საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 15, №4, 2021
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 15, no. 4, 2021

Mathematics

On a New Method of Testing the Hypothesis of Equality of Two Bernoulli Regression Functions for Group Observations

Petre Babilua[*] and Elizbar Nadaraya[**]

[*]*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*

In the paper, the limiting distribution is established for an integral square deviation of estimates of Bernoulli regression functions based on two group samples. Based on this a new test is constructed for the hypothesis testing on the equality of two Bernoulli regression functions. The question of consistency of the constructed test is studied and for some close alternatives the asymptotic of the test power is investigated. *© 2021 Bull. Georg. Natl. Acad. Sci.*

Bernoulli regression function, limiting distribution, consistency, power of the test

Let random variables $Y^{(i)}$, i=1,2 have two values 1 and 0 with probability ("success") p_i and ("failure") 1-p_i, i=1,2 , respectively. Assume that the probability of "success" p_i is the function of an independent variable $x \in [0,1]$, i.e. $p_i = p_i(x)) = \mathbf{P}\left\{Y^{(i)} = 1 \,\middle|\, x\right\}$ (see [1-3]). Assume that x_j , $j = 1,\ldots,n$, are the points of division of the interval $[0,1]$:

$$x_j = \frac{2j-1}{2n}, \quad j = 1,\ldots,n .$$

Let, further, $Y_{ij}^{(k)}$, $j = 1,2,\ldots,m_i^{(k)}$, $i = 1,2,\ldots,n$, $k = 1,2$ be mutually independent Bernoulli random variables with $\mathbf{P}\left\{Y_{ij}^{(k)} = 1 \,\middle|\, x_i\right\} = p_k(x_i)$, $\mathbf{P}\left\{Y_{ij}^{(k)} = 0 \,\middle|\, x_i\right\} = 1 - p_k(x_i)$, $j = 1,\ldots,m_i^{(k)}$, $i = 1,2,\ldots,n$, $k = 1,2$. The problem is to test the hypothesis H_0 : $p_1(x) = p_2(x) = p_0(x)$, $x \in [0,1]$ based on the group samplings $Y_{ij}^{(k)}$, $j = 1,2,\ldots,m_i^{(k)}$, $i = 1,2,\ldots,n$, $k = 1,2$. Such problems arise, for instance, in quantum bioanalysis in pharmacology. There x is the dose of medication and $p_0(x)$ – probability of effectivity on x [4, 5].

It is required to test the hypothesis H_0 , based on statistic

$$T_n = \frac{1}{2}\, nb_n \int\limits_{\Omega_n(\tau)} \left[\hat{p}_{1n}(x) - \hat{p}_{2n}(x)\right]^2 p_n^2(x)dx, \quad \Omega_n(\tau) = \left[\tau b_n, 1-\tau b_n\right], \quad \tau > 0,$$

$$\hat{p}_{kn}(x) = p_{kn}(x)\, p_n^{-1}(x),$$

$$p_{kn}(x) = \frac{1}{nb_n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{b_n}\right)\bar{Y}_i^{(k)}, \quad \bar{Y}_i^{(k)} = \frac{1}{m_i^{(k)}}\sum_{j=1}^{m_i^{(k)}} Y_{ij}^{(k)},$$

$$p_n(x) = \frac{1}{nb_n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{b_n}\right), \ k = 1, 2,$$

where $K(x)$ is some function that satisfies the requirements formulated below, $b_n \to 0$ is a sequence of positive numbers and $\hat{p}_{kn}(x)$ is kernel estimator of regression function (see [6, 7]).

Assumptions and Notations

Assume that the kernel $K(x) \geq 0$ is chosen such that it is a function with finite variation and satisfies the conditions: $K(x) = K(-x)$, $K(x) = 0$ for $|x| \geq \tau > 0$, $\int K(x)dx = 1$. The class of such functions is denoted by $H(\tau)$.

