SovietRxiv

Soviet-era science, translated into English

Reports of the Academy of Sciences of the USSR

1957

SovietRxiv

View the original and related papers at https://sovietrxiv.org/items/ru-195701.85004

Source: Math-Net.Ru and CyberLeninka. Machine translation. Verify with the original.

SovietRxiv

Abstract

Full Text

Reports of the Academy of Sciences of the USSR

1957. Vol. 117, No. 2

MATHEMATICS

G. P. BASHARIN

ON THE USE OF THE χ^2 GOODNESS-OF-FIT TEST AS A TEST FOR INDEPENDENCE OF TRIALS

(Presented by Academician A. N. Kolmogorov on 30 V 1957)

1. Let the random vector $\vec{\xi} = (\xi_1, \dots, \xi_s)$ have a proper normal distribution with mean value $\vec{0} = (0, \dots, 0)$ and matrix of second moments Λ. Then the function

$$f(\vec{\xi}) = \frac{1}{(2\pi)^{s/2}\sqrt{\det \Lambda}} \exp\{-{}^1/_2\, \vec{\xi}\Lambda^{-1}\vec{\xi}'\} \tag{1}$$

(the prime denotes transposition) is the probability density in R_s, and the quadratic form $\vec{\xi}\Lambda^{-1}\vec{\xi}'$ has a χ^2 distribution with s degrees of freedom ([6]), Ch. 24).

2*. Consider a system with a finite number of states A_i $(i = 1, \dots, s)$. Denote the frequency of the state A_i in a sequence of N trials by m_i, and the frequency of transition from state A_i to state A_j, i.e. the number of series A_iA_j, by m_{ij} $(i, j = 1, \dots, s)$. In the statistical analysis of random numbers and in certain other cases, it is of interest to test the simple hypothesis H_0', according to which the trials are independent and all states are equiprobable, against the composite hypothesis according to which the trials are connected in a simple Markov chain (hypothesis H_1).

The limiting matrix of second moments of the normalized quantities

$$x_{ij} = \sqrt{\frac{s^2}{N}}\left(m_{ij} - \frac{N}{s^2}\right) \quad (i, j = 1, \dots, s)$$

has the form

$$M = \left\| \delta_{ij}^{uv} + \frac{1}{s}\left(\delta_j^u + \delta_v^i\right) - \frac{3}{s^2} \right\| \quad (i, j, u, v = 1, \dots, s), \tag{2}$$

SovietRxiv

where $\delta_{ij}^{uv} = \delta_i^u \delta_j^v;\ \delta_i^u = 1,\ i = u;\ \delta_i^u = 0,\ i \neq u.$

The matrix M has rank $s^2 - s$. Removing the last s rows and s columns from the matrix M, we obtain a nonsingular matrix $\widetilde{M}$ $(i, u = 1, \ldots, s-1;\ j, v = 1, \ldots, s)$. It can be shown that $\det \widetilde{M} = 1/s^s$ and that**

$$\widetilde{M}^{-1} = \left\| \delta_{ij}^{uv} + \frac{s-1}{s}\delta_i^u + \delta_j^v + \delta_j^s + \delta_v^s - \delta_v^i - \delta_j^u + \frac{s-1}{s} \right\| \tag{3}$$

$$(i, u = 1, \ldots, s-1;\ j, v = 1, \ldots, s).$$

---

* The results set forth in Sec. 2 were reported at a meeting of the research seminar of the Department of Probability Theory of Moscow State University in December 1955, and the results of Sec. 3—in June 1954.

** In the résumé of my report published in the journal *Theory of Probability and Its Applications*, 2, 141 (1957), misprints were made in this formula, as well as in formulas (6)—(9).

