საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 10, №3, 2016
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 10, no. 3, 2016

Mathematics

On Testing Hypotheses of Equality Distribution Densities

Petre Babilua[*], Elizbar Nadaraya[**], Grigol Sokhadze[*]

[*] *Department of Mathematics, Faculty of Exact and Natural Sciences, Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia*
[**] *Academy Member, Department of Mathematics, Faculty of Exact and Natural Sciences, Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia*

ABSTRACT. In the paper the test of homogeneity and goodness-of-fit for checking the hypotheses of equality distribution densities is constructed. The power asymptotics of the constructed test of homogeneity and goodness-of-fit for certain types of close alternatives is also studied. *© 2016 Bull. Georg. Natl. Acad. Sci.*

Key words: test of homogeneity, goodness-of-fit test, power of the test, Wiener process, close alternatives

Let $X^{(i)}=\left(X_1^{(i)},\dots,X_{n_i}^{(i)}\right)$, $i=1,\dots,p$, be independent samples with sizes $n_1,n_2,\dots,n_p$, from $p\ge 2$ general population with probability densities $f_1(x),\dots,f_p(x)$ and it is required to test two hypotheses based on samples $X^{(i)}$, $i=1,\dots,p$: test of homogeneity

$$H_0:\quad f_1(x)=\cdots=f_p(x) \tag{1}$$

and goodness-of-fit test

$$H_0':\quad f_1(x)=\cdots=f_p(x)=f_0(x), \tag{2}$$

where $f_0(x)$ is a fully defined density function. In case of hypothesis H_0 common density function $f_0(x)$ is unknown.

In this paper the test is constructed for checking hypothesis H_0 and H_0' against a sequence of “close” alternatives [1, 2]:

$$H_1:\quad f_i(x)=f_0(x)+\sim(n_0)\,\text{¨}_i\left(\frac{x-\ell_i}{\text{¥}(n_0)}\right)\qquad \left(\sim(n_0),\text{¥}(n_0)\to 0\right),$$

$$\int \text{¨}_i(x)\,dx=0,\qquad n_0=\min\left(n_1,\dots,n_p\right)\to\infty .$$

We consider criteria for testing H_0 and H_0', based on statistics

$$T\left(n_1,n_2,\dots,n_p\right)=\sum_{i=1}^{p}N_i\int\left[\overset{€}{f}_i(x)-\frac{1}{N}\sum_{j=1}^{p}N_j\overset{€}{f}_j(x)\right]^2 r(x)\,dx, \qquad (3)$$

where $\overset{€}{f}_i(x)$ is Rosenblatt-Parsen kernel estimator of the distribution density $f_i(x)$:

$$\overset{€}{f}_i(x)=\frac{a_i}{n_i}\sum_{j=1}^{n_i}K\left(a_i\left(x-X_j^{(i)}\right)\right),\qquad N_i=\frac{a_i}{n_i},\quad N=N_i+\dots+N_p.$$

Particular case $p=2$ was discussed in papers [3, 4]. In this case statistics T becomes more visual:

$$T\left(n_1,n_2\right)=\frac{N_1N_2}{N_1+N_2}\int\left(\hat{f}_1(x)-\hat{f}_2(x)\right)^2 r(x)\,dx.$$

In this paper the found limiting distribution of statistics (3) was found for hypothesis H_1 in case, where n_i unlimited is increasing so that $n_i=nk_i$, where $n\to\infty$, and k_i are constant. Let $a_1=a_2=\dots=a_p=a_n$, so $a_n\to\infty$ for $n\to\infty$.

For getting limiting distribution of functional $T_n=T\left(n_1,\dots,n_p\right)$ let us introduce conditions for functions $K(x)$, $f_0(x)$, $¨_i(x)$, $i=1,\dots,p$ and $r(x)$:

(i) $K(x)\ge 0$ – function with bounded variation,

$$\int K(x)\,dx=1,\quad x^2K(x)\in L_1(-\infty,\infty);$$

(ii) density function $f_0(x)$ is bounded and positive on $(-\infty,\infty)$ or is bounded and positive on some finite interval $[c,d]$. Besides, it has bounded derivative in the field where it is positive;

(iii) functions $¨_j(x)$, $j=1,\dots,p$, are bounded and have bounded first order derivatives, also $¨_i(x)$ and $¨_i^{(1)}\in L_1(-\infty,\infty)$.