Let us introduce the following notations:

$$U_n = \frac{1}{2}\, nb_n \int\limits_{\Omega_n(\tau)} \left[\tilde{p}_{1n}(x) - \tilde{p}_{2n}(x)\right]^2 dx,$$

$$\tilde{p}_{in}(x) = p_{in}(x) - \mathbf{E}\, p_{in}(x), \quad i = 1, 2,$$

$$Q_{ij} = \psi_n(x_i, x_j), \quad \psi_n(u, v) = \int\limits_{\Omega_n(\tau)} K\left(\frac{x-u}{b_n}\right)K\left(\frac{x-v}{b_n}\right)dx,$$

$$B_n^2 \equiv B_n^2(p_1, p_2) = (nb_n)^{-2}\sum_{k=2}^{n}\left[\frac{p_1(x_k)\left(1-p_1(x_k)\right)}{m_k^{(1)}} + \frac{p_2(x_k)\left(1-p_2(x_k)\right)}{m_k^{(2)}}\right]$$

$$\times \sum_{i=1}^{k-1}\left[\frac{p_1(x_i)\left(1-p_1(x_i)\right)}{m_i^{(1)}} + \frac{p_2(x_i)\left(1-p_2(x_i)\right)}{m_i^{(2)}}\right]Q_{ik}^2,$$

$$\eta_{ij}^{(n)} = \frac{\varepsilon_i \varepsilon_j Q_{ij}}{nb_n B_n}, \quad \varepsilon_i = \varepsilon_{1i} - \varepsilon_{2i}, \quad \varepsilon_{ki} = \bar{Y}_i^{(k)} - p_k(x_i), \quad k = 1, 2, \quad i = 1, \ldots, n,$$

$$\xi_r^{(n)} = \sum_{i=1}^{r-1}\eta_{ir}^{(n)}, \quad r = 2, \ldots n, \quad \xi_1^{(n)} = 0, \quad \xi_r^{(n)} = 0, \quad r > n,$$

$$\mathcal{F}_r^{(n)} = \sigma(\varepsilon_1, \ldots, \varepsilon_r),$$

where $\mathcal{F}_r^{(n)}$ is the σ-algebra generated by random variables $\varepsilon_1, \ldots, \varepsilon_r$, $\mathcal{F}_0^{(n)} = (\varnothing, \Omega)$ (in the sequel, for the sake of simplicity, we will write ξ_r, η_{ij} and $\mathcal{F}_r$ instead of $\xi_r^{(n)}$, $\eta_{ij}^{(n)}$ and $\mathcal{F}_r^{(n)}$).

Auxiliary Results

To proof the theorems formulated below, we are using the following three lemmas.

Lemma 1. A stochastic sequence $\left(\xi_k, \mathcal{F}_k\right)_{k\geq 1}$ is a martingale difference.

Lemma 2 ([8])**.** Let $K(x) \in H(\tau)$ and $p(x)$, $0 \leq x \leq 1$, be a function of bounded variation. If $nb_n \to \infty$, then

$$\frac{1}{nb_n}\sum_{i=1}^{n} K^{\nu_1}\left(\frac{x-x_i}{b_n}\right) K^{\nu_2}\left(\frac{y-x_i}{b_n}\right) p^{\nu_3}(x_i) =$$
$$= \frac{1}{b_n}\int_0^1 K^{\nu_1}\left(\frac{x-u}{b_n}\right) K^{\nu_2}\left(\frac{y-u}{b_n}\right) p^{\nu_3}(u)du + O\left(\frac{1}{nb_n}\right)$$

uniformly with respect to $x, y \in [0,1]$, where $\nu_i \in \mathbb{N} \cup \{0\}$, $i = 1,2,3$.

Lemma 3. Let $K(x) \in H(\tau)$ and $p_k(x) \in C^1[0,1]$, $k = 1,2$. If $nb_n^2 \to \infty$, then

$$b_n^{-1} N_n^2 B_n^2(p_1, p_2) \geq b_n^{-1}\sigma_n^2(p_1, p_2) \to \sigma^2(p_1, p_2) \quad for \quad n \to \infty,$$

where

$$\sigma_n^2(p_1, p_2) = (nb_n)^{-2}\sum_{k=2}^{n} d(x_k)\sum_{i=1}^{k-1} d(x_i)Q_{ik}^2,$$
$$N_n = \max\left(N_n^{(1)}, N_n^{(2)}\right), \quad N_n^{(k)} = \max_{1\leq i\leq n} m_i^{(k)}, \quad k = 1,2,$$
$$d(x_i) = \sum_{r=1}^{2} p_r(x_i)\left(1 - p_r(x_i)\right), \quad i = 1,\ldots,n,$$
$$\sigma^2(p_1, p_2) = \frac{1}{2}\int_0^1 d^2(x)dx \int_{|t|\leq 2\tau} K_0^2(t)dt,$$
$$d(x) = \sum_{r=1}^{2} p_r(x)\left(1 - p_r(x)\right), \quad K_0 = K * K.$$