To verify formula (3), it suffices to multiply M and $\widetilde{M}^{-1}$. The quadratic form of the random variables x_{ij} $(i = 1, \ldots, s-1;\ j = 1, \ldots, s)$ with matrix (3) has the limiting χ^2 distribution with $s^2 - s$ degrees of freedom and, after simple simplifications, is reduced to the form

$$\frac{s^2}{N} \sum_{i,j=1}^{s} \left(m_{ij} - \frac{N}{s^2} \right)^2 - \frac{s}{N} \sum_{i=1}^{s} \left(m_i - \frac{N}{s} \right)^2 = \psi_2^2 - \psi_1^2. \tag{4}$$

The random variables x_{ij} $(i = 1, \ldots, s-1;\ j = 1, \ldots, s)$ have a limiting nondegenerate normal distribution with probability density

$$f(\|x_{ij}\|) \sim \frac{s^{s/2}}{(2\pi)^{(s^2-s)/2}} \exp\left\{ -\frac{1}{2} \left(\psi_2^2 - \psi_1^2 \right) \right\}. \tag{5}$$

Since the logarithm of the likelihood ratio in the case under consideration is proportional to $\psi_2^2 - \psi_1^2$, the best critical region of the hypothesis H_0' with respect to H_1 is determined from the condition $\psi_2^2 - \psi_1^2 > c$, where c is the level constant.

3. In 1939 Kendall and Smith ([5]) proposed using a test of the hypothesis H_0' based on the statistic

$$\psi_2^2 = \frac{s^2}{N} \sum_{i,j=1}^{s} \left(m_{ij} - \frac{N}{s^2} \right)^2, \tag{6}$$

erroneously assuming that this statistic has the χ^2 distribution with $s^2 - s$ degrees of freedom. In the journal literature this error was first pointed out in 1953 by Good ([2]), who considered a test of the hypothesis H_0' against a compound Markov chain, but only for a prime number of states s. For this case Good obtained, by means of a discrete Fourier transform, statistics having a limiting χ^2 distribution. Attempts to extend Good' s method to the case of composite s were not successful.

Since statistic (6) was used for a long time for practical purposes, it is of interest to obtain its distribution not only for the hypothesis H_0', but also for the hypothesis $H_0 = (p_1, p_2, \dots, p_s)$. The limiting matrix of second moments of the normalized quantities

$$y_{ij} = \frac{m_{ij} - Np_ip_j}{\sqrt{Np_ip_j}} \quad (i, j = 1, \dots, s)$$

has the form

$$M = \left\| \delta_{ij}^{uv} + \sqrt{p_ip_v}\,\delta_j^u + \sqrt{p_up_j}\,\delta_v^i - 3\sqrt{p_ip_jp_up_v} \right\| \quad (i, j, u, v = 1, \dots, s). \quad (7)$$

It can be shown that s characteristic roots of matrix (7) are equal to 0; $(s-1)^2$ are equal to 1, and $s - 1$ are equal to 2. Since (7) is the matrix of second moments of the limiting normal distribution of the quantities y_{ij} $(i, j = 1, \dots, s)$, it follows from this ([6], ch. 24.5) that, as $N \to \infty$, the sampling distribution of the statistic

$$\sum_{i,j=1}^{s} \frac{(m_{ij} - Np_ip_j)^2}{Np_ip_j}$$

converges to the distribution of the sum of independent random variables $\chi^2_{(s-1)^2} + 2\chi^2_{s-1}$, where χ^2_m is a random variable with the χ^2 distribution with m degrees of freedom.

4. If a compound hypothesis competes with the hypothesis H_0', according to which the trials are connected into a homogeneous Markov chain of order $\nu - 1$, then the method based on the likelihood ratio leads to statistics that are functions of the frequencies of series of length ν in the original sample of length N ([1,3,4,7]). Since in order that it be possible

were to use limiting distributions, the sample size would have to be of order 10^s, and therefore the possibility of applying the indicated statistics decreases very rapidly as s grows and, in particular, as ν grows. In this connection it becomes necessary to use statistics depending only on the frequencies of certain series. As an example we give a statistic that is a function of the frequencies of s series

SovietRxiv

of length ν, consisting of identical states, and that has the limiting distribution χ^2_s, if the hypothesis H'_0 is true.

