INTERVAL ESTIMATION FOR SHANNON QUASI-NORM

Jakub Šácha*

The paper is concerned with estimation of the probability function of a discrete random variable by minimizing the Shannon quasi-norm while meeting the moment conditions for the probability function estimated. After describing this method in detail, the paper further focuses on deriving confidence intervals for probabilities and possible application of these methods to particular data sets.

Keywords: *divergence, quasi-norm, estimation, confidence interval*

1. Introduction

Estimating the shape of a probability function is among the most important problems of mathematical statistics. The paper is concerned with estimation the probability function of a discrete random variable using the concept of an f-divergence (of the distance) of two distributions [3]. To use this method, however, usually some other conditions must be fulfilled concerning the probability. These are mostly given by the values assigned initially to selected parameters such as mean value and variance. The papers published to date on this topic have been concerned with point estimation of the values of the probability function of an unknown discrete random variable using a chosen quasi-norm and meeting the moment conditions [3], [4], [5], [6], [7] or general linear conditions [8] of the distribution estimated. Another line of research in the field of using quasi-norms focused on estimation of categorial variable probability distribution [9].

This paper is a continuation of a previous one and deals with the derivation of confidence intervals for the values of the probability function for Shannon's quasi-norm. Finally, this method is applied to a particular problem.

2. f-divergence of discrete probability distributions

We denote $\mathbb{R}$ the set of real numbers, and $\mathbb{R}^*$ the set of real numbers extended with the improper elements $-\infty$ and $+\infty$. The set $I(a,b) \subset \mathbb{R}^*$ is open, half-open or closed interval, which can be bounded or unbounded and, furthermore, extended with the improper elements $-\infty$ and $+\infty$. Definitions and Theorems in this section are taken from [3].

A function $f : I(a,b) \to \mathbb{R}^*$ is called *convex at point* $u_0 \in (a,b)$, if f is continuous into interval $(a,b) \subset I(a,b)$, continuous from the right at a and continuous from the left at b if theese points belong to $I(a,b)$, and such a number $\lambda(u_0) \in \mathbb{R}$ exists, that $f(u) \geq f(u_0) + \lambda(u_0)(u - u_0)$ for all $u \in I(a,b)$, $u \neq u_0$. A convex function f is called *strictly convex at the point* u_0, if the above non-strict inequality is replaced by a strict one.

* Ing. J. Šácha, Institute of Mathematics, Faculty of Mechanical Engineering, Brno University of Technology, Technická 2, 616 69 Brno, Czech Republic

A function f is called *convex*, *strictly convex*, into $I(a,b)$, if f is convex, strictly convex, respectively, at every point $u \in I(a,b)$.

If a function $f : [0,\infty] \to \mathbb{R}^*$ is convex in $[0,\infty]$, and strictly convex in $u = 1$, then the limit

$$f(*) = \lim_{u\to\infty} \frac{f(u)}{u} \in \mathbb{R}^*$$

exists, and

$$-\infty < f(1) < f(0) + f(*) \ .$$

Furthermore, for every $v_0 \in (0,\infty)$, we have

$$\lim_{\substack{u\to 0+\\ v\to v_0}} u\, f\left(\frac{v}{u}\right) = v_0\, f(*) \ , \qquad \lim_{\substack{u\to 0+\\ v\to v_0}} v\, f\left(\frac{v}{u}\right) = v_0\, f(0) \ .$$

Consider a discrete probability space (Ω, Σ, P) where Ω is a finite set, Σ is a σ-algebra, and P is a probability measure. Without loss of generality, we can assume that the probability function $\mathbf{p}$ of some discrete random variable X agrees with the probability measure P.

Let a function $f(u)$ be convex in $(0,\infty)$, strictly convex at $u = 1$, and $f(1) = 0$. We define an *f-divergence* of probability distributions $\mathbf{p}$ and $\mathbf{q}$ on the discrete probability space (Ω, Σ, P) as the functional

$$D_{\mathrm{f}}(\mathbf{p},\mathbf{q}) = \sum_x q(x)\, f\left(\frac{p(x)}{q(x)}\right)$$

where we put $0f(0/0) = 0$, and $0f(p/0) = p\, f(*)$ for all $p \in [0,1]$.

The probability functions $\mathbf{p}$ and $\mathbf{q}$ on the discrete probability space (Ω, Σ, P) are *orthogonal*, and we write $\mathbf{p}\perp\mathbf{q}$, if such disjoint sets $E, F \subset \Omega$ exist that $\sum_{x\in E} p(x) = 1$ and $\sum_{x\in F} q(x) = 1$.

Let, for an arbitrary f-divergence, the below inequality holds

$$0 \le D_{\mathrm{f}}(\mathbf{p},\mathbf{q}) \le f(0) + f(*)$$

where not both equalities can be true simultaneously. The left equality is valid if $\mathbf{p} = \mathbf{q}$, and the right equality is valid if $\mathbf{p}$ and $\mathbf{q}$ are orthogonal, and simultaneously $f(0) + f(*) < \infty$.

