Application of the Fourfold Tables Method for Analysis of Dynamics of Social Systems

Shikhalev A. M.

Panasyuk M.V.

Burilov A.R.

Kazan state University, Institute of organic and physical chemistry of the Russian Academy of Sciences, Kazan,Russia

Doi:10.5901/mjss.2014.v5n18p335

Abstract

The problem of comparison of statistics expressed in quantitative scale, when the use of all possibilities of statistical analysis is methodologically limited is discussed. The model of analysis of the dynamics in short time series, including that with barely discernible levels, using standardized fourfold table in terms of linguistic variable is considered. Such approach allows building the mapping of statistical point indexes from quantitative scale to order (rank) scale, and forming groups according the studied parameter with a given discrepancy level – weak, significant, strong and absolute.

Keywords: *statistics, time series, fourfold table, test for concordance, linguistic variables, fuzzy sets, classification.*

1. Problem Statement

In the system of statistics types the particular significance has such their characteristics as quantitative, qualitative, temporal and spatial. And if processes of comparison of any two of statistical indicators for the last three characteristics are not problem, assessment of relative importance (or unimportance) of the quantitative discrepancies (differences) is mainly subjective depending on this or that decision maker attitude (DM). It is clear that the numbers 10.0 and 200.0 in size of index changing of [1.0 - 1000.0] confidently differ by researcher, but with the same level of confidence it is doubtful in the specified range for such a pair of quantitative indicators as, for example, 10.0 and 10.5.

Experts and decision makers can assume that either 10.0 or 10.5 is essentially the same value, or these quantitative characteristics define essential qualitative discrepancy of considered statistic. The same applies to analysis of dynamic processes. The conclusions that during the affected period significant changes in dynamics of the process took place or not are often associated with a quantitative analysis of the trend. Then, in present case, raises the problem of whether the discrepancy in 0.5 is sufficient in making appropriate expert or management decisions.

The solution to this problem on the basis of such methods of preliminary analysis of dynamics as moving average method, method of analytical smoothing, etc. is based on the use of large volume and representative samples observations, setting the system of elements of dynamic rows. Parameters of short-term series calculated as result of the use of autoregressive models, analysis of correlations between different time series in case of analysis of their system and the forecast results are often not statistically valid.

Thus, the use of standard mathematical apparatus for analysis of time series is not valid if investigated time series have the following properties:

1) The number of members of a time series' statistical aggregate is small and would not let to achieve the required level of statistical significance;
2) The parameters of dynamics change slightly, at the level of statistical accuracy, and forecast results in comparison with base observable values of a variable remain outwardly almost unchanged (Table 1, Data).

The peculiarities of time series of socioeconomic indexes which show development processes are widespread, and correct interpretation of the results of dynamics analysis and forecasting is often questionable. This leads to the development of mathematical techniques that allow more detailed analysis of such series.

In the analysis of complex dynamic systems where qualitative conclusions, such as classification, ranking of population of data etc. often are the ultimate purpose of the research, formal mathematical and heuristic procedures for transforming the available quantitative data into the data of qualitative nature are usually used. The complexity of these procedures, for example, system of data mining procedures, many-valuedness and complexity of interpretation of results

of dynamics analysis obtained with their help lead to a search for other methods, which allow solving the mentioned problems by quality characteristics' output on level of statistical observations.

This is also heavily based on problems of data collection. As is known, qualitative information could be got much easier than quantitative and qualitative information is more stable than quantitative. Therefore, in some cases, especially when time series is distinguished by mentioned peculiar properties, it may be useful to transform the numeric (quantitative) indexes to qualitative, for example, by converting the values into discrete, categorical variable, to binary relations, preference relations, etc. [2] In these cases fourfold tables are advisable to use as the data structure and method of analysis.

2. The Formation of Mathematical Support for Analysis on the Basis of a Fourfold Table in T.Saati Linguistic Scale

Currently, there are many works devoted to the use of a fourfold table, but the most complete statement of the substance of question is given in the G. Upton's monograph [9].

Suppose **A** and **B** are dichotomous variables that take two different values of A1 and A2 or B1 and B2 respectively. Therefore, there are four possible types of response (result), which exhausting the following combinations of variables: (A1, B1), (A1, B2), (A2, B1), and (A2, B2).

As statistical variables (or attributes) **A** and **B**, each of which is measured in dichotomous scale of items, almost any dichotomous distribution can be used, which divides into natural statistical partitions, into attributes as well, to measure which dichotomous nominal scale is sufficient for the study. This method under different names (four-cell table, fourfold table, dichotomous variables correlation, the method of alternating variations) is widely used in formalization of process of expert estimations.

