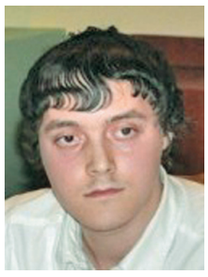


Ninchuk V.S., Chepurko V.A.

ABOUT CHECKING OF RANDOMNESS OF DATA WITH TIES

The paper considers research of various nonparametric methods of checking a hypothesis of randomness of sampled data in a situation when these data contain repeating (tied) observations. The repeating observations named ties lead to ambiguous ranking. The arithmetic-mean rank is usually assigned to elements of the tied group. As a result, statistical decision rules change as the criterion statistics changes its distribution. This paper is devoted to research of capacity of various criteria at presence of ties.

Keywords: *tied group, rank, inversion, Spearman test, Kendall test, integral test.*

Introduction

For obtaining correct estimations and statistical prediction of various complex technical systems' life time, it is necessary to use well-conditioned methods. At the analysis of statistical data failures of recoverable objects the following assumption is often used: the observable failure flow is random, homogeneous in time, failure frequencies are casual, independent and etc. However, these preconditions (initial hypotheses) frequently are not met. Therefore, the accepted final decision about researched object state and all numerical estimations of dependability accepted in this case will be rather doubtful.

Homogeneousness in time is understood as the fact, that failure frequencies will be equally distributed without dependence from where an interval of the fixed length can be found on time axis. Randomness, first of all, means absence in observable frequencies of any simple regularities and trends. For example, it is possible to consider a situation as regularity presence, if the failure flow contains time interval (sometimes not one) with the essentially greater failure rate than it is on the average, in the whole observable time. The simple linear trend points to the presence of close to linear dependence between observable failure frequency and time t.

Some task solution methods in the statistical analysis of the data demand improvements, modernizations, and sometimes development, especially in those cases when the initial statistical information possesses certain types of shortcomings. To such types of shortcomings it is possible to attribute presence of repeated observations (tied data) in initial data. In this situation Kendall test is frequently enough applied to check randomness with P. Sena's amendments to ties [1]. But as is known, in some cases, for example, at small sample, Kendall test is less powerful in comparison with Spearman test. It is possible to expect, that if amendments to Spearman test is taken into account correctly then application of new test becomes more preferable in sense of power. In addition to that it is desirable to perform the comparative analysis of power and the integrated test for randomness offered in [2].

This paper considers the methodology of a hypothesis check about statistical data randomness in view of the amendment to available ties. At the same time it is understood that the accepted hypothesis about randomness will mean actually absence of monotonous trends in an observable time series. Besides, the method of statistical tests is applied to investigate powers of various tests against alternative of monotonous trend presence.

At first we shall make the brief description of randomness hypothesis.

Hypothesis of randomness

The hypothesis about the beginning of time of system approaching limiting state is often enough stands up for alternative to the randomness hypothesis. In this case failure frequencies have the obvious trend to increase. If a new system failures is considered, then by virtue of the known phenomenon of artificial aging, failure frequencies will be rather high in the beginning a researched time interval with the gradual tendency of their reduction. In this case the assumption of negative trends' presence in observable frequencies will be considered as alternative to a zero hypothesis. If the researcher is interested simply in presence of a monotonous trend (without the indication of its decreasing or increasing), then he should consider two-sided alternative hypothesis.

Hypothesis of randomness is, perhaps, the first and fundamental hypothesis used at processing of the numerical random data. It consists also in the assumption of tendencies' absence of determined regularities in these data. Let's formulate it.

In various statistical problems the initial data $\mathbf{X}=(X_1,...,X_n)$ frequently is considered as a random sample of some distribution $L(\xi)$, i.e. it is assumed that components X_i of a data vector $\mathbf{X}$ are independent and equally distributed random variables. As a rule, this assumption is justified and follows from the problem nature, but sometimes it requires check.

