{x=1}^{\infty} E(T_x) &= n \sum_{x=1}^{\infty} \left[\frac{\Gamma(x+k)}{x!\Gamma(k+1)} \omega^{k+1} (1-\omega)^{x-1} \right] \\ &= \frac{n\omega}{k(1-\omega)} \sum_{x=1}^{\infty} Pr(x|k,\omega) = \frac{n\omega}{k(1-\omega)} (1 - Pr(0|k,\omega)) \\ &= \frac{n\omega}{k(1-\omega)} (1-\omega^k). \end{aligned} \tag{82}$$

Hence, the log likelihood is expressed as

$$\begin{aligned} l(\omega,k) = -\frac{n\omega}{k(1-\omega)}(1-\omega^k) + S(k+1)\ln\omega + (n-S)\ln(1-\omega) \\ + \sum_{x=1}^{\infty} t_x (\ln\Gamma(x+k) - \ln\Gamma(k+1)) + c_2 \end{aligned} \tag{83}$$

where c_2 does not depend on ω and k. Log likelihood function (83) was proposed by Hoshino in [9], our role is its detailed derivation based on joint distribution of size indices (79).

4.2.2 Fitting the model by the ML estimation

In this section we will apply log likelihood function (83) to estimating parameters of population model. For this purpose we will use data, which are obtained by simulation of $n = 500$ observations from multinomial distribution with the number of classes $s = 100$. The population probabilities of classes were the members of a geometric sequence

$$p_i = p_0 q^i, \ i = 1, \ldots, 100 \tag{84}$$

Table 1: Fitting of ENB model to simulated data

x	t_x	MLE	x	t_x	MLE	x	t_x	MLE	x	t_x	MLE
1	9	8, 13	11	0	0, 91	21	1	0, 39	$\cdots$	$\cdots$	$\cdots$
2	5	4, 68	12	0	0, 82	22	1	0, 37	34	2	0, 18
3	3	3, 31	13	0	0, 75	23	0	0, 34	$\cdots$	$\cdots$	$\cdots$
4	3	2, 55	14	0	0, 68	24	0	0, 32	48	1	0, 09
5	0	2, 07	15	0	0, 62	25	0	0, 30	$\cdots$	$\cdots$	$\cdots$
6	2	1, 73	16	0	0, 57	26	0	0, 28	50	1	0, 08
7	0	1, 49	17	1	0, 53	27	0	0, 27	$\cdots$	$\cdots$	$\cdots$
8	2	1, 29	18	2	0, 49	28	0	0, 25	62	1	0, 04
9	1	1, 14	19	1	0, 45	29	0	0, 24			$\cdots$
10	3	1, 01	20	1	0, 42	30	1	0, 22	$\sum$	**41**	**41, 00**

where the common ratio $q = 0.9$ and p_0 is the scale parameter. Empirical size indices t_x are given in the Table 1.

To find estimates $\hat{\omega}$ and $\hat{k}$ we have to maximize log likelihood function (83). One way to do it is to find partial derivatives $\partial l/\partial\omega$ and $\partial l/\partial k$ and solve the system of equations

$$\begin{cases} \partial l(\omega, k)/\partial\omega = 0 \\ \partial l(\omega, k)/\partial k = 0 \end{cases}. \tag{85}$$

Hoshino [9] proposed the Newton-Rhapson algorithm for solving system (85). However, unconstrained Newton-Rhapson algorithm has a problem that if bad approximation is given, then iterations may drive us outside of region defined by constraints $k > -1$ and $0 < \omega < 1$. To account for these constraints we have used constrained Newton-Rhapson optimization method available for example in module IML of SAS software. Besides the Newton-Rhapson method, there are several methods available in SAS/IML module. The SAS

code required to solve the optimization problem discussed can be found in Appendix 1.

The ML-estimates of the parameters are $\hat{\omega} = 0.0312$ and $\hat{k} = 0.188$. According to the fitted model, estimated values of $E(T_x)$, found by (77), are exhibited in Table 1 in column named MLE. The same data are plotted in Figure 4.

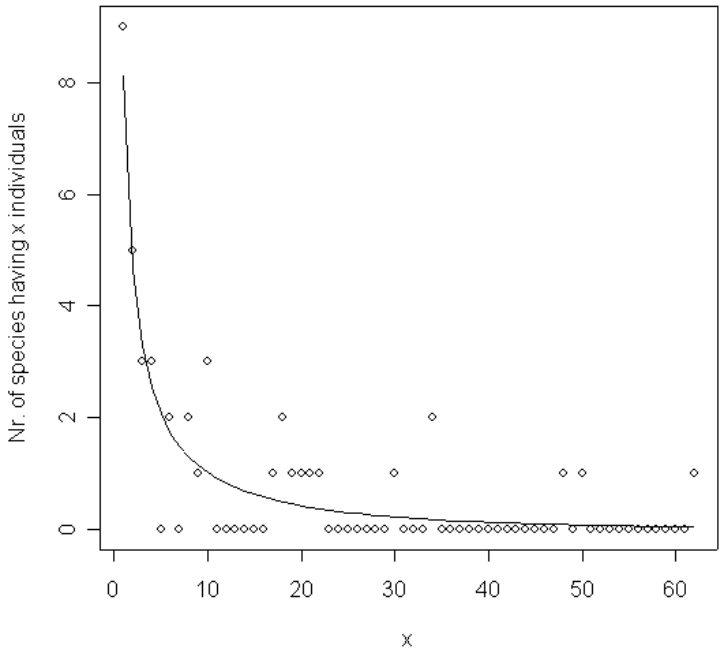

Figure 4: Plot of empirical and estimated size indices

5 Estimation of sample coverage

5.1 Review of literature on estimation of sample coverage

The problem of estimating the coverage was first discussed in Good [7] with application in studies of the literary vocabulary and accident proneness. The estimator that is proposed in the article, was suggested to Good by D. M. Turing and thus is called the Turing estimator. This estimator is given by

$$\hat{C}_{Tur} = 1 - \frac{t_1}{n}, \tag{86}$$

where, as before, t_1 is the number of colors in the sample, which are represented by one ball and n is the sample size. The Turing estimator is derived using the Bayes' theorem (see Appendix 2 for the derivation).

The Turing estimator has been discussed by many authors. Esty [4] has proved a normal limit law for the Turing estimator in the case of multinomial sampling scheme. It means that there exists such $\sigma > 0$, so that

$$\frac{C_n - \hat{C}_{Tur}}{\sigma}\sqrt{n} \to N(0,1) \tag{87}$$

in distribution in process $n \to \infty$. This allows to calculate an assymptotic confidence interval for the sample coverage. Mao and Lindsay [14] proved the similar normal limit law for the Poisson sampling scheme

$$\frac{C_\nu - \hat{C}_{Tur}}{\delta}\sqrt{s} \to N(0,1) \text{ for some } \delta > 0. \tag{88}$$

The difference between limit laws (87) and (88) is that in the first case the sample size n goes to infinity, while in the second case the number s of colors becomes infinite.

