საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 14, №2, 2020
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 14, no. 2, 2020

Mathematics

On the Chi-Square Test of Homogeneity in Case of a Simultaneous Increase of the Number of Observations and the Number of Interval Partitions

Petre Babilua[*], Elizbar Nadaraya[**], Mzevinar Patsatsia[§]

[*] *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*
[§] *Department of Mathematics, Faculty of Mathematics and Computer Sciences, Sokhumi State University, Tbilisi, Georgia*

The limiting distribution of the statistic of the homogeneity test of chi-square is established in case of a simultaneous increase of the number of observations and the number of interval partitions in case of "close" alternatives of Pitman type. Also, it is compared with another test based on the integral square deviation of a nonparametric kernel estimate of density. It is shown that the limiting power of the above-mentioned test is greater than the limiting power of Pearson's Chi-square test. © *2020 Bull. Georg. Natl. Acad. Sci.*

Homogeneity hypothesis, goodness-of-fit test, power of test, Wiener process, test consistency, kernel type estimator of density, histogram

Let us consider two independent samples

$$X^{(1)}=\left(X_1^{(1)},X_2^{(1)},\ldots,X_{n_1}^{(1)}\right) \text{ and } X^{(2)}=\left(X_1^{(2)},X_2^{(2)},\ldots,X_{n_2}^{(2)}\right),$$

of independent and equally distributed random variables, with distribution densities $f_1(x)$ and $f_2(x)$, respectively, concentrated on the interval $[0,1]$. It is required, using the samples $X^{(1)}$ and $X^{(2)}$, to test the hypothesis $H_0: f_1(x)=f_2(x)=f_0(x)$ (homogeneity hypothesis of two samples). The homogeneity hypothesis only states that the distribution densities $f_1(x)$ and $f_2(x)$ coincide and does not fix the form of the general distribution density $f_0(x)$.

Let n_i, $i=1,2$, increase without bound so that $n_i=nk_i$, where $n\to\infty$, and k_i are constants. Let, further, the interval $[0,1]$ be partitioned into s_n equal intervals $\Delta_1,\Delta_2,\ldots,\Delta_{s_n}$ of length $h_n=s_n^{-1}$, $s_n\to\infty$, so that $\frac{s_n}{n}\to 0$ as $n\to\infty$.

To test the hypothesis H_0 against the sequence of "close" Pitman type alternatives H_1

$$f_i(x) = f_0(x) + \alpha_n \varphi_i(x) + o(\alpha_n), \quad \alpha_n \to 0, \quad \int \varphi_i(x)\,dx = 0, \quad i = 1,2$$

we use the statistic

$$\chi^2_{n_1,n_2} = n_1 n_2 \sum_{i=1}^{s_n} \frac{1}{\nu_i + \mu_i} \left(\frac{\nu_i}{n_1} - \frac{\mu_i}{n_2} \right)^2,$$

where ν_i and μ_i are the numbers of observations from the first $X^{(1)}$ and the second $X^{(2)}$ samples, respectively, belonging to the intervals Δ_i, $|\Delta_i| = h_n$, $i = 1,2,\ldots,s_n$.

Theorem 1. Let $f_0(x)$, $\varphi_i(x)$, $i = 1,2$ satisfy the Lipschitz condition of order 1 and $\inf\limits_{0 \le x \le 1} f_0(x) > 0$. If $\alpha_n = n^{-\frac{1}{2}} s_n^{\frac{1}{4}}$ and $n^{-1} s_n^{\frac{9}{2}} \to 0$ as $n \to \infty$, then for the alternative H_1

$$\frac{\chi^2_{n_1,n_2} - s_n}{\sqrt{2s_n}} \xrightarrow{d} N\left(\frac{A(\varphi)}{\sqrt{2}}, 1 \right),$$

$$A(\varphi) = \frac{k_1 k_2}{k_1 + k_2} \int_0^1 \left(\varphi_1(x) - \varphi_2(x) \right)^2 f_0^{-1}(x)\,dx,$$

$\xrightarrow{d}$ denotes convergence in distribution, $N(a,b)$ is a random variable that has normal distribution with mean value and variance b^2.