The hypothesis of randomness H^* consists in the assumption of distribution function symmetry (or its density) [3].

$$H_* : F_{\mathrm{X}}(x_{r_1}, x_{r_2}, ..., x_{r_n}) = F_{\mathrm{X}}(x_1, x_2, ..., x_n), \qquad (1)$$

The simplified alternative of this hypothesis is also frequently considered, assuming additionally presence of components' independence.

$$H_0 : F_{\mathbf{X}}(x_1, ..., x_n) = F(x_1)...F(x_n), \qquad (2)$$

where $F(x)$ is some distribution function. Such hypothesis is called the hypothesis of randomness though actually it states independence and identical distribution of a vector component $\mathbf{X}$. Thus, independent and equally distributed random variables $X_1,X_2,...,X_n$ should be considered for H_0. It is obvious, that $H_0 \subset H^*$, i.e. H_0 is more rigorous assumption of initial data nature than H^*. Nevertheless, just this zero hypothesis H_0 will be checked further in the paper.

In nonparametric problem statement it is expedient to consider the following alternatives [4].

For example:

$H_1^{(+)}$: $F_1(x) < F_2(x) < ... < F_n(x)$ – alternative of increase,

$H_1^{(-)}$: $F_1(x) > F_2(x) > ... > F_n(x)$- alternative of decrease.

Goodness-of-fit test for check of hypothesis H_0 can be constructed, based on various consideration. First, it is supposed, that vector $\mathbf{X}$ has continuous distribution. If the randomness hypothesis really takes place, then components of vector $\mathbf{X}$ "are equal in rights" and consequently the data should not be in any sense are ordered. In other words, the situation corresponding to hypothesis H_0 can be characterized as "total chaos" or "state of total disorder". At deviations from H_0 the initial data have this or that order, ties become apparent or dependences on the order of indexation. Therefore, the distribution test of H_0 can be constructed based on statistics, measuring the degree of initial data "disorder".

Tied data

This section will be devoted to research of various approaches to taking into account of tied data. As is known (see, for example [1]), group of tied observation is called a set of observation having the same value. At the same time several tied groups can be observed in the initial data. If an observation has the value different from other sample units it can be considered as a group of ties of volume (size) 1.

Two methods of manipulation with concurrences were discussed in literature. The first method consists in ordering concurrent observations in random manner. Its advantage is simplicity, and it does not require new theory, but at the same time we, obviously, sacrifice the information contained in observations, and it is possible to expect, that it will be less effective, than the second method, consisting in the fact that to each of group of the concurrent observation the average rank of this group is attributed. Advantages of both methods were investigated not much, but it is shown, that Wilcoxon test at random partitioning of concurrences has smaller AOE in comparison when average ranks are given.

Before appearance of the further information the method of an average rank seems to be more common. For this reason this method of an average rank assignment is applied also in this paper.

Kendall test of independence randomness free from distribution

Known statistics of Kendall T_n can be applied to check of a hypothesis of independence of one data set (let it be $(X_1,...,X_n)$) from another one $(Y_1,...,Y_n)$ [5]. In this case the initial information is the two-dimensional array:

$$\begin{pmatrix} X_1 & X_2 & ... & X_n \\ Y_1 & Y_2 & ... & Y_n \end{pmatrix}. \qquad (3)$$

Столбцы this file are rearranged so that bottom line Y-ов has been ordered (is allowable on increase). If after that and in the top line the tendency to increase (or to decrease) X-ов it is possible

Columns of this array are rearranged in such a manner that the lower of *Y*s has been ordered (let us suppose ascending ordering). If after this, in the upper row tendency to increase (or decrease) *Xs* will be noticeable also, it is possible to

judge about the presence of positive (negative) correlation dependence of X on Y. The degree of this dependence can be measured by the number of inversions T_n. For this kind of data statistics T_n is an estimate of the value τ, which is defined as follows:

$$\tau=2P\left((X_1-X_2)(Y_1-Y_2)>0\right)-1.$$

Let's reduce Kendall test in the form presented in [1]. However, it should be noted, that shown below statistics K as a matter of fact is a number of Kendall inversions T_n for Xs in that case when columns of two-dimensional array (3) are arranged in ascending order of Ys.

