Testing of hypothesis of random variables independence on the basis of nonparametric method of confidence estimation of probability density

A V Lapko[1,2], V A Lapko[1,2], E A Yuronen[1]
[1]Reshetnev Siberian State Aerospace University
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russian Federation
[2]Institute of Computer Modeling Siberian Branch of the RAS
50, Akademgorodok, Krasnoyarsk, 660036, Russian Federation

E-mail: lapko@icm.krasn.ru

Abstract. Algorithmic approach of hypothesis testing of identity of distribution laws of random values is offered. Its basis is made by technique of creation of confidence borders for a probability density.

1. Introduction

Problems of hypothesis testing about random distributions are one of the main in the theory of mathematical statistics. For their decision the Pearson's criterion which does not depend on random distributions and their dimension is widely used [1]. However the formation technique of this criterion contains difficult formalizable stage of splitting a range of values of random values into intervals. This problem is absent in criterion of Kolmogorov-Smirnov which is used at a hypothesis testing about distributions of one-dimensional random values [2]. The technique based on replacement of a problem of hypothesis testing about random distributions by a problem of hypothesis testing about equality of probabilities of an error of pattern recognition to threshold value [3, 4].

Problems of statistical hypotheses testing of random distributions are intimately bound to creation of confidence regions for their probability densities or cumulative distribution functions [1].

In this thesis the algorithmic approach of hypothesis testing of random distributions using confidence borders of their probability densities is offered and investigated. The basis of approach is made by a regression assessment of a probability density.

2. Regression assessment of the probability density and its property

Let there is a selection $V = \left(x^i, i = \overline{1,n}\right)$ from n independent values of a one-dimensional random value x with unknown probability density $p(x)$.

For estimation of a probability density $p(x)$ in the conditions of selections of large volume we will use a technique of "compression" of input statistical data V [5]. Let's break definition range $p(x)$ into N not crossed intervals 2β long and we will create sets of random values X^j, $j = \overline{1,N}$. As characteristics X^j we will accept frequency $\overline{P}^j$ hits of a random value x in the j-th interval and its center z^j. On the basis of the received information we will define

a data array $V_1 = \left(z^j, y^j = \overline{P}^j/(2\beta), j = \overline{1,N}\right)$ made of the centers z^j injected intervals and values of estimates of probability densities y^j corresponding to them.

For a quantity choice N intervals of sampling of a range of values of a random value it is possible to use recommendations of publications [6-10].

As an assessment of a required probability density $p(x)$ we will accept statistics

$$\tilde{p}(x) = c^{-1} \sum_{j=1}^{N} \overline{P}^j \Phi\left(\frac{x - z^j}{c}\right) \tag{1}$$

in which nuclear functions $\Phi(u)$ obey H:

$$\Phi(u) = \Phi(-u), \quad 0 \le \Phi(u) < \infty,$$
$$\int_{-\infty}^{+\infty} \Phi(u)\,du = 1, \quad \int_{-\infty}^{+\infty} u^2 \Phi(u)\,du = 1.$$

Values of diffuseness coefficients of nuclear functions $c(N)$ in expression (1) decrease with body height N.

Asymptotic properties of statistics (1) are defined by the following statement.

Theorem. Let: 1) $p(x)$ is limited and continuous with all the derivatives to the second order inclusive; 2) nuclear functions $\Phi(u)$ obey H; 3) at $N \to \infty$, c and $\beta \to 0$; $\beta/c \to 0$, and $nc \to \infty$. Then the regression assessment of a probability density possesses properties of an asymptotic unbiasedness and convergence in the mean squared:

Asymptotic unbiasedness

$$M\left(\overline{p}(x) - p(x)\right) \sim \frac{p^{(2)}(x)\left(\frac{\beta^2}{3} + c^2\right)}{2}.$$

Convergence in the mean squared

$$M\left(\overline{p}(x) - p(x)\right)^2 \sim \frac{\|\Phi(x)\|^2}{nc} + \frac{\left\| p^{(2)}(x)\right\|^2 \left(\frac{\beta^2}{3} + c^2\right)}{4} +$$
$$+ \frac{\|\Phi(x)\|^2 \beta}{c}\left(2\| p(x)\|^2 + \frac{\beta^4 \left\| p^{(2)}(x)\right\|^2}{18}\right) + \beta c \| u\Phi(u)\|^2 \left(2\left\| p^{(1)}(x)\right\|^2 + \frac{\beta^2 \left\| p^{(2)}(x)\right\|^2}{3}\right) +$$
$$+ \frac{\beta c^3 \left\| p^{(2)}(x)\right\|^2 \left\| u^2 \Phi(u)\right\|^2}{2}.$$

Here the following designations are accepted:

$$\|\Phi(u)\|^2 = \int_{-\infty}^{+\infty} \Phi^2(u)\,du,\ \|p(x)\|^2 = \int_{-\infty}^{+\infty} p^2(x)\,du,\ \left\| p^{(1)}(x) \right\|^2 = \int_{-\infty}^{+\infty} \left(p^{(1)}(x)\right)^2 du,$$

$$\left\| p^{(2)}(x) \right\|^2 = \int_{-\infty}^{+\infty} \left(p^{(2)}(x)\right)^2 du,\ \|u\Phi(u)\|^2 = \int_{-\infty}^{+\infty} u^2\,\Phi^2(u)\,du;$$

$p^{(1)}(x)$, $p^{(2)}(x)$ are the first and flexon of probability density $p(x)$; M - is the sign of expected value.

