The Second International Conference "Problems of Cybernetics and Informatics"
September 10-12, 2008, Baku, Azerbaijan. The plenary paper
www.pci2008.science.az/4/07.pdf

SUFFICIENT EMPIRICAL METHOD (SEM) OF POISSON SAMPLE HYPOTHESIS TESTING

Evgeny Chepurin, I. Chepurina

Moscow State University, Moscow, Russia, *echepurin@mail.ru*

1. Let $y=(y_1,...,y_n)$, y_i - be independent random variables, $i=\overline{1,n}$.Further will test the hypothesis

$$\Gamma_1 : y_i \stackrel{d}{=} POIS(a_i\theta_0),\ \theta_0>0,\ a_i>0 \tag{1}$$

where a_i are the known quantities, $i=\overline{1,n}$. Against the alternative estimation

$$\Gamma_2 : y_i \stackrel{d}{\neq} POIS(a_i\theta_0),\ \theta_0>0,\quad i=\overline{1,n} \tag{2}$$

The problem of testing of Γ_1 against the alternative estimation Γ_2 a $n\to\infty$ was considered in detail in the Bolshev's paper , see [1.p 153-165] We site the main results of this paper

Assume $a=\sum_{i=1}^{n}a_i$, $\pi_i=\frac{a_i}{a}$, $i=\overline{1,n}$, $\vec{\pi}=(\pi_1,...,\pi_n)^T$.

As is known, a family of Poisson distributions for the hypothesis Γ_1, possesses sufficient statistics $S(y)=\sum_{i=1}^{n}y_i$

$$S(y)\stackrel{d}{=}POIS(a\ \theta_0).$$

It is immediately shown that

$$y/(S(y)=s)\stackrel{d}{=}MULTI(s;n,\vec{\pi}),.$$

i.e. conventional distribution of the random variable $y/(S(y)=s)$ is "tree", i.e. it is independent of the interfering parameter θ_0.

As is shown in the paper [1] the relation [2] is characteristically for the hypothesis Γ_1 and remind that is means that is conventional distribution of the random variable $y/(S(y)=s)$ is polynomial of concrete form then $y_1,...,y_n$ are independent identically distributed Poisson random variables of the form [1]. Characteristically property of [2] should be tested. Thereby, the similar criterion will be constructed to test the hypothesis Γ_1. Theory of such criterions is considered in detail in the monograph [1]. In the order to test the hypothesis of polynomial with the known probabilities vector $\overrightarrow{\pi_n}$, usually it is advisable to use the statistics of χ^2 of the form

$$X^2=\sum_{i=1}^{n}\frac{(y_i-s\pi_i)^2}{s\pi_i}=\sum_{i=1}^{n}\frac{y_i^2}{s\pi_i}-s. \tag{3}$$

As it is noted in [1,] as $n\to\infty$, and $s\to\infty$ $\min s\pi_i\geq 10$

$$X^2\stackrel{d}{\approx}\chi^2_{n-1} \tag{4}$$

The case when $n \to \infty$, but $s << n$, i.e when $\theta_0 << 1$. Is in obviously interesting events for application. When analyzing the data on the number of traffic incidents in motor – insurance/One can be faced with such a situation. So, in the volume of brief-case $m = 1749$ general number of traffic incidents equals $s = 57$ or for $n = 2389$, the number $s = 67$. The cited examples of relations between and are characteristically for motor-insurance practice. In this case conventional distribution of statistics $X^2 / \left(S(y) = s\right)$ as $n \to \infty$ will be asymptotically normal with the mean value

$$E\left\{X^2 / \left(S = s\right)\right\} = n - 1 \tag{5}$$

and dispersion

$$D\left\{X^2 / \left(S = s\right)\right\} = 2(n-1) + \frac{1}{s}\left(\sum_{i=1}^{n} \frac{1}{\pi_i} - n^2 - 2n + 2\right) \tag{6}$$

see [1.] Thus, for the statistics may be used for testing polynomial hypothesis allowing for its conventional asymptotic normality and relations (5), (6). It the hypothesis of type is considered as an alternative one, the critical domain should be taken in the form of/

It there are no such a priori notions on alternative estimation the critical domain is chosen bilateral, The constant and are chosen on the basis of values of the first kind, see [1]/

2. The theory considered above decsn"t cover the case of small size and moderate values of n. For testing the hypothesis Γ_1 the use of other statistics unlike the statistics (3), but sensitive to the deviations of data from ten hypothesis Γ_1 of great interest. In particular

$$T(y) = \max_{1 \le i \le n} \left| \frac{y_i}{n} - \pi_i \right| \tag{7}$$

is related to such statistics. However, for $n << \infty$ is not succeeded to find

$$\Psi(u) = P\left\{T(y) < u / S(y) = s\right\} \tag{8}$$

conventional distribution function of the statistics $T(y)$ on the hyper surface $\left\{y : S(y) = s\right\}$since it is a very complicated multivariate problem to overcome. The similar statement is also true for the distribution function of the statistics χ^2. However it is relatively simple to estimate the value

$$a_{abserv}(s) = 1 - \Psi\left(T(y)\right) \tag{9}$$

conventional observed level of significance by efficient empirical over aging method (SEM-method).Really, let $z = \left(z_1, ..., z_s\right)^T$ where z_i are independent identically distributed one dimensional discrete random variables, $z_i \in N, i = 1, s,$ with the distribution function $\Gamma_0(u)$ where $I(A)$is the indicator of the event A .Then having assumed $y_i^* = \sum_{j=1}^{s} I\left(i = z_j\right)$ for $i = \overline{1.n}, y^* = \left(y_1^*, ... y_n^*\right)$we get $y^* \stackrel{d}{=} y / \left(S(y) = s\right)$.The random variable y^* is said to be a variance of data y since unconventional distribution y^* coincides for Γ_1 with the distribution y, i.e. it contains the same y number information about the hypothesis Γ_1. Remind that

$$a_{abserv.}(s) = P\left\{T\left(y^*\right) \ge T(y) / S(y) = s\right\} \tag{10}$$

