Soltan A. ALIEV, Andrey N. BOBYLYAK, Yaroslav I. YELEYKO

CRITERION FOR INDEPENDENCE OF DATA

Abstract

This publication provides a criterion for the independence of data.

The classical theory of statistical conclusions is based on the concept of sampling. By the definition, a sample is a random vector where components are the results of observations of some random variable [3, p.7]. In many cases, we make the assumption that elements of the sample are independent random variables.

In practice, very important is the fact that our data are independent.

Reproduced in this publication criterion can check independent of dates.

We consider some of the random variable ξ. $\xi_1, ..., \xi n, ...$ are observations of ξ. We need to find a subsequence of independent random variables.

Let $\xi^{(n)} = (\xi_1, ..., \xi_n)$ be a sample for this random variable ξ.

Lemma. [2, p. 91-95]. *Let $\xi_0, \xi_1, ...$ be a sequence of independent identically distributed random variables, $\beta = \min\limits_k\{k \in N \mid \xi_k > \xi_0\}$. Then*

$$P\{\beta = k\} = \frac{1}{k(k+1)}, k = 1, 2, ...$$

Theorem 1. *Let $(\xi_1, ..., \xi_n)$ be a sample, $P\{\xi_1 < x\} = F(x)$ be a continuous function, $(\xi_1, ..., \xi_m)$ be a subsample $(m < n)$, $\xi_{(1)}, ..., \xi_{(m)}$ be a variational series of subsample.*

1. If $\beta = \min\{k \in N : \xi_{m+k} > \xi_{(m)}\}$, then $P\{\beta > k\} = \frac{m}{m+k}$.

2. If $\beta = \min\{k \in N : \xi_{m+k} > \xi_{(m-r+1)}\}$, then $P\{\beta > k\} = \frac{C_m^r}{C_{m+k}^r}$ and if $r \geq 2$, $x > 0$ then $P\{\beta \leq mx\} \to 1 - \frac{1}{(1+x)^r}$, $m \to \infty$.

3. If $\beta = \min\{k \in N : \xi_{m+k} \leq \xi_{(1)}\}$, then $P\{\beta > k\} = \frac{m}{m+k}$.

4. If $\beta = \min\{k \in N : \xi_{m+k} < \xi_{(r)}\}$, then $P\{\beta > k\} = \frac{C_m^r}{C_{m+k}^r}$ and if $r \geq 2$, $x > 0$, $P\{\beta \leq mx\} \to 1 - \frac{1}{(1+x)^r}$, $m \to \infty$.

Proof. First note that the first statement is a partial case of the second assertion and the third statement is a partial case of the fourth assertion. So just prove the second assertion.

With the same reasoning as in Lemma, we get that

$$\{\beta > k\} \Leftrightarrow \{\xi_{(r)} \text{ of the subsample } \xi_1, ..., \xi_{m+k} \text{ is } \xi_{(m-r+1)} \text{ of the sample}\} \Leftrightarrow$$

$$\{\xi_{(r)} \text{ of the subsample } \xi_1, ..., \xi_{m+k} \text{ match } \xi_{(r)} \text{ of the subsample } \xi_1, ..., \xi_m\}.$$

Among the $\xi_1, \ldots, \xi_{m+k}$ we can choose r values by C_{m+k}^r ways and among the $\xi_1, \ldots, \xi_m$ we can choose r values by C_m^r ways. Therefore $P\{\beta > k\} = \frac{C_m^r}{C_{m+k}^r}$.

Let $r \geq 2$, then given that β is a discrete random variable that takes only natural values, we obtain

$$\forall\, x > 0 \quad P\{\beta \leq mx\} = P\{\beta \leq [mx]\}.$$

Denote $[mx] = k$, then

$$P\{\beta \leq mx\} = P\{\beta \leq k\} = 1 - \frac{m!}{(m-r)!}\frac{(m+k-r)!}{(m+k)!}.$$

Using the Stirling formula, we obtain that

$$P\{\beta \leq k\} = 1 - \frac{m^m}{e^m}\sqrt{\frac{2\pi m}{2\pi(m-r)}}\left(\frac{e}{m-r}\right)^{m-r} \times$$

$$\times \left(\frac{m+k-r}{e}\right)^{m+k-r}\sqrt{\frac{2\pi(m+k-r)}{2\pi(m+k)}}\left(\frac{e}{m+k}\right)^{m+k} =$$

$$= 1 - \frac{m^m(m+k-r)^{m+k-r}}{(m-r)^{m-r}(m+k)^{m+k}}\sqrt{\frac{m(m+k-r)}{(m-r)(m+k)}} =$$

$$= 1 - \frac{\left(1+\frac{k-r}{m}\right)^{\left(\frac{m}{k-r}+1\right)(k-r)}}{\left(1-\frac{r}{m}\right)^{\left(\frac{m}{r}+1\right)(-r)}\left(1+\frac{k}{m}\right)^{\left(\frac{m}{k}+1\right)k}}\sqrt{\frac{m(m+k-r)}{(m-r)\,(m+k)}}.$$