3. Creation of confidence borders for the probability density

The assessment of a probability density (1) allows to construct confidence borders for $p(x)$ on the basis of confidence estimation of probabilities P^j of events $x \in X^j$, $j = \overline{1, N}$.

Top P_h^j and bottom P_d^j borders of an interval assessment of probability P^j of event $x \in X^j$ with a confidence coefficient γ are defined by expressions

$$P_h^j = \overline{P}^j + \frac{u_{1-\alpha/2}}{\sqrt{n}} \sqrt{\overline{P}^j \left(1 - \overline{P}^j\right)}, \tag{2}$$

$$P_d^j = \overline{P}^j - \frac{u_{1-\alpha/2}}{\sqrt{n}} \sqrt{\overline{P}^j \left(1 - \overline{P}^j\right)}, \tag{3}$$

where $u_{1-\alpha/2}$ - quantile of level $1-\alpha/2$ of a reference normal distribution. Values $u_{1-\alpha/2}$ are determined by tables of quantiles of a normal distribution at $\alpha = 1-\gamma$.

Let's make computing experiment and determine by $\left(\overline{P}^j,\ j = \overline{1, N}\right)$ according to (2), (3), values P_h^j, $j = \overline{1, N}$ and P_d^j, $j = \overline{1, N}$.

On this basis let's carry out synthesis top and bottom confidence borders

$$\tilde{p}_h(x) = c_h^{-1} \sum_{j=1}^{N} P_h^j \Phi\left(\frac{x - z^j}{c_h}\right), \tag{4}$$

$$\tilde{p}_d(x) = c_d^{-1} \sum_{j=1}^{N} P_d^j \Phi\left(\frac{x - z^j}{c_d}\right) \tag{5}$$

for probability density $p(x)$.

Let's define optimum coefficients of a diffuseness of nuclear functions in expressions (4), (5) with use of a method of "the sliding examination". For example, for an upper bound (4) choice of optimum diffuseness coefficient c_h of nuclear functions is carried out as a result of expression minimization

$$\sum_{i=1}^{N}\left(\frac{P_h^i}{2\beta}-\tilde{\tilde{p}}_h\left(z^i\right)\right)^2 .$$

Here $\frac{P_h^i}{2\beta}$ - us value of the upper confidence bound $p(x)$ within the i -th interval of sampling of a random value x, and

$$\tilde{\tilde{p}}_h\left(z^i\right)=\frac{1}{c_h}\sum_{\substack{j=1\\ j\neq i}}^{N}P_h^j\Phi\left(\frac{z^i-z^j}{c_h}\right)$$

represents its assessment on this interval.

4. Technique of hypothesis testing of identity of distribution laws of random values

There are two independent selections $V_1=\left(x^i, i=\overline{1,n_1}\right)$ and $V_2=\left(x^i, i=\overline{1,n_2}\right)$ the one-dimensional random values taken from universes X_1, X_2, which are characterized by unknown probability densities $p_1(x)$, $p_2(x)$. It is necessary to confirm or disprove a hypothesis H_0 of identity of their distribution laws.

Let Ω_t, $t=1,2$ confidence regions for probability densities $p_t(x)$, $t=1,2$ at the given confidence coefficient γ.

Then realization of a hypothesis H_0 is defined by the following rule

$$m_{12}(\gamma):\begin{cases}\textit{hypothesis } H_0 \textit{ is fair if}\\ \Omega_1\subseteq\Omega_2 \quad \textit{or} \quad \Omega_2\subseteq\Omega_1,\\ \textit{otherwise the hypothesis } H_0 \textit{ is rejected}.\end{cases}$$

For implementation of this rule according to the technique offered above let construct for probability densities $p_1(x)$, $p_2(x)$ confidence borders

$$\tilde{p}_h^t(x)=\frac{1}{c_h^t}\sum_{j=1}^{N_t}P_h^j(t)\Phi\left(\frac{x-z^j}{c_h^t}\right),\ \tilde{p}_d^t(x)=\frac{1}{c_d^t}\sum_{j=1}^{N_t}P_d^j(t)\Phi\left(\frac{x-z^j}{c_d^t}\right),\ t=1,2,$$

where $P_h^j(t)$, $P_d^j(t)$ - interval estimates $P^j(t)$ of an event $x\in X_t^j$ with a confidence coefficient γ; N_t - quantity of not crossed intervals of areas X_t; c_h^t, c_d^t - diffuseness coefficients of nuclear functions in statisticians $\tilde{p}_h^t(x)$, $\tilde{p}_d^t(x)$.