In particular, if $m_i^{(1)} = m_i^{(2)} = N_n$, $i = 1,\ldots,n$, then

$$\lim_{n\to\infty} \frac{N_n^2 B_n^2(p_1, p_2)}{b_n} = \sigma^2(p_1, p_2),$$

also, if $p_1(x) = p_2(x) = p_0(x)$, then

$$\lim_{n\to\infty} \frac{N_n^2 B_n^2}{b_n} = \sigma^2(p_0) = 2\int_0^1 p_0^2(x)\left(1 - p_0(x)\right)^2 dx \int_{|x|\leq 2\tau} K_0^2(x)dx.$$

Asymptotic Normality of Statistic U_n

The following assertion holds true.

Theorem 1. Let $K(x) \in H(x)$ and $p_k(x) \in C^1[0,1]$, $k = 1,2$. If $\dfrac{N_n^4}{nb_n^2} \to 0$ and $N_n^4 b_n \to 0$ as $n \to \infty$, then

$$\frac{U_n - \mathbf{E}U_n}{B_n} \xrightarrow{d} N(0,1), \quad B_n = B_n(p_1, p_2),$$

where $\xrightarrow{d}$ denotes convergence in distribution, and $N(0,1)$ is a random variable having a standard normal distribution $\Phi(x)$.

Corollary 1. Let $K(x) \in H(\tau)$, $p_k(x) \in C^1[0,1]$, $k = 1,2$. Further, let, $m_i^{(1)} = m_i^{(2)} = N_n$, $i = 1,\ldots,n$. If $\dfrac{N_n^4}{nb_n^2} \to 0$ and $N_n^4 b_n \to 0$ as $n \to \infty$, then

$$b_n^{-\frac{1}{2}} \left(\frac{N_n U_n - \mathbf{E}N_n U_n}{\sigma(p_1, p_2)} \right) \xrightarrow{d} N(0,1).$$

Theorem 2. Let $K(x) \in H(x)$ and $p_k(x) \in C^1[0,1]$, $k = 1,2$. Further let $m_i^{(1)} = m_i^{(2)} = N_n$, $i = 1,\ldots,n$. If $\dfrac{N_n^4}{nb_n^2} \to 0$ and $N_n^4 b_n \to 0$ as $n \to \infty$, then

$$b^{-1/2} \left(\frac{N_n U_n - \Delta(p_1, p_2)}{\sigma(p_1, p_2)} \right) \xrightarrow{d} N(0,1),$$

where

$$\Delta(p_1, p_2) = \frac{1}{2} \int_0^1 d(x)dx \int_{|t| \le \tau} K^2(t)dt.$$

Corollary 2. Let all conditions of Theorem 3.2 be fulfilled. Let the hypothesis $H_0: p_1(x) = p_2(x) = p_0(x)$ be true. Then for $n \to \infty$

$$b_n^{-1/2} \left(\frac{\hat{T}_n - \Delta(p_0)}{\sigma(p_0)} \right) \xrightarrow{d} N(0,1),$$

$$\hat{T}_n = N_n T_n = \frac{1}{2} nb_n N_n \int_{\Omega_n(\tau)} \left[\hat{p}_{1n}(x) - \hat{p}_{2n}(x) \right]^2 p_n^2(x)dx,$$

$$\Delta(p_0) = \int_0^1 p_0(x)\left(1 - p_0(x)\right)dx \int_{|t| \le \tau} K^2(t)dt$$

$$\sigma^2(p_0) = 2\int_0^1 p_0^2(x)\left(1 - p_0(x)\right)^2 dx \int_{|t| \le 2\tau} K_0^2(t)dt.$$

Corollary 3. Let $K(x)$ and $p_i(x)$ satisfy conditions of Theorem 2. Further, let $m_i^{(1)} = m_i^{(2)} = N_0$, $i = 1,2,\ldots,n$, $1 \le N_0 < \infty$, and the hypothesis be true

$$H_0: \quad p_1(x) = p_2(x) = p_0(x).$$

If $nb_n^2 \to \infty$, then for $n \to \infty$

$$b^{-1/2}\left(\bar{T}_n - \Delta(p_0)\right)\sigma^{-1}(p_0) \xrightarrow{d} N(0,1), \qquad \bar{T}_n = N_0 \hat{T}_n .$$

Application of the Statistic $\hat{T}_n$ for Hypothesis Testing

As an important application of the Corollary 2 let us construct the testing of ***simple*** hypothesis H_0, according to which $p_1(x) = p_2(x) \equiv p_0(x)$, where the function $p_0(x)$ is fully defined. The critical region is established by the inequality

$$\hat{T}_n \ge d_n(\alpha) = \Delta(p_0) + b_n^{1/2}\sigma(p_0)\lambda_\alpha , \tag{1}$$

where $\Phi(\lambda_\alpha) = 1-\alpha$, $\Phi(\lambda)$ – standard normal distribution.