The most widely used f-divergences are listed in Table 1.

$f(u)$	Notation Name	$D_{\mathrm{f}}(\mathbf{p},\mathbf{q})$
$u \ln u$	$I(\mathbf{p},\mathbf{q})$ I-divergence	$\sum_x p(x) \ln \frac{p(x)}{q(x)}$
$(u^{1/2}-1)^2$	$D_{1/2}(\mathbf{p},\mathbf{q})$ Hellinger distance	$1 - \sum_x (p(x)\, q(x))^{1/2}$
$(u-1)^2$	$\chi^2(\mathbf{p},\mathbf{q})$ χ^2-divergence	$\sum_x \frac{(p(x)-q(x))^2}{q(x)}$

Tab.1: The most widely used f-divergences

3. Quasi-norms of discrete probability distributions

Let $D_{\mathrm{f}}(\mathbf{p},\mathbf{q})$ be a f-divergence of probability distributions $\mathbf{p} = (p_1,\ldots,p_m)$ and $\mathbf{q} = (q_1,\ldots,q_m)$ where $m > 1$ on a finite probability space (Ω,Σ,P) and $S = \{\mathbf{p} \in \mathbb{R}^m : \forall p_j \geq 0, \sum_{j=1}^m p_j = 1\}$. We designate

$$V(\mathbf{q}) = \int_S D_{\mathrm{f}}(\mathbf{p},\mathbf{q})\,\mathrm{d}S$$

the integral of f-divergences of all distributions $\mathbf{p}$ from a some stationary selected distribution $\mathbf{q} \in S$ [5]. If the function $V(q)$ and the function

$$G(q_j) = \frac{\partial V(\mathbf{q})}{\partial q_j} = \int_S \frac{\partial D_{\mathrm{f}}(\mathbf{p},\mathbf{q})}{\partial q_j}\,\mathrm{d}S$$

exist (i.e. the both integrals converge), further if the derivation $G'(q_j)$ exists in $[0,1]$, and the function f has continuous second derivation in $(0,\infty)$, then $V(\mathbf{q})$ take the absolute minimum into S for the probability distribution

$$\mathbf{q} = \mathbf{p}_0 = \left(\frac{1}{m},\ldots,\frac{1}{m}\right) .$$

(S is the set of all probability distributions on (Ω,Σ,P), and the both mentioned integrals are of the first type on the hypersurface S with dimension $m-1$ in $\mathbb{R}^m$.)

In line with the above proposition, we choose the probability function $\mathbf{p}_0 = (1/m, \ldots, 1/m)$ to estimate the unknown probability function of the observed discrete random variable X. The probability function $\mathbf{p}_0$ is closest to all the probability functions on (Ω,Σ,P) as measured by the integral $V(\mathbf{q})$ of the distances (f-divergences), and, in addition, $\mathbf{p}_0$ has also the maximum vagueness expressed by the Shannon entropy. By analogy to the introduction of an induced norm in a linear space with metric by means of a neutral element, we will define the following concept.

Let $\mathbf{p} = (p_1,\ldots,p_m)$, and $\mathbf{p}_0 = (1/m,\ldots,1/m)$ where $m > 1$ be discrete probability distributions on probability space (Ω,Σ,P), and D_{f} a f-divergence defined on given space. Then the f-divergence $D_{\mathrm{f}}(\mathbf{p},\mathbf{p}_0)$ is called the *quasi-norm* of the probability distribution $\mathbf{p} = (p_1,\ldots,p_m)$ on (Ω,Σ,P).

Any quasi-norm $D_{\mathrm{f}}(\mathbf{p},\mathbf{p}_0)$ has the following properties:

a) $D_{\mathrm{f}}(\mathbf{p},\mathbf{p}_0) = (1/m)\sum_{j=1}^m f(m\,p_j)$,

b) $D_{\mathrm{f}}(\mathbf{p},\mathbf{p}_0)$ is a non-negative symmetric function of variables p_j, $j = 1,\ldots,m$.

In the following sections we deal with *Shannon quasi-norm* $S(\mathbf{p},\mathbf{p}_0)$ which is implied by I-divergence $I(\mathbf{p},\mathbf{q})$

$$S(\mathbf{p},\mathbf{p}_0) = \sum_{j=1}^m \left(p_j\,\ln p_j - \frac{1}{m}\,\ln\frac{1}{m}\right) = \sum_{j=1}^m p_j\,\ln p_j + \ln m\ .$$