For example, suppose you want to compare two numbers 10.0 and 10.5, which will be brought together as initial data in the cross-tabulation as model example shows (Table 1).

Table 1. The way of presenting of basic values

Attributes	$\Delta x = \lvert x_{i+1} - x_i \rvert$	x_i
Significant discrepancy	a = 0.5	b = 10
Insignificant discrepancy	c = 0	d = 10

In this case, it can be noted that the Yule association factor Q = (ad - bc) / (ad + bc) is equal to 1, which indicates the existence of a complete positive correlation between attributes A1 and B1, i.e. the discrepancy equal to 0.5 between 10.5 and 10.0 is significant and, obviously, as a hypothesis Q = 1, it will always be present.

After the value of Q is determined, it is necessary to define the degree of confidence of found discrepancies (correlation between attributes A and B), for which Pearson goodness-of-fit test $\aleph^2$ is applied [3], which, in contrast to classical recordings, for fourfold distribution has the form $\aleph^2 = n \bullet \varphi^2$, where $\varphi = (ad\text{-}bc)/((a+c) \bullet (b+d) \bullet (a+b) \bullet (c+d))^{1/2}$ is Pearson association factor, expressing in contrast to Yule association factor Q a measure of bilateral relation and therefore, for our case, its calculation is only the formal operation.

When b = D and C = 0, the formula for calculating the coefficient will be:

$\varphi = \Delta x / (2\Delta x \bullet (\Delta x + b))^{1/2}$

The parameters a, b, c and d are analogues of frequencies of compared attributes, the sums (a + c) and other sums are analogues of frequency arrays sums, and n is the total sum of the four parameters.

Then obtained empirical value $\aleph^2$ is compared to tabulated (critical) value $\aleph^2_{crit}$, on the basis of which a judgment about the choice between statistical hypothesis (null hypothesis - H_0) – a priori recognition that discrepancies between empirical fact and the expected theoretical characteristics are random, and a working hypothesis H which is showing opposite fact, is produced. In our study, the adoption of H_0 will mean no significant discrepancy of Δx in relation to basic value of x. Then the distribution of values-frequencies (a, b, c, d) is random and does not represent relationship between them.

The alternative hypothesis H is based on the assumption of existence of this relationship. However, when approximate inequality of the $\aleph^2 \geq \aleph^2_{crit}$ form is valid, base values shown in form of Table 1 does not correspond to the H_0 hypothesis that is rejected. Then alternative hypothesis H about the existence of significant discrepancies between adjacent elements of time series can be confirmed with a given degree of statistical significance. To achieve this, it is

necessary to use Pearson method and modification of $\aleph^2$ proposed by Yates, which, in his opinion, gives results that are better correlated with the distribution of $\aleph^2$ [9].

Yates statistics $\aleph^{*2}$ differs from Pearson $\aleph^2$ statistics only by form of numerator in the formula of φ calculation. In both cases empirical $\aleph^{*2}$ and $\aleph^2$ are compared with critical values of considered levels of significance by values of theoretical calculation tables for one degree of freedom, which will be considered in more detail.

3. Definition Area of Confidence Limits

Graphical representation of Chi-square theoretical values gives the opportunity to establish, that in interval of degree of statistical confidence P from 0 to 70% (or for risk parameter from 1.0 to 0.3) the curve $\aleph^2_{crit} = \aleph^2_{crit}(P)$ has a clear linear relationship, asymptotically approaching to 100%. Therefore intermediate values of P in the computer model of proposed method of time series analysis rightly be obtained using linear interpolation by equation of line that passing through the two given points, which are fixed Chi-square values and tabulated significance levels or probabilities of P (%).

Not quite clear nature of the obtained results for arbitrary values of n and the fact that analysis of degree of dependence between A and B for sufficiently large values of n leads to any conclusions [9], have necessitated the creation of authoring tools (author's programs) that realize calculation of Q and φ coefficients, as well as other calculated values of the indexes specified in the author's algorithm.

Indeed, the expression $\aleph^2 = n \bullet \varphi$ shows it more than expressive. Empirical value $\aleph^2$ depends on n in direct proportion, and inequality $\aleph^2 \geq \aleph^2_{crit}$ will be more valid than more is n. Therefore, with the aim of comparability of obtained results, basic data should be normalized, i.e. to result, for example, in a notional 100%. However, this operation creates some doubts regarding the possibility of applying Yates criterion $\aleph^{*2}$ in whole spectrum of socio-economic research due to nonmonotonicity of function in the case of representation of time series data. Therefore, in a case of analysis of socioeconomic dynamics, preference should be given to the Pearson criterion.

