SAMPLING FROM POPULATIONS WITH LARGE NUMBER OF CLASSES

Mihhail Juhkam, Kalev Pärna

Problem formulation

- Consider a population divided into mutually exclusive classes, the total number of classes is unknown
- We wish to draw a sample which contains at least one object from each class
- Increasing the sample and identification of membership of objects is often costly or time-consuming
- We may limit ourselves with discovering those classes, which represent a dominating part (e.g. 99%) of the population

Coverage of a sample

- Let s be the number of classes (s is unknown)
- Let the relative frequencies of classes be $p_1 \geq p_2 \geq \ldots \geq p_s$, $\sum_{i=1}^{s} p_i = 1$.
- We call the set $\{p_i\}_{i=1}^{s}$ a ***class distribution*** of population
- Define the ***coverage*** of a sample as the sum of relative frequencies of classes which are represented in the sample
- Two questions arise:
  - Is the coverage of a given sample large enough?
  - If not, then, how many additional objects we have to draw into the sample to achieve the given coverage?

Sampling schemes

- ***Multinomial scheme*** — drawing a sample of size n from urn with replacement
- ***Poisson scheme*** — s simultaneous Poisson processes with itensities $p_1 \geq p_2 \geq \ldots \geq p_s$, sample is drawn during time ν
- These schemes are approximately identical if p_i's are small
- Further we will discuss only Poisson scheme

Coverage as a random variable

- Introduce random indicators

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

$i = 1, 2, \ldots, s$.

- The coverage of the sample then equals

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

- Distribution of I_i^ν:
$P\left\{I_i^\nu = 0\right\} = e^{-\nu p_i}$,
$P\left\{I_i^\nu = 1\right\} = 1 - e^{-\nu p_i}$

Mean and variance of coverage

- Mean value of the coverage

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

- Variance of the coverage

$$DC_\nu = \sum_{i=1}^{s} p_i^2 e^{-\nu p_i} (1 - e^{-\nu p_i})$$

- Example: equiprobable classes $p_i = 1/s$, $i = 1, \ldots, s$

$$EC_\nu = 1 - e^{-\nu/s}$$

$$DC_\nu = \frac{1}{s} e^{-\nu/s} (1 - e^{-\nu/s})$$

Defining class distribution directly

- Class distribution can be defined directly by a nonincreasing function $\pi(i, \vec{\theta})$ of class number i and a vector of parameters $\vec{\theta}$

$$p_i = \pi(i, \vec{\theta})$$

- Examples:
  - Uniform class distribution $\pi(i) = 1/s$, $i = 1, \ldots, s$
  - Linearly decreasing class distribution $\pi(i, \alpha) = p_0 - \alpha i$, $\alpha > 0$, $i = 1, \ldots, s$
  - Exponentially decreasing class distribution $\pi(i, \beta) = p_0 \beta^i$, $0 < \beta < 1$, $i = 1, \ldots, s$

Defining class distribution by density function

- Class distribution can be defined by a density function $f(p)$ which satisfies two conditions
  - $f(p) = 0$ for $p \leq 0$
  - $\int_0^\infty \frac{f(p)}{p} dp < \infty$
- Algorithm:
  - Let $g(p) = f(p)/p$
  - Find points $0 = \xi_s < \xi_{s-1} < \ldots < \xi_1 < \xi_0 = \infty$ such that