In [5] Esty has shown that the Turing estimator is assymptotically the same efficient as the best coverage estimator developed under the strong hypothesis that all colors have equal probabilities

$$p_1 = p_2 = \ldots = p_s = 1/s,$$

even when the hypothesis is true. It is worth noting that in the case of equiprobable colors the ML estimate of the number s of colors in population was obtained by Lewontin and Prout [12]. This estimate is given by the equation

$$\frac{n}{\hat{s}} = \sum_{j=\hat{s}-v_n+1}^{\hat{s}} \frac{1}{j},$$

where v_n is the number of colors in the sample or, equivalently, the sample colority. In the case when all colors are equiprobable, the sample coverage C_n equals to the sample colority V_n divided by the number of classes s. Thus it is natural to estimate the sample coverage by $v_n/\hat{s}$.

In contrast to the nonparametric Turing estimator, Engen [3] used a parametric Poisson-Gamma model to estimate the sample coverage. This parametric model was introduced in Chapter 4 and the estimator of the sample coverage is derived in the following section. The Engen's estimator of sample coverage performs well for different kinds of population structures.

5.2 Estimation of sample coverage in the case of ENB model

Next we will estimate the sample coverage under ENB model, defined by (77). In the case of Poisson sampling scheme the mean of the sample coverage can be expressed as

$$E(C_\nu) = \sum_{i=1}^{s} p_i(1 - e^{-\lambda_i \nu}) \tag{89}$$

(see paragraph 2.3.1). Each summand of the sum in the right side of (89) can be approximated (see Comment 1) by

$$p_i(1-e^{-\nu p_i}) \approx \int_{\xi_i}^{\xi_{i-1}} p(1-e^{-\nu p})g(p)dp.$$

The coverage is then approximated as follows

$$\begin{aligned} E(C_\nu) &\approx \sum_{i=1}^{s} \int_{\xi_i}^{\xi_{i-1}} p(1-e^{-\nu p})g(p)dp \\ &= \int_0^\infty p(1-e^{-\nu p})g(p)dp = 1 - \int_0^\infty e^{-\nu p} f(p)dp \\ &= 1 - \frac{\alpha^{k+1}}{(\nu+\alpha)^{k+1}} \int_0^\infty f(p)dp = 1-\omega^{k+1}, \end{aligned} \tag{90}$$

where

$$\omega = \frac{\alpha}{\nu+\alpha}.$$

Hence, we have

$$E(C_\nu) \approx 1-\omega^{k+1}. \tag{91}$$

Formula (91) is proposed by Engen [3], we have only derived it here in full details. The coverage of the sample can be estimated by

$$\hat{C}_\nu = 1-\hat{\omega}^{\hat{k}+1}, \tag{92}$$

where the parameters ω and k are replaced by their ML estimates $\hat{\omega}$ and $\hat{k}$, given by the log likelihood function (83).

Due to (70)

$$ET_1 \approx \alpha(1-\omega)\omega^k = \frac{\alpha\nu}{\alpha+\nu}\omega^k = \nu\omega^{k+1}.$$

It means that

$$\omega^{k+1} \approx \frac{ET_1}{\nu}.$$

Therefore the expectation (91) may be written as

$$E(C_\nu) \approx 1 - \frac{ET_1}{\nu}. \tag{93}$$

The expression in the right side of (93) does not depend on the parameters ω and k of Gamma distribution. After replacing the expectation ET_1 by the realization t_1 of size index T_1 and replacing the mean sample size ν by the actual sample size n, then we obtain the Turing estimator (86) of the sample coverage. This implies that both the Turing estimator (86) and the estimator (92) derived by Engen, are approximately the same.

Example 9. Estimation of sample coverage. Here we will continue working with simulated dataset described in paragraph 4.2.2. We calculate the Engen's coverage estimate (92)

$$\hat{C}_\nu = 1 - \hat{\omega}^{\hat{k}+1} = 1 - 0.0312^{1+0.188} = 0.984$$

and the Turing estimate (86)

$$\hat{C}_{Tur} = 1 - \frac{t_1}{n} = 1 - \frac{9}{500} = 0.982.$$

We see that the estimates are very close. For comparison, the actual sample coverage of our sample was

$$\sum_{i=1}^{100} p_i I(F_i > 0) = 0.971,$$

thus, the relative error of coverage estimate is

$$\frac{0.982 - 0.971}{0.971} = 1.1\%.$$

The estimate of the total number s of colors according to (74) is

$$\hat{s} = \frac{\hat{\alpha}}{\hat{k}} = \frac{n\hat{\omega}}{\hat{k}(1-\hat{\omega})} = \frac{500 \cdot 0.0312}{0.188(1-0.0312)} = 86,$$

which is not a precise estimate, since the actual number of colors in the population was $s = 100$.

5.3 Inspection of goodness of ENB model: a Monte-Carlo experiment

In this subsection an experiment will be conducted to investigate the goodness of the coverage estimator (91), which was obtained under assumptions of the ENB model. Different kinds of populations structures will be inspected. In order to check the performance of the ENB model for the estimation of the sample coverage, multinomial samples were simulated from populations having following structures.

1. Uniform color probabilities described in Example 2.
2. Two different exponentially decreasing color distributions described in Example 5 with common ratio $q = 0.95$ and $q = 0.98$.
3. Inverse color probabilities described in Example 6.
4. Quadratically decreasing probabilities described in Example 7.

In each of five color distributions the number of colors was $s = 200$. With each distribution 25 multinomial samples of four different sizes ($n = 200$, 300, 500 and 1000) were generated. Maximum likelihood estimates of the parameters ω and k (given by the log likelihood (83)) of the ENB model were obtained. Next the sample coverage was estimated using the formula (92)

$$\hat{C}_\nu = 1 - \hat{\omega}^{\hat{k}+1}.$$

Since the color distributions are known, the actual sample coverage C_n can be calculated directly, using the definition (5). The averages of actual C_n and

the averages of estimates $\hat{C}_n$ for each sample size and each color distribution are presented in Table 2.

Table 2: Actual coverages and estimated coverages of simulated samples (averaged over 25 samples)

	C_{200}	$\hat{C}_{200}$	C_{300}	$\hat{C}_{300}$	C_{500}	$\hat{C}_{500}$	C_{1000}	$\hat{C}_{1000}$
EXP95	0.904	0.908	0.934	0.941	0.959	0.968	0.982	0.983
EXP98	0.778	0.779	0.857	0.855	0.918	0.920	0.965	0.968
CONST	0.628	0.635	0.780	0.769	0.924	0.914	0.995	0.993
INV	0.771	0.772	0.825	0.819	0.885	0.875	0.951	0.939
SQR	0.772	0.773	0.860	0.873	0.930	0.943	0.979	0.981

From the table Table 2 we see, that the estimates are precise (in average) enough even for small sample sizes. To have a better look at the quality of coverage estimation, averages and standard deviation of relative errors $\rho = |\hat{C} - C|/C$ were also computed. Calculated characteristics of relative errors are presented in the Table 3. From Table 3 we see that the coverage estimates have small relative errors for all population distributions, the biggest error being 6.1%. Hence we have demonstrated that the ENB model fits well different kinds of population structures.