Since $m \to \infty$, then

$$P\{\beta \leq k\} \to 1 - \frac{\left(1+\dfrac{k-r}{m}\right)^{k-r}}{\left(1-\dfrac{r}{m}\right)^{-r}\left(1+\dfrac{k}{m}\right)^{k}} = 1 - \left(\frac{1+\dfrac{k}{m}-\dfrac{r}{m}}{1+\dfrac{k}{m}}\right)^{k}\left(\frac{1-\dfrac{r}{m}}{1+\dfrac{k}{m}-\dfrac{r}{m}}\right)^{r}.$$

Given that $\dfrac{k}{m} = \dfrac{[mx]}{m} = x - \dfrac{\{mx\}}{m} \to x$, $\dfrac{r}{m} \to 0$, we obtain

$$P\{\beta \leq mx\} \to 1 - \frac{1}{(1+x)^r},\ m \to \infty.$$

The proof of parts 3 and 4 is completely analogous to that of parts 1 and 2 for the random variable $\beta = \min\{k \in N : \xi_{m+k} < \xi_{(r)}\}$, which is a symmetrical analog of the already studied random variable $\beta = \min\{k \in N : \xi_{m+k} > \xi_{(m-r+1)}\}$.

Theorem is proved.

Remarks. Let the absolutely continuous random variable η_r be given by its distribution function

$$F_{\eta_r}(x) = \begin{cases} 1 - \frac{1}{(1+x)^r}, & x > 0; \\ 0, & x \leq 0. \end{cases}$$

Then according to the proven theorem, for large m, the random variable $\dfrac{\beta}{m}$ converges to the random variable η_r in distribution. This is equivalent to the following:

$$\forall\, a,\, b \in R \;\; P\left\{a < \frac{\beta}{m} < b\right\} \to P\left\{a < \eta_r < b\right\},\ m \to \infty.$$

Practical calculations show that if $r \leq 0.4m$, the value is quite precise.

Later we will use the following notation

$$\beta_{m,r} = \min\{k \in N : \xi_{m+k} > \xi_{(m-r+1)}\},$$

$$\beta^r_{m,r} = \min\{k \in N : \xi_{m+k} < \xi_{(r)}\}.$$

For the subsample $(\xi_1, \ldots, \xi m)$ the serial statistics are denoted by $\xi_{(1)}, \ldots, \xi_{(m)}$.

If a sub-samples is not uniquely determined the ordinal statistics will be denoted by $\xi_{(1)_m}, \ldots \xi_{(m)_m}$.

Theorem 2. *The characteristic function of a random variable η_r is equal to*

$$\psi_r(t) = 1 + \frac{i}{r-1}t + \ldots + \frac{i^k(r-k-1)!}{(r-1)!}t^k + ... + \frac{i^{r-1}}{(r-1)!}t^{r-1} + Er(t),$$

where $E_r(t)$ is the remainder term of t^{r-1} order, i.e.

$$\forall t \in R \quad \mid E_r(t) \mid \leq \frac{2 \mid t \mid^{r-1}}{(r-1)!}$$

and

$$E_r(t) = \begin{cases} \frac{ite^{-it}}{(r-1)!}\left(Ci(\mid t \mid) + isi(\mid t \mid) \cdot sign(t)\right) t^{r-1}, & t \neq 0; \\ 0, & t = 0, \end{cases}$$

where $Ci(t) = -\int\limits_1^\infty \frac{\cos(y)}{y}dy$, $si(t) = -\int\limits_1^\infty \frac{\sin(y)}{y}dy$.