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

  - Define the class probabilities by

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

Example: Defining class distribution by density function

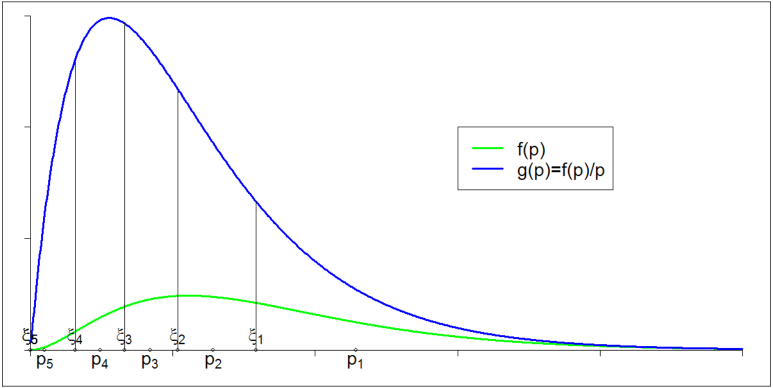

Figure: Defining class distribution by Gamma density

Finding density that produces a given set of class probabilities

- The set of class probabilities is defined directly by function $p_i = \pi(i)$
- How to find such density function f that produces the same set of probabilities?
- The approximate density (when s is large) is

$$f(x) = -x\left(\pi^{-1}(x)\right)', \quad x > 0.$$

- For example: $p_i = p_0\beta^i$, $i = 1, \ldots, s$, $0 < \beta < 1$ then

$$f(x) = -x\left(\frac{\ln x - \ln p_0}{\ln \beta}\right)' = -\frac{1}{\ln \beta}, \quad x \in [p_0\beta^s, p_0\beta]$$

Estimation of sample coverage: nonparametric approach

Turing estimator proposed by I.J. Good (1953)

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

where t_1 is the number of classes in sample, which are represented by exactly one object. Normal limit law for this estimator has been proved

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

as $s \to \infty$

Estimation of sample coverage: parametric approach

- Gamma-Poisson model was used by S. Engen (1974) (probabilities are defined by Gamma density function, Poisson sampling scheme)
- The idea is to
  - estimate the parameters of Gamma distribution
  - approximate the expression of mean coverage

$$
\begin{aligned}
EC_\nu &= \sum_{i=1}^{s} p_i(1-e^{-\nu p_i}) \approx \sum_{i=1}^{s} \int_{\xi_i}^{\xi_{i-1}} p(1-e^{-\nu p})g(p)dp \\
&= \int_0^\infty p(1-e^{-\nu p})g(p)dp = 1 - \int_0^\infty e^{-\nu p} f(p)dp,
\end{aligned}
$$

  where f is density of Gamma distribution and $g(p) = f(p)/p$
  - integrate the last expression and get the estimate of sample coverage in terms of parameters of Gamma distribution

Estimation of required sample size for exponentially decreasing class distribution

- Consider the exponentially decreasing class distribution which is given by

$$p_i = p_0 \beta^i, \quad 0 < \beta < 1, i = 1, \ldots, s$$

- Corresponding density function which defines the same class distribution is given by

$$f(x, \beta) = \begin{cases} -\frac{1}{\ln \beta}, & x \in [p_0 \beta^s, p_0 \beta], \\ 0, & \text{otherwise}, \end{cases}$$

- We are interested in estimating the sample size $\nu_{1-\eta}$, required to achieve the given coverage $1 - \eta$
- Denote by T_x the number of classes which are represented in sample by exactly x objects. The random variables T_x are called ***size indices***

Estimation of required sample size for exponentially decreasing class distribution

- The expectations ET_x, $x = 1, 2, \ldots$ express as follows

$$ET_x = \sum_{i=1}^{s} \frac{(\nu p_i)^x}{x!} e^{-\nu p_i}$$

- It can be shown that $ET_x \to -\frac{1}{x \ln \beta}$ as $s, \nu \to \infty$ so that $s/\nu = const$
- After substituting the expectations ET_x by the realizations t_x of size indices T_x we obtain the system of equations

$$-\frac{1}{x \ln \beta} = t_x, \quad x = 1, 2, \ldots. \tag{1}$$

- The least squares estimate of β from the system (1) (if we use first m equations) is

$$\hat{\beta} = \frac{\sum_{x=1}^{m} \frac{1}{x^2}}{\sum_{x=1}^{m} \frac{t_x}{x}}$$

Estimation of required sample size for exponentially decreasing class distribution