Table 3: Characteristics of relative errors of coverage estimates

	$n = 200$		$n = 300$		$n = 500$		$n = 1000$	
	$\bar{\rho}$	$\mathrm{sd}(\rho)$	$\bar{\rho}$	$\mathrm{sd}(\rho)$	$\bar{\rho}$	$\mathrm{sd}(\rho)$	$\bar{\rho}$	$\mathrm{sd}(\rho)$
EXP95	0.029	0.017	0.018	0.015	0.010	0.007	0.004	0.003
EXP98	0.043	0.042	0.033	0.023	0.018	0.011	0.006	0.003
CONST	0.052	0.043	0.039	0.026	0.020	0.013	0.005	0.003
INV	0.039	0.033	0.024	0.016	0.020	0.015	0.015	0.009
SQR	0.061	0.043	0.025	0.019	0.021	0.019	0.006	0.003

6 Estimation of sample size required for achieving given coverage

In the previous sections we have estimated the sample coverage. The purpose of estimating the sample coverage is to answer the question "is it worthwhile to extend the sample?". If the coverage of the sample is large enough (99.9%, for example), then we can stop drawing objects into the sample, since the structure of the population is well established already. But if the coverage is not sufficient yet, then we wish to know how many additional objects must be drawn to achieve the given coverage. Thus, the problem is to estimate the sample size $n_{1-\eta}$, required for achieving the given coverage $1-\eta$ [1].

With this problem we started already in the thesis [11] where estimation of required sample size in the special case of uniform color distribution was discussed. We bring some results from [11] and provide improvements of the methods introduced there for estimating the required sample size.

It should be emphasized that in all the methods of this section the estimates of required sample size are calculated using only the points

$$V_1 \leq V_2 \leq \ldots \leq V_n$$

of the sample colority curve. No other additional information about the sample is used. Thus, the information about sample frequencies of colors is not used during estimation.

[1] Note that in this section $n_{1-\eta}$ denotes the total sample size, i.e. it includes objects that have been already drawn. Thus, if n_0 objects are already drawn then we have to draw $n_{1-\eta} - n_0$ additional objects to achieve the coverage $1-\eta$.

6.1 Uniform color distribution

Our main results in [11] were two methods for estimating the sample size $n_{1-\eta}$, required for achieving the given coverage $1-\eta$ in the case of uniform color distribution. These methods are Method 1 and Method 2 below. In addition we propose a new Method 3.

6.1.1 Method 1: Estimating of required sample size by the "two-point" method of moments

In the case of uniform color distribution the color probabilities are $p_i = 1/s$, $i = 1, \ldots, s$. The expectations of sample colority V_n and coverage C_n can be approximately expressed as

$$E(V_n) \approx s\left(1 - e^{-\frac{n}{s}}\right), \quad E(C_n) \approx 1 - e^{-\frac{n}{s}},$$

(see formulae (24) and (25)). Recall that the approximation is good, provided that $p_i = 1/s$ is small, and hence, provided that s is large. Suppose that the colority curve

$$V_1 \leq V_2 \leq \ldots \leq V_n$$

is known. If we choose two points of colority curve, say V_t and V_{2t}, then the mean sample colorities at this points are approximately

$$\begin{array}{l} E(V_t) \approx s(1 - e^{-\frac{t}{s}}) \\ E(V_{2t}) \approx s(1 - e^{-\frac{2t}{s}}) \end{array}. \tag{94}$$

Thus

$$\frac{E(V_{2t})}{E(V_t)} \approx \frac{1 - e^{-\frac{2t}{s}}}{1 - e^{-\frac{t}{s}}} = 1 + e^{-\frac{t}{s}}. \tag{95}$$

Next we express s from (95):

$$s \approx -\frac{t}{\ln\left(\frac{E(V_{2t})}{E(V_t)} - 1\right)}. \tag{96}$$

In order to estimate s, we substitute the sample values of V_t and V_{2t} in the place of their expectations, and get an estimate for the number s of classes in population

$$\hat{s} \approx -\frac{t}{\ln\left(\frac{V_{2t}}{V_t} - 1\right)}. \tag{97}$$

Then we estimate the required sample size $n_{1-\eta}$ needed to achieve coverage $1 - \eta$. For this purpose, we substitute in the expression of mean sample coverage

$$E(C_{n_{1-\eta}}) \approx 1 - e^{-\frac{n_{1-\eta}}{s}}$$

the required coverage $1 - \eta$ into the place of its expectation to obtain

$$1 - \eta \approx 1 - e^{-\frac{n_{1-\eta}}{s}} \Rightarrow n_{1-\eta} \approx -s \ln \eta.$$

Finally, by substituting the estimate $\hat{s}$ (97) into the latter approximate equation, we get

$$\hat{n}_{1-\eta} \approx \frac{t \ln \eta}{\ln\left(\frac{V_{2t}}{V_t} - 1\right)}. \tag{98}$$

We call the latter estimate of the required sample size the "two-point" estimate, since two points of colority curve are used.

6.1.2 Method 2: Estimating of required sample size by nonlinear regression

In this method we estimate the regression model, where the required sample size (or its appropriate transformation) is the response variable and colority values $V_{t_1}, \ldots, V_{t_k}$ (maybe appropriately transformed) are the explanatory variables. In the thesis [11] two regression models were evaluated: one model for $n_{0.99}$ and other model for $n_{0.999}$. This method gives estimates with smaller biases and standard errors, compared to the method of the moments

(Method 1). However, the method of regression is of very limited use, because each regression equation is usable only for certain value of required coverage and for certain interval of sample sizes.

6.1.3 Method 3: "One-point" method of moments

Here we demonstrate that the Method 1 can be improved by using only one point of a colority curve instead of two points (V_t and V_{2t}). The "one-point" method of moments that we will next propose, gives a smaller relative error of additional sample size, compared to the "two-point" method. Consider the approximate expression of the mean of the sample colority

$$E(V_t) \approx s(1 - e^{-\frac{t}{s}}). \tag{99}$$

If we substitute the realization of colority in place of the colority expectation, then we get

$$V_t \approx s(1 - e^{-\frac{t}{s}}). \tag{100}$$

To estimate number of colors s, the equation (100) can be solved (for s) numerically, for example by Newton's method. When an estimate $\tilde{s}$ is obtained, then the required sample size is obtained by

$$\tilde{n}_{1-\eta} = -\tilde{s} \ln \eta.$$

6.1.4 Monte-Carlo comparison of Method1 and Method 3

The comparison of Method 1 and Method 2 in one particular case is provided in Table 4. The data in the table were obtained by simulating 50 samples of size $n = 250$ from population with $s = 500$ and uniform distribution of colors. From simulated values of colority curve, the required sample size $n_{0.99}$

Table 4: Comparison of Method 1 and Method 3 for estimating the sample size $n_{0.99}$ required to reach 99%-coverage

	Method 1	Method 3
Average relative error $\bar{\rho}$	0.366	0.196
Standard deviation sd(ρ)	0.438	0.160
Average estimated $n_{0.99}$	2598	2353
Actual value of $n_{0.99}$	2294	
Average estimated s	564.1	510.9

to achieve 99% coverage was estimated using both Method 1 and Method 3. In Method 3 the only the last point V_{250} of colority curve was used to obtain the estimate of $n_{0.99}$. In Method 1, the points V_{250} and V_{125} were used to calculate the estimate of $n_{0.99}$. The actual value of sample size $n_{0.99}$, when the coverage achieves 99%, was also evaluated for each sample.