4. Estimation of discrete probability distribution

Let (Ω, Σ, P) where $\Omega = \{x_1^*, \ldots, x_m^*\} \subset \mathbb{R}$, $m > 1$, and Σ set of all subsets of Ω be a discrete probability space. We assume that a discrete random variable X on (Ω, Σ, P) takes mutually different values x_j^* with unknown probabilities

$$p_j = P(X = x_j^*) \ , \quad j = 1, \ldots, m \ , \quad m > 1 \ .$$

By observing the random variable X, we receive a sample $(x_1, \ldots, x_n)$ and, processing it, we get the grouped sample data

$$\left(\left(x_1^*, \frac{f_1}{n} \right), \ldots, \left(x_m^*, \frac{f_m}{n} \right) \right)$$

where f_j is the absolute frequency of observed value x_j^*. Next we assume that $n > m$ and $f_j > 0$ for any $j = 1, \ldots, m$. If, after n trials, we receive frequency $f_j = 0$, then we skip the data class with index number j. We denote

$$M_k = \sum_{j=1}^{m} \frac{f_j}{n} \, x_j^{*k} \ , \quad k = 0, \ldots, K$$

the $K + 1$ first empirical general moments of sample $(x_1, \ldots, x_n)$.

An *estimation* $\mathbf{p}(\boldsymbol{\lambda}) = (p_1(\boldsymbol{\lambda}), \ldots, p_m(\boldsymbol{\lambda}))$ of the probability distribution of the observed discrete random variable X where $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_K)$ is a real vector of parameters has a *minimum quasi-norm satisfying the* $K < m - 1$ *moment constraints*

$$\sum_{j=1}^{m} p_j \, x_j^{*k} = M_k \ , \quad k = 0, \ldots, K$$

if the quasi-norm $D_{\mathrm{f}}(\mathbf{p}(\boldsymbol{\lambda}), \mathbf{p}_0)$ is minimum for given values M_k.

For the estimation of discrete probability $\mathbf{p} = (p_1, \ldots, p_m)$ using moments M_k, we choose Shannon quasi-norm.

For $K < m - 1$, we have [2], [4]

$$p_j(\boldsymbol{\lambda}) = \exp\left(-1 - \sum_{k=0}^{K} \lambda_k \, x_j^{*k} \right) \ , \quad j = 1, \ldots, m$$

where λ_k, $k = 0, \ldots, K$, are Lagrange multipliers for Lagrange function

$$\Lambda(\mathbf{p}, \boldsymbol{\lambda}) = S(\mathbf{p}, \mathbf{p}_0) + \sum_{k=0}^{K} \lambda_k \left(\sum_{j=1}^{m} p_j \, x_j^{*k} - M_k \right) \ .$$

The estimates of parameters λ_k for Shannon quasi-norm have a maximum likelihood [2]. If we denote $S_K = \min S(\mathbf{p}(\boldsymbol{\lambda}), \mathbf{p}_0)$, then

$$S_K = \ln m - \sum_{j=1}^{m} \left(1 + \sum_{k=0}^{K} \lambda_k \, x_j^{*k} \right) \exp\left(-1 - \sum_{k=0}^{K} \lambda_k \, x_j^{*k} \right) \ .$$

For $K = 0$, we have $p_j = \frac{1}{m}$, $j = 1, \ldots, m$ and $S_0 = 0$. Specifically, for $K = m - 1$, this is an interpolation $p_j = f_j/n$, $j = 1, \ldots, m$ and $S_{m-1} = (1/n) \sum_{j=1}^{m} f_j \ln f_j + \ln(m/n)$. We have $S_0 \leq \cdots \leq S_{m-1}$.

The Lagrange multipliers λ_k can be calculated from the system of non-linear equations that matches the zero gradient of Lagrange function or by directly applying some non-linear optimizing method.

5. Confidence intervals of probabilities p_j for Shannon's quasi-norm

5.1. Deriving the distribution law for the vector of Lagrange multipliers

For Shannon's quasi-norm, estimates of the parameters λ_k are maximally reliable as this is also estimation of parameters using a modified method of minimum chi-square [1]. The estimate of the vector $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_K)$ has an asymptotically $(K+1)$-dimensional normal distribution $N(\boldsymbol{\lambda}_0, n^{-1}[\mathbf{J}(\boldsymbol{\lambda}_0)]^{-1})$ where $\boldsymbol{\lambda}_0$ is the theoretical vector of the parameters of the distribution of random variable X and $\mathbf{J}(\boldsymbol{\lambda}_0)$ is Fisher's information matrix [1].

Specifically, the estimate λ_k of the theoretical value of a fixed multiplier λ_{0k}, $k = 0, \ldots, K$ has an asymptotically normal distribution $N(\lambda_{0k}, 1/(n\, J(\lambda_{0k})))$. Denote $\bar{\boldsymbol{\lambda}} = \left(\bar{\lambda}_0, \ldots, \bar{\lambda}_K\right)$ the estimate of the vector of Lagrange multipliers obtained by solving the minimization problem for a particular observation. To calculate a confidence interval of λ_{0k} using this normal distribution, we put $\lambda_{0k} = \bar{\lambda}_k$. Fisher's measure of information is

$$J(\lambda_{0k}) = E\left(\left[\frac{\partial \ln p(x, \boldsymbol{\lambda}_0)}{\partial \lambda_{0k}}\right]^2\right) ,$$

where E stands for the mean value, $J(\lambda_{0k}) = J(\bar{\lambda}_k)$ a $x = x_j^*$. Simplifying, we get

$$J(\bar{\lambda}_k) = \sum_{j=1}^{m} p_j\, x_j^{*2k} \approx \sum_{j=1}^{m} \frac{f_j}{n}\, x_j^{*2k} = M_{2k} ,$$

where the probabilities p_j can be determined by minimizing Shannon's quasi-norm.