- Next we use the approximative expression of mean sample coverage
$$EC_\nu \approx \int_0^\infty p(1 - e^{-\nu p}) f(p) dp,$$
- After substituting $\nu = \nu_{1-\eta}$, $EC_{\nu_{1-\eta}} = 1 - \eta$ and $f(p) = -\frac{1}{\ln \hat{\beta}}$ we obtain
$$\eta \approx -\frac{1 - \beta^{\nu_{1-\eta}}}{\nu_{1-\eta} \ln \beta} \qquad (2)$$
- The estimate of required sample size $\nu_{1-\eta}$ can be obtained by numerical solving of the equation (2)

Numerical example: estimation of required sample size

- We simulated 100 multinomial samples of size $\nu = 500$ from 4 populations with exponentially decreasing class distribution with $\beta = 0.95, 0.97, 0.98$ and 0.99.
- We estimated required sample sizes $\nu_{0.99}$, $\nu_{0.995}$ and $\nu_{0.999}$ for acieving coverages 0.99, 0.995 and 0.999 using method proposed before
- We obtained the actual requred sample sizes $\nu_{0.99}$, $\nu_{0.995}$ and $\nu_{0.999}$ by continuing simulation
- The relative errors of estimate were calculated and averaged over 100 samples

$$\rho = AVG\left(\frac{\hat{\nu}_{1-\eta} - \nu_{1-\eta}}{\nu_{1-\eta}}\right)$$

Class distribution used in example

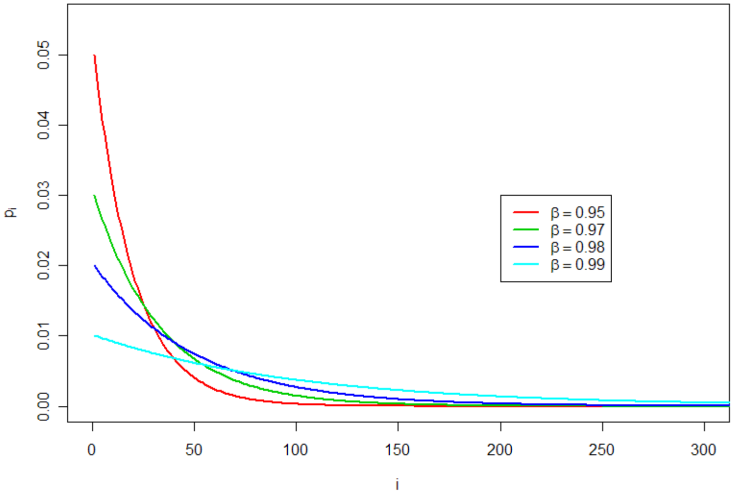

Figure: Exponentially decreasing class distribution used in example

Results of numerical example: estimation of required sample size

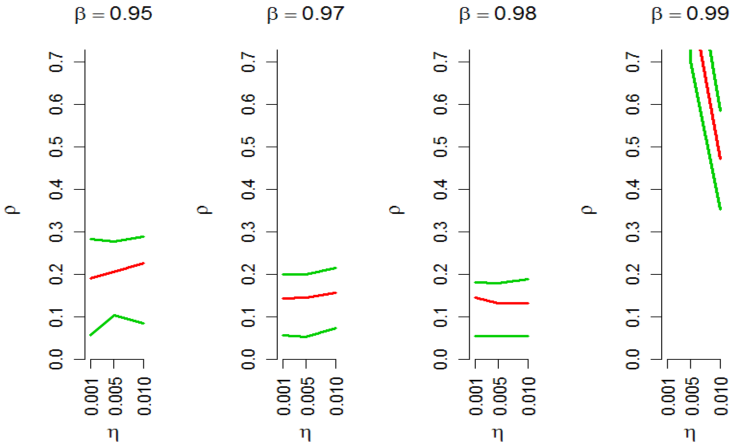

Figure: Average relative error of estimated sample size

Conclusion

- The proposed metod for estimation of the required sample size works for populations with large number of classes and vor values of β wich are not very close to 1
- How to check assumption that the class distribution is exponentially decreasing?
- Is there any simple estimate for some other family of class distributions?