In the table, there are shown characteristics of estimates for both methods, averaged over 50 samples:

(a) averages $\bar{\rho}$ of relative errors of estimated sample sizes $n_{0.99}$,

(b) standard deviations sd(ρ),

(c) averages of estimated sample sizes,

(d) average of actual sample size,

(e) the averages of estimated number s of colors (these averages can be obtained by dividing averages in (c) by a constant $-\ln \eta$).

From the Table 4 we conclude that the Method 3 outperforms the Method 1 by all presented characteristics. In the Method 3 both the average and the

standard deviation of relative errors are about two times smaller than in the Method 1.

Though the computation of estimates is simpler in the case of Method 1, we recommend to use the Method 3, which provides more precise estimates. Furthermore, the optimization problem required in the Method 3 can easily be solved by most of statistical or mathematical software packages.

The case of uniform color distribution discussed in this subsection is quite unusual in practice. More usual is the situation where we have a small number of dominating classes and a large number of rare classes, which are represented by a small number of objects. In two following subsections we will model such populations by two types of color distributions: linearly decreasing color distribution and exponentially decreasing color distribution.

6.2 Linearly decreasing color distribution

In the case of linearly decreasing color distribution the color probabilities are given by formula

$$p_i = p_0 - ai, \quad a > 0, \ i = 1, \ldots, s.$$

The constant p_0 is uniquely defined from the constraint $\sum_i (p_0 - ai) = 1$. We will express the color probabilities by introducing the parameter r, which equals to the relation of the biggest and the smallest color probability, i.e.

$$r = \frac{\pi(1)}{\pi(s)} = \frac{p_0 - a}{p_0 - as}.$$

If the value of the parameter r is given, the coefficients p_0 and a can be calculated as follows

$$p_0 = \frac{2(rs-1)}{s(s-1)(r+1)}, \tag{101}$$

$$a = \frac{2(r-1)}{s(s-1)(r+1)}. \tag{102}$$

Hence, the color probabilities are given by equation

$$p_i = \frac{2}{s(s-1)(r+1)}((rs-1)-(r-1)i), \quad r \geq 1,\ i = 1, \ldots, s. \tag{103}$$

Note that the value $r = 1$ corresponds to the uniform color distribution. Next we will find an estimate $\hat{n}_{1-\eta}$ of the sample size $n_{1-\eta}$ required for achieving the coverage $1-\eta$ based on the sequence $V_{t_1}, \ldots, V_{t_k}$ of sample colorities. We will use the estimation method that consists of two steps: (1) estimation of population parameters s and r and (2) estimation of the required sample size $n_{1-\eta}$ based on estimates found in (1).

6.2.1 Method 4 for estimation of required sample size

(1) Consider the approximate expression of the mean colority (13)

$$E(V_n) \approx \sum_{i=1}^{s}(1-e^{-np_i}). \tag{104}$$

We then simplify the sum in (104), by substituting the expression $p_0 - ai$ of color probabilities in the place of p_i and using the formula of the sum of geometric sequence:

$$\begin{aligned}\sum_{i=1}^{s}(1-e^{-np_i}) &= s - \sum_{i=1}^{s} e^{-n(p_0-ai)} = s - e^{-np_0}\sum_{i=1}^{s} e^{nai} \\ &= s - e^{-n(p_0+a)}\frac{e^{-nas}-1}{e^{-na}-1}\end{aligned}$$

Now we have

$$E(V_n) \approx s - e^{-n(p_0+a)} \frac{e^{-nas} - 1}{e^{-na} - 1}. \tag{105}$$

We can express the right side of (105) only in terms of s, n and r using equalities (101) and (102). Hence, we consider the right side of (105) as a function $h_n(s, r)$, of s and r, getting

$$E(V_n) \approx h_n(s, r).$$

Substitution of the mean colority $E(V_n)$ by corresponding sample colority V_n gives us the equation in s and r

$$V_n \approx h_n(s, r).$$

Using all given points $V_{t_1}, \ldots, V_{t_k}$ of colority curve we obtain k equations

$$V_{t_j} \approx h_{t_j}(s, r), \quad j = 1, \ldots, k. \tag{106}$$

Generally, the system (106) of equations have no exact solution. To estimate the parameters s and r we can solve corresponding least-squares optimization problem, i.e. we find values $\hat{s}$ and $\hat{r}$ that minimize the following function

$$\Theta(s, r) = \sum_{j=1}^{k} (V_{t_j} - h_{t_j}(s, r))^2. \tag{107}$$

We use estimates $\hat{s}$ and $\hat{r}$ in the following step to estimate required sample size.

(2) Consider the approximate expression of the mean coverage (16)

$$E(C_n) \approx \sum_{i=1}^{s} p_i(1 - e^{-np_i}), \tag{108}$$

and substitute the mean coverage $E(C_n)$ by the sample coverage C_n to obtain

$$C_n \approx \sum_{i=1}^{s} p_i(1 - e^{-np_i}). \tag{109}$$

Consider (109) in the case $n = n_{1-\eta}$:

$$C_{n_{1-\eta}} \approx \sum_{i=1}^{s} p_i(1 - e^{-n_{1-\eta}p_i}). \tag{110}$$

The coverage $C_{n_{1-\eta}}$ approximately equals to $1-\eta$. We also replace probabilities p_i by their estimates $\hat{p}_i$, which are calculated from (103) using the estimates $\hat{s}$ and $\hat{r}$, found at the step (1). Thus, we obtain

$$1 - \eta \approx \sum_{i=1}^{\hat{s}} \hat{p}_i(1 - e^{-n_{1-\eta}\hat{p}_i}). \tag{111}$$

Since

$$\sum_{i=1}^{\hat{s}} \hat{p}_i = 1,$$

the (111) simplifies to

$$\eta \approx \sum_{i=1}^{\hat{s}} \hat{p}_i e^{-n_{1-\eta}\hat{p}_i}. \tag{112}$$

The only unknown in equation (112) is the required sample size $n_{1-\eta}$. By solving the equation (112) for $n_{1-\eta}$, using some numerical algorithm, we obtain the estimate $\hat{n}_{1-\eta}$.

6.2.2 Monte-Carlo experiment: evaluation of Method 4

For evaluation of the Method 4 the following Monte-Carlo experiment was conducted. One hundred samples of size $n = 250$ were simulated for 19 different populations with linearly decreasing distribution of colors with $s = 500$ colors and with 19 different values of parameter r.

From simulated colority curves, sample sizes $n_{0.99}$ required to achieve 99% coverage were estimated by the Method 4, using points (V_{50}, V_{100}, V_{150}, V_{200} and V_{250}) of colority curves. The actual value of sample size $n_{0.99}$, when the coverage achieves 99%, was also registered for each sample.