Similarly, for an entry of Fischer's information matrix, we get J_{ts} for $t, s = 0, \ldots, K$

$$J_{ts} = E\left(\frac{\partial \ln p(x, \boldsymbol{\lambda}_0)}{\partial \lambda_{0t}}\, \frac{\partial \ln p(x, \boldsymbol{\lambda}_0)}{\partial \lambda_{0s}}\right) .$$

If we use $p_j \approx f_j/n$ as an initial estimate, we get

$$J_{ts} = \sum_{j=1}^{m} \frac{f_j}{n}\, x_j^{*(t+s)} = M_{t+s} .$$

The above reasoning can be summarized in the following theorem:

Theorem 5.1 *The vector of Lagrange multipliers $\boldsymbol{\lambda}$ has an asymptotically $(K+1)$-variate normal distribution $N(\bar{\boldsymbol{\lambda}}, [n\,\mathbf{J}]^{-1})$ where the vector of mean values $\bar{\boldsymbol{\lambda}} = (\bar{\lambda}_0, \ldots, \bar{\lambda}_K)$ is obtained by solving a minimization problem and*

$$\mathbf{J} = \begin{pmatrix} M_0 & M_1 & M_2 & \ldots & M_K \\ M_1 & M_2 & M_3 & \ldots & M_{K+1} \\ M_2 & M_3 & M_4 & \ldots & M_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ M_K & M_{K+1} & M_{K+2} & \ldots & M_{2K} \end{pmatrix}$$

is an estimate of Fisher's information matrix.

5.2. Deriving the distribution law for the vector of probabilities

Now we will be concerned with confidence intervals of the probabilities p_j. To do this we need to determine the form of the distribution of $\mathbf{p} = (p_1, \ldots, p_m)$. As the vector $\mathbf{p}$ depends on the vector $\boldsymbol{\lambda}$, which has an asymptotically normal distribution [1], it is clear that the distribution of $\mathbf{p}$ is again asymptotically normal. The dependency has the form

$$p_j\left(\boldsymbol{\lambda}\right) = \exp\left(-1 - \sum_{k=0}^{K} \lambda_k \, x_j^{*k}\right) \;, \quad j = 1, \ldots, m \;.$$

The above non-linear relationship can be linearized to $\mathbf{p} = \mathbf{a} + \mathbf{B}\,\boldsymbol{\lambda}$ using first order Taylor polynomial. We choose the point $\bar{\boldsymbol{\lambda}} = (\bar{\lambda}_0, \ldots, \bar{\lambda}_K)$ as the centre of this Taylor expansion. We put

$$\overline{p}_j = \exp\left(-1 - \sum_{k=0}^{K} \bar{\lambda}_k x_j^{*k}\right) \;, \quad j = 1, \ldots, m \;.$$

It is in the neighbourhood of $\bar{\mathbf{p}} = (\bar{p}_1, \ldots, \bar{p}_m)$ that we will be most interested in the values of the probability function of vector $\mathbf{p}$. The Taylor polynomial then has the following form

$$T(p_j)(\boldsymbol{\lambda}) = p_j(\bar{\boldsymbol{\lambda}}) + \sum_{i=0}^{K} \frac{\partial p_j}{\partial \lambda_i}\left(\bar{\lambda}_i\right)\left(\lambda_i - \bar{\lambda}_i\right) \;.$$

Substituting for

$$\frac{\partial p_j(\bar{\lambda}_i)}{\partial \lambda_i} = \exp\left(-1 - \sum_{k=0}^{K} \bar{\lambda}_k \, x_j^{*k}\right) x_j^{*i} \;, \quad j = 1, \ldots, m \;,$$

we get

$$T(p_j)(\boldsymbol{\lambda}) = \exp\left(-1 - \sum_{k=0}^{K} \bar{\lambda}_k \, x_j^{*k}\right) + \sum_{i=0}^{K} x_j^{*i} \, \exp\left(-1 - \sum_{k=0}^{K} \bar{\lambda}_k \, x_j^{*k}\right)\left(\lambda_i - \bar{\lambda}_i\right) \;.$$

After some simplification,

$$T(p_j)(\boldsymbol{\lambda}) = \bar{p}_j\left(1 - \sum_{i=0}^{K} \bar{\lambda}_i x_j^{*i}\right) + \bar{p}_j \sum_{i=0}^{K} \lambda_i \, x_j^{*i} \;.$$