Proof. By the definition of the characteristic function,

$$\psi_r(t) = \int\limits_{-\infty}^{+\infty} e^{itx}dF_{\eta_r}(x) = \int\limits_0^\infty \frac{re^{itx}}{(1+x)^{1+r}}dx = re^{-it}\int\limits_1^\infty \frac{e^{itx}}{x^{r+1}}dx = re^{-it}I_{r+1}(t).$$

Integrating r times by parts we obtain

$$\psi_r(t) = 1 + \frac{i}{r-1}t + \ldots + \frac{i^k(r-k-1)!}{(r-1)!}t^k + \ldots + \frac{i^{r-1}}{(r-1)!}t^{r-1} + \frac{i^r e^{-it}}{(r-1)!}I_1(t)t^r. \quad (1)$$

That formula proved up to the remainder term. Then when $t \neq 0$ we obtain

$$E_r(t) = \frac{i^r e^{-it}}{(r-1)!}I_1(t) = \frac{i^r t^r e^{-it}}{(r-1)!}\left(\int\limits_1^\infty \frac{\cos(tx)}{x}dx + i\int\limits_1^\infty \frac{\sin(tx)}{x}dx\right) =$$

$$= \frac{ite^{-it}}{(r-1)!}\left(Ci\left(|t|\right) + isi\left(|t|\right)\right)t^{r-1}, \quad (2)$$

where $Ci(t) = -\int\limits_1^\infty \frac{\cos(y)}{y}dy$, $si(t) = -\int\limits_1^\infty \frac{\sin(y)}{y}dy$.

Since $\psi_r(0) = 1$ [1, p. 120], we have $E_r(0) = 0$.

To prove the theorem, it remains to show that the remainder term $E_r(t)$ is of order t^{r-1}, i.e. to prove that

$$\forall t \in R \mid E_r(t) \mid \leq \frac{2 \mid t \mid^{r-1}}{(r-1)!}.$$

When $t = 0$ the statement is obvious.
Assuming that $t \neq 0$.
According to (1)

$$\psi_1(t) = 1 + ite^{-it} I_1(t). \tag{3}$$

Substituting (3) in (2), we obtain

$$E_r(t) = \frac{i^{r-1}}{(r-1)!}(-1 + \psi_1(t))t^{r-1}.$$

Since $\forall\, t \in R \mid \psi_1(t) \mid \leq 1$ we have $\mid E_r(t) \mid \leq \dfrac{2 \mid t \mid^{r-1}}{(r-1)!}$.

The theorem is proved.

Corollary 1. *The characteristic function of* η_r , $\forall r > 1$, *is*

$$\psi_r(t) = 1 + \frac{i}{r-1}t + ... + \frac{i^k \prod\limits_{l=0}^{[r]-k-2}(r-k-l-1)}{\prod\limits_{l=0}^{[r]-2}(r-l-1)}t^k + ...+$$

$$+... + \frac{i^{r-1}}{\prod\limits_{l=0}^{[r]-2}(r-l-1)}t^{r-1} + E_r(t), \tag{4}$$

where $E_r(t)$ *is the remaining member of the order* t^{r-1}, *namely*

$$\forall\, t \in R \mid E_r(t) \mid \leq \frac{2 \mid t \mid^{r-1}}{\prod\limits_{l=0}^{[r]-2}(r-l-1)}.$$

Proof. Proof of Corollary can be conducted on a similar scheme as the proof of the theorem, in view of the difference that the parameter r is real and that the product does not convolves in the classic definition of factorial.

Therefore,

$$\psi_r(t) = 1 + \frac{i}{r-1}t + ... + \frac{i^k \prod\limits_{l=0}^{[r]-2-k}(r-k-l-1)}{\prod\limits_{l=0}^{[r]-2}(r-l-1)}t^k + ...$$

$$... + \frac{i^{r-1}}{\prod_{l=0}^{[r]-2}(r-l-1)}t^{r-1} + \frac{i^r e^{-it}}{\prod_{l=0}^{[r]-2}(r-l-1)}I_r(t).$$

Then we obtain that

$$\forall\, t \in R \mid E_r(t) \mid = \left| \frac{i^{r-1}}{\prod_{l=0}^{[r]-2}(r-l-1)}(-1+\psi_{[r]}(t))t^{r-1} \right| \leq \frac{2 \mid t \mid^{r-1}}{\prod_{l=0}^{[r]-2}(r-l-1)}.$$

The result is proved.

Corollary 2. *Let* $\eta_{r_1}, ..., \eta_{r_s}$ $(\forall\, j \quad r_j \in \{2, 3, ...\})$ *be independent random variables,* ψ_Σ *be the characteristic function of their sum* $\eta_\Sigma = \eta_{r_1} + ... + \eta_{r_s}$, $p = 1 + \frac{1}{\sum_{j=1}^{s}\frac{1}{r_j-1}}$. *Then* $M\left[\eta_p\right] = M\left[\eta_\Sigma\right]$.