In order to understand, which subset of points V_{50}, V_{100}, V_{150}, V_{200}, V_{250} we have to take as input information for estimation, we have compared the goodness of estimates for all possible subsets. The result was that the estimates based on only the last colority point V_{250} had the smallest relative errors. Further we propose only the estimates based on this one point V_{250}.

In Table 5, characteristics of estimates for 19 used values of r, averaged over 100 samples, are given:

(a) averages $\bar{\rho}$ of relative errors of estimated sample sizes $n_{0.99}$,

(b) standard deviations of relative errors $\mathrm{sd}(\rho)$,

(c) averages of estimated sample sizes,

(d) averages of actual sample sizes.

In Figures 5 and 6, there are shown the estimated and the actual sample size for $r < 20$ and $r \geq 20$, accordingly.

From Table 5 we observe that the estimated and the actual sample size become closer as the value of the population parameter r increases. For values $r \geq 20$ the relative error of estimate stabilizes at the value of about 0.12. For small values of r ($r < 10$) the relative error is very large. The reason is that the parameter r is strongly overestimated. Since the case $r = 1$ corresponds to the uniform color distribution, we can use the Method 3 for estimation of required sample size if r is close to 1. Note that the estimate of the sample size by the Method 3 in Table 4 is very close to the actual sample size in Table 5 in the case $r = 1.1$.

The conclusion is that the Method 4 gives acceptable estimates of required sample size for populations with values of parameter $r \geq 10$. For $r < 10$ it is recommended to use Method 3 to get more precise estimate.

Table 5: Comparison of sample size estimates obtained by Method 4 with simulated sample size

r	Estimated $n_{0.99}$	Actual value of $n_{0.99}$	Average relative error $\bar{\rho}$	sd(ρ)
1.1	4764.7	2298.6	1.088	0.292
1.5	4669.3	2333.7	1.018	0.321
2	4630.7	2416.2	0.934	0.328
5	4190.3	2884	0.466	0.249
10	3903.2	3187.4	0.249	0.181
15	3807.3	3339.2	0.186	0.141
20	3675.9	3427.3	0.134	0.123
25	3708.9	3460.9	0.145	0.136
50	3637.6	3513.9	0.135	0.096
100	3616.6	3539.9	0.121	0.109
200	3512.4	3467.9	0.129	0.094
300	3559.8	3568.7	0.141	0.102
400	3594.9	3572.1	0.108	0.088
500	3543.5	3598.5	0.106	0.084
600	3567	3605.8	0.131	0.097
700	3503.1	3487.3	0.121	0.098
800	3592.3	3594.3	0.114	0.084
900	3512.2	3559.2	0.123	0.087
1000	3591.8	3483.1	0.113	0.090

6.3 Exponentially decreasing color distribution

In the case of exponentially decreasing color distribution the color probabilities are given by formula

$$p_i = p_0 q^i, \quad 0 < q \leq 1, \ i = 1, \ldots, s.$$

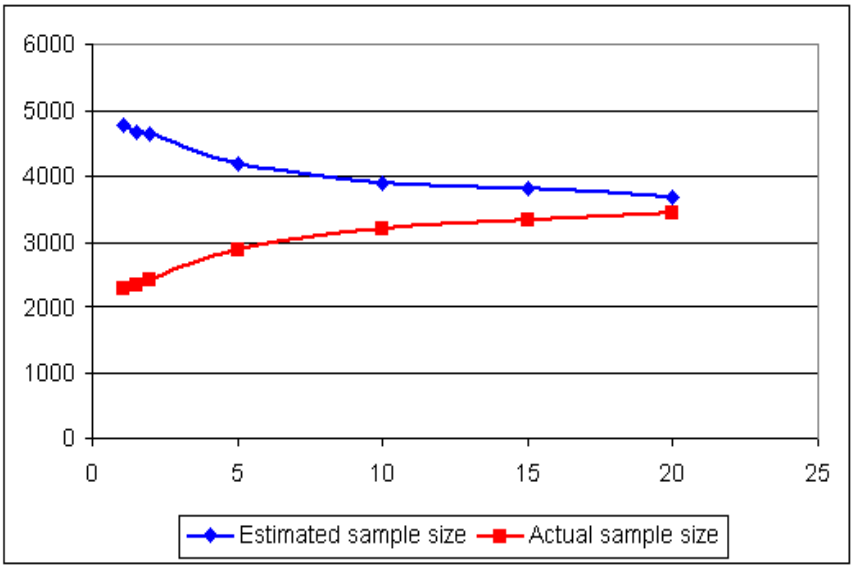

Figure 5: Comparison of required sample size, estimated by Method 4 and the actual sample size for values of $r \leq 20$

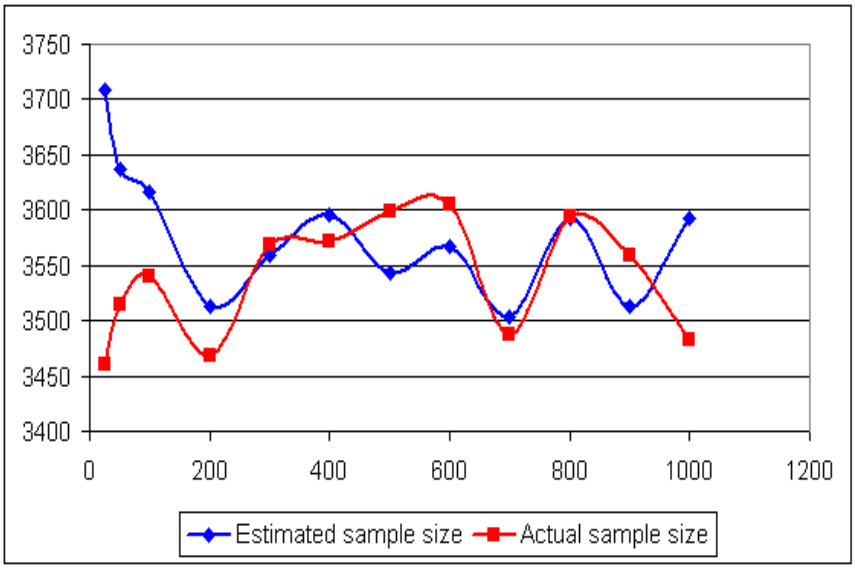

Figure 6: Comparison of required sample size, estimated by Method 4 and the actual sample size for values of $r > 20$

The constant p_0 is uniquely defined by q and s from the constraint $\sum_i p_0 q^i = 1$. Note that the value $q = 1$ corresponds to the uniform color distribution. Next we will find an estimate $\hat{n}_{1-\eta}$ of the sample size $n_{1-\eta}$ required for achieving the coverage $1-\eta$ based only on the sequence $V_{t_1}, \ldots, V_{t_k}$ of sample colorities. We will use the estimation method that consists of two steps: (1) estimation of number of colors s and the population parameter q, and (2) estimation of the required sample size $n_{1-\eta}$ based on estimates found in (1).