(iv) weighed function $r(x)$ is piece-continuous, bounded and integrable, besides $r(\ell_k)\ne 0$, $k=1,\dots,p$, where ℓ_k is some fixed points of continuity of $r(x)$.

The following is true:

Theorem 1. *Le us fulfill the conditions* (i)–(iv), *also* $f_i(x)\ge 0$, $x\in(-\infty,\infty)$. *If*

$na_n^{-1/2}\sim_n^2¥_n\to c_0\ne 0$, $a_n¥_n\to\infty$, $\sim_n¥_n=o\left(n^{-1/2}\right)$ $\left(\sim_n=\sim(n_0),\ ¥_n=¥(n_0)\right)$,

$na_n^{-2}\to\infty$ and $a_n^2\sim_n¥_n\to\infty$, then random variable $a_n^{1/2}\left(T_n-«\right)$ under hypothesis H_1 has normal limit distribution $\left(A(¨),±^2\right)$, where

$$A(¨)=c_0\sum_{i=1}^{p}\left(k_i-\frac{k_i^2}{\overline{k}}\right)r(\ell_i)\int ¨_i^2(x)\,dx,$$

$$±^2=2(p-1)\int f_0^2(x)r(x)\,dx\,R(K_0),\quad K_0=K*K,$$

$$«=(p-1)\int f(x)r(x)\,dx\,R(K),\quad R(g)=\int g^2(x)\,dx,$$

$$\overline{k}=k_1+\dots+k_p,\quad p\ge 2.$$

Conditions of Theorem 1 about a_n, $\sim_n$ and $¥_n$ are fulfilled, for example, if: $a_n=n^{¢}$, $\sim_n=n^{-\sim}$, $¥_n=n^{-}$ at $\frac{¢}{2}=1-2\sim-$, $\sim+\ >\frac{1}{2}$, $0<¢<\frac{1}{2}$, $0<\ <¢$, and conditions about $\sim$, and $¢$ are fulfilled, for example, if

$$\text{¢} = \frac{1}{4}, \qquad = \frac{1}{5}, \quad \sim = \frac{27}{80}; \quad \text{¢} = \frac{2}{9}, \qquad = \frac{1}{6}, \quad \sim = \frac{13}{36};$$

$$\text{¢} = \frac{1}{5}, \qquad = \frac{1}{6}, \quad \sim = \frac{1}{30} \text{ etc.}$$

From Theorem 1 we will state two Corollaries:

Corollary1. *Let the conditions* (i), (ii) *and* (iv) *be fulfilled under* $K(x)$, $f_0(x)$ *u* $r(x)$. *If* $na_n^{-2} \to \infty$, *then random variable* $a_n^{1/2}(T_n - \text{«})$ *under hypothesis* H_0' *has a normal limit distribution* $\left(0, \pm^2\right)$.

By Corollary1 a test for checking hypothesis H_0' can be constructed; critical region for checking hypothesis can be defined by inequality

$$T_n \ge d_n(\sim), \tag{4}$$

where

$$d_n(\sim) = \text{«} + a_n^{-1/2} \pm \text{ª}_\sim ,$$

$\text{ª}_\sim$ is the quantile of level $1-\sim$ $(0 < \sim < 1)$ of the standard normal distribution $\Phi(x)$.

Corollary 2. *Under conditions of Theorem* 1 *local behavior of the power* $P_{H_1}\left(T_n \ge d_n(\sim)\right)$ *is as follows*

$$P_{H_1}\left(T_n \ge d_n(\sim)\right) \to 1 - \Phi\left(\text{ª}_\sim - \frac{A(\text{¨})}{\pm}\right),$$

when $n \to \infty$.

Let introduce

$$f_n^*(x) = \frac{1}{k}\sum_{j=1}^{p} k_j \, \text{€}f_j(x),$$

$$\overline{\text{«}}_n = \int f_n^*(x) r(x)\, dx,$$

$$\Delta_n^2 = \frac{1}{k}\sum_{i=1}^{p} k_i \Delta_{in}^2, \qquad \Delta_{in}^2 = \int \text{€}f_i^2(x) r(x)\, dx .$$

The Theorem is true.