The limiting matrix of second moments of the normalized frequencies $\sqrt{\frac{s^\nu}{N}}\left(m_{i_1 i_2 \dots i_\nu} - \frac{N}{s^\nu}\right)$ has the form

$$\left\| \delta^{j_1 \dots j_\nu}_{i_1 \dots i_\nu} + \sum_{k=1}^{\nu-1} \frac{1}{s^k} \left(\delta^{j_1 \dots j_{\nu-k}}_{i_{k+1} \dots i_\nu} + \delta^{j_{k+1} \dots j_\nu}_{i_1 \dots i_{\nu-k}} \right) - \frac{2\nu - 1}{s^\nu} \right\|, \tag{8}$$

$$i_1, \dots, i_\nu,\ j_1, \dots, j_\nu = 1, \dots, s; \qquad \nu \geq 1.$$

Therefore the limiting matrix of second moments of s normalized frequencies $\sqrt{\frac{s^\nu}{N}}\left(m_{i \dots i} - \frac{N}{s^\nu}\right)$ has the form

$$M_s = \left\| c_\nu \delta^j_i - \frac{2\nu - 1}{s^\nu} \right\|, \qquad i, j = 1, \dots, s; \qquad c_\nu = \frac{s^\nu + s^{\nu-1} - 2}{(s-1)s^{\nu-1}}; \qquad \nu \geq 1. \tag{9}$$

Inverting the matrix (9) for $\nu \geq 2$, we obtain

$$M_s^{-1} = \frac{1}{c_\nu} \left\| \delta^j_i + d_\nu \right\|, \qquad i, j = 1, \dots, s; \qquad d_\nu = \frac{2\nu - 1}{\frac{s}{s-1}(s^\nu + s^{\nu-1} - 2) - s(2\nu - 1)}. \tag{10}$$

It follows from this that the quadratic form of the random variables $\sqrt{\frac{s^\nu}{N}}\left(m_{i \dots i} - \frac{N}{s^\nu}\right)$ with matrix (10), having the limiting distribution χ^2_s, is written in the form

$$\frac{1}{c^\nu} \frac{s^\nu}{N} \left\{ \sum_{i=1}^{s} \left(m_{i \dots i} - \frac{N}{s^\nu} \right)^2 + d_\nu \left(\sum_{i=1}^{s} m_{i \dots i} - \frac{N}{s^{\nu-1}} \right)^2 \right\}. \tag{11}$$

In the case $\nu = 2$ one can construct an analogous statistic for testing the hypothesis H_0. In this case

$$M_2 = \left\| (1 + 2p_i)\delta^j_i - 3p_i p_j \right\|, \qquad i, j = 1, \dots, s.$$

It can be shown that

SovietRxiv

$$M_2^{-1} = \left\| \frac{1}{1+2p_i}\delta_i^j + \frac{3}{1-3\sum_{i=1}^{s}\frac{p_i^2}{1+2p_i}} \frac{p_i}{1+2p_i}\frac{p_j}{1+2p_j} \right\|; \qquad i,j=1,\dots,s. \tag{12}$$

The quadratic form of the variables $\frac{m_{ii}-Np_i^2}{\sqrt{N}p_i}$ $(i=1,\dots,s)$ with matrix (12), which as $N\to\infty$ has the limiting distribution χ_s^2, is written in the form

$$\frac{1}{N}\sum_{i=1}^{s}\frac{(m_{ii}-Np_i^2)^2}{p_i^2(1+2p_i)} + \frac{3}{\left(1-3\sum_{i=1}^{s}\frac{p_i^2}{1+2p_i}\right)N}\left[\sum_{i=1}^{s}\frac{m_{ii}-Np_i^2}{1+2p_i}\right]^2. \tag{13}$$