Let now $p_0(x)$ be not defined by hypothesis (i.e. testing **complicated** hypothesis). Then directly using (1) is impossible. Previously unknown parameters $\Delta(p_0)$ and $\sigma^2(p_0)$ should be replaced by their consistent estimates $\hat{\Delta}$ and $\hat{\sigma}^2$, respectively. As an estimate of $\Delta(p_0)$ and $\sigma^2(p_0)$ we consider the following statistics

$$\hat{\Delta}_n = \int\limits_{\Omega_n(\tau)} \lambda_n(x)dx \int\limits_{|x|\le\tau} K^2(x)dx ,$$

$$\hat{\sigma}_n^2 = 2\int\limits_{\Omega_n(\tau)} \lambda_n^2(x)dx \int\limits_{|x|\le 2\tau} K_0^2(x)dx$$

$$\lambda_n(x) = \frac{1}{2}\left[p_{1n}\left(p_n(x) - p_{1n}(x)\right) + p_{2n}(x)\left(p_n(x) - p_{2n}(x)\right)\right]$$

and we will show that

$$b_n^{-1/2}\left(\hat{\Delta}_n - \Delta(p_0)\right) \xrightarrow{\mathbf{P}} 0, \quad \hat{\sigma}_n^2 \xrightarrow{\mathbf{P}} \sigma^2(p_0) \tag{2}$$

Theorem 3. Let $K(x) \in H(\tau)$ and $p_1(x) = p_2(x) = p_0(x) \in C^1[0,1]$. Let $m_i^{(1)} = m_i^{(2)} = N_n$, $i = 1,\ldots,n.$ If $\dfrac{N_n^4}{nb_n^2} \to 0$ and $N_n^4 b_n \to 0$ as $n \to \infty$, then

$$b_n^{-1/2}\left(\frac{\hat{T}_n - \hat{\Delta}_n}{\hat{\sigma}_n}\right) \xrightarrow{d} N(0,1).$$

Theorem 3 enables us to construct the asymptotic test for testing the complicated hypothesis $H_0: \ p_1(x) = p_2(x)$, $x \in [0,1]$. The critical domain is established by the inequality

$$\hat{T}_n \ge \tilde{d}_n(a) = \hat{\Delta}_n + b_n^{1/2}\hat{\sigma}_n\lambda_\alpha , \qquad \Phi(\lambda_\alpha) = 1-\alpha . \tag{3}$$

Now, let us consider the question whether the test based on (3) is consistent. The following assertion is true.

Theorem 4. *Let* $K(x)\in H(\tau)$, $p_1(x), p_2(x)\in C^1[0,1]$. *If* $\frac{N_n^4}{nb_n^2}\to\infty$ *and* $N_n^4 b_n\to 0$ *as* $n\to\infty$, *then*

$$\gamma_n(p_1,p_2)=\mathbf{P}_{H_1}\left(\hat{T}_n\geq\tilde{d}_n(\alpha)\right)\longrightarrow 1.$$

An alternative hypothesis H_1 here is any pair $\left(p_1(x),p_2(x)\right)$, $0\leq p_i(x)\leq 1$, $p_i(x)\in C^1[0,1]$, $i=1,2$, such that $p_1(x)\neq p_2(x)$, $x\in[0,1]$.