Theorem 2. *Let all the conditions of Theorem* 1 *be fulfilled. Then*

$$a_n^{1/2}(T_n - \text{«}_n)\pm_n^{-1}$$

under hypothesis H_1 has a normal limit distribution $\left(A(\text{¨})\pm^{-1}, 1\right)$, where

$$\text{«}_n = (p-1)R(K)\overline{\text{«}}_n, \qquad \pm_n^2 = 2(p-1)R(K_0)\Delta_n^2 .$$

Proof. It is obvious

$$a_n^{1/2}(T_n - \text{«}_n)\pm_n^{-1} = a_n^{1/2}(T_n - \text{«})\pm^{-1}\left(\pm\pm_n^{-1}\right) + a_n^{1/2}(\text{«} - \text{«}_n)\pm_n^{-1} .$$

As it is enough to show

$$a_n^{1/2}\left(\overline{\text{«}}_n - \int f_0(x) r(x)\, dx\right) = o_p(1) \tag{5}$$

and

$$\Delta_n^2 - \int f_0^2(x) r^2(x)\,dx = o_p(1). \tag{6}$$

But (6) follows from Theorem 2.1 Bhattacharyya G. K., Roussas G. G. [5] (see also [6], [2]).

Let us proove (5). We have

$$a_n^{1/2} E\left|\int f_n^*(x) r(x)\,dx - \int f_0(x) r(x)\,dx\right| \le$$

$$\le a_n^{1/2} E\left|\int \left(f_n^*(x) - Ef_n^*(x)\right) r(x)\,dx\right| + a_n^{1/2} \int \left|Ef_n^*(x) - f_0(x)\right| r(x)\,dx =$$

$$= A_{1n} + A_{2n}.$$

It is not difficult to show

$$E\text{€}_i(x) = f_0(x) + O\left(\frac{1}{a_n}\right) + \sim_n \int K(t)\,\text{¨}_i\left(\frac{x-\ell_i}{\text{¥}_n} - \frac{t}{a_n\text{¥}_n}\right) dt,$$

$O(\,\cdot\,)$ uniformly in $x \in (-\infty, \infty)$. So

$$Ef_n^*(x) = f_0(x) + O\left(\frac{1}{a_n}\right) + \sim_n \frac{1}{\overline{k}} \sum_{j=1}^{p} k_j \int K(t)\,\text{¨}_j\left(\frac{x-\ell_j}{\text{¥}_n} - \frac{t}{a_n\text{¥}_n}\right) dt$$

it follows,

$$A_{2n} \le c_1 a_n^{-1/2} + c_2 a_n^{1/2} \sim_n \text{¥}_n.$$

Next, we have

$$A_{1n} \le a_n^{1/2} E^{1/2} \left(\int \left(f_n^*(x) - Ef_n^*(x)\right) r(x)\,dx\right)^2 \le$$

$$\le c_3 a_n^{1/2} \max_{1 \le j \le p} \left\{ \frac{1}{n} \int f_j(u)\,du \left(\int K(t)\, r\left(u - \frac{t}{a_n}\right) dt\right)^2 \right\}^{1/2} \le c_4 \left(\frac{a_n}{n}\right)^{1/2}.$$

It follows,

$$A_{1n} + A_{2n} \le c_4 \left(a_n^{-1/2} + \sqrt{a_n} \sim_n \text{¥}_n + \left(\frac{a_n}{n}\right)^{1/2} \right) \to 0$$

as $\sqrt{a_n} \sim_n \text{¥}_n \le a_n^2 \sim_n \text{¥}_n \to 0$ and $\dfrac{a_n}{n} \to 0$.

From Theorem 2 we will state two corollaries.

Corollary 3. *Random variable*

$$a_n^{1/2}(T_n - \text{«}_n)\pm_n^{-1}$$

under hypothesis H_0 has a normal limit distribution $(0,1)$.

This result can be used for constructing an asymptotic test for checking hypothesis $H_0: \ f_1(x) = \cdots = f_p(x)$ (test of homogeneity); critical region can be defined by inequality:

$$T_n \ge \tilde{d}_n(\sim) = \text{«}_n + a_n^{-1/2} \pm_n \text{ª}_\sim, \tag{7}$$

where $\text{ª}_\sim$ is the quantile of level $1 - \sim$ of the standard normal distribution $\Phi(x)$.