It we generate independent choice of versions of data $\left(y_1^*(1),...y_n^*(B)\right)$ of volume B where $y^*(j)\overset{d}{=}y/\left(S(y)=s\right)$ for $j=\overline{1,B}$ then for the hypothesis Γ_1 the quantity

$$a_{abserv.}(s)=\frac{1}{B}\sum_{j=1}^{B}I\left(T\left(y^*(j)\right)\geq T(y)\right) \tag{11}$$

Is a consistent unbiased estimate for $a_{abserv.}(s)$, [3] .What is more since are random variables, then it follows from (11) that

$$Ba_{abserv}(s)\overset{d}{=}BIN\left(B;a_{abserv}(s)\right) \tag{12}$$

Where BIN (B,q) is binomial random variable, i.e the number of success in Bernoulli 's B tests with probability of success in the separate test equal to $q,0<q<1$ notice that the value may be chosen arbitrarily large, its choice is restricted only by computer of computations and power of data samples of psevdo random identically distributed numbers.

But neither the first nor the second reasons are not obstacles for estimating $a_{abserv.}(s)$

On the other hand, there exist the data samples of random

On the other hand, there exist data samples of random values of practically unrestricted power see [4] .On the other hand the problem on estimation of finding $a_{abserv.}(s)$is ideally structured for conducting computations in powerful multi-processor computers. Is for the choice of the quantity B , it may be realized either on the basis of two-stage procedure of construction of γ – continental interval of reassigned width for $a_{abserv.}(s)$ see[5]. Or on the basis of obvious asymptotic equality following from Mouvr –Laplace theorem,

$$P\left\{\sqrt{B}\left|\frac{a_{abserv}(s)-a_{abserv}(s)}{\sqrt{a_{abserv}(s)(1-a_{abserv}(s))}}\right|\leq u_{\frac{1+\gamma}{2}}\right\}=\gamma+O\left(\frac{1}{\sqrt{B}}\right) \tag{13}$$

where , $\Phi(u_\varepsilon)=\varepsilon,0<\varepsilon<1,0<\gamma<1,\Phi(u)=\int_{-\infty}^{u}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz$, see[6] It follows from the last equality thert the error in estimation of $a_{abserv.}(s)$ with probability γ doesn't exceed $\Delta=\frac{u_{\frac{1+\gamma}{2}}}{2\sqrt{B}}$. Having given Δ and γ we get

$$B\approx\left(\frac{u_{\frac{1+\gamma}{2}}}{2\Delta}\right)^2.$$

Notice that the stated methodology is not "fastened" to concrete statistics of criterion. It can be realized for any statistics of criterion not being S measurable

For constructing criterion of dimension (approximately) α_1 the following

$$\lambda(y)=\begin{cases} accept\ \Gamma_1\ for\ \hat{a}_{abserv} \geq \alpha_1 \\ acject\ \Gamma_1\ for\ \hat{a}_{abserv} < \alpha_1 \end{cases}$$

Should be accepted.

It is possible to estimate the power of the accepted criterion for a number of alternative estimations (including complicated ones) The SEM- method essary for this purpose is stated in [3] and [7].

3.The method of construction of criterion considered in 2 is transferred in a natural way, to the problem on solution of hypothesis testing on homogeneity and stochastic ordering of traffic incidents floras. Namely these problems are faced invage rates distribution of different contingent of policy-hollers.

4. The Poisson flow is base stochastic model in a number of suvestigations on insurance theory. Accept ion of Poisson hypothesis (or mixed Poisson property) very simplifies mathematical simulation of this important component of wage rates system of insurance . However given analysis of statistical data in particular, Russian testifies in favour of outcast it ОСАГО both of Poisson hypothesis and a number of widely used mixed Poisson family of distributions.

References

1. Bolshev L.N. Probability theory and mathematical statistics. Selected papers. M. Nauka, 1987.
2. Leman E. Statistical hypothesi testing. Moscow Nauka, 1979.
3. Chepurin E.V. On Analitic- Computer Methods of Statistical Inferences of Small Size Data Samples//Proceedings of the International Conference Probabilistic Analysis of Rare Events.Riga, Aviation University, p.180-194.
4. Deng L.Y. Random number generation for the New Century.The American Statistican, May 2000, v. 54, 2, pp. 145-150.
5. Zax Sh. Theory of statistical conclusions. Moscow Mir, 1975.
6. Gnedenko B.V, Course of probability theory. Moscow URSS, 2001.
7. Andronov A.M. Hajiyev A.H., Chepurin E.V. see. Abstracts PCI'2006: pci2006.science.az