UDC 621.3:622:519.24

Application of nonparametric statistics methods for evaluation of some peculiarities of mining production

Boris Kobilyanskiy

PhD.,
Department of health and environmental safety
Teaching and research professional education institute of
Ukrainian engineering and pedagogical academy

Anatoly Mnukhin

Dr. Eng.,
Head of the research laboratory,
Zaporizhzhya State Engineering Academy

Abstract
At the present stage of technical development all complex systems of industrial production and particularly the coal industry require automating of entire process including the work of mining and transport equipment, which can not be done without a detailed mathematical description of specific phenomena characterizing the specific production.
Therefore, for the first time to solve the problem of complex technology management, multifactorial systems, it is suggested to use non-parametric statistics, which was not reflected in the national literature. Nevertheless, such mathematical approaches allow to perform correct forecast of the phenomena in question beyond the available experimental data with sufficient accuracy to solve the most practical problems, especially in those cases where the amount of data is relatively small.
Keywords: COAL INDUSTRY, ASSESSMENT, A NONPARAMETRIC STATISTICS

At present, coal industry enterprises are equipped with automated lines and installations, modern technological systems and sites, flexible manufacturing systems are being introduced. All this facilitate a solution to two interrelated problems: output of better products and improving the safety of the production process. Health, as a system of special knowledge, is intended as a means of ensuring the safety of technological processes and production. [1] Ukrainian scientists constantly pay attention to the improvement of safety management system of various companies [3-6]. One of the main ways of assessing of technology-related risk at the enterprises of higher risk, coal in particular, is a study of the level of possible injuries.

Chemical and mining production, military science, major planning of processes and phenomena – all they use mathematical statistics methods of high level, based on the theory of distributions, and, above all, normal and quasi-normal distributions. However, it is evident that as the complexity of character of processes under consideration, i.e., by stimulated use of other skew distributions, Student or Kolmogorov in particular, to describe the array of experimental data or their samples, traditional statistical methods are insufficient for this, and even with their full compliance to the solution of a number of mining electrical engineering problems [2], their use for forecast of the ergonomics systems, such as coal industry enterprises , they do not allow to give correct assessment of the phenomenon under consideration and forecast the system behavior in the conditions different from those under consideration.

In order to properly understand the idea of *non-parametric statistics* (the term was first used by Wolfowitz, 1942), one should get acquainted with the ideas of *parametric statistics*. So initially one should be familiar with the concept of statistical significance of criterion based on the distribution of certain statistics. In short, if you know the distribution of the observed variable, you can predict how in repeated samples of equal volume will “behave” used statistics – i.e. how it will be distributed. Let us suppose, for example, there are 100 random samples from a single population by 100 adults each. We calculate the average height (age or employment period) of subjects in each sample, i.e., we build the sample mean. Then distribution of sample means can be well approximated by a normal distribution (more precisely, *t*- Student’s distribution with 99 degrees of freedom). Now imagine that there randomly removed one more sample of the inhabitants of a city where, according to your ideas, there live people with above-average height. If the average height of people in the sample falls into the top 95% of the *t* distribution critical region, it can be reasonable to conclude that the inhabitants of the city, in fact, on average, higher (than in the general population), i.e. it is really a city of tall people.

The question is: “Are most variables have a normal distribution? In this example, we used the fact that the repeated samples of equal volume mean values (human height) will be t-distribution (with a certain mean and variance). However, this is true only if the variable (height) has a normal distribution, i.e., that people of a certain distribution of growth is normally distributed (Fig.1)

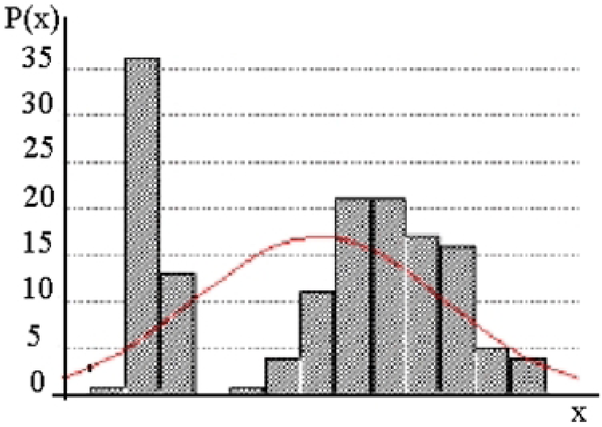

Figure 1. Comparison of the distribution laws

One of the factors limiting the application of criteria based on the assumption of normality is the sample size. As long as the sample is large enough (for example, 100 or more observations), we can assume that the sampling distribution is normal, even if you are not sure that the distribution of the variable in the population is normal. However, if the sample is small, these criteria should be used only if there is certainty that the variable is indeed a normal distribution. However, there is no way to test this hypothesis on a small sample.

Using criteria based on the assumption of normality, moreover, limited by scale measurements. Such statistical methods as *t*-test, regression and so on, suggest that the original data are continuous. However, there are situations where data are rather ranked (measured on an ordinal scale) than measured accurately.