Thus, we can write

$$\mathbf{p} \approx \mathbf{a} + \mathbf{B}\,\boldsymbol{\lambda} \;,$$

where

$$\mathbf{a} = \begin{pmatrix} \bar{p}_1\left(1 - \sum\limits_{i=0}^{K} \bar{\lambda}_i \, x_1^{*i}\right) \\ \vdots \\ \bar{p}_m\left(1 - \sum\limits_{i=0}^{K} \bar{\lambda}_i \, x_m^{*i}\right) \end{pmatrix}$$

and

$$\mathbf{B} = \begin{pmatrix} \bar{p}_1\, x_1^{*0} & \bar{p}_1\, x_1^{*1} & \dots & \bar{p}_1\, x_1^{*K} \\ \bar{p}_2\, x_2^{*0} & \bar{p}_2\, x_2^{*1} & \dots & \bar{p}_2\, x_2^{*K} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{p}_m\, x_m^{*0} & \bar{p}_m\, x_m^{*1} & \dots & \bar{p}_m\, x_m^{*K} \end{pmatrix} .$$

Using the theorem on a linear transformation of a normally distributed random vector [1]. We will obtain the parameters of the normal distribution of the vector $\mathbf{p}$. It is clear that the mean value of $E(\mathbf{p}) = \mathbf{a} + \mathbf{B}\,\boldsymbol{\lambda} = \mathbf{p} = (\bar{p}_1, \dots, \bar{p}_m)$. The above reasoning can be summarized in the following theorem:

Theorem 5.2 *The vector of probabilities* $\mathbf{p}$ *has an asymptotically* m*-variate normal distribution* $N(\bar{\mathbf{p}}, \mathbf{B}\,[n\mathbf{J}]^{-1}\,\mathbf{B}^{\mathrm{T}})$ *where the vector of mean values* $\bar{\boldsymbol{\lambda}} = (\bar{\lambda}_0, \dots, \bar{\lambda}_K)$ *and the vector* $\bar{\mathbf{p}} = (\bar{p}_1, \dots, \bar{p}_m)$ *can be obtained by minimizing Shannon's quasi-norm under the given conditions,*

$$\mathbf{J} = \begin{pmatrix} M_0 & M_1 & M_2 & \dots & M_K \\ M_1 & M_2 & M_3 & \dots & M_{K+1} \\ M_2 & M_3 & M_4 & \dots & M_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ M_K & M_{K+1} & M_{K+2} & \dots & M_{2K} \end{pmatrix}$$

and

$$\mathbf{B} = \begin{pmatrix} \bar{p}_1\, x_1^{*0} & \bar{p}_1\, x_1^{*1} & \dots & \bar{p}_1\, x_1^{*K} \\ \bar{p}_2\, x_2^{*0} & \bar{p}_2\, x_2^{*1} & \dots & \bar{p}_2\, x_2^{*K} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{p}_m\, x_m^{*0} & \bar{p}_m\, x_m^{*1} & \dots & \bar{p}_m\, x_m^{*K} \end{pmatrix} .$$

Using this theorem, it is easy to construct confidence intervals (both individual and coupled) of probabilities p_j for Shannon's quasi-norm. We will do this in the following example.

6. Example

Observing a random variable X assuming the values $x = 1, \dots, 7$, which can indicate the number of defects in a microscope window, we produced a sample sized $n = 180$. After sorting the data, we obtained a discrete empirical distribution of X in the below table 2, x_j^* are middles of classes and f_j are the observed absolute frequencies:

x_j^*	1	2	3	4	5	6	7
f_j	15	36	21	15	27	42	24

Tab.2

Let us determine 95 % individual confidence intervals of probabilities p_j meeting minimum Shannon's quasi-norm under secondary conditions given by the general moments

$$M_0 = \frac{1}{n} \sum_{j=1}^{m} f_j = 1 \ ,$$

$$M_1 = \frac{1}{n} \sum_{j=1}^{m} f_j\, x_j^* = 4.25 \ ,$$

$$M_2 = \frac{1}{n}\sum_{j=1}^{m} f_j\, x_j^{*2} = 21.95\ ,$$

$$M_3 = \frac{1}{n}\sum_{j=1}^{m} f_j\, x_j^{*3} = 125.05\ .$$

This is a non-linear optimization problem, with a solution whose existence and uniqueness is guaranteed by the convexity of the objective function and linearity (in probabilities p_j) of the moment constraints. A classical analytical approach with Lagrange multipliers was used to solve it. The solution to the corresponding nonlinear system of equations and other numerical calculations were done in Maple. Let us choose, step by step, the number of moment constraints $K = 1, 2, 3, 4$. The results are shown in the following summary with graphics indicating the 95 % confidence intervals in figures 1, 2, 3, 4.