Let's designate through a V_t^{γ} - a range of a random value of an $x^j\in V_t$, $j=\overline{1,N_t}$ getting to a confidence region of a probability density of a $p_t(x)$, $t=1,2$ at the given confidence coefficient γ. Formation of a set V_t^{γ} is carried out on the basis of the following decisive rule

$$m_t\left(x^j\right): x^j \in V_t^{\gamma} \text{ if } \tilde{p}_h^t\left(x^j\right) \geq \tilde{p}^t\left(x^j\right) \geq \tilde{p}_d^t\left(x^j\right),\ j=\overline{1,n_t},\ t=1,2.$$

It is easy to notice that the set V_t^{γ} represents an assessment of a confidence region $\boldsymbol{\Omega}_t$, $t=1,2$. Therefore there is an opportunity instead of $m_{12}(\boldsymbol{\gamma})$ use the rule

$$\overline{m}_{12}(\boldsymbol{\gamma}): \begin{cases} \textit{hypothesis } H_0 \textit{ is fair if} \\ \tilde{p}_d^2\left(x^j\right) \geq \tilde{p}_d^1\left(x^j\right) \quad \textit{and} \quad \tilde{p}_h^2\left(x^j\right) \leq \tilde{p}_h^1\left(x^j\right) \forall\ x^j \in V_2^{\gamma} \\ \textit{or} \\ \tilde{p}_d^1\left(x^j\right) \geq \tilde{p}_d^2\left(x^j\right) \quad \textit{and} \quad \tilde{p}_h^1\left(x^j\right) \leq \tilde{p}_h^2\left(x^j\right) \forall\ x^j \in V_1^{\gamma}\ . \end{cases}$$

This rule is based on replacement of operations with confidence regions $\boldsymbol{\Omega}_t$, $t=1,2$ set by sets V_t^{γ}, $t=1,2$ on check of ratios between their borders $\tilde{p}_h^t(x)$, $\tilde{p}_d^t(x)$, $t=1,2$.

5. Acknowledgment

The structure of a regression assessment of a probability density allows to solve a problem of confidence estimation of a probability density on its basis. The idea of the offered approach consists in decomposition of input statistical data and the subsequent analysis of probabilistic characteristics of the received sets of random values. On this basis, using a nonparametric assessment of a regression curve, synthesis of confidence borders of a probability density is carried out. The area sizes determined by confidence borders depend on quantity of intervals of sampling of random values, their probabilistic characteristics and volume of input statistical data.

If the confidence region of a probability density includes a confidence region of other compared density, they are identical. Otherwise the hypothesis of identity of distribution laws of random values is rejected.

The offered approach opens possibility of generalization of the received results on a hypothesis testing about distributions of many-dimensional random values.

This work was supported within the basic part of the State Task of the Ministry of Education and Sciences of the Russian Federation for higher educational institutions in 2014–2016 (SibGAU No. B121/14) and Program of Siberian Branch of the Russian Academy of Sciences IV.35.1 «Theoretical bases and technologies of creation and use of the integrated information systems for problem solving of support of a decision making».

References

[1] Pugachev V S Probability Theory and Mathematical Statistics (Nauka, Moscow, 1979).
[2] Smirnov N V 1930 Estimation of the Difference between the Distribution Curves in Two Independent Samples *Bull. Mosk. Univ.*, **2** pp3–14.
[3] Lapko A V , Lapko V A 2010 Nonparametric algorithms of pattern recognition in the problem of testing a statistical hypothesis on identity of two distribution laws of random variables *Opt. Instrum. Data Proc.*, **46** pp 545-550.
[4] Lapko A V, Lapko V A 2012 Comparison of empirical and theoretical distribution functions of a random variable on the basis of a nonparametric classifier *Opt. Instrum. Data Proc.*, **48** pp37-41.

[5] Lapko A V , Lapko V A 2014 Regression estimate of the multidimensional probability density and its properties *Opt. Instrum. Data Proc.*, **50** pp148-153.
[6] Heinhold I, Gaede K W *Ingenieur Statistik* (Springer-Verlag, Munich, Vienna, 1964).
[7] Freedman D, Diaconis P 1981 On the histogram as a density estimator: L2 theory *Probability Theory and Related Fields*, **57** pp 453–476.
[8] Scott D 1979 On optimal and data-based histograms *Biometrika*, **66** pp 605–610.
[9] Sturges H A 1926 The Choice of a Class Interval *J. Amer. Stat. Association*, **21** pp 65–66.
[10] Lapko A V, Lapko V A 2013 Optimal selection of the number of sampling intervals in domain of variation of a one-dimensional random variable in estimation of the probability density *Measur. Techn.*, **56** pp 763 – 767.