6.3.1 Method 5 for estimation of required sample size

(1) Consider the approximate expression of the mean colority (13)

$$E(V_n) \approx \sum_{i=1}^{s} (1 - e^{-np_i}). \tag{113}$$

By substituting p_i by $p_0 q^i$ into the (113) we get

$$E(V_n) \approx s - \sum_{i=1}^{s} e^{-np_0 q^i} \tag{114}$$

The right side of (114) depends only on s, n and q. Hence, we consider the right side of (114) as a function $h_n(s, q)$, of s and q, getting

$$E(V_n) \approx h_n(s, q).$$

Substitution of the mean colority $E(V_n)$ by corresponding sample colority V_n gives us the equation in s and q

$$V_n \approx h_n(s, q).$$

Using all given points $V_{t_1}, \ldots, V_{t_k}$ of colority curve we obtain k equations

$$V_{t_j} \approx h_{t_j}(s, q), \quad j = 1, \ldots, k. \tag{115}$$

To obtain estimates $\hat{s}$ and $\hat{q}$ of the parameters s and q, we find the values of s and q that minimize the following least-squares function

$$\Theta(s,q) = \sum_{j=1}^{k} (V_{t_j} - h_{t_j}(s,q))^2. \tag{116}$$

We use $\hat{s}$ and $\hat{q}$ in the following step to estimate the required sample size.

(2) Consider the approximate expression of the mean coverage (16)

$$E(C_n) \approx \sum_{i=1}^{s} p_i(1 - e^{-np_i}). \tag{117}$$

Substitute the mean coverage $E(C_n)$ by the sample coverage C_n to obtain

$$C_n \approx \sum_{i=1}^{s} p_i(1 - e^{-np_i}). \tag{118}$$

Consider the (118) in the case $n = n_{1-\eta}$:

$$C_{n_{1-\eta}} \approx \sum_{i=1}^{s} p_i(1 - e^{-n_{1-\eta}p_i}). \tag{119}$$

The coverage $C_{n_{1-\eta}}$ approximately equals to $1 - \eta$. We also replace the probabilities p_i by their estimates $\hat{p}_i$, using the estimates $\hat{s}$ and $\hat{q}$, found at the step (1). After this, we obtain

$$1 - \eta \approx \sum_{i=1}^{\hat{s}} \hat{p}_i(1 - e^{-n_{1-\eta}\hat{p}_i}). \tag{120}$$

Knowing that

$$\sum_{i=1}^{\hat{s}} \hat{p}_i = 1,$$

the (120) simplifies to

$$\eta \approx \sum_{i=1}^{\hat{s}} \hat{p}_i e^{-n_{1-\eta}\hat{p}_i}. \tag{121}$$

By solving the equation (121) numerically for $n_{1-\eta}$ we obtain the estimate $\hat{n}_{1-\eta}$ of the required sample size.

6.3.2 Monte-Carlo experiment: evaluation of Method 5

For evaluation of Method 5 the following Monte-Carlo experiment was conducted. One hundred samples of size $n = 250$ were simulated for 22 different populations with exponentially decreasing distribution of colors with $s = 500$ colors and with 22 different values of parameter q in interval $[0.95, 0.999]$.

From values V_{50}, V_{100}, V_{150}, V_{200}, V_{250} of simulated colority curves, the sample size $n_{0.99}$, required to achieve 99% coverage was estimated using Method 5. The actual value of sample size $n_{0.99}$, when the coverage achieves 99%, was also registered for each sample.

Like in the Method 4, the estimates based on only the last colority point V_{250} had the smallest relative errors. Further we propose only the estimates based on this one point V_{250}.

In Table 6, characteristics of estimates for, averaged over 100 samples, are given:

(a) averages $\bar{\rho}$ of relative errors of estimated sample sizes $n_{0.99}$,

(b) standard deviations of relative errors $\mathrm{sd}(\rho)$,

(c) averages of estimated sample sizes,

(d) averages of actual sample sizes.

Note that in Table 6 there are shown estimates only for the 16 first values of q in interval $[0.95, 0.993]$, since relative errors become extremely large for greater values of q and so the estimates have no practical meaning.

The Figure 7 illustrates the Table 6.

From Table 6 and Figure 7 we observe that the estimated and the actual sample size are very close for values of q in $[0.95, 0.98]$. The closer is q to 1, the bigger is the difference between estimated and actual sample sizes. For values of q greater than 0.99 the estimates become useless. We also observe that the actual sample size increases for values of q in $[0.95, 0.987]$, achieves the maximum somewhere around the point $q = 0.987$ and then starts to decrease. This property of the actual sample size is understandable, since exponentially decreasing color distribution converges to the uniform color distribution in the process $q \to 1$. Therefore, for fixed number of colors

$$n_{1-\eta}^{exp}(q) \xrightarrow{q \to 1} n_{1-\eta}^{const},$$

where $n_{1-\eta}^{exp}(q)$ is the required sample size in the case of exponentially decreasing color distribution with parameter q and $n_{1-\eta}^{const}$ is the required sample size in the case of uniform color distribution. In our case we see that the average value of actual sample size for $q = 0.999$, which equals to 2433 is very close to the corresponding average value 2294 in the case of uniform color distribution (see Table 4).

We conclude that the Method 5 gives good estimates for values of q less than 0.985 and it is not recommended to use this method if the value of the population parameter q is greater than 0.99. If q is 0.997 and greater then the Method 3 can be used. For intermediate values ($q \in (0.99, 0.997)$) the Method 5 overestimates and the Method 3 underestimates the required smple size.

Table 6: Comparison of sample size estimates obtained by Method 5 with simulated sample size

q	Estimated $n_{0.99}$	Actual value of $n_{0.99}$	Average relative error $\bar{\rho}$	sd(ρ)
0.95	1954	1853.4	0.189	0.189
0.955	2244	2122.4	0.219	0.200
0.96	2446	2437.8	0.158	0.137
0.965	2838	2791	0.151	0.109
0.97	3323	3236.4	0.136	0.130
0.975	3948	3990.4	0.120	0.101
0.98	4954	4999.2	0.132	0.118
0.985	6749	6208.6	0.147	0.143
0.986	7043	6616.4	0.124	0.086
0.987	7518	6832.4	0.143	0.137
0.988	8285	6746.2	0.243	0.151
0.989	9077	6733	0.364	0.211
0.99	10099	6519.4	0.570	0.235
0.991	10730	6125.8	0.768	0.254
0.992	11973	5492.8	1.198	0.317
0.993	13578	4888.8	1.800	0.386
0.994		4342.8		
0.995		3817.4		
0.996		3283.6		
0.997		2877.6		
0.998		2582.6		
0.999		2433		