Corollary 1. For $n \to \infty$

$$P_{H_1}\left(\chi^2_{n_1,n_2} \ge s_n + \lambda_\alpha \sqrt{2s_n} \right) \longrightarrow 1 - \Phi\left(\lambda_\alpha - \frac{k_1 k_2}{k_1 + k_2} \frac{1}{\sqrt{2}} \int_0^1 \varphi^2(x) f_0^{-1}(x)\,dx \right),$$

where $\varphi(x) = \varphi_1(x) - \varphi_2(x)$.

Before proving the theorem we make

Remark 1. As is known, the limit distribution of the statistic $\chi^2_{n_1,n_2}$ for $n_1, n_2 \to \infty$ and the fixed number of partitioned intervals does not depend on the general unknown $f_0(x)$. If, however, with the growth of the numbers of observations n_1 and n_2 the number of partitioned intervals grows with the same rate, then the limit distribution of the statistic will essentially depend on $f_0(x)$ [1].

Remark 2. In the option of the authors, the result has most likely been known for $A(\varphi) = 0$.

Let us give an example of comparison of the power of the test $\chi^2_{n_1,n_2}$ with test $\hat{T}_{n_1,n_2}$ [2]:

$$\hat{T}_{n_1,n_2} = \frac{m_1 m_2}{m_1 + m_2} \int_0^1 \left(\hat{f}_1(x) - \hat{f}_2(x) \right)^2 r_n^*\,dx,$$

$$\hat{f}_i(x) = \frac{a_i}{n_i} \sum_{j=1}^{n_i} K\left(a_i \left(x - X_j^{(i)} \right) \right), \quad n_i = k_i n,$$

$$m_i = \frac{n_i}{a_i}, \quad a_i = s_n, \quad i = 1,2, \quad r_n^*(x) = \left[f_n^*(x) \right]^{-1},$$

$$f_n^*(x) = \frac{1}{\bar{k}} \sum_{j=1}^{2} k_j \hat{f}_j(x), \quad \bar{k} = k_1 + k_2.$$

Theorem 2 ([2]).

(i) Let $f_0(x)$ and $\varphi_i(x)$, $i=1,2$ satisfy the conditions of Theorem 1.

(ii) vanishes outside the finite interval $[-A;A]$ and, together with its derivatives, is continuous on this interval. If $\alpha_n = n^{-\frac{1}{2}} s_n^{\frac{1}{4}}$ and $n^{-1} s_n^{\frac{9}{2}} \ln n \to 0$, then for the alternative H_1

$$s_n^{\frac{1}{2}} \left(\hat{T}_{n_1,n_2} - \mu_0 \right) \xrightarrow{d} N\left(A(\varphi), \sigma_0^2 \right),$$

$$A(\varphi) = \frac{k_1 k_2}{k_1 + k_2} \int_0^1 \left(\varphi_1(x) - \varphi_2(x) \right)^2 f_0^{-1}(x)\, dx,$$

$$\mu_0 = \int K^2(u)\, du, \quad \sigma_0^2 = 2\int K_0^2(u)\, du, \quad K_0 = K * K.$$

Corollary 2. Let the conditions (i) and (ii) be fulfilled for $f_0(x)$ and $K(x)$, respectively. If $n^{-1} s_n^{\frac{9}{2}} \ln n \to 0$, then for the hypothesis H_0

$$s_n^{\frac{1}{2}} \left(\hat{T}_{n_1,n_2} - \mu_0 \right) \xrightarrow{d} N\left(0, \sigma_0^2 \right).$$

Using this corollary we may construct the test for the hypothesis H_0: $f_1(x) = f_2(x)$. The critical domain for testing this hypothesis is defined by the inequality

$$\hat{T}_{n_1,n_2} \geq d_n(\alpha) = \mu_0 + s_n^{-\frac{1}{2}} \lambda_\alpha \sigma_0,$$

where λ_α is the quantile of the level $1-\alpha$, $0<\alpha<1$ of the standard normal distribution $\Phi(x)$.