5. Let s shot pellets be thrown successively into a square box Q, consisting of $s\times s$ square cells. After in the cell-

cell (i,j), the next pellet may with equal probability fall into any of the remaining cells except those cells which are located in row i and column j of the square Q. The result of a series of s throws can be represented in the form of a substitution matrix A_μ, and the result of N independent series–in the form of the matrix W

$$W=\sum_{\mu=1}^{N}A_\mu=\|m_{ij}\| \quad (i,j=1,\dots,s); \qquad \sum_{j=1}^{s}m_{ij}=\sum_{i=1}^{s}m_{ij}=N. \tag{14}$$

It can be shown that, as $N\to\infty$, the random variables

$$z_{ij}=\sqrt{\frac{s}{N}}\left(m_{ij}-\frac{N}{s}\right) \quad (i,j=1,\dots,s)$$

have a limiting normal distribution of rank $(s-1)^2$ with matrix of second moments

$$M=\frac{1}{s^2(s-1)}\left\|s^2\delta_{ij}^{uv}-s\left(\delta_i^u+\delta_j^v\right)+1\right\|, \qquad i,j,u,v=1,\dots,s. \tag{15}$$

The random variables z_{ij} $(i,j=1,\dots,s-1)$ have a limiting nondegenerate normal distribution with matrix of second moments which, in cell notation, has the form

SovietRxiv

$$\widetilde{M} = \frac{1}{s^2(s-1)} \left\{ s^2 \begin{pmatrix} E_{s-1} & \cdots & 0 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 0 & \cdots & E_{s-1} \end{pmatrix} - s \begin{pmatrix} E_{s-1} & \cdots & E_{s-1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ E_{s-1} & \cdots & E_{s-1} \end{pmatrix} - s \begin{pmatrix} c_{s-1} & \cdots & 0 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 0 & \cdots & c_{s-1} \end{pmatrix} + \begin{pmatrix} c_{s-1} & \cdots & c_{s-1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ c_{s-1} & \cdots & c_{s-1} \end{pmatrix} \right\} \tag{16}$$

where E_{s-1} is the identity matrix of order $s-1$; c_{s-1} is a matrix of order $s-1$, all of whose elements are equal to 1. It can be shown that

$$\det \widetilde{M} = \frac{1}{s^{2(s-1)}(s-1)^{(s-1)^2}},$$

and that the quadratic form in the variables z_{ij} $(i, j = 1, \ldots, s-1)$ with matrix $\widetilde{M}^{-1}$ reduces to the form

$$\frac{s-1}{N} \sum_{i,j=1}^{s} \left(m_{ij} - \frac{N}{s} \right)^2. \tag{17}$$

According to § 1, this quadratic form has the limiting distribution $\chi^2_{(s-1)^2}$, and the function

$$f(\|z_{ij}\|) \sim \frac{s^{s-1}(s-1)^{(s-1)^2/2}}{(2\pi)^{(s-1)^2/2}} \exp\left\{ -\frac{1}{2} \frac{s-1}{N} \sum_{i,j=1}^{s} \left(m_{ij} - \frac{N}{s} \right)^2 \right\} \tag{18}$$

is the limiting probability density of the variables z_{ij} $(i, j = 1, \ldots, s-1)$. Formula (17) may be used for statistical purposes.

In conclusion I express my deep gratitude to my scientific adviser, Prof. V. Ya. Kozlov, and also to Acad. A. N. Kolmogorov and B. A. Sevastyanov for valuable advice.

Received
25 V 1957

References Cited

1. M. S. Bartlett, Proc. Cambr. Phil. Soc., **1**, 86 (1951).
2. I. J. Good, Proc. Cambr. Phil. Soc., **49**, 2765 (1935).
3. I. J. Good, Biometrika, **42**, 531 (1955).
4. P. Hoel, Biometrica, **41**, 430 (1954).

SovietRxiv

5. M. G. Kendall, B. Babington Smith, J. Roy. Stat. Soc., suppl. **6**, 51 (1939).

6. G. Cramér, *Mathematical Methods of Statistics*, Moscow, 1948.

7. N. V. Smirnov, *Vestn. LGU*, No. 11 (1955).

Note: Figure translations are in progress. See original paper for figures.

Source: Math-Net.Ru and CyberLeninka. Machine translation. Verify with the original.