Proof. From the properties of the characteristic function [1] it follows

$$\psi_\Sigma(t) = \left(1 + \frac{i}{r_1-1}t + ... + \frac{i^k(r_1-k-1)!}{(r_1-1)!}t^k + ... + \frac{i^{r_1-1}}{(r_1-1)!}t^{r_1-1} + E_{r_1}(t)\right) \times$$

$$\times \left(1 + \frac{i}{r_s-1}t + ... + \frac{i^k(rs-k-1)!}{(r_s-1)!}t^{k-1} + ... + \frac{i^{r_1-'1}}{(r_s-1)!}t^{r_s-1} + E_{r_s}(t)\right).$$

Since $\forall j \quad r_j > 1$, the characteristic function of sum has the form

$$\psi_\Sigma(t) = 1 + it\sum_{j=1}^{s}\frac{1}{rj-1} + ...$$

As $\psi_p(t)$ has a similar structure and

$$\psi_p(t) = 1 + \frac{i}{p-1}t + ... + \frac{i^k\prod_{l=0}^{[p]-k-2}(p-k-l-1)}{\prod_{l=0}^{[p]-2}(p-l-1)}t^k + ... + \frac{i^{p-1}}{\prod_{l=0}^{[p]-2}(p-l-1)}t^{p-1} + E_p(t),$$

then for equality $M\left[\eta_p\right] = M\left[\eta_\Sigma\right]$ necessarily $\frac{1}{p-1} = \sum_{j=1}^{s}\frac{1}{r_j-1}$, i.e. $p = 1 + \frac{1}{\sum_{j=1}^{s}\frac{1}{r_j-1}}$.

The result is proved.

Theorem 3. *Let* $\eta_1, ..., \eta_{r_s}$, $\eta_{p_{(1)}}, ..., \eta_{p(s)}$ $(\forall\, l \quad r_l \in \{2, 3, ...\}, p(l) = 1 + \frac{1}{\sum_{j=1}^{l}\frac{1}{r_j-1}})$ *be independent random variables and*

$$P\{\eta_{r_j} < x\} = \begin{cases} 0, & x \le 0, \\ 1 - \dfrac{1}{(1+x)^{r_j}}, & x > 0, \end{cases}$$

$$P\{\eta_{p(j)} < x\} = \begin{cases} 0, & x \le 0, \\ 1 - \dfrac{1}{(1+x)^{p(j)}}, & x > 0. \end{cases}$$

Then $\dfrac{1}{s}\sum\limits_{j=1}^{s}\eta_{r_j} \xrightarrow{p} \dfrac{\eta_{p(s)}}{s}$.

Proof. At the first we consider the random variable $\overline{\eta_s} = \dfrac{1}{s}\sum\limits_{j=1}^{s}\eta_{r_j} - \dfrac{\eta_{p(s)}}{s}$. According to (4)

$$\psi_{\eta_{r_j}}(t) = 1 + \frac{i}{r_j - 1}t + o(t)$$

at $t \to 0$. Fixing an arbitrary t, we get that for $\dfrac{\eta_{r_j}}{s}$

$$\psi\left(t\right) = 1 + \frac{i}{r_j - 1}\frac{t}{s} + o\left(\frac{1}{s}\right)$$

at $s \to \infty$.
The characteristic function for the sum $\sum\limits_{j=1}^{s}\dfrac{\eta_{r_j}}{s}$ is

$$\psi_{\eta_{r_j/s}}(t) = \prod_{j=1}^{s}\left(1 + \frac{i}{r_j - 1}\frac{t}{s} + o\left(\frac{1}{s}\right)\right) = 1 + \frac{it}{s}\sum_{j=1}^{s}\frac{1}{r_j - 1} + o(1)$$

at $s \to \infty$.

Since $\sum\limits_{j=1}^{s}\dfrac{\eta_{r_j}}{s}$ and $\eta_{p(s)}$ are independent random variables, the characteristic function for $\overline{\eta_s} = \dfrac{1}{s}\sum\limits_{j=1}^{s}\eta_{r_j} - \dfrac{\eta_{p(s)}}{s}$ is

$$\psi(t) = \left(1 + \frac{it}{s}\sum_{j=1}^{s}\frac{1}{r_j - 1} + o(1)\right)\left(1 - \frac{1}{p(s) - 1}\frac{t}{s} + o\left(\frac{1}{s}\right)\right) = 1 + o(1)$$

at $s \to \infty$.

The theorem is proved.