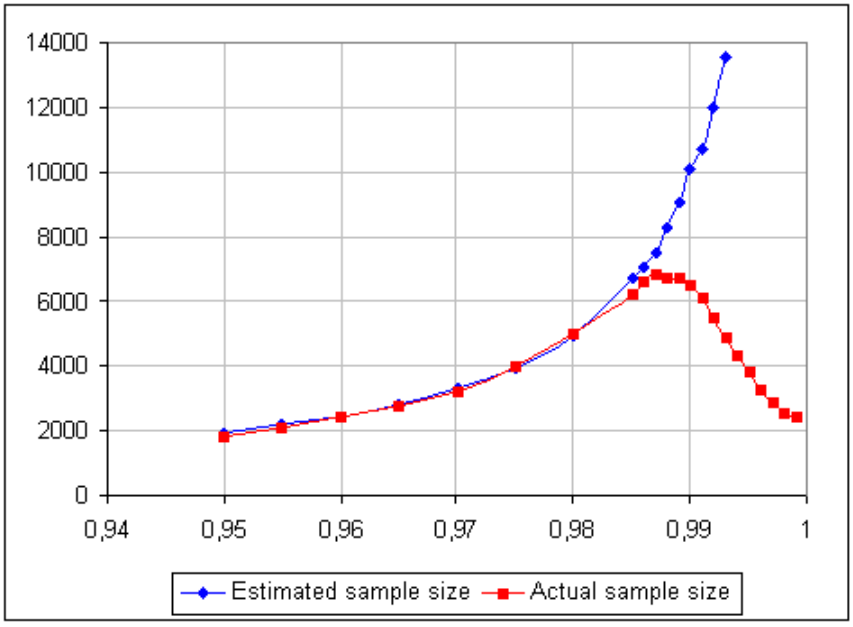

Figure 7: Comparison of required sample size, estimated by Method 5 and the actual sample size

7 Summary

In this work we provided several methods for estimation of the sample coverage and estimation of the sample size, required to achieve given coverage.

For estimation of the sample coverage we have used Engen's ENB (Extended Negative Binomial) method. This method is based on describing the color distribution by a Gamma-density. We have shown that this method provides precise estimates of the sample coverage for very different color distributions.

For estimation of the required sample size we have proposed methods for 3 types of color distributions. All the proposed methods use only the points colority curve as input data. These distributions are the following: (1) uniform color distribution, (2) linearly decreasing color distribution, (3) exponentially decreasing color distribution.

Evaluation of estimation methods for these 3 distributions gave the following results.

In the Method 3, provided for the case of uniform color distribution, the relative error of the sample size was 20%, which is acceptable.

The relative error of estimate in the Method 4 for the case of linearly decreasing color distribution depends on the value of parameter r of the color distribution. Error is small for values of $r > 20$ and it increases when r closes to 1 (this corresponds to the uniform color distribution). For $r < 10$ it is not recommended to use the proposed method.

The relative error of estimate in the Method 5 for the case of exponentially decreasing color distribution the relative error is acceptable for values of parameter $q < 0.988$. For the greater values of q the relative error is so big, that the estimates have no use. The case $q = 1$ corresponds to uniform color

distribution. If $q > 0.996$ then we can use the method for uniform distribution (Method 3) instead.

The conclusion is that if linearly or exponentially decreasing distribution is sufficiently different from the uniform distribution, then methods proposed for these distributions provide precise estimates of the required sample size.

References

[1] M.G. Bulmer. On fitting the poisson lognormal distribution to species-abundance data. *Biometrics*, 30(1):101–110, 1974.

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

[3] S. Engen. The coverage of a random sample from a biological community. *Biometrics*, 31(1):201–208, 1975.

[4] W. W. Esty. A normal limit law for a nonparametric estimator of the coverage of a random sample. *The Annals of Statistics*, 11(3):905–912, 1983.

[5] W. W. Esty. The efficency of Good's nonparametric coverage estimator. *The Annals of Statistics*, 14(3):1257–1260, 1986.

[6] R.A. Fisher, A.S. Corbet, and C.B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12(1):42–58, 1943.

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

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

[9] N. Hoshino. Engen's extended negative binomial model revisited. *Annals of the Institute of Statistical Mathematics*, 57(2):369–387, 2005.

[10] N. Hoshino and A. Takemura. Relationship between logarithmic series model and other superpopulation models useful for microdata disclosure risk assessment. *Journal of the Japan Statistical Society*, 28(2):125–134, 1998.

[11] M. Juhkam. Valimi värvilisus ja katvus multinomiaalse ning Poissoni skeemi korral, 2002.

[12] R.C. Lewontin and T. Prout. Estimation of the number of different classes in a population. *Biometrics*, 12(2):211–223, 1956.

[13] R. H. MacArthur. On the relative abundances of bird species. *Proceedings of the National Academy of Science*, 43:293–295, 1957.

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

Üldkogumid suure arvu klassidega: mudelid ning valimi katvuse ja valimimahu hindamine

Mihhail Juhkam

Resümee

Kaasaaegse loodusteaduse mitmed ülesanded tingivad selliste üldkogumite matemaatilist käsitlemist, kus klasside arv on väga suur, nätieks sajad või isegi tuhanded klassid. Sellises situatsioonis on üheks tähtsaks ülesandeks teada saada, kui suurt osa (tõenäosuse mõttes) esindab üldkogumist võetud juhuslik valim ehk kui suur on juhusliku valimi katvus.

Käesolevas töös on esmalt vaadeldud mudeleid klassi tõenäosuste jaotuste kirjeldamiseks. Tähtsaim nendest on Gamma-jaotusel põhinev Gamma-Poisson mudel.

Seejärel on välja pakutud mitmeid meetodeid valimi katvuse hindamiseks suure klasside arvuga üldkogumi korral ning meetodid vajaliku valimimahu hindamiseks etteantud katvuse saavutamiseks.

Läbiviidud Monte-Carlo katsete põhjal selgus, et Engeni poolt [3] väljapakutud ENB (Extended Negative Binomial) mudel annab piisavalt täpseid valimi katvuse hinnanguid erinevate üldkogumi värvijaotuste korral. Piisavalt head katvuse hinnangud on saadud nii suurte kui ka väikeste valimimahtude korral. Katvuse hinnangu keskmine suhteline viga ei ületanud 6.1% ning suhtelise vea standardhälve ei ületanud 4.3%.

Etteantud katvuse saavutamiseks vajaliku valimimahu hindamiseks on pakutud välja mitmed meetodid kolme erineva värvijaotuse jaoks: (1) ühtlane

värvijaotus, (2) lineaarselt kahanev värvijaotus, (3) eksponentsiaalselt kahanev värvijaotus.

Meetodite täpsuse hindamise eesmärgil läbiviidud simuleerimiskatse näitas, et

(1) ühtlase värvijaotuse korral on hinnangu keskmine suhteline viga 20%,

(2) lineaarselt kahaneva värvijaotuse korral sõltub hinnangu suhteline viga värvijaotuse parameetrist r ning $r > 20$ korral ei ületa keskmine suhteline viga 14.5% (ühele lähedaste r väärtuste korral on parem kasutada meetodit ühtlase jaotuse jaoks),

(3) eksponentsiaalselt kahaneva värvijaotuse korral sõltub hinnangu suhteline viga värvijaotuse parameetrist q ning $q < 0.985$ korral ei ületa keskmine suhteline viga 14.7% ($q > 0.997$ korral on parem kasutada meetodit, mis on välja töötatud ühtlase jaotuse jaoks).