In addition to that it should be noticed, that if $Y_i=i$ is replaced for every i then independence test will transform to randomness test in sense of the previous section. The alternative $\tau>0$ ($\tau<0$) will mean presence of a positive (negative) trend of Xs.

For check of a hypothesis about independence of random variables X and Y (from whence it follows that $\tau=0$), namely

$$H_0: P\left(\{X\le a\}\cap\{Y\le b\}\right)= \\ = P(x\le a)\cdot P(Y\le b)\quad \forall a,b, \tag{4}$$

it is necessary to do the following.

To put

$$K=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\operatorname{sgn}\left(\left(X_i-X_j\right)\left(Y_i-Y_j\right)\right),$$

$$\text{where } \operatorname{sgn}(x)=\begin{cases}1, & x>0,\\ 0, & x=0,\\ -1, & x<0.\end{cases} \tag{5}$$

Thus to each pair indexes (i, j) it is attributed +1, for $i<j$, if

$(X_i-X_j)(Y_i-Y_j)>0$, and otherwise it is attributed 1. Adding up these figures (with their signs), we obtain the sum K.

It is known, that for the large samples it is necessary to use

$$K^*=\frac{K-E_0(K)}{\sqrt{D_0(K)}}=\frac{K}{\sqrt{n(n-1)(2n+5)/18}}, \tag{6}$$

where $E_0(K)$ и $D_0(K)$ is an expectation and a dispersion of statistics K at correct zero hypothesis.

At realization of H_0 statistics K^* asymptotically *(for* $n\to\infty$*)* has standard normal distribution $N(0, 1)$.

The approximate test in this case will be the following:

To reject H_0, if если $K^*\ge u(\alpha)$, (7)

To accept H_0, if если $K^*<u(\alpha)$,

where the constant $u(\alpha)$ is a fractile of standard normal law, i.e. the value satisfying the equation

$$P_0[K^*\le u(\alpha)]=\alpha.$$

If among n observations of X or among n observations Y there are ties then the dispersion $D_0(K)$ in the definition of K^* should be replaced as follows:

$$D_0(K)=\frac{5}{12}\frac{\left(\begin{array}{l}n(n-1)(2n+5)-\\ -\sum_{i=1}^{g}t_i(t_i-1)(2t_i+5)-\\ -\sum_{j=1}^{h}u_j(u_j-1)(2u_j+5)\end{array}\right)}{18}+ \\ +\frac{\left\{\sum_{i=1}^{g}t_i(t_i-1)(t_i-2)\right\}\left\{\sum_{j=1}^{h}u_j(u_j-1)(u_j-2)\right\}}{9n(n-1)(n-2)}+ \\ +\frac{\left\{\sum_{i=1}^{g}t_i(t_i-1)\right\}\left\{\sum_{j=1}^{h}u_j(u_j-1)\right\}}{2n(n-1)}, \tag{8}$$

where g is a number of groups of tied observations X; ti is the size of i-th group of observations X; h is a number of groups of tied observations Y; u_j is the size of j-th group of observations Y. The amendment (8) for the first time has been obtained by P. Sena.

Rank Spearman test

Spearman in his studies for checking hypothesis H_0 (1) has offered the following measure of linear tie between random variables.

$$R=\frac{12}{n(n^2-1)}\sum_{i=1}^{n}\left(r_i^x-\frac{n+1}{2}\right)\left(r_i^y-\frac{n+1}{2}\right), \tag{9}$$

where r_i^x and r_i^y are ranks of Xi and Y_i. Arrays are ordered separately. R refers to as *factor of Spearman rank correlation* [6].