As a mean of discretization of obtained results it is appropriate to use T. Saati dual tetradic scale (nine-position table) [6], in which measure of discrepancy of binary comparisons is assessed as "no discrepancy", "weak discrepancy", "substantial discrepancy", "strong discrepancy", "absolute discrepancy"; elements of antisymmetric expert matrix numbers 1, 3, 5, 7 and 9 are assigned respectively (numbers 2, 4, 6, 8 - for expression of some expert intermediate preferences).

The possibility of using T. Saati table not only in expert, but in formal-mathematical terms, is motivated by the known similarity of null hypothesis verification process of the general theory of statistics with the process of obtaining of membership functions on some basis - not necessarily an expert - like measure of extent to which a measure of discrepancy between neighboring members of time series corresponds to concept, the meaning of which can be formalized as a fuzzy set C, in which discrepancies may be absent, weak, significant, strong or absolute.

The content of such fuzzy set can be considered in terms of the linguistic variable [1, 5, and 6]. Its formalized exposition is <b, T, X, G, M>,

Where b is the measure of discrepancy of two neighboring elements of time series and is the name of the linguistic variable (LV);

T = {α_1, α_3, α_5, α_7, α_9,} = {no discrepancy, weak discrepancy, significant discrepancy, strong (obvious) discrepancy, absolute discrepancy} is base terminal set of LV, nouns of fuzzy variables, range of definitions of each of them is set X;

G – Syntactic procedure (grammar) allowing operating with elements of term-set T;

M - Semantic procedure, which allows transforming any new value of LV formed by the procedure G into fuzzy variable.

It should be noted that unlike syntactically dependent LVs for which a procedure of formation of new values G depends on many basic values of T, there are LVs for which a procedure of formation of new values does not depend on a set of basic values of T, but on range of X definition, i.e. G=G(x). Therefore, it would be logical to assume that arbitrary values of syntactically independent LV are one-to-one identified by some values of x from the definition area of X [5].

A range of definitions in the proposed method is established according to the following considerations. It is known that in analysis of socioeconomic processes *0.01, 0.05, and 0.10* levels of significance are most commonly used. They correspond to probabilities P equal to 99, 95, and 90%, respectively. As is known, the adequate measurement involves not only general representation of studied object and its parties, but also right choice of scale, for which we used its determination as algorithm by means of which each observed object is associated with some number. Scale values for these objects will be designated as assigned numbers of objects [7].

In our case, the object is a part of formed linguistic variable, precisely – range of definition (existence) of set X. To establish boundaries of its existence let's associate P = 99% with the first value of term-set α_9 = {absolute discrepancy}; P = 95% - with the value of α_7 = {strong discrepancy}; P = 90% - with the value of α_5 = {significant discrepancy}; P = 50% - with the value α_1 = {no discrepancy}, or so: whether there are discrepancies (differences), whether there are no discrepancies (or absence of discrepancies is verisimilar).

You can draw an analogy with upper and lower quintiles' boundaries under normal distribution curve, where exactly half of frequency ratio gets into. Then the value of LV α_3 = {weak discrepancy} will be quite appropriate to assign scale value as average between 50 and 90% as P = 70% that is quite close in value to fraction of frequencies in the "plus or minus sigma" range under the curve of normal distribution, equal to 68, 26% [7].

Although the 70, 00% and 68, 26% values are very close, the discrepancy between them is still fundamental, although *practically imperceptible*. Therefore, the range of X definition can be considered with a sufficient degree of confidence installed as X = [50, 99]. It could also be possible to fill with content procedures G and M using logical operations and modifiers like AND, OR, MUCH, NOT, SLIGHTLY etc. Then semantic procedure can be set by rules on operations with elements of fuzzy set C - union, intersection, and other.

The urgent need for this is not yet apparent, since, according to the statement of problem, here we have to deal with the analysis of already small values of growth rate of statistical indexes in time. However, it would be necessary to provide the comparison of experimental (actual) growth with the so-called "critical" (by analogy with statistics) increase Δx corresponding to $\aleph^2_{crit}$ for values of X in the range from 50% to 99% (Table 2), which, as our testing of the proposed method demonstrated, is quite sufficient for obtaining a stable, uniquely interpretable results.

To illustrate the proposed method (analysis of short time series dynamics that characterize social processes, including that with barely discernible levels, using normalized fourfold tables in terms of linguistic variable) let's refer to data of model example, where x_i = 10.0 and x_{i+1} = 10.5.

The content of the listing of the author's computer program is presented in Table 2.