From the above facts it follows that can be taken as the statistic criterion the random variable

$$T_n = \frac{1}{s + s'}\left(\sum_{j=1}^{s}\frac{\beta_{m_j,r_j}}{m_j} + \sum_{j=1}^{s}\frac{\beta_{m_j,r_j}^{r_j}}{m_j}\right)$$

interval around the expectation

$$\Delta_n = \left(\frac{1}{s + s'}\left(-\frac{\epsilon}{2} + \left(1 + \frac{1}{p - 1}\right)^{-p}\right)^{-\frac{1}{p}} - \frac{1}{s + s'};\right.$$

$$\frac{1}{s+s'}\left(\frac{\varepsilon}{2}+\left(1+\frac{1}{p-1}\right)^{-p}\right)^{-\frac{1}{p}}-\frac{1}{s+s'}\Bigg)$$

as the decision making region.

There remains the problem of choosing "optimal" sequences $(m_j, r_j)_{j=1}^{s}$ and $(m_j, r_j)_{j=1}^{s'}$ such that, firstly, $\frac{\beta_{m_1,r_1}}{m_1},\dots,\frac{\beta_{m_s,r_s}}{m_s},\frac{\beta^{r_1}_{m_1,r_1}}{m_1},\dots,\frac{\beta^{r_{s'}}_{m_{s'},r_{s'}}}{m_{s'}}$ be independent and, secondly, that the value $s+s'$ be large.

From the above facts it follows that one of the possible sequences of rational choice $(m_j, r_j)_{j=1}^{s}$ and $(m_j, r_j)_{j=1}^{s'}$ is as follows:

For a sequence $\xi 1, ..., \xi_n$ we fix a subsample of length m_1, $\xi_1, ..., \xi_{m_1}$.

$\forall \tilde{r}_j \in \{2, ..., [0.4 \cdot m_1]\}$we obtain sequences $\beta_{m_1,\tilde{r}j}$ or $\beta^{\tilde{r}_j}_{m_1,\tilde{r}_j}$ $j = \overline{1, s'_1}$ sorted ascending $\beta_{m_1,\tilde{r}j}$ or $\beta^{\tilde{r}_j}_{m_1,\tilde{r}_j}$ and in the extent of their values in $\tilde{r}_j$.

Then as r_1 we take the average value $r_1 = \tilde{r}_{1+\left[\frac{\overline{s_1}-1}{2}\right]}$, and the corresponding value of $\beta_{m_1,\tilde{r}j}$ or $\beta^{\tilde{r}_j}_{m_1,\tilde{r}_j}$, which is denoted by β_1.

The following values m_2, r_2 can be chosen a similar procedure the sequence $\xi_{m_1+\beta 1+1}, ..., \xi n$ etc.

Remarks. The independence in total of the sequence of random variables

$$\frac{\beta_{m_1,r_1}}{m_1},\dots,\frac{\beta_{m_s,r_s}}{m_s},\frac{\beta^{r_1}_{m_1,r_1}}{m_1},\dots,\frac{\beta^{r_s}_{m_s,r_s}}{m_s}$$

is obviously, as they belong to different groups of mutually independent sub-samples of the sample $\xi_1, ..., \xi_n$.

The coefficient 0.4 in the choice of $\tilde{r}_j \in \{2, ..., [0.4 \cdot m_1]\}$ is recommended for practical application of the criterion, which is caused by approximation of the random variable $\frac{\beta_{m,r_1}}{m}$ by the random variable η_r.

Conclusion

The constructed criterion is the first attempt at assessing the unknown data sample for independence. The advantage of it that does not require knowledge of data distribution. But it is clear that this is a foundation to false conclusions, so more detailed assessment should be done certain assumptions about the nature of the data and apply other more narrow criteria of independence, which generally allows a comprehensive approach to the problem.

References

[1]. Gihman I.I., Skorohod A.V., Yadrenko M.Jo. *The Theory of Probability and Mathematical Statistics.* Kyiv, "Vyshcha shkola", 1988, 439 p.

[2]. Dynkin Ye. B., Yushkevich A.A. *Theorems and Problems on Markov processes.* Moskow, "Nauka", 1967, 232 p.

[3]. Ivchenko G.I., Medvedev Yu. I. *Mathematical Statistics.* Moskow, "Vyshchaya shkola", 1984, 248 p.

Soltan A. Aliev,
Institute of Mathematics and Mechanics of NAS of Azerbaijan
9, B.Vahabzade str., AZ1141, Baku, Azerbaijan
Tel.: (99412) 539 47 20 (off.)

Andrey N. Bobylyak, Yaroslav I. Yeleyko
I. Franko Lvov National Universitety
1, Universitet str., 79000, Lvov, Ukraine

Received: January 11, 2012; Revised: April 18, 2012.