As from the alteration of summands the sum is not changed change it is possible array columns (3) to order in such a manner that Y grew and then to define r_i that is obtained (relative) ranks of X_i. In this case it is possible to receive the following form of rank Spearman statistics [1].

$$R=\frac{12}{n(n^2-1)}\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)\left(r_i-\frac{n+1}{2}\right). \tag{11}$$

Sometimes also more convenient form for calculation is used [5]:

$$\rho=1-\frac{6}{n(n^2-1)}\sum_{i=1}^{n}(i-r_i)^2. \tag{12}$$

Let's show, that $R=\rho$ *at absence of ties:*

$$R=\frac{12}{n^3-n}\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)\left(r_i-\frac{n+1}{2}\right)= \\ =\frac{12}{n^3-n}\left(\sum ir_i-n\frac{(n+1)^2}{4}\right).$$

$$\rho=1-\frac{6}{n^3-n}\sum_{i=1}^{n}(r_i-i)^2=\frac{12}{n^3-n}\left(\sum ir_i-n\frac{(n+1)^2}{4}\right).$$

At presence of one tied group with size t:

$$R = \frac{12}{n^3 - n}\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)\left(r_i - \frac{n+1}{2}\right) = \\ = \frac{12}{n^3 - n}\left[\sum ir_i - n\frac{(n+1)^2}{4}\right]. \quad (10)$$

$$\rho = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}(r_i - i)^2 = \\ = \frac{12}{n^3 - n}\left(\sum ir_i - n\frac{(n+1)^2}{4} + \frac{t^3 - t}{24}\right). \quad (11)$$

In addition to the above it is possible to draw a conclusion, that at absence of ties R=ρ, but at presence of other ties it is not quite correctly to use factor ρ. The error, first of all, will be connected with the fact that the mathematical form of statistics R in more degree corresponds to estimation of rank correlation than statistics ρ convenient for calculations.

Therefore, having taken for a basis the statistics R we shall calculate amendments to dispersion, similar to P. Sena's amendments for statistics of the normalized statistics of Kendall test K * (6).

Dispersion of Spearman statistics at presence of connected groups

At the big sample sizes it is better to use normalized Spearman test, i.e. test, which statistics at H_0 will have approximately the normal law.

The fact that Spearman statistics at correct H_0 has asymptotically normal distribution is proved, for example, in the monograph [1]. Standardizations of the normal law can be achieved similarly to (6).

$$R^* = \frac{R - E_0(R)}{\sqrt{D_0(R)}} = \sqrt{n-1}\cdot R, \quad (12)$$

because the expectation $E_0(R)$=0 and dispersion $D_0(R)$= 1/(n–1).

At performance of hypothesis H_0 the statistics R* has asymptotic distribution N (0,1) (for $n\to\infty$).

Let's find the dispersion of Spearman test statistics at presence of connected groups. Next, the dispersion of Spearman test statistics calculations R without taking into account factor 12/(n^3–n), i.e. i.e. without statistics $\tilde{R} = \frac{R(n^3 - n)}{12}$.

Let's designate $c_i = i - \frac{n+1}{2}$.

$$D_0\left(\tilde{R}\right) = D_0\left[\sum_{i=1}^{n} c_i r_i\right] = \\ = \sum_{i=1}^{n} c_i^2 D_0\left(r_i\right) + 2\sum_{i<j} c_i c_j cov_0\left(r_i, r_j\right) = \\ = \sum c_i^2\left(D_0\left(r_i\right) - cov_0\left(r_i, r_j\right)\right)$$

Rank dispersion:

$$D_0\left(r_i\right) = E_0\left(r_i^2\right) - E_0^2\left(r_i\right) = \frac{n^2 - 1}{12} - \frac{t(t^2 - 1)}{12(n-1)},$$

Rank covariation r_i and r_j:

$$cov_0(r_i, r_j) = E_0\left(r_i r_j\right) - E_0\left(r_i\right)E_0\left(r_j\right) = \\ = \frac{n^2(n+1)^2}{12} - \frac{n(n+0.5)(n+1)}{3n(n+1)} - \frac{(n+1)^2}{4}.$$