A typical example is given by the group of data: the first position is occupied by a group with maximum number of workers of particular specialty, the second position is occupied by a group with a maximum number of workers among the remaining groups (among the groups, from which the first group is removed), etc. Knowing the ratings and the workers of one of the groups more than the number of workers of other one, but how much more it is impossible to say. Imagine you have 5 groups: A, B, C, D, E, which are located on the first 5 places. Let this month we had the following setup: A, B, C, D, E, and in the previous month: D, E, A, B, C. The question is, have there happened significant changes in the rankings of groups? In this situation, obviously, we can not use the t-test to compare these two sets of data, and go to specific probabilistic calculations (and any statistical test contains a probabilistic calculation!). We may suppose: what's the likelihood that the difference in the two arrangements of groups is caused by accidental causes, or this difference is too large and can not be explained by mere chance. In these arguments, we use only the grades or permutation of groups and do not use the specific form of the visitors number distribution.

Essentially, for each parametric criterion there is at least one nonparametric alternative.

In general, these procedures fall into one of these ca- tegories:

- criteria of differences for independent samples;
- criteria for dependent samples differences;
- assessment of the degree of dependence between variables.

In general, the approach to statistical criteria in the analysis of data has to be pragmatic and not to be loaded by unnecessary theoretical arguments. With access to the computer system of STATISTICA, you can easily apply several criteria to the data. Knowing about some problems of the methods you will select the right decision by simple experimentation. Development of the plot is quite natural: if it is necessary to compare the values of two variables, one should use the *t*-test. However, it should be mentioned that the test is based on the assumption of normality and equality of variances in each group. Exemption from these assumptions leads to a non-parametric tests, which are particularly useful for small samples.

Further, there are two situations related to the initial data: dependent and independent samples, which use the *t*-test for dependent and independent samples, respectively.

Development of the *t*-test leads to the analysis of variance, which is used when the number of comparison groups is greater than two. The corresponding deve- lopment of the non-parametric procedures leads to non-parametric analysis of variance, however, significantly poorer than the classical analysis of variance.

To evaluate the dependence, or the degree of the connection closeness, one calculates the Pearson correlation coefficient. Strictly speaking, its application has limitations, such as those associated with the type of scale, in which the data is measured and the nonlinearity function, that is why as an alternatively there also may be used non-parametric, or so-called rank correlation coefficient, for example, for ranked data. If the data are measured in nominal scale, they are of course presented in contingency tables that use the chi-square test of Pearson with different variations and adjustments for accuracy.

So, in essence, there are only a few types of criteria and procedures that one needs to know and to be able to use depending on the specific data. It is necessary to determine what criteria should be applied in a particular situation.

Nonparametric methods are most appropriate when the sample size is small. If a lot of data (eg, n> 100), often it does not make sense to use nonparametric statistics.

If the sample size is very small (for example, n = 10 or less), than the levels of significance for nonparametric tests, which use the normal approximation can be regarded only as rough estimates.

Differences between independent groups

If there are two samples (eg, men and women), which should be compared with respect to a mean value, for example, medium-pressure or the number of leukocytes in the blood, one can use the t-test for independent samples. Nonparametric alternatives to this test are the criterion of Val'd-Volfovits, Mann-Whitney [7th test and two-sample Kolmogorov-Smirnov criterion.**]**

Differences between dependent groups

If you want to compare two variables relating to the same sample, e.g., medical records of the same patients before and after medication, there is usually used a *t*-test for dependent samples.

Alternative non-parametric tests are the sign test and Wilcoxon test.

If we consider more than two variables related to the same sample, it is commonly used analysis of variance (ANOVA) with repeated measurements.

In order to evaluate the dependence between two variables, there usually calculated the Pearson correlation coefficient. Nonparametric counterparts of Pearson correlation coefficient are Spearman's rank correlation coefficient R, Kendall statistics, Gamma coefficient.

The coefficient of rank correlation (rank correlation coefficients') estimates the amount of dependence between the variables measured in ordinal scale, i.e., between ordinal variables.

A transparent method of building a pair of correlation coefficients of the generalized correlation coefficient is suggested by Daniels (Daniels NE 1948, Biometrika, v. 35, p. 416-417).

The generalized correlation coefficient is determined by the formula:

$$\Gamma = \frac{\sum a_{ij} b_{ij}}{\sqrt{\left(\sum a_{ij}{}^2\right)\left(\sum b_{ij}{}^2\right)}} \qquad (1)$$

where $a_{ij} = a(X_i X_j)$, $b_{ij} = b(Y_i, Y_j)$ —are some functions of pairs of observations of X and Y, respectively, the sum is taken over all pairs i, j.

Note that when $a_{ij} = X_j - X_i$, $b_{ij} =$ Ij- Y_i. We get the usual Pearson correlation coefficient. If the variables are ranked, we are working with the rank. Let us order the X_i values in ascending order, that is build a variation number of these values. The order of value X_i in this series is called its rank and denoted Ri.

Then let's arrange Y_i values in ascending order. Number of Y_{ii} value in this series is called its rank and denoted S_i.

Spearman's rank correlation is calculated as generalized correlation coefficient with the replacement of observations by their ranks. Formally for the generalized correlation coefficient it is necessary to put $a_{ij} = R_j - R_i$, $b_{ij} = S_j – S_i$.