Corollary 3. *Under conditions of Theorem2 local behavior of the power* $P_{H_1}\left(T_n \ge \tilde{d}_n(\sim)\right)$ *as follows*

$$P_{H_1}\left(T_n \geq \tilde{d}_n\left(\sim\right)\right) \to 1-\Phi\left(\text{ª}_{\sim} - A\left(\text{¨}\right)\pm^{-1}\right),$$

when $n \to \infty$.

Remark 1. *Under hypothesis* H_1 we have

$$F_i\left(x\right) = F_0\left(x\right) + \sim_n \text{¥}_n U_i\left(\frac{x-\ell_i}{\text{¥}_n}\right), \quad U_i\left(u\right) = \int_{-\infty}^{u} \text{¨}_i\left(x\right)dx$$

and according to Theorem 1 $\sim_n \text{¥}_n = o\left(\frac{1}{\sqrt{n}}\right)$. So it can be written

$$\sup_x \left|F_i\left(x\right) - F_0\left(x\right)\right| = o\left(\frac{1}{\sqrt{n}}\right). \tag{8}$$

It is well-known that the test based on deviation between empirical distribution functions, for example, criterion Kolmogorov-Smirnov and test Cramer-Mises-Smirnov (analogues to these criteria for $p \geq 2$ was constructed by **Kiefer, J** [7]) differs local close alternatives from null hypothesis, if $F_i\left(x\right) - F_0\left(x\right) = O\left(\frac{1}{\sqrt{n}}\right)$ uniformly in $x \in \left(-\infty, \infty\right)$, in case of (8) the above mentioned test cannot asymptotically distinguish such hypotheses from null hypothesis (limiting power will be equal with level of the test). However, tests (4) and (7) based on estimators of distribution density are more powerful asymptotically (under hypothesis H_1) than tests based on empirical distribution functions (analogues questions for one sample considered in paper of Rosenblatt [1]).

Remark 2. Tests (4) and (7) for checking hypotheses H_0' and H_0, against alternatives H_1 are asymptotically strictly unbiased as $A\left(\text{¨}\right) > 0$ and equal to 0 if and only if $\text{¨}_i\left(x\right) = 0$, almost everywhere, $i = 1, \ldots, p$.

მათემატიკა

განაწილების სიმკვრივეთა ტოლობის ჰიპოთეზათა შემოწმების შესახებ

პ. ბაბილუა*, ე. ნადარაია, გ. სოხაძე***

**ივანე ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი, ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი, მათემატიკის დეპარტამენტი, თბილისი, საქართველო*
***აკადემიის წევრი, ივანე ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი, ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი, მათემატიკის დეპარტამენტი, თბილისი, საქართველო*

ნაშრომში აგებულია ერთგვაროვნების და თანხმობის ჰიპოთეზათა შემოწმების კრიტერიუმები. მოძებნილია აგებული კრიტერიუმების ზღვარითი სიმძლავრე დაახლოებადი ალტერნატივებისთვის.

REFERENCES

1. *Rosenblatt M.* (1975) Ann. Statist. **3**: 1-14.
2. *Nadaraya E. A.* (1989) Nonparametric estimation of probability densities and regression curves. Mathematics and its Applications (Soviet Series), 20. Kluwer Academic Publishers Group, Dordrecht.
3. *Nadaraya E. A.* (1975) Soobshch. Akad. Nauk Gruz. SSR **78**: 25-28 (in Russian).
4. *Anderson N. H., Hall P.,. Titterington D. M* (1994) J. Multivariate Anal., **50**, 1: 41-54.
5. *Bhattacharyya G. K., Roussas G. G.* (1970) Skand. Aktuarietidskr. **1969**: 201-206.
6. *Mason D. M., Nadaraya E. A., Sokhadze G. A.* (2010) Integral functionals of the density. Inst. Math. Statist. Collect. 7, Inst. Math. Stat. Beachwood, OH. 153-168.
7. *Kiefer J.* (1959) Ann. Math. Statist. **30**: 420-447.

Received May, 2016