Corollary 3. The local behavior of the power

$$P_{H_1}\left(\hat{T}_{n_1,n_2} \geq d_n(\alpha) \right) \longrightarrow 1 - \Phi\left(\lambda_\alpha - \frac{A(\varphi)}{\sigma_0} \right). \tag{2}$$

Let $\alpha_n = n^{-\frac{1}{2}+\frac{\delta}{4}}$, $0<\delta<\frac{2}{9}$, and

$$K(x) = \begin{cases} 1, & |x| \leq \frac{1}{2}, \\ 0, & |x| > \frac{1}{2}. \end{cases}$$

Then

$$\sigma_0^2 = 2\int K_0^2(u)\, du = 2 \int_{|x|\leq 1} \left(1 - |x| \right)^2 dx = \frac{4}{3}.$$

From (2), we have

$$P_{H_1}\left(\hat{T}_{n_1,n_2} \geq d_n(\alpha) \right) \longrightarrow 1 - \Phi\left(\lambda_\alpha - \frac{\sqrt{3}}{2} \frac{k_1 k_2}{k_1 + k_2} \int_0^1 \varphi^2(x) f_0^{-1}(x)\, dx \right),$$

$$\varphi(x) = \varphi_1(x) - \varphi_2(x).$$

Comparing (1) and (3) we conclude that the asymptotic test $\hat{T}_{n_1,n_2}$ is more powerful than the test $\chi^2_{n_1,n_2}$ with respect to a relatively alternative hypothesis H_1.

მათემატიკა

ერთგვაროვნების ხი-კვადრატ კრიტერიუმის შესახებ, როცა ერთდროულად იზრდება დაკვირვებათა და ინტერვალის დაყოფათა რიცხვი

პ. ბაბილუა*, ე. ნადარაია**, მ. ფაცაცია§

** ივანე ჯავახიშვილის თბილისის სახელმწიფო უნივერსიტეტი, ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი, მათემატიკის დეპარტამენტი, თბილისი, საქართველო*
*** აკადემიის წევრი, ივანე ჯავახიშვილის თბილისის სახელმწიფო უნივერსიტეტი, ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი, მათემატიკის დეპარტამენტი, თბილისი, საქართველო*
§ სოხუმის სახელმწიფო უნივერსიტეტი, მათემატიკის და კომპიუტერულ მეცნიერებათა ფაკულტეტი, მათემატიკის დეპარტამენტი, თბილისი, საქართველო

ნაშრომში დადგენილია პირსონის ერთგვაროვნების კრიტერიუმის ზღვარითი სიმძლავრე პიტმანის ტიპის დაახლოებადი ალტერნატივების მიმართ, როდესაც შერჩევის მოცულობა და ინტერვალის დაყოფათა რიცხვი ერთდროულად მიისწრაფვის უსასრულობისკენ. გარდა ამისა, ის შედარებულია სხვა კრიტერიუმთან, რომელიც დაფუძნებულია სიმკვრივის არაპარამეტრულ გულოვან შეფასებათა ინტეგრალურ კვადრატულ გადახრაზე. ნაჩვენებია, რომ აღნიშნული კრიტერიუმის ზღვარითი სიმძლავრე უფრო მეტია ვიდრე პირსონის ხი-კვადრატ კრიტერიუმის ზღვარითი სიმძლავრე.

REFERENCES

1. Ivchenko G. I., Levin V. V. (1976) Asimptoticheskaya normal`nost` odnogo klassa statistik v polinomial`noj skheme. *Teor. veroiatnost. i primenen*. **21**, 1: 190-195 (in Russian).
2. Babilua P., Nadaraya E. (2018) On the homogeneity test based on the kernel-type estimators of a distribution density. *Trans. A. Razmadze Math. Inst*. **172**, 3: 318-331.

Received December, 2019