Because $\sum c_i^2 = \frac{n(n-1)(n+1)}{12}$, then after some simple calculations we obtain the following:

$$D_0\left(\tilde{R}\right) = \frac{n^2(n+1)^2(n-1)}{144}\left(1 - \frac{t(t^2 - 1)}{n(n^2 - 1)}\right).$$

Thus, if the number of connected groups is equal to k, then we have

$$D_0\left(\tilde{R}\right) = \frac{n^2(n+1)^2(n-1)}{144}\left(1 - \frac{\sum_{i=1}^{k} t_i(t_i^2 - 1)}{n(n^2 - 1)}\right).$$

The final form of the normalized Spearman test statistics has the following expression:

$$R^* = \frac{\tilde{R}}{\sqrt{D_0(\tilde{R})}} = \frac{\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)\left(r_i - \frac{n+1}{2}\right)}{\sqrt{\frac{n^2(n+1)^2(n-1)}{144}\left(1 - \frac{\sum_{i=1}^{k} t_i(t_i^2 - 1)}{n(n^2 - 1)}\right)}}. \quad (13)$$

where k is the number of connected groups, t_i is the size of i-th connected group.

At absence of ties we obtain (13). Thus, approximate Spearman test at presence of ties will be the following:

To reject H_0, if $R^* \geq u(\alpha)$, (14)

To accept H_0, if R^* if R* < $u(\alpha)$,

where the constant $u(\alpha)$ is a fractile of standard normal law, i.e. the quantity satisfying the equation

$P_0[R^* \leq u(\alpha)] = \alpha$.

Integral test

The integral test has been offered in [2] constructed on statistics of the following form:

$$In^{(++)} = \frac{\sum_{i=1}^{n}\left(\sum_{1}^{i} r_j - \overline{r^{(+)}}\right)\left(\sum_{1}^{r_i} j - \overline{j^{(+)}}\right)}{S_r^{(+)} S_j^{(+)}};$$

$$In^{(-+)} = \frac{\sum_{i=1}^{n}\left(\sum_{i}^{n} r_j - \overline{r^{(-)}}\right)\left(\sum_{1}^{r_i} j - \overline{j^{(+)}}\right)}{S_r^{(-)} S_j^{(+)}}; \quad (15)$$

$$\text{Where } \overline{r^{(+)}} = \frac{1}{n}\sum_{i=1}^{n}(n+1-i)r_i;\ \overline{r^{(-)}} = \frac{1}{n}\sum_{i=1}^{n} ir_i;$$

$$S_r^{(+)} = \sqrt{\sum_{i=1}^{n}\left(\sum_{j=1}^{i} r_j - \overline{r^{(+)}}\right)^2};\ S_r^{(-)} = \sqrt{\sum_{i=1}^{n}\left(\sum_{j=i}^{n} r_j - \overline{r^{(-)}}\right)^2};$$

$$\overline{j^{(+)}} = \frac{(n+1)(n+2)}{6} - \frac{1}{24n}\sum_{l=1}^{s} t_l\left(t_l^2-1\right);$$

$$S_j^{(+)} = \sqrt{\sum_{i=1}^{n}\left(\frac{r_i\left(r_i+1\right)}{2} - \overline{j^{(+)}}\right)^2}.$$

In order to avoid distribution asymmetry it is offered to use linear combination of initial data (15) as the final version of statistics, for example:

$$In^{(\circ+)} = 0,5(In^{(++)} - In^{(-+)}), \qquad (16)$$

In this case any statistics will be unbiased under condition of zero hypothesis performance. For statistics of the type (19) tabular support is constructed and it was proved, that in asymptotics the given statistics has the same distribution, as Spearman statistics [2]. For the size of observation $n > 7$ critical values were obtained with the help of the improved normal approximation for distribution of Spearman statistics [1]:

$$In_\alpha = \frac{u(\alpha)}{\sqrt{n-1}}\left\{1 - \frac{0.19}{n-1}\left[u^2(\alpha) - 3\right]\right\},$$

where the constant $u(\alpha)$ is a fractile of standard normal law.