- $K = 0$

  Here are the point estimates of the Lagrange multipliers

  $$\boldsymbol{\lambda} = (0.94591)\ ,$$
  $$\mathbf{p} = (0.14286, 0.14286, 0.14286, 0.14286, 0.14286, 0.14286, 0.14286)\ .$$

  And here the 95 % confidence intervals for probabilities p_j

  $$P(p_j \in (0.14263, 0.14308)) = 0.95$$

  for all $j = 1, 2 \ldots, 7$.

- $K = 1$

  Here are the point estimates of the Lagrange multipliers and probabilities

  $$\boldsymbol{\lambda} = (-1.20458, 0.06270)\ ,$$
  $$\mathbf{p} = (0.11743, 0.12503, 0.13313, 0.14174, 0.15091, 0.16068, 0.17108)\ .$$

  And the 95 % confidence intervals of the probabilities p_j are the following

  $$P(p_1 \in (0.11688, 0.11799)) = 0.95\ ,$$
  $$P(p_2 \in (0.12464, 0.12543)) = 0.95\ ,$$
  $$P(p_3 \in (0.13285, 0.13340)) = 0.95\ ,$$

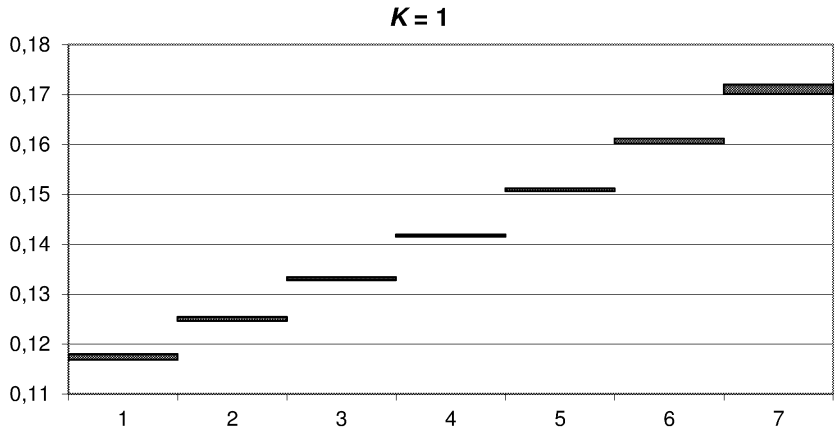

Fig.1: 95 % confidence intervals for $K = 1$

$$P(p_4 \in (0.14152, 0.14196)) = 0.95 ,$$
$$P(p_5 \in (0.15063, 0.15120)) = 0.95 ,$$
$$P(p_6 \in (0.16018, 0.16118)) = 0.95 ,$$
$$P(p_7 \in (0.17014, 0.17201)) = 0.95 .$$

- $K = 2$

  Here are the point estimates of the Lagrange multipliers and probabilities

$$\boldsymbol{\lambda} = (-1.28410, 0.11356, -0.00621) ,$$
$$\mathbf{p} = (0.11340, 0.12469, 0.13542, 0.14525, 0.15387, 0.16100, 0.16637) .$$

  And the 95 % confidence intervals of the probabilities p_j are

$$P(p_1 \in (0.11239, 0.11441)) = 0.95 ,$$
$$P(p_2 \in (0.12430, 0.12508)) = 0.95 ,$$
$$P(p_3 \in (0.13497, 0.13586)) = 0.95 ,$$
$$P(p_4 \in (0.14463, 0.14587)) = 0.95 ,$$
$$P(p_5 \in (0.15331, 0.15444)) = 0.95 ,$$
$$P(p_6 \in (0.16049, 0.16150)) = 0.95 ,$$
$$P(p_7 \in (0.16479, 0.16796)) = 0.95 .$$

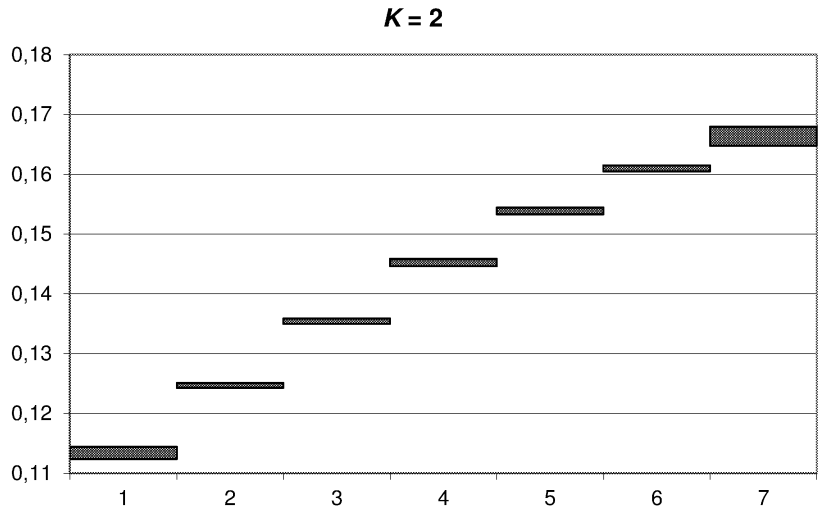

Fig.2: 95 % confidence intervals for $K = 2$

- $K = 3$. The point estimates of the Lagrange multipliers and probabilities are

$$\boldsymbol{\lambda} = (-1.17237, 0.01058, -0.02944, -0.293) ,$$
$$\mathbf{p} = (0.11574, 0.12255, 0.13288, 0.14496, 0.15634, 0.16378, 0.16376) .$$