Table 2. The scale of formation of the linguistic variable

Measure of discrepancy of two indexes - P (%)	$\aleph^2_{crit}$	Critical value of discrepancy (%)	Measure of discrepancy	Identifier	Actual value of discrepancy (%)
1	2	3	4	5	6
P > 99			absolute		
P = 99	6.635	15.30	absolute		
95 < P < 99			strong		
P = 95	3.841	8.32	strong		
90 < P < 95			significant		
P = 90	2.706	5.72	significant		
70 < P < 90			weak	*	5.00
P = 70	1.074	2.20	weak		
50 < P < 70			no discrepancy		
P = 50	0.455	0.92	no discrepancy		
P < 50			no discrepancy		

Information in columns 1 - 4 of Table 2 is formed continuously, whereas the result of specific decision for each paired comparison of time series elements is displayed in columns 5 and 6. Although the measure of discrepancy between the x_i and x_{i+1} is referred to the value of term-set α_3 = {weak discrepancy}, empirical value of discrepancy is equal to 5.00%, which is much closer to the critical value of discrepancy when **P** = 90%, i.e. to 5.72 than to 2.20 for **P** = 70%, below which there comes the scope of the term-set value α_1 = {no discrepancy}. Thus it is clear that the measure of discrepancy between 10.0 and 10.5 in this case is closer to "significant", visually and formally be as "weak".

Finally, the cited above example once again testifies to the fact that compromising on inaccuracies in discretization (dichotomization, categorization, scaling, and abstraction) [4] you can get a generic, argumentative variant of analysis, when the possibility of using standard statistical methods is very limited. The advantages of this method are determined also by the fact that process of using point and ranking scales in general is characterized by expert subjectivity, and use ranking scale, obtained by the proposed method, removes the specified element of subjectivity (Table 2) by replacing it with results of applying the set of proposed formal operations.

4. An Example of Using the Method. Conclusions

We analyze the demographic dynamics of the Russian Federation for 1970-1999 and form hypotheses about possible causes of its change over that period. Initial data on a ten-year periods are presented in the Table 3 [8].

Table 3. The population of the Russian Federation

Population \ years	1970	1979	1989	1999
Population, thousand people	130079	137551	147400	147105

The results of calculations by the proposed method are presented in Table 4.

Table 4. Quantitative estimators (Table 3) transformation into ranking

Years	The population of the Russian Federation, thousand people	Measure of discrepancy	Probability of discrepancy - P, %	Group
1970 1979	130.1 137.6	Weak	89.0	1 1
1979 1989	137.6 147.4	Significant	92.7	1 2
1989 1999	147.4 147.1	No discrepancy	24.0	2 2

Therefore, according to the Tables 4 and 5 data, it is possible to determine two groups of changes in years: the period of 1970-1989, during which the dynamics of the Russian population had changed in positive way and the period of 1989-1999 when from 1989 (the last census in the USSR) until 1999 the population had decreased, but only slightly. This is due to the fact that during the transition period, accompanied by a reduction in social security, the population of the Russian Federation subjectively "voted" the decline in fertility, which was shown statistically as imbalance between birth rates and mortality, when the birth rate was almost linearly decreased, and mortality steadily (and almost linearly) grew.

The proposed method of analysis of short time series by determination their measures of discrepancy in the adopted linguistic scale only ascertains a measure of discrepancy between adjacent elements of time series, and thus give guidelines for informal interpretation of results. Therefore, the proposed method of generalization (one-dimensional structurization, classification, and grouping) can be determined as non-domain-specific but method-oriented. It allows classifying of considered socio-economic phenomena by analyzed attribute and thus ensuring semantic stability of information between members of formed groups.

References

Abdullah L., Najib L. A new type-2 fuzzy set of linguistic variables for the fuzzy analytic hierarchy process. *Expert Systems with Applications,* 41(7), 3297-3305

Berkson J. Limitations of the application of fourfold table analysis to hospital data. *International Journal of Epidemiology,* 43(2), 511-515

Eliason S.R., Stryker R. Goodness-of-fit tests and descriptive measures in fuzzy-set analysis. *Sociological Methods and Research,* 38(1), 102-146

Kuss O. The danger of dichotomizing continuous variables: A visualization. *Teaching Statistics,* 35(2), 78-79

Malyshev N.G., Bershtein L.S., Bozhenyuk. A.V. *Nechetkie modeli dlya ekspertnykh sistem SAPR*, Moscow: Enrgoatomizdat, 1991, 136pp.

Saati T., Kerns K. *Analytical Planning. Systems Organizing* /Transl. from Engl. Moscow: Radio and connection, 1991, 224pp.

Sociologist's Workbook, Moscow: Nauka, 1983, 478pp.

The Russia's Population (2014) // www.gks.ru [Russian Federal State Statistics Service Site]. 3rd of August (http://www.gks.ru/free_doc/new_site/population/demo/demo11.xls)

Upton G. The Analysis of Cross-tabulated Data /Transl. from Engl. Moscow: Financy i Statistika, 1982, 143pp.