Thus, for any fixed alternative the power of the test for testing the hypothesis based on T_n tends to 1. However, if with the change of n the alternative changes and approaches the basic hypothesis H_0, then the test power will not necessarily converge to 1. As an example let us consider the sequence of Pitman type alternatives close to the hypothesis H_0:

$$H_{1n}:\quad p_1(x)=p_0(x),\qquad p_2^{(n)}(x)=p_0(x)+\alpha_n\varphi(x),\qquad \alpha_n\to 0.$$

Theorem 5. $K(x)\in H(\tau)$, $p_0(x)$ and $\varphi(x)\in C^1[0,1]$. If $b_n=n^{-\delta}$, $N_n=n^{\beta}$, $\alpha_n=n^{-\frac{1+\beta}{2}+\frac{\delta}{4}}$ $\left(4\beta<\delta,\quad 4\beta+2\delta<1,\quad 1+\beta>\delta/2\right)$, then statistic $b_n^{-1/2}\left(\hat{T}_n-\hat{\Delta}_n\right)\hat{\sigma}_n^{-1}$ for the alternative H_{1n} has a normal limiting distribution with the parameters

$$\left(\frac{1}{\sigma(p_0)}\int_0^1\varphi^2(u)du,1\right),$$

i.e. the limits power of the test (3) is equal to

$$1-\Phi\left(\lambda_\alpha-\frac{1}{\sigma(p_0)}\int_0^1\varphi^2(u)du\right).$$

Remark 1. Analog of the Theorems 1 and 2 are true for the statistic $\overline{T}_n$.

Remark 2. It should be emphasized that the behavior of the estimate $\hat{p}_{kn}(x)$, $k=1,2$, near the boundary of the interval $[0,1]$ is worse than within interval $[\tau b_n,1-\tau b_n]$ (see [9]). That is why to avoid difficulties associated with this boundary effect, we consider the integral mean-square deviation on $\Omega_n(\tau)$.

Remark 3. Let x_i be the partition points of the interval $[0,1]$, chosen so that $H(x_j)=\frac{2j-1}{2n}$, $j=1,\ldots,n$, where $H(x)=\int_0^x h(u)du$, $h(u)$ is some continuous distribution density on $[0,1]$. In that case, the generalization of the results of the paper can be obtained by arguments analogous to those used above.

მათემატიკა

ორი ბერნულის რეგრესიის ფუნქციის ტოლობის ჰიპოთეზის შემოწმების ახალი მეთოდის შესახებ დაჯგუფებული მონაცემებისთვის

პ. ბაბილუა* და ე. ნადარაია**

*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 დადგენილია ბერნულის რეგრესიის ფუნქციის შეფასებათა ინტეგრალური კვადრატული გადახრის ზღვარითი განაწილების კანონი ორი დაჯგუფებული შერჩევისთვის. ამის საფუძველზე აგებულია ახალი კრიტერიუმი ორი ბერნულის რეგრესიის ფუნქციის ტოლობის ჰიპოთეზის შემოწმების შესახებ. შესწავლილია აგებული კრიტერიუმის ძალდებულების საკითხი, გარდა ამისა, შესწავლილია კრიტერიუმის სიმძლავრის ასიმპტოტიკა გარკვეული ტიპის დაახლოებადი ალტერნატივებისთვის.

REFERENCES

1. Efromovich S. (1999) Nonparametric curve estimation. Methods, theory, and applications. Springer Series in Statistics. Springer-Verlag, New York.
2. Copas J. B. (1983) Plotting p against x . *Appl. Statist*. **32**, 2: 25-31.
3. Okumura H. and Naito K. (2004) Weighted kernel estimators in nonparametric binomial regression. The International Conference on Recent Trends and Directions in Nonparametric Statistics. *J. Nonparametr. Stat*. **16**, 1-2: 39-62.
4. Müller H.G. and Schmitt T. (1988) Kernel and probit estimates in quantal bioassay. *J. Amer. Statist.* Assoc. **83**, 403: 750-759.
5. Aerts M. and Veraverbeke N. (1995) Bootstrapping a nonparametric polytomous regression model. *Math. Methods Statist*. **4,** 2: 189-200.
6. Nadaraya E. A. (1964) On a regression estimate. *Teor. veroiatnost. i primenen.* **9 :** 157-159 (in Russian).
7. Watson G. S. (1964) Smooth regression analysis. *Sankhya Ser.* A **26** : 359-372.
8. Nadaraya E., Babilua P. and Sokhadze G. (2010) Estimation of a distribution function by an indirect sample. *Ukr. Mat. Zh*. **62,** 12: 1642-1658; translation in Ukr. *Math. J*. **62**, 12: 1906-1924.
9. Hart J. D. and Wehrly Th. E. (1992) Kernel regression when the boundary region is large, with an application to testing the adequacy of polynomial models. *J. Amer. Statist. Assoc*. **87**, 420: 1018-1024.

Received September, 2021