Kendall's coefficient is calculated, if in the formula for generalized coefficient to put the $f_{ij} = 1$ if $R_i < R_j$ and $f_{ij} = -1$ at $R_i > R_j$. Values b_{ij} are given by similar relations with the replacement of ranks R_{ij} grades S_i observations Y. So, we can clearly see that the idea of correlations arises from the same source.

If there are more than two variables, there used Kendall's concordance coefficient. For example, it is used to assess the consistency of the views of independent experts (judges), for example, points gived to the same participants.

If there are two categorical variables, to assess the degree of dependence there used standard statistics and the relevant criteria for contingency tables: Chi-square, phi-coefficient, Fisher's exact test.

Classical statistics Pearson's chi-square is remarkable in that its distribution is approximated by the distribution of chi-square test, for which there are detailed tables. Percentage points of distribution of chi-square can be effectively calculated in STATISTICA system using probability calculator.

Property of chi-square test (accuracy of distribution approximation of chi-square) for 2 × 2 tables with small expected frequencies can be improved by reducing the absolute value of differences between expected and observed frequencies by 0.5 before squaring. This so-called Yates' correction for continuity for frequency tables 2 × 2, which is typically used when the cells contain only few low frequencies and become less than 5 (or even less than 10).

If the sum of frequencies is small, it is better to use Fisher exact test instead of the chi-square test.

There are recommendations of Kokren for tables 2 × 2: if the sum of all the frequencies in the table is less than 20, than it is necessary to use the Fisher's exact test. If the sum of frequencies is greater than 40, it is possible to use the chi-square test with continuity correction.

However, these recommendations are not universal [7].

Since the data usually have cells with low frequencies (2 and 3), to improve the accuracy of the chi-square test using Yates correction. Since we are interested in one-sided alternative, we divide the level of p = 0.0012 in half and get 0.0006.

It is not easy to give simple advice concerning the use of non-parametric procedures. Each nonparametric procedure in the module has its advantages and disadvantages. For example, two-sample Kolmogorov-Smirnov test is sensitive not only to the difference in the position of two distributions, such as the differences in the average, but also sensitive to the form of distribution. Wilcoxon pairwise comparisons suggest that it is possible to rank the differences between the compared observations. If it is not, it is better to use a sign test. In general, if the result of this study is an important observation (for example, one answers the question - Does expensive and painful drug therapy help people?), than it is always advisable to test non-parametric tests. Perhaps the results of testing (by various tests) will be different. In this case, we try to understand why different tests gave different results. On the other hand, nonparametric tests have less power than their parametric competitors, and if it is important to detect even weak effects (for example, when finding out whether a given food additive dangerous to health) multiple tests should be conducted and test statistic should be carefully chosen.

When the samples are very large, then the sample means are subjected to the normal law, even if the original variable is not normal, or measured with an error. Thus, parametric methods, which are more sensitive (have a greater statistical power) are always suitable for large samples. Most significance tests of a lot of nonparametric statistics, are based on asymptotic theory (large samples) so the appropriate tests are often not implemented, if the sample size is too small.

Thus, it follows from the above that the processing of sample sizes in hundreds of data typical for coal industry in Ukraine, methods of nonparametric statistics are the most suitable.

References

1. Minko V. M (2012). *Okhrana truda v mashinostroenii* [Occupational safety in mechanical engineering] Moscow, Academy.
2. *Typove polozhennya pro poryadok provedennya navchannya i perevirky znan z pytan ohoroni pratsi* [Typical provisions on training and knowledge tests on safety] (2005): NPAOP 0.00-4.12-05. Kharkiv, Fort.
3. Stupnizka N.V. (1999) *Pidvishchennya efektyvnosti planuvannya zahodiv zapobigannya vyrobnychomu travmatyzmu na pidpryyemstvakh mashynobuduvannya* [Improved planning measures to prevent occupational injuries in the mechanical engineering]. Extanded abstract of Doctor`s thesis. Lviv.
4. Kruzhylko O.Ye. (2001)*Udoskonalennya kompleksnoyi otsinku stanu ohorony pratsi na pidpryyemstvah*[Improving comprehensive assessment of safety in enterprises]. Extanded abstract of Doctor`s thesis. Kyiv..
5. Hunchenko O. M. (2007) *Udoskonalennya systemy upravlinnya ohoronoyu pratsi na mashynobudivnykh pidpryyemstvakh* [Improving safety management in engineering enterprises]. Extanded abstract of Doctor`s thesis. Luhansk.
6. Kasyanov M.A., Meduanik V.O., Hunchenko O.M., Vyshchnyevskiy D.A. (2008) Problemy stanu I neobkhidnosti vdoskonalennya systemy upravlinnya okhoronou pratsi v haluzi mashinobuduvannya [The problems of the state and the need to improve safety management in engineering]. *Journal of East Ukrainian National University named after V. Dahl.*, 6, 3–9.
7. Elyseeva I.I. (1996) *Obshchaya teoriya statistiki* [General Theory of Statistics]. Moscow.