The approximate integral test at presence of ties will be the following.

$$\text{To reject } H_0, \text{ if } In^{(\circ+)} \geq In_\alpha, \qquad 17)$$
$$\text{To accept } H_0 \text{ if } In^{(\circ+)} < In_\alpha$$

The further researches will concern the comparative analysis of Spearman, Kendall tests and integral test at presence of ties.

Comparison of tests

As is known, one of the basic methods of comparison of statistical tests is the analysis of their power – of conditional probability to accept alternative provided that it is true. Power function for each test depends on the chosen significance level and some parameter Δ (or a vector of parameters $\bar{\Delta}$ and from data size). In our case the test comparison algorithm for checkup of randomness hypothesis according to power has been based on usual Monte Carlo method it was constructed as follows.

Gaussian (Normal) pseudo-random sample X_1, X_2, ..., X_n with the fixed size n, with the specified structure of ties and with gradually increasing positive trend Δ is repeatedly modeled under the following formula:

$$X_i = \xi_i + i\cdot\Delta,$$

Ties of the specified ties are modeled. Steps connected with m time. Power $P_1(\Delta)$ is estimated by effective assessment.

$$P_i(\Delta) = \frac{\nu_i(\Delta)}{m}.$$

Next the trend angle increases. The steps connected to generation of sample m time and calculations of power function are repeated.

Fig. 1 shows diagrams of four tests':

1. Kendall test with amendment for Sena's tie (7), (8) – Kendall;
2. Spearman test constructed on statistics without taking into account ties (12) – Spearman old;
3. Spearman test (14) with amendment for tie (13) – Spearman new;
4. Integral test (17).

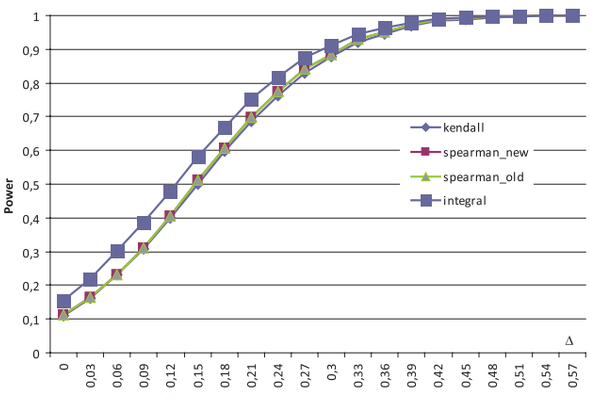

Fig. 1. Comparison of tests

Sample sizes n=10.

It is apparent from the diagram that at the small sample sizes the integral test possesses the greater power, but has a small shift caused by an error of normal approximation. At increase of sample size it is natural to expect reduction of this shift.

The conclusion

The paper considers the issue of checking randomness hypothesis by Spearman test at presence of ties' identical values in initial data. In particular the following problems have been solved:

1. Normalized Spearman test with amendment for tie has been developed.
2. Experiments on comparison of test power have been carried out.

References

1. **Hollender M., Wolf D.** Non-parametric methods of statistics. M.: Finance and statistics, 1983.-518 p.

2. **Antonov A.V., Chepurko V.A.** Integral test for checking trend hypothesis. Dependability, 2006. No, 2. Pp. 17-27.

3. **Gaek Ia., Shidak Z.** The theory of rank tests. M.: Science, 1971.-376 p.

4. **Burtaev J.F., Ostreykovsky V.A.** Statistical analysis of dependability of objects with limited information. M.: Energoatom izdat, 1995,-240 p.

5. **Kendall M.** Rank correlations.Foreign statistical studies. M., «Statistics», 1975.216 p.

6. **Van der Varden.** Mathematical statistics. M.: Foreign literature edition, 1960.435 p.