Appendix

A1. SAS/IML functions for solving nonlinear optimization problems (NLP)

Suppose we wish to maximize the log likelihood function given by (83) where the constraints

$$k > -1,\ 0 < \omega < 1$$

are satisfied. We are given the array

$$t = (t_1, t_2, \ldots, t_{62})$$

of size indices and the sample size $n = 500$. Next we propose the SAS program code that solves the optimization task. Program consists of two main blocks: definition of log likelihood function *Loglik(x)* and the call of the optimization algorithm.

```
proc iml;
/* Giving input parameters */
  n=500;
  t={9 5 3 3 0 2 0 2 1 3 0 0 0 0 0 0 1 2 1 1
     1 1 0 0 0 0 0 0 0 1 0 0 0 2 0 0 0 0 0 0
     0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 1};

/* Definition of the log likelihood function */
/*   Function has two parameters:            */
/*     x[1]=OMEGA, x[2]=K                    */
  start Loglik(x) global(n,t);
    nu=sum(t);
```

```
    y = j(1,2,0.);
    s=0;
    do i=1 to ncol(t);
      s=s+t[i]*(Lgamma(i+x[2])-Lgamma(1+x[2]));
    end;
    f =-n*(x[1])##(x[2]+1)*(((x[1])##(-x[2])-1)/((1-x[1])*x[2]))
    +(n-nu)*log(1-x[1])+nu*(x[2]+1)*log(x[1])+s;
    return(f);
  finish Loglik;

/* Giving initial values to parameters OMEGA and K */
  x = {0.99 -0.5};
/* Giving options for optimization task                    */
/* The meaning of options:                                 */
/*      opt[1]=1 shows that this is the maximization task */
/*      opt[2]=2 specifies the amount of printed output   */
  optn = {1 2};
/* Defining constraints for parameters OMEGA and K */
/*   Constraint 1: 1e-6<=OMEGA<=0.999999          */
/*   Constraint 2: -0.999999<=K                   */
  con={1e-6     -0.999999,
       0.999999         .};
/* Running the nonlinear optimization by Newton-Raphson method     */
/*   Description of parameters:                                    */
/*   rc - variable, that indicates the reason for the termination  */
/*       of the optimization process                               */
/*   xr - contains the optimal point if termination was successful */
```

```
/*   "Loglik" - specifies an IML module that defines the          */
/*       objective function                                       */
/*   x -  represents a starting point for the iterative           */
/*       optimization process                                     */
/*   optn - indicates an options vector that specifies details    */
/*       of the optimization process                              */
/*   con - specifies a constraint matrix that defines lower       */
/*       and upper bounds for the n parameters                    */

  call nlpnra(rc,xr,"Loglik",x,optn,con);
quit;
run;
```

After running the program code with SAS we get the solution of optimization task $\hat{\omega} = 0.0312$ and $\hat{k} = 0.188$.

A2. Derivation of Turing estimator of sample coverage

We are interested in finding the distribution of the probability q_r of a color that is represented r times in a sample. Suppose that all color probabilities

$$p_1, p_2, \ldots, p_s$$

have different values (if some of p_i's are equal, then these probabilities can be adjusted microscopically so as to be made unequal). The possible set of values of q_r is a set $\{p_i\}$.

Let the prior probability function of q_r be

$$\mathbf{P}\{q_r = p_i\} = 1/s, \quad i = 1, \ldots, s.$$

If we introduce the following notations

$$A = \{\text{color is represented } r \text{ times in the sample}\},$$

$$B_i = \{\text{probability of given color is } p_i\}, \quad i = 1, \ldots, s,$$

then the probability $\mathbf{P}\{q_r = p_i\}$ expresses as $\mathbf{P}(B_i|A)$ and is obtainable by the Bayes' formula:

$$\mathbf{P}\{q_r = p_i\} = \mathbf{P}(B_i|A) = \frac{\mathbf{P}(A|B_i)\mathbf{P}(B_i)}{\sum_{j=1}^{s} \mathbf{P}(A|B_j)\mathbf{P}(B_j)}.$$

Here, $\mathbf{P}(A|B_i)$ is the probability of that the color i is represented r times in the sample. This probability equals

$$\mathbf{P}(A|B_i) = \binom{n}{r} p_i^r (1-p_i)^{n-r}.$$

If the probabilities of choosing the color for consideration are equal, then

$$\mathbf{P}(B_i) = 1/s, \quad i = 1, \ldots, s.$$

Therefore, we can find the probability $\mathbf{P}\{q_r = p_i\}$ of interest

$$\mathbf{P}\{q_r = p_i\} = \mathbf{P}(B_i|A) = \frac{\frac{1}{s}\binom{n}{r} p_i^r (1-p_i)^{n-r}}{\frac{1}{s}\sum_{j=1}^{s} \binom{n}{r} p_j^r (1-p_j)^{n-r}} = \frac{p_i^r (1-p_i)^{n-r}}{\sum_{j=1}^{s} p_j^r (1-p_j)^{n-r}}.$$

The mean value of q_r is then

$$E(q_r) = \sum_{i=1}^{s} p_i \mathbf{P}\{q_r = p_i\} = \frac{\sum_{i=1}^{s} p_i^{r+1}(1-p_i)^{n-r}}{\sum_{j=1}^{s} p_j^r (1-p_j)^{n-r}}. \tag{122}$$

If we denote, as before, the number of colors, which are represented by exactly r colors in the sample by T_r $(r = 0, 1, 2, \ldots)$ and the sample frequency of the ith color by F_i then

$$E(T_r) = E\left[\sum_{i=1}^{s} I(F_i = r)\right] = \sum_{i=1}^{s} \mathbf{P}(F_i = r) = \sum_{i=1}^{s} p_i^r (1-p_i)^{n-r}.$$

The expectation (122) may be expressed through expectations of size indices T_r as follows

$$\begin{aligned} E(q_r) &= \frac{\sum_{i=1}^{s} p_i^{r+1}(1-p_i)^{n-r}}{\sum_{j=1}^{s} p_j^{r}(1-p_j)^{n-r}} = \frac{r+1}{n+1}\frac{\sum_{i=1}^{s}\binom{n+1}{r+1} p_i^{r+1}(1-p_i)^{n-r}}{\sum_{j=1}^{s}\binom{n}{r} p_j^{r}(1-p_j)^{n-r}} \\ &= \frac{r+1}{n+1}\frac{E_{n+1}(T_{r+1})}{E_n(T_r)}, \end{aligned}$$

where the subscript near the mean operator shows the size of the sample. If we replace the expectations $E_{n+1}(T_{r+1})$ and $E_n(T_r)$ by observed values t_{r+1} and t_r then we obtain an approximation

$$E(q_r) \approx \frac{r+1}{n+1}\frac{t_{r+1}}{t_r}.$$

The product $E(q_r)t_r$ approximately equals to the total probability of colors, that are represented by r balls in the sample

$$E(q_r)t_r \approx \frac{r+1}{n+1}t_{r+1}.$$

Since the colority of a sample equals to the total probability of colors, that are represented by at least 1 ball in the sample, we obtain the following estimator for the sample coverage

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

When the sample size is large then $n+1$ can be replaced by n:

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

The latter estimator is the Turing estimator of the sample coverage.