  And the 95 % confidence intervals of the probabilities p_j are the following

$$P(p_1 \in (0.11421, 0.11726)) = 0.95 ,$$
$$P(p_2 \in (0.12207, 0.12303)) = 0.95 ,$$
$$P(p_3 \in (0.13222, 0.13353)) = 0.95 ,$$

$$P(p_4 \in (0.14433, 0.14559)) = 0.95 ,$$
$$P(p_5 \in (0.15558, 0.15710)) = 0.95 ,$$
$$P(p_6 \in (0.16306, 0.16450)) = 0.95 ,$$
$$P(p_7 \in (0.16177, 0.16575)) = 0.95 .$$

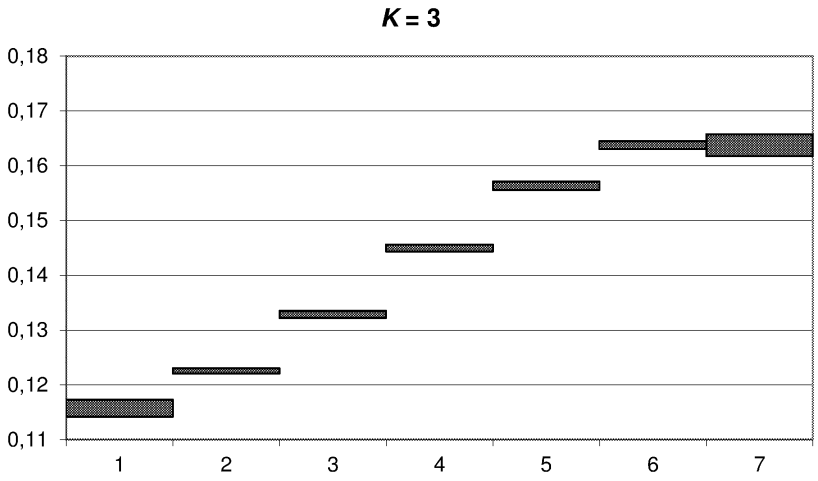

Fig.3: 95 % confidence intervals for $K = 3$

- $K = 4$. The point estimates of the Lagrange multipliers are

$$\boldsymbol{\lambda} = (-6.01720, 7.47108, -3.52618, 0.64806, -0.04032) ,$$
$$\mathbf{p} = (0.08505, 0.19390, 0.12147, 0.09085, 0.13391, 0.24395, 0.13087) .$$

and the 95 % confidence intervals of the probabilities p_j are the following

$$P(p_1 \in (0.08412, 0.08597)) = 0.95 ,$$
$$P(p_2 \in (0.19218, 0.19562)) = 0.95 ,$$
$$P(p_3 \in (0.12089, 0.12206)) = 0.95 ,$$
$$P(p_4 \in (0.09038, 0.09132)) = 0.95 ,$$
$$P(p_5 \in (0.13328, 0.13454)) = 0.95 ,$$
$$P(p_6 \in (0.24159, 0.24630)) = 0.95 ,$$
$$P(p_7 \in (0.12949, 0.13225)) = 0.95 .$$

Pearson test was used to check whether the number of moment constraints $K = 4$ is sufficient for the estimate to comply with the observation at a significance level of 0.05 [6].

7. Conclusion

The traditionally used methods to estimate the probability distribution and the software tools for implementing them usually only concentrate on the most ordinary distribution types. This can be restrictive in some application area. Application of the present method

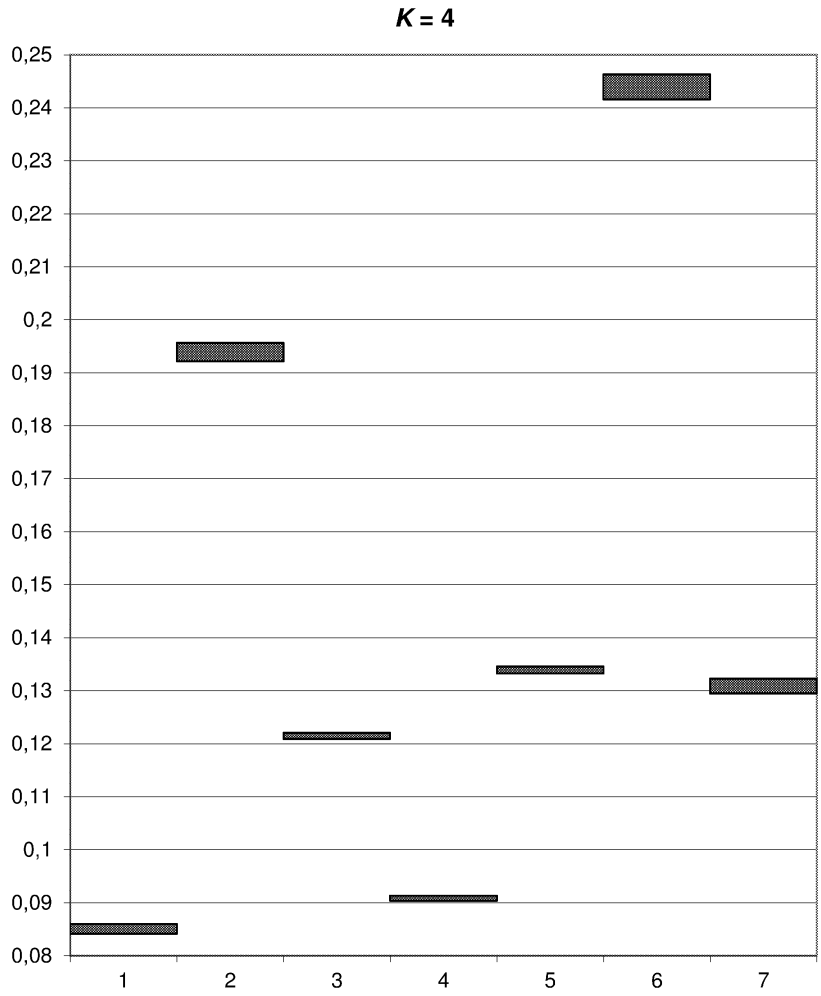

Fig.4: 95 % confidence intervals for $K = 4$

prefers no particular distribution type. It only requires the knowledge of estimates of the distribution characteristics chosen (such as their mean values, variances, and higher moments) and, taking them into consideration, seeks the most uncertain probability distribution $\mathbf{p}$ of the probability space in question. This basic concept is the reason why the estimates are sometimes called 'pessimistic'. The method can also be applied to multi-modus distributions as our example demonstrates, multivariate populations and discretized continuous random variables.

The construction of confidence intervals for p_i is based on the maximum confidence of Lagrange multiplier estimates λ_i of the optimisation problem under consideration and, thus, an asymptotic normality of the distribution of the vector $\boldsymbol{\lambda}$. As the vector $\mathbf{p}$ is a non-linear function of the vector $\boldsymbol{\lambda}$, the assumed normality of the vector $\mathbf{p}$ is, thus, only approximate. This presents, however, no problem with large sizes.

The paper is only concerned with individual confidence intervals of the probabilities p_i with coupled interval confidences being a further possible extension.

References

[1] Anděl J.: Matematická statistika, Praha, Bratislava: SNTL -ALFA, 1978
[2] Karpíšek, Z., Jurák, P.: Modeling of Probability Distribution with Maximum Entropy, MENDEL '01, 7th International Conference on Soft Computing, Brno, 2001, pp. 232–239, ISBN 80-214-1894-X
[3] Vajda I.: Theory of statistical inference and information, Bratislava: Kluwer Academic Publishers, 1989
[4] Karpíšek Z., Jurák P.: Estimate of Discrete Probability Distribution by Means of Hellinger Distance, MENDEL'02, 8th International Conference on Soft Computing. Brno, 2002, pp. 301–306, ISBN 80-214-2135-5
[5] Jurák P., Karpíšek Z.: Hellinger Quasi-norm and Shannon Quasi-norm in N-dimensional space, MENDEL '04, 10th International Conference on Soft Computing, Brno, 2004, pp. 210–215, ISBN 80-214-2676-4
[6] Šácha J., Karpíšek Z., Jurák P.: Qusi-Norms for Discrete Probability Distribution Estimation, Risk, quality and reliability, Ostrava, Czech republic: VŠB – Technical University of Ostrava, 2007, s. 179–185. ISBN: 978-80-248-1575-6
[7] Karpíšek Z., Jurák P., Šácha J.: Divergences for Discrete Probability Distribution Estimations, Proceedings of Summer School DATASTAT '06, Brno: Masaryk University Brno, 2007, s. 109–120, ISBN: 978-80-210-4493-7
[8] Šácha J., Karpíšek Z.: Odhad rozdělení pravděpodobnosti s obecnými lineárními podmínkami, Informační bulletin České statistické společnosti, roč. 22 (2), Praha, 2011, pp. 192–199, ISSN 1210-8022
[9] Lacinová V., Karpíšek Z., Sadovský Z.: Pesimistické odhady rozdělení pravděpodobnosti kategoriální veličiny, Informační bulletin České statistické společnosti, roč. 22 (2), Praha, 2011, pp. 138–145, ISSN 1210-8022

Received in editor's office: August 1, 2012
Approved for publishing: September 20, 2012